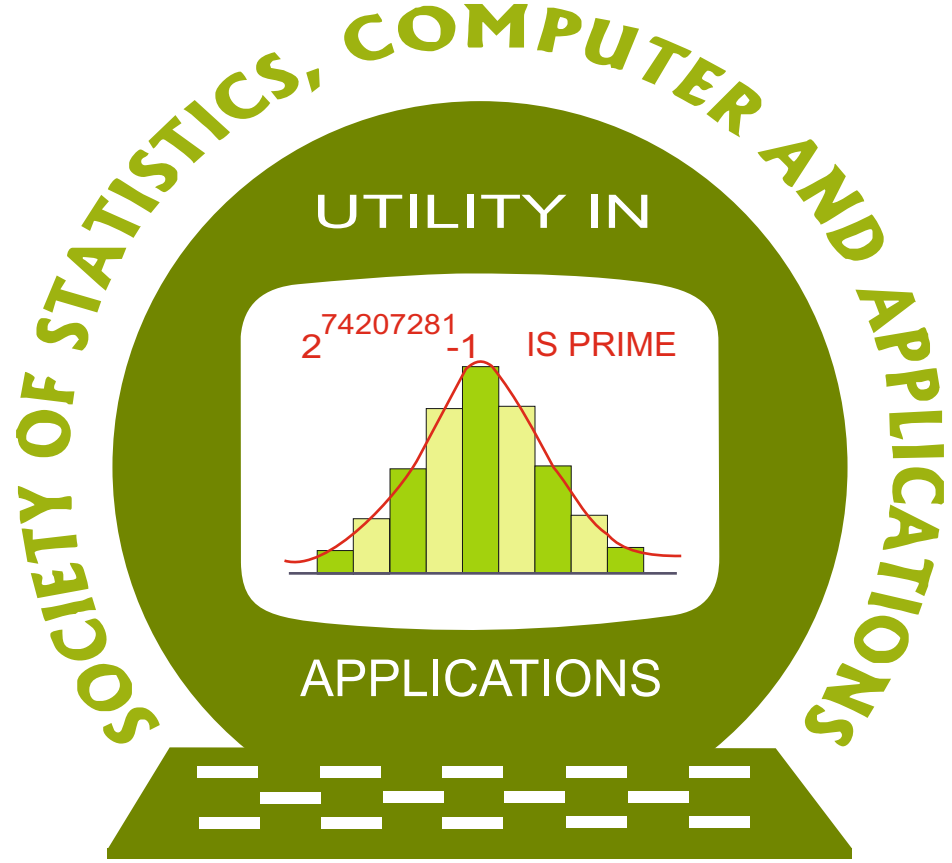
ISSN 2454-7395(online)

STATISTICS AND APPLICATIONS



FOUNDED 1998

Journal of the Society of Statistics, Computer and Applications

https://ssca.org.in/journal.html

Volume 22, No. 3 (C R Rao Special Issue), 2024 (New Series)

Society of Statistics, Computer and Applications

Council and Office Bearers

Founder President
Late M.N. Das

President
V.K. Gupta

Executive President
Rajender Parsad

Patrons

A.C. Kulshreshtha, A.K. Nigam, Bikas Kumar Sinha, D.K. Ghosh
G.P. Samanta, K.J.S. Satyasai, Pankaj Mittal, Prithvi Yadav
R.B. Barman, R.C. Agrawal, Rahul Mukerjee, Rajpal Singh

Vice Presidents

A. Dhandapani, Manish Kumar Sharma, Manisha Pal, P. Venkatesan
Praggya Das, Ramana V. Davuluri, S.D. Sharma, V.K. Bhatia

Secretary
D. Roy Choudhury

Foreign Secretary
Abhyuday Mandal

Treasurer
Ashish Das

Joint Secretaries

Aloke Lahiri, Shibani Roy Choudhury, Vishal Deo

Council Members

B. Re. Victor Babu, Banti Kumar, Imran Khan, Mukesh Kumar, Parmil Kumar
Piyush Kant Rai, Rajni Jain, Rakhi Singh, Raosaheb V. Latpate, Renu Kaul
Sapam Sobita Devi, Shalini Chandra, V. Srinivasa Rao, V.M. Chacko, Vishnu Vardhan R.

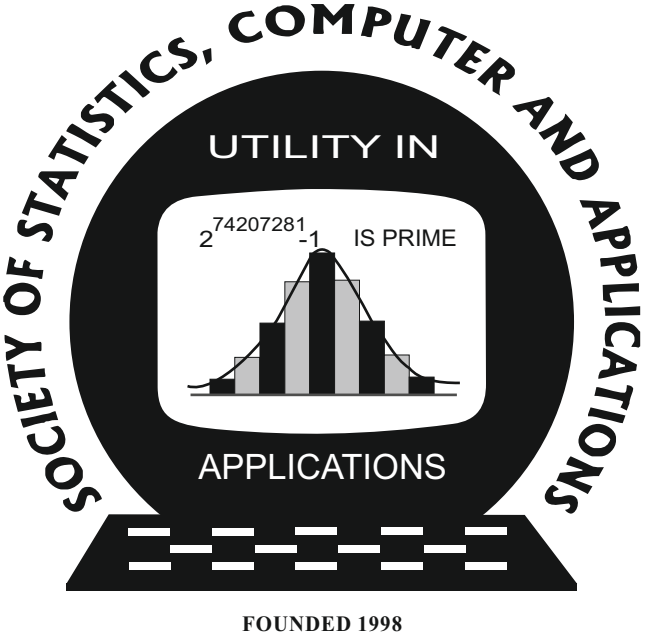
Ex-Officio Members (By Designation)

Director, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
Chair Editor, Statistics and Applications
Executive Editors, Statistics and Applications

Society of Statistics, Computer and Applications
Registered Office: I-1703, Chittaranjan Park, New Delhi- 110019, INDIA
Mailing Address: B-133, Ground Floor, Chittaranjan Park, New Delhi-110019, INDIA

Statistics and Applications

ISSN 2454-7395(online)



Journal of the Society of Statistics, Computer and Applications

https://ssca.org.in/journal.html

Volume 22, No. 3 (C R Rao Special Issue), 2024 (New Series)

Special Issue on "Life and Work of C R Rao (1920-2023): The Revolutionary of Statistical Sciences" in fond memory and honour of Late Prof C R Rao who left for his heavenly abode on August 22, 2023

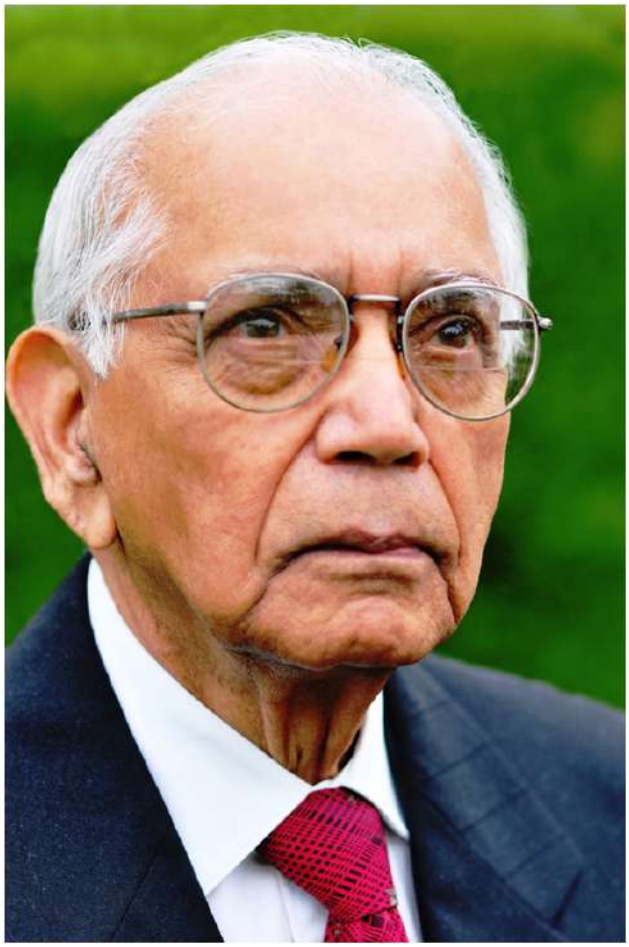
Statistics and Applications

Volume 22, No. 3 (C R Rao Special Issue), 2024 (New Series)

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series)
http://www.ssca.org.in/journal

CONTENTS

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series)
http://www.ssca.org.in/journal



Professor Calyampudi Radhakrishna Rao
(September 10, 1920 – August 22, 2023)

"He [My Dad] loved Nrityero Tale Tale. He made me learn that dance at the Manipuri dance school that was on the Indian Statistical Institute (ISI) campus in the Manipuri style although I was a Bharatanatyam and Kuchipudi dancer. Years later I taught it to one of my students here in Buffalo and she performed it during her graduation show. It continues to be one of my favorites too. Dad took me to several Bengali Plays in the theatres in Shyambazar. I think some of the ISI faculty used to act in those plays. You* probably know that he also spoke Bengali."

... ***Tejaswini***

[**indicates – an Email correspondence to Bikas Kumar Sinha on September 23, 2024*].

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series)
http://www.ssca.org.in/journal

From Chair Editor's Desk....

When the world of statistical fraternity, and the world of scientists across all disciplines of sciences, art and commerce, got this sad news of this noble and pious soul leaving the body of Professor Calyampudi Radhakrishna Rao and seeking shelter in the paradise, everyone was moved and touched and was in tears. The world had lost an all-time great visionary, a researcher and a gem with humane values. Everybody felt that this loss has created a void that will not be easy to fill.

Professor C. R. Rao, a Doyen in the domain of Statistical Sciences, was blessed by a super power to ignite the lamp of his soul while he descended on the Earth! He was extraordinarily talented and he roamed over various upcoming areas in the domain of statistical sciences with extreme ease and command and served the broader statistical world in a wide variety of leadership positions. Known to most of us as simply 'CRR', his name remains a byword for excellence.

The present Chair Editor of Statistics and Applications (S&A), Vinod Kumar Gupta, expressed a strong desire to Bikas Kumar Sinha and twin brother Bimal Kumar Sinha to bring out a Special Issue (SI) of the journal S&A in the honour of and to pay rich and befitting tributes to the fond memory of CRR. The twin brothers immediately responded in affirmation and the three decided that the December Issue of 2024 will be a SI to commemorate and pay our huge respect to this doyen of statistical sciences by bringing out a special issue having research papers echoing the research interests of CRR by those who have been his family, his students, his colleagues and collaborators and his friends and well wishers. The editorial board of the S&A resolved to appoint a team of Guest Editors (GEs) led by Bikas K. Sinha and twin brother Bimal K. Sinha and supported by the Chair Editor Vinod K. Gupta for this SI covering different areas of expertise and succeeded in having some of them from around the globe with direct personal contacts with CRR. The other important GEs of the team comprised of Shyamal Peddada, Thomas Mathew, Tapan K. Nayak, Bhramar Mukherjee, Nairanjana Dasgupta, and S. Ejaz Ahmed.

The team of GEs in their wisdom decided that this SI will be called as "***Life and Work of CR Rao (1920-2023): The Revolutionary of Statistical Sciences.***"

When this project [publication of special issue of S&A as a tribute to the fond memory of Late CRR] was to be launched, the GEs got spontaneous responses from several dozen contributors across the continents who were forthcoming with contributions related to the research interests of CRR. A good number of the contributors also had their personal

experiences to narrate.

We decided to add a special feature to this SI. CRR lived for 19 days short of 103 years. To highlight this significant event, we have compiled a list of 103 research papers of CRR that had impacted the furtherance of research in those times and in the chosen areas. The selection of 103 research papers is a representative sample of the total number of research papers of CRR, including some of his early-life path breaking research. We could trace his earliest paper in 1941 (perhaps among the earliest papers published at the age of 21 years). We made an effort to have at least one paper for each of the years from 1941 till 2020. Unfortunately, we could not find any journal publications for the years 2011, 2015 and 2019, respectively. Towards this compilation, we received tremendous help and useful documents from T. J. Rao, B. L. S. Prakasa Rao and T. Krishna Kumar. This help is gratefully acknowledged.

The **37** papers that appear in this SI have been classified into two broad categories. The Part I comprises of eight **(8)** papers describing the "**Facets of Professor C. R. Rao.**" These papers are mostly non-technical in nature. We start with an account of 'THE FAMILY MAN' - as CRR had been! This comes from his daughter Tejaswini. She has also obliged us by providing some unique family pictures. Then we have accounts of personal experiences of CRR's associates, collaborators, students [direct/indirect].

The Part II comprises of **29** "**Regular Research Papers**" covering a wide range of theoretical topics and applicational areas. The contributors had freedom to make their own choice of the topics and deal with their contribution in the best possible way leading to innovative research. The selection of the topics was essentially dictated by the research interests of CRR and every paper in a way addressed one of these research interests. Of course, in the name of CRR, the motivation has been research-oriented and every contributor tried to focus on the use of latest available tools and techniques – at times developing new techniques.

Upon our request, BLS Prakasa Rao spontaneously agreed to write the Preface for this CRR Memorial SI. We are deeply obliged to him for his contribution. The Preface itself is his experience and vision about the life and work of CRR.

We are highly indebted to the GEs for rendering their valuable support in the preparation of this SI. Their choice of topics and contributors in different areas is praiseworthy. They maintained the timeline very strictly and this is the reason we have been able to bring out this SI in time.

Our heartfelt thanks and gratitude go to all the contributors, the distinguished authors, who have enriched the volume with their thought-provoking and illuminating contributions across various topics in theory and applicational areas. Their contributions have added value to this SI. Their unflinching support to maintain time schedule in submitting their original contribution, then preparing the revised version based upon the reviewers suggestions and formatting the paper as per the journal requirements are unparalled. We are grateful to all the contributors for this.

The reviewers, an unobservable layer without whom the process of journal publication cannot function at all, have also been prompt and thorough. Their suggestions helped in improving the quality and presentation of the contents. We are indebted to all the reviewers and thank them sincerely for their support. We would like to place on record our highest admiration for the Executive Council of SSCA and the Editorial Board of S&A for their support and for entrusting their faith on the GEs for bringing out this special issue as a tribute to Professor C. R. Rao. The help received from Baidya N. Mandal and R. Vishnu Vardhan, Managing Editors, Jyoti Gangwani, Production Executive and Siva G., who helped with the latex template, for bringing the papers in the format of the journal is highly appreciated. This SI contains papers of high academic standards covering a wide spectrum of statistical research. We are confident that the readers would find these papers enjoyable and a resource for generating newer ideas for advancing research in statistical sciences. We reiterate that the 36 papers in this SI are by CRR's family, former students, colleagues, collaborators, friends, and others who have been influenced by his research, teaching, mentoring, and generous friendship.

This SI is our sincere endeavor to pay homage and rich tributes to this giant doyen of statistical sciences with a towering stature, filled with traits of humanity like gentleness, kindness, humbleness and gratitude along with passion, vigor, enthusiasm, and zeal to enrich statistical sciences with fresh and fragrant ideas.

It may not be out pf place to mention here that the Society of Statistics, Computer and Applications (SSCA) was founded in 1998 to honour great legendary Professor M. N. Das. Since then, the SSCA has been organizing National / International Conferences every year along the length and breadth of the country. It has organized thus far 26 conferences. The SSCA, among other scientific activities, also brings out this journal called Statistics and Applications (S&A). The journal is available at https://ssca.org.in/journal.html. The Issue (No. 3) of the Volume 22 of S&A has been brought out as a tribute to an all-time great legendry and father of modern statistics CRR, who at the age of 18 days short of 103 years (10 September 1920 – 22 August 2023), left for his heavenly abode. We fervently hope that this endeavor serves as a rich tribute to this dynamic and passionate researcher who revolutionized statistical sciences with his original path-breaking research. The statistical fraternity is poorer in his loss.

Bikas Kumar Sinha
Bimal Kumar Sinha
December 2024 *Vinod Kumar Gupta*

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series)
http://www.ssca.org.in/journal

Guest Editors Panel

The Guest Editors (GEs) panel appointed for bringing out the Volume 22, No. 3, December 2024 (Special Issue) of Statistics and Applications on "**Life and Work of C R Rao (1920-2023): The Revolutionary of Statistical Sciences**" in fond memory and honour of Late Prof C R Rao who left for his heavenly abode on August 22, 2023, comprises of the following:

1. *Bikas Kumar Sinha – Leader*
2. *Bimal Kumar Sinha – Co-Leader*
3. *Shyamal D. Peddada – Member*
4. *Thomas Mathew – Member*
5. *Tapan K. Nayak – Member*
6. *Bhramar Mukherjee – Member*
7. *Nairanjana Dasgupta – Member*
8. *Syed Ejaz Ahmed – Member*
9. *V. K. Gupta – Member and Chair Editor 'Statistics and Applications'*

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series)
http://www.ssca.org.in/journal

PREFACE

Professor Bikas Kumar Sinha, who is the Leader of the Guest Editors Panel for this special issue of "Statistics and Applications", invited me to write a preface for this special issue of the journal to be released in memory of Late Professor C. R. Rao. I am honored to be a part of the group as I was a student of Professor C. R. Rao in the very first batch of M. Stat. program at the Indian Statistical Institute, Calcutta (now Kolkata) during the years 1960-62 and was his colleague at the Indian Statistical Institute, New Delhi during the years 1976-79 before he left for USA. The contributors to this volume are well known specialists in their chosen branches or areas of statistics and their contributions reflect the areas to which Professor C. R. Rao has made significant contributions.

Calyampudi Radhakrishna Rao (aka) as C. R. Rao needs no introduction to statisticians, mathematicians, scientists or communication engineers. In the volume "Glimpses of Indian Statistical Heritage", edited by J. K. Ghosh, S. K. Mitra and K. R. Parthasarathy (Wiley Eastern Limited, New Delhi (1992)), who are themselves distinguished statisticians and probabilists, C. R. Rao wrote an autobiographical account highlighting the circumstances and influences that led him to a career in statistics and probability. He titled his autobiographical account as "Statistics as a Last Resort". It is appropriate to mention that he came into statistics by chance. By spending a life time putting chance to work, he has built an inspiring legacy.

C. R. Rao was born on September 10, 1920 in Huvvina Hadagalli, then in the integrated Madras province and now in the state of Karnataka. His father C. Doraiswamy Naidu was an Inspector of Police and his mother was A. Laxmikanthamma and Rao grew up in a family environment. Rao was admitted in class 2 (second grade) in 1925 when he was only 5 years old. Since Rao's father was an inspector of police, the job required the family to move from place to place once in every two or three years.

Rao completed his classes 2 and 3 in a town named Gudur, classes 4 and 5 in Nuzvid and first and second forms in Nandigama all in the present state of Andhra Pradesh. At this stage, his father retired and decided to settle down in Visakhapatnam. Rao finished his high school and joined the Andhra University for obtaining his first college degree in Visakhapatnam. Rao's early childhood involved frequent moves from one place to another but that did not affect his studies . His parents provided him guidance and environment conducive to studying and instilled in him work ethics that endowed him to achieve higher

goals in life. As a student, his ambition was to keep on learning. He said that he has inherited his father's analytical ability and his mother's zeal and industry.

Rao said that his mother was instrumental in instilling a sense of discipline in him. In his book on "Statistics and Truth: Putting Chance to work", Rao acknowledges her contribution to his life with the dedication "For instilling in me the quest for knowledge, I owe to my mother, A. Laxmikanthamma, who, in my younger days, woke me up every day at four in the morning and lit the oil lamp for me to study in the quiet hours of the morning when the mind is fresh".

Rao graduated with the B.A. (Hons) degree in Mathematics at the Andhra University in Vizag. It was at the Andhra University as a seventeen year old that Rao developed research interest in mathematics. His most inspiring teacher was a Cambridge trained mathematician Dr. Vommi Ramaswami who was the head of the Department of Mathematics. Rao finished the B.A. (Hons) course at the age of 19 and wanted to pursue a research career in mathematics. With a first class and first rank in B.A. (Hons) degree, Rao thought he would qualify for a scholarship for doing research in mathematics. He did not get the scholarship for bureaucratic reasons. He was in search of a job and saw an advertisement for a mathematician for the army survey unit. He went to Calcutta to appear for an interview for the job but was not successful. During his stay in Calcutta, he met one Subramanian who was employed in Bombay but had been sent to Calcutta for training in statistics at the Indian statistical institute (ISI). Rao joined ISI at his suggestion. As they say "Rest is history".

Rao obtained his Ph.D. degree from the Cambridge University and became a professor at ISI at the age of 29 years. After retiring from ISI in 1980, he moved to USA and worked for another forty three years and was a research professor at the University of Buffalo in USA till the time of his passing away. He received several awards including the Bhatnagar award, India Science award from the Government of India, National Medal of Science from USA and elected as Fellow of several academies in India and abroad. He received 39 honorary doctorates from universities in India and abroad. Several students received Ph.D. under his guidance.

As they say "Statistics is the poetry of sciences". Statistics is the soul of scientific inquiry. It is applied by researchers across a spectrum of science, engineering, business, technology, medical, government and financial settings to name some. These applications lead ultimately to tangible benefits that improve the well being of humanity. With the increasing role of information technology, the society has been inundated by a data deluge and statisticians are the society's experts for extracting usable information from the mass of noise in those data sets. Statistics and statisticians make the science better. It is an invisible science. It is said that "A physicist solves a problem in physics using the available knowledge in physics, a chemist does the same thing in chemistry, so also a biologist and an engineer. There is nothing like a statistical problem a statistician is trying to solve with the available knowledge of statistics. His or her job is to help the scientists to solve problems in their discipline by applying available statistical methodology, but more often by developing

appropriate new statistical methodology".

C. R. Rao was among the world wide leaders in statistical science over the last several decades.

Rao's career in statistics is dotted with remarkable achievements. The first result in statistics to bear Rao's name was proven by him, while still at ISI, at the age of 25 and came to be known as the Cramer-Rao inequality. In his remarkable 1945 paper published in the *Bulletin of the Calcutta Mathematical Society*, Rao demonstrated three fundamental results that paved the way for the modern field of statistics and provided statistical tools heavily used in science. The first now known as the Cramer-Rao lower bound provides a means of knowing when a method for estimating a quantity is as good as any method can be. The second result named as the Rao-Blackwell theorem provides a means of transforming an estimate into a better, in fact optimal, estimate. Together, these results form a foundation on which much of statistics is built. And the third result provides insights that pioneered a new interdisciplinary field that has come to be known as *information geometry*. Combined, these results help scientists extract information from data efficiently. The monumental work by Rao has not only revolutionized statistical thinking in its time but also continues to exert influence on human understanding of sciences across wide spectrum of disciplines according to the Chair of the International Prize in Statistics which Rao received . Rao made distinct and extensive contributions to several branches of the subject of statistics and its applications leading to efficient methods of statistical analysis.

Once a doctor examining him for some stomach ailment told Rao that the food for each individual in stomach would be a variable and normally distributed..(a term familiar to the statisticians). Rao told "the doctor was trying to give me a lecture in statistics, which I had been teaching to my students for over 25 years ...(at that time)." Rao lost his baggage during one of his international travels. One of the agents of the airlines called Rao next day and said "Good News Mr. Cramer Rao, we found your baggage" thinking that Cramer Rao is Rao's name but it is the lower bound named after him and Professor Cramer who have discovered the result.

In multivariate analysis, one has to deal with extraction of information from a large number of measurements made on each sample unit. Not all measurements carry independent information. It is possible that a subset of measurements may lead to procedures which are more efficient than using the whole set of measurements. Rao developed a test to ascertain whether or not the information contained in a subset is the same as that given in the complete set. He also developed a method for studying clustering and other inter-relationships among individuals or populations. Using general diversity measures applicable to both qualitative and quantitative data, the method of analysis of diversity was developed by Rao for which he introduced the concept of quadratic entropy in the analysis of diversity.

Combinatorial arrangements known as orthogonal arrays were introduced by Rao for use in the design of experiments. These arrangements are widely used in multi-factorial experiments to determine the optimum combinations of factors to solve industrial problems.

These have also applications in coding theory. An important result of practical interest resulting from this novel approach is the Hamming-Rao bound associated with orthogonal arrays.

Rao's work was done in India and his intellect shaped statistics worldwide. He was among the worldwide leaders in statistical science. His research, scholarship and professional service had a profound influence in the theory and applications of statistics and are incorporated into standard references for statistical study and practice.

When Rao joined the ISI in early forties, statistics was not considered as an independent subject and no university offered courses at the Masters level. Rao developed numerous courses in statistics over the years which were later converted into bachelor's and masters degree at ISI when ISI was declared as an Institute of National Importance by an act of Parliament in 1959. Rao also initiated the Ph.D. program in theoretical statistics and probability. Rao guided the research work of over fifty students for Ph.D.

As the Head of Research and Training School at the ISI, Rao developed a variety of courses to train statisticians to work in different applied areas. Rao established research units in ISI to work on special projects in subjects such as economics, sociology, psychology, genetics, anthropology, geology and related areas. The idea of establishing these applied research units is to provide interaction between statisticians and scientists to promote the application of statistical methods in research in other areas and to develop new statistical methods motivated by real problems.

Pandit Jawaharlal Nehru, who was the Prime Minister at that time, was greatly interested in development of statistics. He visited ISI a number of times at the invitation of Professor Mahalanobis and Rao had the opportunity of discussing with him the national statistical system and training of statisticians to work in state statistical bureaus. Nehru moved a resolution in the parliament in 1959 declaring ISI as an Institute of National Importance.

Rao was the author of 14 books. Two of his books were translated into several European, Japanese and Chinese languages. Rao received 39 honorary doctorates from universities in 19 countries spanning over all continents. Rao received several awards and medals. Some of them are the National Medal of Science, the highest award given to a scientist in USA in 2002, India Science award in 2009, the highest award given to a scientist in India and the Guy Medal in Gold from the Royal Statistical Society in 2011, the highest award given to a statistician in UK.

Rao has received the Bhatnagar award in 1963 and International Mahalanobis prize in 2003 for lifetime achievement in statistics and the promotion of best statistical practice from the International Statistical Institute. The Ministry of Statistics and Program Implementation (MOSPI), Government of India has instituted a National award in honor of C. R. Rao. He was elected as a Fellow of the Royal Society (FRS)in UK, Fellow of the Indian National Science Academy (FNA), Fellow of the Indian Academy of Sciences (FASc), Fellow of the National Academy of Sciences (FNASc) in India, and Fellow of the Third world Academy of Sciences besides several others. Rao celebrated his 102nd birthday on September 10, 2022.

C. R. Rao, a professor whose work more than 75 years ago continued to exert a profound influence on science, has been awarded the 2023 International Statistics Prize in his 102nd year. Awarded biennially at the World Congress of International Statistical Institute, the International Statistics Prize in Statistics is managed by a foundation consisting of five major statistical societies: American Statistical Association, Institute of Mathematical Statistics, International Biometric Society, International Statistical Institute and the Royal Statistical Society.

A scientist visiting ISI from the former Soviet Union went to meet Dr. Rao (as he is known to all the workers at ISI) when he was in Calcutta at his residence. He was told that Rao was repairing his car. He met him in the office room later when Dr. Rao was with his students, then saw him playing badminton outdoors in the evening and had dinner with him in the night. The scientist remarked that "I have seen the mechanic, the athlete, the scholar and the perfect host, all in one day." He was an enthusiastic photographer and was very much interested in spreading dance forms such as Kuchipudi.

I have also graduated from the Department of Mathematics from the Andhra University in 1960 as Professor C. R. Rao and later joined ISI as a student during the years 1960-62 for my Masters program in Statistics. I met Professor Rao as a student at the age of 17 and attended courses given by him. I was his colleague at the Indian Statistical Institute, New Delhi during the years 1976-79 and later joined the CR Rao Advanced Institute of Mathematics, Statistics and Computer science, Hyderabad as Ramanujan chair Professor at his invitation. I had the privilege of participating in a Zoom meeting honoring him across continents during his centenary year. ISI Retired Employees Association has released a book entitled "A Tribute to the Legend of C. R. Rao, The Centenary Volume". Professor T. J. Rao, who is a well-known survey sampling expert, and I were both students of Professor C. R. Rao at the ISI and all the three of us are alumni from the Department of Mathematics at the Andhra University, Vizag.

Rao passed away on August 22, 2023 at the age of 102 just about two weeks before his 103rd birthday on September 10, 2023. He was a Research Professor at the University at Buffalo, USA till the last day. India has lost a distinguished statistician and a great scientist.

I am happy to note that the journal "Statistics and Applications" has taken the initiative to bring out this special issue in memory of Late Professor C. R. Rao and happy to be a part of this activity to pay my homage as his student, his colleague and his admirer."

B. L. S. Prakasa Rao
Hyderabad, India

August 12, 2024

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 1–2
http://www.ssca.org.in/journal

Calyampudi Radhakrishna Rao – As a Family Man

Tejaswini Rao
Daughter of C.R. Rao, Pittsburg, US.

Received: 28 May 2024; Revised: 30 May 2024; Accepted: 01 June 2024

Calyampudi Radhakrishna Rao, popularly known as CR Rao and henceforth addressed as Rao in this text, was first and foremost a loving and devoted family man. He married Bhargavi in 1948 and since that day their relationship was one of mutual admiration and respect, and unwavering support. Rao took great pride in his wife's competence, academic achievements, her intelligence and ability to articulate her opinion on a wide variety of topics. Bhargavi was his constant companion, traveling the world over with him, creating the supportive atmosphere needed for pursuing his academic activities, and providing the social environment to foster his relationships with his colleagues and students. Many remember evenings at the Rao's home filled with lively conversation and debates, sumptuous meals, and warm hospitality. Bhargavi was truly the love of his life in many ways. The Rao family grew to include their daughter Tejaswini and son-in-law Vincent O'Neill, son Veerendra and daughter-in-law Malini, grandsons Amar, and his wife Mitra along with a great grandson Khai, and Rohith and his partner Hannah Garfield. Rao, despite his busy days filled with teaching, research, administration, and travels took personal interest in overseeing the activities of his children. His children thought of him as being quiet and serious when they were young but soon as young adults realized and recognized his sensitive, caring, and humorous nature that nurtured them. His grandchildren were a source of boundless joy whom he entertained with trips to the park and conversation involving interesting and amusing questions. They remember their grandfather not as the renowned scientist or the stoic face seen in many of his photographs but as the kind, generous, funny, and mischievous man that he really was. Rao and his great grandson had a special bond, Rao a 102-year-old man and Khai less than a year old. Their faces lit up when they saw each other and they enjoyed each other's company in a beautiful silence filled with affectionate smiles and gurgling laughs. Rao was very fond of children and the feelings were mutual. He would regale them with riddles and jokes that many of them remember and talk about until today as adults. He was a Pied Piper as his colleague's children would follow him and Bhargavi during their evening walks with the added attraction of finding candy in Uncle Rao's pockets.

Rao was a creature of habit, cherishing daily walks with his wife and the morning ritual of reading his newspaper. After moving to the US, every evening, like clockwork, he would indulge in a glass of wine at exactly 7 pm while watching Jeopardy and Wheel of Fortune or his favorite sport, tennis with his family. Rao enjoyed watching all forms of sports and considered himself quite the badminton player. One of his favorite pastimes while living on the Indian Statistical Institute campus in Kolkata was playing badminton with colleagues

Corresponding Author: Tejaswini Rao
Email: tejarao@gmail.com

and friends after his evening walks. Rao loved the arts, especially dance. He saw Ram Gopal perform in London when he was a student in Cambridge in the early 1940s. This event started his lifelong interest in dance. He promoted many Indian classical dancers as a member of various cultural associations in Calcutta and Delhi, and he also served as president of the Kuchipudi Academy in Delhi. In addition, he enrolled Tejaswini in dance classes as a five year old and she became an accomplished performer of two Indian classical dance styles, Bharata Natyam and Kuchipudi. He fostered his love for the arts in his children, enriching their lives. Rao also had a passion for photography, capturing his international travels with Bhargavi in film from the early 1950's, taking hundreds of treasured photographs of his family, colleagues, and friends. He loved giving his friends enlargements of photographs he had taken of them. Many of his photographs have appeared in magazines and newspapers. Rao's other interests were gardening, writing humorous articles, and a little bit of cooking as well. He had a green thumb and could grow any flower or vegetable from seeds. Mothering each of the plants and watching them spring to life was a source of relaxation and pleasure. His subtle sense of humor that he is known for is also reflected in published articles that he has written on everyday incidents that he has observed. So the powerful sense of observation that made him the scientist that he was extended to every aspect of his life. Rao enjoyed occasionally giving instructions on how to prepare a dish, doing all the prep work and took delight in teasing Bhargavi that he taught her to cook.

Rao valued the importance of education and established educational scholarships in many universities for talented and disadvantaged students. He also instituted medals at various Institutions to recognize accomplished Statisticians. Acute sense of observation and discipline made Rao the remarkable scientist and statistician but family and friends best remember him for his humility, affectionate and compassionate heart, gentle nature, and a mischievous sense of humor that never failed to bring a smile.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 3–6
http://www.ssca.org.in/journal

Calyampudi Radhakrishna Rao – As a Teacher in Calcutta School

Kirti R. Shah
University at Waterloo, Canada.

Received: 31 May 2024; Revised: 03 June 2024; Accepted: 05 June 2024

I first heard of Calyampudi Radhakrishna Rao (hereafter addressed as Rao) when our Professor at University of Bombay recommended two texts, one by Herold Cramer and the other by C. R. Rao. I bought both the books. I did not know that the two names were connected by the famous Cramer-Rao inequality.

In 1959 I applied for admission to the Research and Training School (RTS) of the Indian Statistical Institute (ISI) and I was very excited to receive a letter of admission signed "C. R. Rao". That signature is still fresh in my mind.

I joined RTS in August 1959. That was before the Parliament passed the Indian Statistical Institute Act, 1959, which gave ISI authority to confer degrees. So, the RTS had full freedom to innovate and to experiment. We could even set an exam where all the students failed. And these were very bright students. Professor R. R. Bahadur set an exam on Sophistication and asked me to mark it to find out if the students were more sophisticated than the teachers! All this freedom vanished the following year when the degree programs were started.

Professor P. C. Mahalanobis had a very broad vision of Statistics. He wanted to explore applications of statistics across all sciences. Rao was entrusted with the task of incorporating this vision in undergraduate courses when the ISI started degree courses in 1960.

It was an exciting place full of enthusiasm. World renowned scientists visited ISI. I remember that to accommodate the travel plans of a famous scientist who was passing through Calcutta, Dr. Rao arranged a seminar at 11:00 PM and it was well attended.

The place had world renowned statisticians such as Mahalanobis, Haldane, Bahadur and Basu, to name a few. Any teacher or research student would announce a seminar for researchers interested in the topic. Varadarajan gave a series of lectures on Metric Topology, and it was attended only by Varadhan and me.

Research students did not have any prescribed course work nor any comprehensive exams. Students studied a topic and had discussions with other students and faculty. They would even give a series of seminars on the topic they were studying. They would find their own research topic and choose a faculty member to supervise the thesis

Corresponding Author: Kirti R. Shah
Email: dakshakirti@gmail.com

Rao was a great leader. He gave research students complete freedom. Faculty and research students could work on any area of mathematics or statistics. Research units were created for scientists working in special areas of applied statistics.

Rao took teaching very seriously. It was regarded as a sacred duty. Research students also took teaching responsibilities. They were groomed by senior teachers when they assisted by playing various roles. They set term tests and marked these. When they became more experienced, they would even teach a course independently.

Typically, at the RTS the teacher did not take any notes to assist in teaching. I too developed this habit when I started to teach.

During my first year at the RTS, I was asked to assist Rao in his Design of Experiments class. I attended the classes. I thought that I knew the subject, but I found that I knew only the mathematics associated with the subject but not the subject. He was a great teacher. He would call a student and ask him to work through a problem on the blackboard and help the student solve it on the blackboard.

During my second year he asked me to teach a one term course on Linear Estimation in six weeks saying that it could be done. When I agreed with this, he asked me to go ahead and do this. He gave me ten minutes to prepare and then to go and teach. After six weeks, he took over and taught the full course. The students learnt the niceties of the subject and I learnt how to improve my teaching.

I was intimidated by the brilliance of my fellow research students though nobody did anything to make me feel this way. I mentioned this to Rao and he did his best to allay my fears and concerns. He said that many research students go through such lean phases and eventually come out of it. This went a long way in building my confidence and getting me started.

To further build up my confidence, he assigned me some administrative tasks. ISI had just started a summer institute for Statistics teachers at the various Universities to expose them to current research in many areas of Statistics. Initially, he asked me to help organize this and later gave me the designation of Program Director to run this. He also gave me more latitude in later years.

Rao also encouraged me to join the tea club at the RTS. I did not know that research students could join. I looked forward to the teatime in the afternoon. The atmosphere was very relaxed, and the research students got to know the faculty better.

Kolmogoroff's visit in 1962 was an exciting event. We celebrated his 60th birthday during his visit. He came to Bangalore for the next summer course. Since he did not wish to stay at a hotel, he occupied the main guest room in the ISI building at 4, Richmond Road. Rest of us, including Rao and Adhikary, and I, occupied small rooms there. Since Kolmogoroff did not speak English, Adhikary interpreted his lectures.

Staying with these dignitaries under the same roof was an exciting time. Kolmogoroff's comments on the social structure and on individuals he came across were very insightful.

Next Summer Course was held on the campus of Andhra University in Waltair. I paid a courtesy call to the Vice Chancellor who introduced me to another person present as

an outstanding statistician. Obviously he mistook me for Rao but it was embarrassing to correct him. Rao, Basu, Varadarajan and my family occupied individual cottages in Motel Ocean View. It was a very pleasant time. He got to know Daksha (my wife) much better and took fatherly interest in her. This continued through the years. My daughter Swati was sixteen months old and did not yet walk. Rao would encourage us to accompany him and others to join them in evening walk on the beach saying that he would carry Swati, which he initially did and later on passed on the burden to others. During this stay at Waltair, Basu became seriously ill. Fortunately, he recovered soon.

Because of these two summer courses, Rao and my family got to know each other better.

I started working with J. Roy at the RTS and this continued for a year and a half. We wrote a couple of research papers jointly. After that he took charge of the new computer unit and would have no time to see me. I carried on with my work, found my thesis problem and solved it. When I showed this to Rao, he was quite pleased.

Throughout his career in India, he combined research with heavy administrative duties. He could do this because he had the ability to move from one to the other seamlessly. I would knock at his door and enter when he was deeply involved in a research problem. My purpose was to seek his guidance on some administrative issues. He would listen to me and give his advice. Next minute he would be back studying the research problem where he left it.

Rao took interest in students who graduated from ISI even after they left the Institute. In many cases he did this all through their career and helped them get appropriate positions.

I left ISI in 1964 to join Michigan State University as a visiting faculty. I returned to ISI in 1967 and then left for Waterloo in 1968 where I stayed till 2005 even though I formally retired in 2002.

Rao set up a branch of ISI in New Delhi where Graduate degree courses were given. I visited there during one of the early years of this and stayed a full year. He nurtured this branch in the same way as he did nurture the RTS in Kolkata.

A few years later he retired from ISI and spent the rest of his career in the U.S.

Rao visited Michigan State University during my stay there and visited Waterloo several times when I was there. He and Mrs. Rao stayed with us during one of these visits. Additionally, I met him at various conferences. Hardly a year went by during which we did not meet. His interest in me and his guidance continued.

He was a professor at various universities in the U.S before his retirement at Buffalo where he stayed with his daughter Tejaswini. I visited him there several times. The painful expression on his face when he saw me with my left arm amputated, touched me deeply.

Tejaswini did a great job in preserving and organizing photos of Rao with several dignitaries throughout his illustrious career. She arranged these and other documents associated with his career like a museum.

She also did a superb job in arranging 100th birthday celebrations for Rao. Unfortu-

nately, because of covid this had to be done on Zoom. It was a memorable event.

In conclusion, I must say that I feel very lucky to have known him. He touched so many lives and mine was one of these.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 7–11
http://www.ssca.org.in/journal

Calyampudi Radhakrishna Rao - A Collaborator and a Statistician for the Ages

A. S. Hedayat[1] **and John Stufken**[2]
[1]*Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago*
[2]*Department of Statistics, George Mason University*

Received: 05 June 2024; Revised: 08 June 2024; Accepted: 09 June 2024

Abstract

C. R. Rao has long been one of the most visible names in statistics and beyond. Even now that he is no longer with us, his work will provide inspiration for researchers for many years to come. After a few broad observations about C. R. Rao's incredible impact on statistics and science, we focus on two areas of contribution by C. R. Rao that are perhaps not as broadly known: Orthogonal arrays and sampling plans for excluding contiguous units. These areas, which overlap with our own interests, demonstrate how C. R. Rao's search for better solutions to statistical problems led him towards defining and studying combinatorial structures.

Key words: Orthogonal arrays; Sampling plans excluding contiguous units

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Statistical science houses many distinguished researchers who have made highly influential contributions. If a hundred experienced statistical researchers were asked to make a list of 20 giants in the field, past or present, these lists would probably all look very impressive. But, most likely, there would only be few names that appeared on virtually every list. Those would be the names of the superstars, the visionaries, the trailblazers, the titans of the discipline. One of those very few names would be that of Calyampudi Radhakrishna Rao, better known to most as C. R. Rao (shortened to CR Rao from hereon).

With his many seminal contributions to statistics, CR Rao has been an inspiration to uncountable number of researchers, and will continue to be so for many more years to come. The staying power of his contributions, many of which were far ahead of their time, is astonishing and a testament to the depth and influence of the contributions. These contributions will live on and many will be taught to future generations of statisticians and scientists. CR Rao will undoubtedly be a statistician for the ages.

Corresponding Author: A. S. Hedayat
Email: hedayat@uic.edu

As others have noted, the impact of CR Rao's contributions were not limited to the field of statistics alone. This is, for example, crystal clear from the citation for his National Medal of Science awarded by US President George W. Bush. It reads:

"For his pioneering contributions to the foundations of statistical theory and multivariate statistical methodology, and their applications, enriching the physical, biological, mathematical, economic and engineering sciences."

It is therefore no surprise that journals like Science (Banks and Clarke, 2023) and Nature (Peddada and Khattree, 2023) are among the many that paid tribute to CR Rao shortly after his death. CR Rao's devotion to science, along with his perspective on statistics, is also captured in the quotation featured on the website of the CR Rao Advanced Institute of Mathematics, Statistics and Computer Science (AIMSCS) on the University of Hyderabad Campus:

"We study physics to solve problems in physics, chemistry to solve problems in chemistry, and botany to solve problems in botany. There are no statistical problems that we solve using statistics. We use statistics to provide a course of action with minimum risk in all areas of human endeavor with unavailable evidence."

For those interested to learn more about the life of CR Rao, about the person that he was, about his major contributions, and about his many awards, the AIMSCS web pages at `https://crraoaimscs.res.in/` and various publications (*e.g.*, Bera and Ghosh, 2021, and O'Grady, 2023) provide stories and perspectives that are more interesting and informed than anything we can offer on these aspects. Therefore, while we have benefited greatly and drawn enormous inspiration from many of CR Rao's contributions, we reminisce briefly on two areas of common research interest.

2. Orthogonal arrays

Just as for various other seminal contributions by CR Rao, he introduced orthogonal arrays (OAs) in the 1940s. In fact, as communicated by CR Rao in his Foreword in Hedayat, Sloane and Stufken (1999), he introduced a subclass of OAs in a chapter of his MS thesis in 1943. Rao (1946) reports on this subclass, which he called hypercubes of strength d. This was followed by Rao (1947, 1949), in which the complete class of OAs was studied even though the name orthogonal arrays had not yet been introduced.

Based on CR Rao's writing in the aforementioned Foreword, when he joined ISI in 1941 to study statistics, he was surprised to see a broad research interest in combinatorial structures, which was primarily fueled by the interests of RC Bose and his collaborators. With a strong background in mathematics, CR Rao was quickly able to become an important contributor in this arena, leading to his work on hypercubes of strength d (Rao, 1946) and a collaboration with K.R. Nair (Nair and Rao, 1948). He formulated the more general definition of OAs not until 1947, after having moved to study in Cambridge, UK.

Formally, an OA of strength t based on s symbols, N runs and k factors is an $N \times k$ array with entries from the set of s symbols so that for every $N \times t$ subarray every possible t-tuple based on the s symbols appears equally often as a row of the subarray. Given that there are s^t possible t-tuples, this common number must be N/s^t, which is also known as the

index of the OA. Such an array is often denoted by $\mathrm{OA}(N, k, s, t)$, while the index is often written as λ.

CR Rao introduced OAs because of their properties for use in fractional factorial experiments. This interest is visible in Nair and Rao (1948), as also in Rao (1947). The bounds for the number of runs in an OA given by Rao (1947) are easily understood if one understands the relationship between the strength of an OA and models for which all parameters are estimable when using the runs of an OA to form a fractional factorial. While statistical properties motivated CR Rao, he was also interested in combinatorial aspects of OAs, as is clearly shown by Rao (1949). Various methods of construction are discussed there, including a simplified construction of OAs that already appeared in Rao (1946). A similar idea appeared shortly afterwards in Hamming (1950) for the construction of error-correcting codes, which led Hedayat, Sloane, and Stufken (1999) to refer to these arrays as Rao-Hamming OAs.

CR Rao returned to his interest in OAs on several later occasions, such as in Rao (1961, 1973).

After so many years since their introduction, OAs remain an active area of research, both in statistics and mathematics. They also continue to be used frequently in factorial experiments.

3. Sampling plans excluding contiguous units

When sampling from a finite population $U = \{1, \ldots, N\}$, a fixed-size n sampling plan, where $n < N$, can be presented as $\{(s_\ell, p_\ell), \ell = 1, 2, ..., m\}$, where the s_ℓ's are distinct subsets of size n from U, p_ℓ is the probability that s_ℓ is the sample outcome, and m is the support size of the sampling plan. The first-order inclusion probability π_i for unit $i \in U$ is defined as the probability that unit i is in the selected sample, and can be computed as

$$\pi_i = \sum_{s_\ell \ni i} p_\ell, \quad i = 1, \ldots, N.$$

Similarly, for two distinct units i and j, the second-order inclusion probability π_{ij} is the probability that both the units are in the selected sample. Thus,

$$\pi_{ij} = \sum_{s_\ell \ni i,j} p_\ell, \quad i, j = 1, \ldots, N, i \neq j.$$

A fixed-size n sampling plan for which $\pi_i = n/N$ and $\pi_{ij} = n(n-1)/(N(N-1))$ has, with respect to the first- and second-order inclusion probabilities, the same characteristics as simple random sampling (SRS). The plan is identical to SRS if $m = \binom{N}{n}$, so that the s_ℓ's consist of all subsets of U of size n, and $p_\ell = 1/m$ for every ℓ.

A common problem would assume that unit $i \in U$ has value Y_i for a certain characteristic (*e.g.*, income or years of secondary education). A study might be interested in the population total $T = \sum_{i=1}^{N} Y_i$ or the population mean T/N. For a large population, observing all Y_i's (a census) might be too time consuming or expensive. Moreover, it would be unnecessary since accurate results can be obtained from a random sample based on a

sampling plan with relatively small sample size n. A design-based unbiased estimator of T based on a sample s_ℓ is in that case given by the Horvitz-Thompson estimator

$$\hat{T} = \sum_{i \in s_\ell} \frac{Y_i}{\pi_i}.$$

An unbiased estimator for the variance of $\hat{T}$ exists if and only if $\pi_{ij} > 0$ for all units $i, j \in U, i \neq j$ (*cf.* Hedayat and Sinha, 1991).

If the units can be thought of as naturally being placed on a straight line or circle to depict their proximity (*e.g.*, housing units on a street or street block), then it is conceivable for some characteristics (*e.g.*, household income) that neighboring units provide similar information. One might get a better estimate for the population total by avoiding the selection of contiguous units. This idea was explored in Hedayat, Rao and Stufken (1988a, 1988b). Considering the population units to be ordered on a circle, so that each unit has two contiguous units, these authors define a fixed-size n sampling plan to be a balanced sampling plan without contiguous units if all first-order inclusion probabilities are equal (this common value must be n/N), the second-order inclusion probabilities for contiguous units are 0, and all other second-order inclusion probabilities are equal (this common value must be $n(n-1)/(N(N-3))$. They show that, in terms of the variance of the Horvitz-Thompson estimator, balanced sampling plans without contiguous units are more efficient than SRS if the first-order serial correlation of the Y_i's exceeds $-1/(N-1)$ (which it will if there is any validity to the premise that contiguous units have similar Y_i's).

One may also wonder when such balanced sampling plans without contiguous units exist. Hedayat, Rao and Stufken (1988a) show by construction that such plans exist for every value of $N \geq 3n$ when $n = 3$ or 4.

In later years, there have been various extensions of these initial results. This includes, for example, requiring second-order inclusion probabilities to be 0 not only for immediate neighbors (*cf.* Stufken, 1993), considering population units to be ordered in a 2-dimensional layout (*cf.* Wright, 2008), and additional existence results, including in the combinatorial literature (*cf.* Guo, Wang, and Feng, 2022). Also, alternative methods have been developed for spatial populations (*cf.* Deville and Tillé, 1998).

4. In conclusion

While C. R. Rao's passing is an enormous loss for the scientific community, his widespread influential contributions provide assurance that his work will be with us for many years to come. In fact, we have no doubt that his contributions will remain an inspiration for budding statistical researchers. In that way, CR Rao is truly a statistician for the ages.

References

Banks, D. and Clarke, J. L. (2023). C. R. Rao (1920–2023). *Science*, **382**, 771–771.

Bera, A. and Ghosh, P. (2021). Glimpses from the life and work of Dr. C. R. Rao: A living legend in statistics. *A Tribute to the Legend of Professor CR Rao: The Centenary Volume*, **1**, 49–65.

Deville, J.-C. and Tille, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, **85**, 89–101.

Guo, C., Wang, X., and Feng, T. (2022). Cyclic balanced sampling plans excluding contiguous units with block size four. *Discrete Mathematics*, **345**, 112899.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell System Technical Journal*, **29**, 147–160.

Hedayat, A., Rao, C., and Stufken, J. (1988a). Designs in survey sampling avoiding contiguous units. *Handbook of Statistics*, **6**, 575–583.

Hedayat, A., Rao, C., and Stufken, J. (1988b). Sampling plans excluding contiguous units. *Journal of Statistical Planning and Inference*, **19**, 159–170.

Hedayat, A. and Sinha, B. K. (1991). *Design and Inference in Finite Population Sampling*. John Wiley & Sons.

Hedayat, A. S., Sloane, N. J. A., and Stufken, J. (2012). *Orthogonal Arrays: Theory and Applications*. Springer Science & Business Media.

Nair, K. and Rao, C. (1948). Confounded designs for asymmetrical factorial experiments. *Journal of the Royal Statistical Society. Series B (Methodological)*, **10**, 109–131.

O'Grady, C. (2023). "A life to enlighten us": Remembering C. R. Rao, 1920–2023. *Significance*, **20**, 8–12.

Peddada, S. D. and Khattree, R. (2023). C. R. Rao, statistician who transformed data analytics (1920–2023). *Nature*, **622**, 691–691.

Rao, C. R. (1946). Hypercubes of strength'd' leading to confounded designs in factorial experiments. *Bulletin of the Calcutta Mathematical Society*, **38**, 67–78.

Rao, C. R. (1947). Factorial experiments derivable from combinatorial arrangements of arrays. *Supplement to the Journal of the Royal Statistical Society*, **9**, 128–139.

Rao, C. R. (1949). On a class of arrangements. *Proceedings of the Edinburgh Mathematical Society*, **8**, 119–125.

Rao, C. R. (1961). Combinatorial arrangements analogous to orthogonal arrays. *Sankhyā: The Indian Journal of Statistics, Series A*, **23**, 283–286.

Rao, C. R. (1973). Some combinatorial problems of arrays and applications to design of experiments. In *A Survey of Combinatorial Theory*, pages 349–359. Elsevier.

Stufken, J. (1993). Combinatorial and statistical aspects of sampling plans to avoid the selection of adjacent units. *Journal of Combinatorics, Information and System Sciences*, **18**, 81–92.

Wright, J. H. (2008). Two-dimensional balanced sampling plans excluding adjacent units. *Journal of Statistical Planning and Inference*, **138**, 145–153.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 13–17
http://www.ssca.org.in/journal

The Importance of C. R. Rao to the Graduate Student

Ezra Becker and Dibyen Majumdar
Department of Mathematics, Statistics, and Computer Science
University of Illinois Chicago

Received: 06 June 2024; Revised: 08 June 2024; Accepted: 09 June 2024

Abstract

This article discusses many of the contributions of C. R. Rao to the education of the statistics graduate student. We submit that Rao's foundational results are not only critical elements of the graduate statistician's analytical toolkit, but also shape the graduate statistician's philosophical approach to statistical research and real-world problem solving.

Key words: Rao-Cramer lower bound; Rao-Blackwell theorem; Linear statistical inference; Graduate student education; C. R. Rao.

AMS Subject Classifications: 01A70, 62F25, 62J05, 97D20

1. Introduction

One might approach the task of honoring the memory of C. R. Rao by selecting one of his myriad contributions to modern statistics and expanding upon why it is so important. We imagine many contributors to this edition of Statistics and Applications might do just that, demonstrating the breathtaking scope and diversity of Rao's foundational results with subtle and nuanced derivations that are sure to be appreciated by the seasoned professors and other established scholars in this journal's readership. However, we believe that one does not need a doctorate and a strong publication history to appreciate what Rao did - and continues to do - for statistics. Indeed, our focus in this article is the immense importance of Rao to beginning graduate students, those just starting out on their journey toward a master's degree or doctorate.

2. In the classroom

STAT 401 is the introductory master's-level probability course at the University of Illinois Chicago, and is taken by most of the first-semester graduate students in the statistics program. We would guess that it is more-or-less similar in content and approach to the beginning master's level probability course at universities across the globe. The course includes an introduction to five major theorems/concepts: the Kolmogorov probability axioms, the Rao-Cramér lower bound, the Central Limit Theorem, Sufficiency and Completeness of

Corresponding Author: Ezra Becker
Email: ebecke1@uic.edu

sample statistics (including Basu's Theorem), and the Rao-Blackwell Theorem. Of course there are several other ideas introduced, such as various discrete and continuous probability distributions, parameter estimation, confidence intervals, and hypothesis testing. But of the five foundational concepts of the course, Rao developed two of them. That fact alone should convince any skeptic how important Rao is to the budding statistician. A closer look at these two concepts yields even more insight into why they are so foundational for beginning graduate students.

Consider the Rao-Cramér lower bound. This elegant relation provides a lower bound for the variance of any unbiased estimator $\widehat{\theta}$ of a parameter of interest θ to the Fisher information for that parameter $\mathcal{I}(\theta)$. In the one-dimensional case, it may be expressed as

$$Var(\widehat{\theta}) \geq \mathcal{I}^{-1}(\theta). \tag{1}$$

In the multidimensional case, where one considers a vector of parameters $\boldsymbol{\theta}$ and a vector of estimable functions of those parameters $\boldsymbol{f(\theta)}$, where the covariance matrix of the estimator $\widehat{\boldsymbol{f(\theta)}}$ is denoted by $\boldsymbol{\Sigma}(\widehat{\boldsymbol{f(\theta)}})$ and the Fisher information matrix is denoted by $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$, it can be shown via a first-order Taylor expansion that

$$\boldsymbol{\Sigma}(\widehat{\boldsymbol{f(\theta)}}) \geq \left(\frac{\partial \boldsymbol{f(\theta)}}{\partial \boldsymbol{\theta}^T}\right) \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}) \left(\frac{\partial \boldsymbol{f(\theta)}}{\partial \boldsymbol{\theta}^T}\right)^T, \tag{2}$$

which approaches equality asymptotically for any maximum likelihood estimator $\widehat{\boldsymbol{f(\theta)}}$ under certain regularity conditions. Thus one may loosely say that "maximizing" the information matrix for $\boldsymbol{\theta}$ is equivalent to "minimizing" the variance of $\widehat{\boldsymbol{f(\theta)}}$. In other words, this lower bound says that one's ability to minimize the uncertainty of a parameter estimate is limited in a simple and direct way by the information in the data available regarding that parameter. And note, this lower bound does not come into consideration directly when one is learning about confidence intervals, or hypothesis testing. Why, then, is it so critically important to the beginning graduate student? The answer is that it teaches the beginning graduate student a critical fact about both the theory and the practice of statistics, namely that we are limited by our data in meaningful ways. But it also teaches the student that important and often fruitful avenues of inquiry might be found through attempts to maximize how that available information is utilized.

The Rao-Blackwell theorem provides an elegant application of that concept. The theorem states that an unbiased parameter estimator conditioned on a sufficient statistic for the parameter under consideration will have less uncertainty than an unconditional estimator. And while sufficiency may not be an entirely glamorous concept in this age of computational methods and big data, the lesson of seeking sound theoretical justification for a proposed "better" estimator or classification method certainly endures. In this sense, Rao's work provides not just a technical foundation for how beginning graduate students engage in the practice of statistics, but also a philosophical starting point for what can be achieved with statistics.

Let us next touch upon the course on linear statistical inference that we assume every graduate statistics program offers. At UIC, this course is listed as STAT 521 at the Ph.D. level. Rao's classic *Linear Statistical Inference and its Applications* (Rao, 1965), and in

particular the 1973 edition, is considered by many to be the bible of linear regression. It is the textbook that many programs use to teach the course, either as the primary text or as an important reference text. In this book Rao builds a coherent foundation for how to think about and approach statistical modeling. But what's also important is that this course - and by extension, Rao's framework for statistical inference - is the anchor point for a whole field of statistical modeling; for example, it is usually (although not always) the case that nonlinear regression is introduced as an extension of linear regression concepts. And, nonparametric models are often not introduced until linear statistical inference has been taught, again so that nonparametric approaches may be explored in contrast to parametric linear methods. In short, Rao's presentation of linear statistical inference is often the foundation for how statistics students think about and approach modeling; how they seek to describe, explain, and anticipate the world around them in mathematical terms; and how they make sense of complex dynamics.

3. Rao's students and their descendants

There is another measure available that speaks to Rao's influence on the emerging statistics graduate student. The Mathematics Genealogy Project ("MGP"), established by Harry Coonce and currently coordinated by the Department of Mathematics (1996), aspires to inventory as many mathematics Ph.D. holders worldwide (or those with equivalent degrees), and as far back historically, as possible. As of May 15, 2024, the MGP held data on 308,994 mathematicians, with the subset identified as statisticians containing 16,596 degree holders. The MGP lists Rao as having 52 direct Ph.D. students, with total descendants - *i.e.*, students of his direct students, and their students, *etc.*–of 822. That is, in a sense Rao is the direct patriarch of almost 5% of all the Ph.D. statisticians listed in the Project. And while the MGP is not comprehensive in its coverage, its coordinators are confident that it represents a comfortable majority of Ph.D. holders worldwide, certainly for degrees granted more than five years ago. The point is that Rao had a direct influence on the development of many, many graduate students, and a one- or two-step removed influence on many more.

It is also worth pointing out that many of Rao's students were themselves hugely influential in the development of diverse fields of statistics, such as Debabrata Basu (author of the famous Basu's Theorem referenced above), T. E. S. Raghavan (game theory), S. R. S. Varadhan (probability theory and large deviations), and others. Thus not only did Rao help forge the statistical perspective of many students directly, in particular he did so for a set of extraordinary students who in turn helped shape statistics – and how graduate students learn and approach the field – in important ways.

4. A walk down memory lane

The Indian Statistical Institute (ISI) was established by P. C. Mahalanobis. C. R. Rao joined the Institute, and in 1972 succeeded Mahalanobis as Director. Under Mahalanobis and Rao's leadership, ISI became one of the earliest, and most eminent, statistics institutes in the world, achieving great advances in both theory and applications. The Institute created an atmosphere of excellence in research, where the mind could roam and attain great heights. There was no boundary – many areas of research, even those not really a part of statistics, were encouraged to blossom. For example, as mentioned above T. E. S. Raghavan got his Ph.D. under the supervision of C. R. Rao in game theory! Graduate students were inspired to

learn and discover, but the unique atmosphere created by Rao put the onus on the student to advance. There were faculty and postdocs who helped excite and advance minds, but learning and problem solving was entirely the responsibility of the student. In the early days, there were essentially no classes for Ph.D. students.

When Dibyen Majumdar (DM) joined ISI Kolkata as a Ph.D. student, he got an opportunity to meet Sujit Kumar Mitra (SKM), a statistician of eminence and a strong collaborator of C. R. Rao. DM wanted to work on linear models. He was told by SKM to read several chapters of Rao (1965) *Linear Statistical Inference and Its Applications*, read all chapters of Rao and Mitra (1971) *Generalized Inverse of Matrices and its Applications*, solve all the exercise problems in those texts and related problems, and to move to ISI New Delhi (Rao's and SKM's campus at the time) only if he succeeded in this learning project. There were no classes for DM, but ISI Kolkata had a galaxy of star researchers who helped DM enormously, even though they were not solving the problems themselves. One person who helped DM was the postdoctoral fellow P. Bhimasankaram, a former student of SKM who had great knowledge of g-inverses.

Here is one example of an exercise that DM had to solve at the start of his graduate student life at ISI Kolkata, for which Bhimasankaram helped him derive the proof.

Exercise 1: Let A $(n \times p)$ and B $(m \times p)$ be two (real) matrices such that

$$\mathcal{M}(A') \cap \mathcal{M}(B') = \{0\}. \tag{3}$$

Then

$$\begin{aligned} A'A(A'A + B'B)^- A'A &= A'A & (4)\\ B'B(A'A + B'B)^- B'B &= B'B & (5)\\ A'A(A'A + B'B)^- B'B &= 0. & (6) \end{aligned}$$

Proof: Since

$$\mathcal{M}(A'A) \subset \mathcal{M}(A'A + B'B), \tag{7}$$

$$\begin{aligned} A'A(A'A + B'B)^-(A'A + B'B) &= A'A,\\ \text{i.e., } A'A(A'A + B'B)^- A'A - A'A &= A'A(A'A + B'B)^- B'B. \end{aligned} \tag{8}$$

It follows from (7) and $\mathcal{M}(B'B) \subset \mathcal{M}(A'A + B'B)$ that we can assume with no loss of generality that $(A'A + B'B)^-$ is symmetric. Taking transposes in (8) we get

$$A'A(A'A + B'B)^- A'A - A'A = B'B(A'A + B'B)^- A'A.$$

In this expression,

$$\begin{aligned} \mathcal{M}(A'A(A'A + B'B)^- A'A - A'A) &\subset \mathcal{M}(A'A) = \mathcal{M}(A'),\\ \mathcal{M}(B'B(A'A + B'B)^- A'A) &\subset \mathcal{M}(B'B) = \mathcal{M}(B'). \end{aligned}$$

Hence condition (3) implies

$$A'A(A'A + B'B)^- A'A - A'A = B'B(A'A + B'B)^- A'A = 0.$$

This establishes (4) and (6). Clearly, (5) will have a similar derivation. □

These results have two lessons for a graduate student. The first, demonstrated by equations (4) and (5), is the richness of the family of generalized inverses that was introduced by Rao (1967), going far beyond the Moore-Penrose inverse. In other words, $A'A$ has a wide variety of generalized inverses with potentially desirable properties. The second, demonstrated by equation (3), starts with the fact that $\mathcal{M}(A')$ and $\mathcal{M}(B')$ are disjoint but not necessarily orthogonal under the inner product $< x, y > = y'x$. However, it is possible to find a positive definite matrix, say Q, that is a generalized inverse of $(A'A + B'B)$; $Q = (A'A+B'B)^-$. Then $\mathcal{M}(A')$ and $\mathcal{M}(B')$ are orthogonal under the inner product $< x, y >$ $= y'Qx$, *i.e.*, $AQB' = 0$.

5. Conclusion

Many beginning statistics graduate students have trouble wrapping their arms around Rao's results the first time they are exposed to them (or even the second or third time!). And, not every beginning graduate student has the wherewithal to understand the more nuanced ramifications of Rao's results, let alone the ability to prove those results. Deeper insights and the ability to "connect the dots" of Rao's results usually come with time and experience. But what all aspiring graduate statisticians have, whether they realize it at the time or not, is the influence of Rao's work laying the foundation for how they perceive the nature and function of statistics. It is in this regard that a hugely important contribution of Rao endures, and will continue to do so as long as graduate students study classical statistics.

Acknowledgements

We are grateful to the editors and the Journal for the opportunity to contribute these thoughts on the impact of C. R. Rao.

References

Department of Mathematics, N. D. S. U. (1996). Mathematics genealogy project. `https://genealogy.math.ndsu.nodak.edu/index.php`. Accessed May 15, 2024.

Rao, C. R. (1965). *Linear Statistical Inference and its Applications*. John Wiley & Sons.

Rao, C. R. (1967). Calculus of generalized inverses of matrices Part I. General theory. *Sankhya-Series A*, **29**, 317–342.

Rao, C. R. and Mitra, S. K. (1971). *Generalized Inverse of Matrices and its Applications*. John Wiley and Sons, New York.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 19–27
http://www.ssca.org.in/journal

Employing Rao Theorems in Mixed Effects Growth Curves

Samaradasa Weerahandi
X-Techniques, 23 Chestnut Street, Edison, NJ 08817

Received: 09 April 2024; Revised: 10 June 2024; Accepted: 12 June 2024

Abstract

This article is motivated by the author's pleasant experience when late Professor Rao helped validate an assertion made in Weerahandi and Berger (1999). Additional implications of Rao (1967) in Growth Curve Models under compound symmetric covariance structure are also presented. The inferences are made using the generalized p-value approach. Desired further research are also discussed.

Key words: Rao's covariance structure; Generalized inference; Compound symmetry; Generalized Hotelling T^2; Parametric bootstrap.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

The author of this article had the pleasure of having a short chat with Professor Rao many years ago on the sidelines at few conferences in New Jersey and in Europe. At that time I had no idea that this world class researcher would care about a favor requested by a mediocre researcher like me. Now, to briefly describe the experience, first consider the following problem in the context of Mixed Effects models in Growth Curves, which is a particular problem involving repeated measures.

Consider the linear mixed effects growth curve model based on observations from n subjects

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \qquad \text{for } i = 1, \ldots, n, \tag{1}$$

where $\mathbf{y}_i$ is the $T \times 1$ vector of responses from ith subject, $\mathbf{X}_i$ and $\mathbf{Z}_i$ are known design matrices of dimension $T \times p$ and $T \times q$, respectively, $\boldsymbol{\beta}_i$ is a vector of fixed effects, and the random effects, $\mathbf{b}_i$ and the error vector $\boldsymbol{\epsilon}$, jointly and independently distributed as

$$\mathbf{b}_i \sim \mathbf{N}_q(\mathbf{0}, \boldsymbol{\Psi}) \tag{2}$$

and

$$\boldsymbol{\epsilon}_i \sim \mathbf{N}_T(\mathbf{0}, \boldsymbol{\Lambda}_i) \ ,$$

Corresponding Author: Samaradasa Weerahandi
Email: Weerahandi@aol.com

where $\mathbf{\Lambda}_i$ is a within-subject covariance matrix of dimension $T \times T$ and $\mathbf{\Psi}$ is usually a between-subject covariance matrix of dimension $q \times q$. The model can also be rewritten in the form of a structured covariance matrix as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta_i} + \mathbf{e}_i, \tag{3}$$

where

$$\mathbf{e}_i \sim \mathbf{N}_T(\mathbf{0}, \mathbf{\Lambda}_i + \mathbf{Z}_i\mathbf{\Psi}\mathbf{Z}_i').$$

When the covariance matrix of a growth curves model has a special structure, classical approaches do not provide exact solutions to inference problems even for a situation of a single growth curve. In this article we concentrate on the case of one group of subjects when the covariances follow compound symmetric structure, which is also known as the intraclass correlation structure.

2. Case of intraclass correlation structure

Weerahandi and Berger (1999) considered the particular case of one group of subjects when the covariance matrix is compound symmetric. In this section, we will concentrate on the distributional results providing details of Professor Rao's contribution. To do so, consider the simple growth curve model

$$Y_{it} = \alpha_i + \mathbf{X}_t'\boldsymbol{\beta} + \epsilon_{it}, \tag{4}$$

where $\mathbf{X}_t'$ is the $p \times 1$ design vector, $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters common for all subjects, α_i is a random effect due to subjects, and ϵ_{it} is the error term. In particular, when one deals with polynomial growth curves, the design matrix is of the form

$$\mathbf{X}_t' = \left(1, t, , t^2, \ldots, t^{p-1}\right)$$

If random effects are all normally distributed, we get

$$\alpha_i \sim N(0, \sigma_\alpha^2) \tag{5}$$

and

$$\epsilon_{it} \sim N(0, \sigma_e^2),$$

where σ_α^2 and σ_e^2 are variance components of the model. Moreover, α_i and all ϵ_{it} terms are assumed to be independently distributed. Collecting data from ith subject, the model for the $T \times 1$ vector of responses, $\mathbf{Y}_i$, can be written in vector form in terms of the $T \times p$ design matrix $\mathbf{X}$ as

$$\mathbf{Y}_i = \alpha_i\mathbf{1}_T + \mathbf{X}\boldsymbol{\beta} + \epsilon_i \ , \tag{6}$$

where $\mathbf{1}_T$ is a $T \times 1$ vector of 1s. It is easily seen from (5) that $Var(Y_{it}) = \sigma_\alpha^2 + \sigma_e^2$ and that $Cov(Y_{it}, Y_{it'}) = \sigma_\alpha^2$, and hence

$$\mathbf{Y}_i \sim N_T(\mathbf{X}\boldsymbol{\beta}, \mathbf{\Sigma}) \text{ with the covariance matrix } \mathbf{\Sigma} = \sigma_\alpha^2\mathbf{1}_T\mathbf{1}_T^{'} + \sigma_e^2\mathbf{I}_T \tag{7}$$

This means that the covariance matrix of the observations vector has the intraclass structure. The model (6) is a special case of model (1) with

$$\mathbf{Z}_i\boldsymbol{\Psi}_i\mathbf{Z}_i' = \sigma_\alpha^2\mathbf{1}_T\mathbf{1}_T^{'} \qquad \text{and} \qquad \boldsymbol{\Lambda}_i = \sigma_e^2\mathbf{I}_T$$

Being a matrix with intraclass structure, the inverse of $\boldsymbol{\Sigma}$ is also an intraclass matrix. More specifically,

$$\boldsymbol{\Sigma}^{-1} = \sigma_e^{-2}\left[\mathbf{I}_T - \frac{\sigma_\alpha^2}{\sigma_e^2 + T\sigma_\alpha^2}\mathbf{1}_T\mathbf{1}_T^{'}\right]. \tag{8}$$

The problem is to make inferences about the unknown parameters β and the variance components σ_α^2 and σ_e^2. It follows from (7) that the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ is the weighted least-squares estimate (WLSE)

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\overline{\mathbf{Y}}, \tag{9}$$

which is also known as the generalized least squares estimate (GLSE) of β, where $\overline{\mathbf{Y}}=\sum \mathbf{Y}_i/N$ is a $T \times 1$ vector, where N is the number of subjects, who were observed over time.

Rao (1967) and Rao (1973) showed that, if the columns of $\boldsymbol{\Sigma}\mathbf{X}$ is a subspace of the vector space spanned by the columns of $\mathbf{X}$, then the GLSE reduces to the ordinary least-squares estimate (OLSE), regardless of what $\boldsymbol{\Sigma}$ is. When $\boldsymbol{\Sigma}$ is as in (7) and the first column of $\mathbf{X}$ is a vector of 1's (i.e., an intercept term is present in the growth curve model), this condition is satisfied and consequently (9) reduces to the OLSE, A covariance matrix satisfying this condition is referred to as Rao's covariance structure; see also Ghosh and Gokhale (1987). Then, the point estimator of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\overline{\mathbf{Y}} \tag{10}$$

3. Controversy about GLSE reducing to OLSE

When the author of this article submitted the manuscript underlying Weerahandi and Berger (1999) for publication in Biometrics, a referee disputed the validity of the distributional results outlined in the above section. The referee thought that Rao (1967) results do not imply that GLSE reduces to the OLSE under the compound symmetric covariances structure. When I referred to McElroy (1967) the referee did not concede and recommended rejection of manuscript. To overcome this dilemma, then I provided my own algebraic derivation, which is simpler to understand, but similar to McElroy (1967), even then the editor did not reconsider the manuscript.

Then, in desperation, I wrote to Professor Rao seeking help. To my surprise, in two weeks I received a letter in regular mail from Professor Rao stating something like "Weerahandi, not only your assertion is correct, but also it is valid under milder conditions and for greater class of covariance structures". When I sent the letter to the editor, she conceded and accepted the manuscript with some minor modifications. So, I am extremely grateful to late professor Rao for his support getting the article published.

4. Generalized inference

Before we review Weerahandi and Berger (1999) results of relevance, let us briefly describe the generalized tests introduced by Weerahandi (1987) and Tsui and Weerahandi (1989. In one-liner, generalized tests are based on random quantities known as Generalized Test Variables (GTV) that are functions of (i) observable random quantities, (ii) their observed values, and (iii) unknown parameters, defined in such a way that

(a). the distribution of GTV is free of unknown parameters, and

(b). at the observed sample points, the observed value of GTV will contain no unknown parameters under the null hypothesis. If a GTV is also monotonic for deviations from the null hypothesis, then it can be used to define extreme regions, on which generalized p-values can be based.

Often GTVs can be derived based on what is known as Generalized Pivotal Quantities (cf. Weerahandi (1993)), abbreviated as GPQs. To be specific, a GPQ of a parameter is also a function of (i) observable random variables, (ii) their observed values, and (iii) unknown parameters, defined in such a way that

(a). its distribution does not depend on nuisance parameters, and

(b). at the observed sample points, its observed value becomes equal to the parameter of interest.

Now getting back to the current problem, although GLSE reduce to OLSE under the compound symmetric covariance structures, even for models involving just one group of subjects, classical approach to inference fails to provide classical confidence bounds or tests of hypothesis concerning even a single parameter of the model. This is because the distribution of OLSE involves nuisance parameters. To be specific, despite the fact that GLSE is the same as the OLSE, the distribution $\widehat{\boldsymbol{\beta}}$ given by

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}/N) \tag{11}$$

involves the unknown variance components.

Nevertheless, Weerahandi and Berger (1999) demonstrated, how generalized tests can be constructed for testing hypotheses concerning one or more component of $\boldsymbol{\beta}$. To be specific. they considered the hypotheses on individual components of the form

$$H_0 : \beta_j \leq \beta_0$$

and provided a generalized test based on the independent sufficient statistics

$$\begin{aligned} \hat{\beta}_j &\sim N(\beta_j, \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}_{jj}/N) \quad j = 1, \cdots p \\ S_e^2 &= \sum_i \sum_t (Y_{it} - \mathbf{X}_t'\widehat{\beta} - (\overline{Y_i} - \overline{Y}))^2, \\ S_w^2 &= T \sum_i (\bar{Y}_i - \bar{Y})^2 \end{aligned} \tag{12}$$

due to Lehman (1986), where $\bar{Y}_i$ is the sample mean for i^{th} subject, $\bar{Y}$ is the sample mean of all the subjects,and $(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}_{jj}$ is the jjth element of the covariance matrix $(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}$.

Let

$$S_j(\sigma_e^2, \sigma_w^2) = \frac{1}{\sqrt{N}}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})_{jj}^{1/2} \text{ and } \sigma_w^2 = \sigma_e^2 + T\sigma_\alpha^2. \tag{13}$$

The sums of squares appearing in (11) are distributed as

$$\begin{aligned} W_1 &= \frac{S_e^2}{\sigma_e^2} \sim \chi^2_{\nu_1}, \qquad \text{where } \nu_1 = N(T-1) - p + 1, \text{ and} \\ W_2 &= \frac{S_w^2}{\sigma_w^2} \sim \chi^2_{\nu_2}, \qquad \text{where } \nu_2 = N - 1. \end{aligned} \tag{14}$$

Then by taking the generalized approach to inference, Weerahandi and Berger (1999) showed that

$$p = \Pr\left(\frac{Z}{\sqrt{W/\nu}} \geq \sqrt{\nu}\frac{(b_0 - \beta_j)}{S_j(\frac{s_e^2}{B}, \frac{s_w^2}{1-B})}\right), \tag{15}$$

is a generalized p-value appropriate for testing the above null hypothesis, where

$$B \sim Beta(\nu_1/2, \nu_2/2) \text{ and } W = W_1 + W_2 \sim \chi^2_\nu : \quad \nu = \nu_1 + \nu_2 = NT - p$$

Although, they did not address the problem of interval estimation, one can construct generalized confidence Intervals on any single component using the Generalized Confidence interval approach suggested by Weerahandi (1993). Using the notion of Generalized Pivotal Quantity, one can also construct generalized confidence ellipsoids for few components of interest, as we demonstrate in the next section.

Taking that approach one can tackle problems involving more complicated compound symmetric covariance structures and number of groups of subjects, in a one-way layout setting as Chi and Weerahandi (1998) did. The Weerahandi and Berger (1999) results itself can be extended to make inferences on a number of regression coefficients, as we further discuss in the following sections.

5. Generalized inference on a vector of coefficients

Weerahandi and Berger (1999) results can be extended way beyond the problem they considered. Confining to the distributional results concerning Rao's covariance structure, consider the problem of constructing confidence regions on a subset of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}_j$, a sub vector of $\boldsymbol{\beta}$, or $\boldsymbol{\beta}$ itself. The generalized inference on $\boldsymbol{\beta}_j$ can be constructed based on the foregoing distributional results along with the following:

$$\widehat{\boldsymbol{\beta}_j} \sim N(\boldsymbol{\beta}_j, (\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})_{jj}^{-1}/N), \tag{16}$$

where $(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})_{jj}$ is the jj^{th} subset of $(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})$ corresponding to the $\boldsymbol{\beta}$ coefficients of interest. Assuming positive definite covariance matrices, we can standardize (16) as

$$\mathbf{Z} = \sqrt{N}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})_{jj}^{1/2})\left(\widehat{\boldsymbol{\beta}_j} - \boldsymbol{\beta}_j\right) \sim N(\mathbf{0}, \mathbf{I}), \tag{17}$$

Now it is evident that, if the covariance matrix $\mathbf{\Sigma}$ were known, testing of hypotheses concerning $\boldsymbol{\beta}_j$ or confidence ellipsoids can be constructed using the χ^2 statistic,

$$\begin{aligned}\tilde{H} &= N\left(\widehat{\boldsymbol{\beta}_j}-\boldsymbol{\beta}_j\right)'\left(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})_{jj}\right)\left(\widehat{\boldsymbol{\beta}_j}-\boldsymbol{\beta}_j\right)\\ &= \left(\widehat{\boldsymbol{\beta}_j}-\boldsymbol{\beta}_j\right)'(S_j^2(\sigma_e^2,\sigma_w^2)_{jj})\left(\widehat{\boldsymbol{\beta}_j}-\boldsymbol{\beta}_j\right)\sim\chi^2_{p_j}\end{aligned} \tag{18}$$

where p_j is the dimension of $\boldsymbol{\beta}_j$

5.1. Hypothesis testing

Typically the covariance matrix is unknown, and in that case, the generalized inferences can be performed based on the *generalized Hotelling* T^2 statistic

$$H=\left(\widehat{\boldsymbol{\beta}_j}-\boldsymbol{\beta}_j\right)'(S_j^2(s_e^2/W1,s_w^2/W2)_{jj})\left(\widehat{\boldsymbol{\beta}_j}-\boldsymbol{\beta}_j\right), \tag{19}$$

because

$$\frac{s_e^2}{W_1}\text{ is a GPQ for }\sigma_e^2\text{ and }\frac{s_w^2}{W_2}\text{ is a GPQ for }\sigma_w^2. \tag{20}$$

First, to perform hypotheses testing concerning sub-vectors of coefficients $\boldsymbol{\beta}_j$, consider null hypotheses of the form

$$H_0:\boldsymbol{\beta}_j=\boldsymbol{\beta}_0,$$

where $\boldsymbol{\beta}_0$ is a certain hypothesized value. Under the null hypothesis, we get from (17)

$$\mathbf{Z}\sqrt{N}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})_{jj}^{1/2})\left(\widehat{\boldsymbol{\beta}_j}-\boldsymbol{\beta}_0\right)=\mathbf{Z}S_j(\sigma_e^2,\sigma_w^2)_{jj},\text{ where }\mathbf{Z}\sim N(\mathbf{0},\mathbf{I}). \tag{21}$$

By taking advantage of the two results (21) and (18), we can define a potential GTV, a *generalized Hotelling* T^2 as

$$\begin{aligned}T^2 &= \left(\widehat{\boldsymbol{\beta}_j}-\boldsymbol{\beta}_0\right)'S_j(\sigma_e^2,\sigma_w^2)_{jj}\left((S_j^2(\frac{s_e^2}{W1},\frac{s_w^2}{W2})_{jj}\right)^{-1}S_j(\sigma_e^2,\sigma_w^2)_{jj}\left(\widehat{\boldsymbol{\beta}_j}-\boldsymbol{\beta}_0\right)\\ &= \mathbf{Z}'\left(S_j^2(\frac{s_e^2}{W1},\frac{s_w^2}{W2})_{jj}\right)^{-1}\mathbf{Z}.\end{aligned} \tag{22}$$

The above random quantity is indeed a GTV, because (i) it is distributed free of unknown parameters, (ii) being a Hotelling T^2 type statistic, it tends to increase for deviations from the null hypothesis, (iii) its observed value $\left(\widehat{\boldsymbol{\beta}_j}-\boldsymbol{\beta}_j\right)'\left(\widehat{\boldsymbol{\beta}_j}-\boldsymbol{\beta}_0\right)$ is free of nuisance parameters, namely the unknown variances. Therefore, the random quantity defined by (22) is indeed a valid GTV. Therefore, the hypothesis can be tested based on the generalized p-value

$$p = Pr(\mathbf{Z}'(S_j^2(\frac{s_e^2}{W_1},\frac{s_w^2}{W_2})_{jj}^{-1}\mathbf{Z})>\left(\widehat{\boldsymbol{\beta}_j}-\boldsymbol{\beta}_0\right)'\left(\widehat{\boldsymbol{\beta}_j}-\boldsymbol{\beta}_0\right)).$$

The p-value is easily computed by numerical integration or Monte Carlo integration, as we further describe below.

5.2. Confidence regions

Generalized confidence ellipsoids for $\boldsymbol{\beta}_j$ are easily computed based on the GPQ corresponding to the above GTV. For example, the $100\gamma\%$ generalized regions is constructed as follows. First, find the cdf of T^2 as

$$\begin{aligned} F_T(t) &= Pr\left(\left(\widehat{\boldsymbol{\beta}_j}-\boldsymbol{\beta}_j\right)' S_j^2((\sigma_e^2,\sigma_w^2)_{jj}^{1/2}(S_j^2(\frac{s_e^2}{W_1},\frac{s_w^2}{W_2})_{jj}^{-1})S_j^2((\sigma_e^2,\sigma_w^2)_{jj}^{1/2}\left(\widehat{\boldsymbol{\beta}_j}-\boldsymbol{\beta}_j\right)\leq t\right) \\ &= \left(\mathbf{Z}'(S_j^2(\frac{s_e^2}{W1},\frac{s_w^2}{W2})_{jj}^{-1})\mathbf{Z}\leq t\right). \end{aligned} \tag{23}$$

Then, find the quantile q_γ such that $F_T(q_\gamma)=\gamma$.

The generalized confidence ellipsoid for $\boldsymbol{\beta}_j$ implied by the above results is

$$\left(\widehat{\boldsymbol{\beta}_j}-\boldsymbol{\beta}_j\right)'\left(\widehat{\boldsymbol{\beta}_j}-\boldsymbol{\beta}_j\right)\leq q_\gamma,$$

because at the observed sample points, mid terms of (23) cancel out, The computation is carried out as follows:

(a). Generate large number, say M, samples from $\mathbf{Z}\sim N(\mathbf{0},\mathbf{I})$,

(b). Generate M random numbers from $\chi^2_{\nu_1}$ and $\chi^2_{\nu_2}$,

(c). Compute and sort the values of $\left(\mathbf{Z}'(S_j^2(\frac{s_e^2}{W1},\frac{s_w^2}{W2})_{jj})\mathbf{Z}\right)$,

(d). Estimate the quantile q_γ as the $M\gamma^{th}$ value of the sorted data.

(e). Construct the generalized ellipsoid using the above formula.

6. Discussion

Further research is necessary to extend forgoing results to more complicated models and hypotheses. Of particular interest is RANOVA (repeated measures ANOVA) and RMANOVA (repeated measures MANOVA) type models involving a number of groups of subjects. Growth curve models involving a number of groups of subjects is a particular case of RMANOVA. Chi and Weerahandi (1998) provided some preliminary results on RMANOVA and provided some guidance on how to handle such problems as multiple comparisons, but did not directly address them. Moreover, there is a need to extend such results to Two-Way RMANOVA, when there are two factors of interest, say treatments groups and groups of subjects characterized by some subject attributes.

One may also consider other approaches to inference, such as the Parametric Bootstrap (PB) approach and the Generalized Fiducial (GF) approach. However, it should be noted, as argued by Ananda et al. (2022), that in most applications, these two methods tend to be subsets of the generalized inference approach. In other words, the latter can reproduce or beat PB based tests and GF based tests, as shown by Ananda et al. (2022).

Kurata (1998) provided a generalization of Rao's Covariance Structure. The results in that article provided distribution theory necessary to tackle greater class of applications combined with generalized approach to inference to handle nuisance parameters.

In a slightly different context, Ghosh and Sinha (1980) studied the criterion robustness of the standard likelihood ratio test (LRT) under the multivariate normal regression model and also the inference robustness of the same test under the univariate set up for certain non-normal distributions of errors. Restricting attention to the normal distribution of errors in the context of univariate regression models, they derived conditions on the design matrix under which the usual LRT of a linear hypothesis (under homoscedasticity of errors) remains valid if the errors have an intraclass covariance structure. The conditions hold in the case of some standard designs. For further related results, the reader is referred to Rao (1967), Zyskind (1967), and Mukhopadhyay and Sinha (1980).

Acknowledgements

I am grateful to the Editors for their guidance and counsel. I also wish to thank the reviewers for their valuable comments and suggestions.

References

Ananda, M. A., Dag, O., and Weerahandi, S. (2023). Heteroscedastic two-way ANOVA under constraints. *Communications in Statistics-Theory and Methods*, **52**, 8207–8222.

Chi, E. M. and Weerahandi, S. (1998). Comparing treatments under growth curve models: exact tests using generalized p-values. *Journal of Statistical Planning and Inference*, **71**, 179–189.

Ghosh, M. and Sinha, B. K. (1980). On the robustness of least squares procedures in regression models. *Journal of Multivariate Analysis*, **10**, 332–342.

Ghosh, S. and Gokhale, D. V. (1987). Estimation and tests for departures from rao-structured covariance matrices. *Biometrical Journal*, **29**, 269–275.

Kurata, H. (1998). A generalization of Rao's covariance structure with applications to several linear models. *Journal of Multivariate Analysis*, **67**, 297–305.

Lehman, E. L. (1986). *Testing Statistical Hypothesis*. John Wiley and Sons, Inc., New York.

McElroy, F. W. (1967). A necessary and sufficient condition that ordinary least-squares estimators be best linear unbiased. *Journal of the American Statistical Association*, **62**, 1302–1304.

Mukhopadhyay, B. B. and Sinha Bikas, K. (1980). A note on the result of M. Ghosh and Bimal K. Sinha: "On the robustness of least squares procedures in regression models". *Calcutta Statistical Association*, **29**, 169–171.

Rao, C. R. (1967). Least squares theory using an estimated dispersion matrix and its application to measurement of signals. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 355–372. Berkeley.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. Wiley, New York.

Tsui, K. and Weerahandi, S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of the American Statistical Association*, **84**, 602–607.

Weerahandi, S. (1987). Testing regression equality with unequal variances. *Econometrica: Journal of the Econometric Society*, **1**, 1211–1215.

Weerahandi, S. (1993). Generalized confidence intervals. *Journal of the American Statistical Association*, **88**, 899–905.

Weerahandi, S. (2004). *Generalized Inference in Repeated Measures: Exact Methods in MANOVA and Mixed Models*. Wiley.

Weerahandi, S. and Berger, V. W. (1999). Exact inference for growth curves with intraclass correlation structure. *Biometrics*, **55**, 921–924.

Xu, L. and Tian, M. (2016). Parametric bootstrap inferences for panel data models. *Communications in Statistics-Theory and Methods*, **46**, 5579–5594.

Zyskind, G. (1967). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. *The Annals of Mathematical Statistics*, **38**, 1092–1110.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 29–30
http://www.ssca.org.in/journal

Reflections on the Life of CR Rao

James L. Rosenberger
Professor Emeritus of Statistics, Pennsylvania State University, USA and Former Director, NISS, Washington DC, USA

Received: 24 July 2024; Revised: 26 July 2024; Accepted: 27 July 2024

In 1988 when Professor CR Rao joined the statistics department at Penn State University, marked the beginning of a dramatic increase in the number of interactions with the world's greatest statisticians. As described in the article published in the Notices of the American Mathematical Society (Vol. 69, No. 6, pp 678-692), CR was invited to join Penn State after his collaborator at University of Pittsburgh passed.

Department Head Thomas P. Hettmansperger of Penn State invited him to accept the inaugural Eberly Family Chair in Statistics, funded in 1986 by the Eberly Family Charitable Trust. CR and his wife Bhargavi moved to State College, PA, purchased a town home a short walk to a bus stop and a park and became active in the community for the next 21 years. Although his initial appointment was for three years, as he approached his 70th birthday the administration waived the mandatory retirement age, and Rao remained active another 10 years and retired in 2001 at age 80.

During his tenure at Penn State, he was regularly recognized for his groundbreaking work in statistics and in addition to his earned Ph.D. and Sc.D. from Cambridge University, UK, he received 27 honorary doctoral degrees from universities in 16 countries around the world.

He was elected Fellow of the National Academy of Sciences, was awarded the Wilks Medal from the American Statistical Association, the Guy Medal in Silver from the Royal Statistical Society, the Megnadh Saha Medal of the Indian National Science Academy, and J.C. Bose Gold Medal of Bose Institute, and the Mahalanobis Centenary Gold Medal of the Indian Science Congress. He has been the president of the International Statistical Institute, the International Biometric Society, and the Institute of Mathematical Statistics, USA.

The Government of India honored him with the second highest civilian award, Padma Vibhushan for "outstanding contributions to Science and Engineering/Statistics" and instituted a cash award in honor of C. R. Rao, "to be given once two years to a young statistician for work done during the preceding three years in any field of statistics."

Shortly after his retirement, in 2002 he was awarded the National Medal of Science, conferred by US President George W. Bush, on June 12, 2002, at the White House. His citation read: "for his contributions to the foundations of statistical theory and multivari-

Corresponding Author: James L. Rosenberger
Email: jrosenberger@niss.org

ate statistical methodology and their applications, enriching the physical, biological, and mathematical, economic, and engineering sciences."

Along with these many achievements, as the Eberly chaired professor of statistics at Penn State, he continued an active research program as Director of the Centre for Multivariate Analysis at Penn State, and he continued to teach courses in multivariate analysis and direct graduate students Ph.D. research.

Speaking as the department head during 15 years of his tenure at Penn State, CR Rao was the catalyst for bringing many international visitors from among the world's leading statisticians to Penn State, which benefited our faculty and graduate students and increased the stature of the department. He endowed several lectureships, the C. G. Khatri and P. R. Krishnaiah Memorial Lectureships, and the CR and Bhargavi Rao Prize which brought renowned speakers to Penn State.

Long after his retirement, CR continued to attend lectures and colloquia events at Penn State and continue to engage in discussions with graduate students during their Ph.D. defense presentations. His impact on the scientific vibrancy of the department was immense and long lived.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 31–35
http://www.ssca.org.in/journal

CR Rao's Shadows on Our Academic Journey

Sumanta Basu[1] **and Jyotishka Datta**[2]
[1] *Cornell University, Ithaca, United States*
[2] *Virginia Polytechnic Institute and State University, Blacksburg, Virginia*

Received: 20 July 2024; Revised: 28 July 2024; Accepted: 31 July 2024

"*One could say if Europe is the mother of differential calculus based on deterministic analysis, India could be called the mother of statistics. When I think of modern statistics, Dr. C.R. Rao features on the top of the list. He once said that "statistics is the technology of finding the invisible and measuring the immeasurable.*"

- Abdul Kalam, Bharat Ratna (past president of India)

We, the authors of this short note, met Prof. C. R. Rao when we were students at the Indian Statistical Institute (ISI), Kolkata. Prof. Rao spent the morning of his visit at the hostel dining room, having breakfast with us, and sharing anecdotes and stories. Fifteen years have passed since that morning, and we have spent these years almost entirely in the United States, pursuing higher education and academic careers. Undoubtedly, our education and careers owe a great deal to the heritage of Statistics in India, and the growth of Statistics over the better part of the last century, both of which were influenced greatly by Prof. C. R. Rao. In what follows, we attempt to capture a few things we remember as lessons from his life and anecdotes that continue to guide and shape us today.

Integration of research and teaching

When Prof. Rao joined Indian Statistical Institute around 1942, he was one of the 15 or so technical workers at the institute, led by P. C. Mahalanobis, known as the 'Professor' at ISI, who did teaching and some research. There were not many textbooks on Statistics yet, and the teachers labored tirelessly turning original research papers into teaching materials, translating the state-of-the-art into classroom materials. During this time as a 'technical apprentice', Rao discovered some of the foundational results in classical statistics, that are still taught in any undergraduate statistics inference course anywhere in the world. One of these was the famous Cramér-Rao inequality that provides a lower bound on the variance of unbiased estimators, indicating the minimum possible variance that any unbiased estimator of a parameter can achieve. As Rao recounts[1], he was presenting a large sample result by Fisher regarding the lower limit of the error of an estimate, and a student in his class[2]

[1]The authors of the present article were fortunate to hear this story from Prof. Rao when he visited the Boys' hostel at ISI Kolkata.

[2]DasGupta (2024) identifies this student as V. M. Dandekar.

Corresponding Author: Jyotishka Datta
Email: jyotishka@vt.edu

asked if a similar result would hold true for small sample size, which is often the case in real application. Rao went home and worked out the solution the same night and answered the student next morning (Champkin, 2011). Due to wartime restrictions, it took two years for Rao's paper to finally appear in the *Bulletin of the Calcutta Mathematical Society* (Rao, 1945). In Sweden, Harold Cramér had derived an analogous result, and Neyman linked the two scientists' name. The beauty of the result is that it holds true for any data distribution under mild regularity conditions.

This story serves as a reminder of why cutting-edge research questions should be blended with teaching statistics, and why one should always encourage students to critically engage with the subject and ask good and possibly difficult questions.

Do not bury the lede

The 1945 paper Rao (1945) has another seminal result – Rao-Blackwellization – which provides a simple way to improve an estimator by conditioning on a sufficient statistics. This, too, was being discovered contemporaneously by David Blackwell in 1947, and the names were combined by Joseph Berkson. Interestingly, as Champkin (2011) points out, Rao did not mention this result in the Introduction, which probably contributed towards it being discovered later. Jokingly, Rao said "It was my first paper, and I was not aware that the introduction is generally written for the benefit of those who do not want to read the paper." This story has now become part of the folklore in our community, and it has a profound implication. For us, it is a reminder of one of the basic rules of journalism: "do not bury the lede", *i.e.*, writers should present the most important information at the beginning of an article or news story. However, we should note here that missed attribution is not an uncommon phenomenon in statistics or machine learning, and while Rao got his credit for the seminal results, Stiegler's law of eponymy (Stigler, 1980) is still commonplace (and perhaps worsening?).

Geometric intuition

It is worth noting that the seminal 1945 paper was written when Rao was only 25 years old, and yet to obtain a PhD. The paper not only introduced Cramér-Rao lower bound and Rao-Blackwellization, it 'introduced differential geometry to statistical inference' and opened the field of information geometry. While presenting a geometric interpretation of the parametric probability densities, Rao defined a 'population space' where Fisher information is used as a distance between densities and the invariant measure turns out to be the square root of the information matrix: an idea containing the essence of Jeffreys' prior. As Efron notes in 'C. R. Rao's Century' (Efron *et al.*, 2020), 'A notable characterization of Rao's work, and Fisher's too, is its reliance on geometric intuition, substituting what, for me, are vivid pictures in place of rote algebra and analysis.' Such 'geometric intuition' has probably been a distinguished characteristic of both the authors' education: the best parts of our theoretical or methodological pursuits were influenced by the geometric intuition about the low-dimensional structures in high-dimensional spaces.

LSI and Impact on ISI education

A major legacy of Prof. C. R. Rao is his iconic Wiley textbook 'Linear Statistical Inference and Its Applications' (LSI) (Rao, 1965). Its encyclopedic breadth aside, what makes this book special is that Rao managed to concisely present and contextualize all the abstract mathematical machinery required, not just to *learn*, but to also *develop*, statistical methods. If we view Statistics as a vehicle that researchers use to advance scientific knowledge, LSI can play the role of not only a driver's manual, but also a mechanic's manual. The book starts with vector spaces and covers linear algebra and probability before introducing statistical theory. Over the last 50 years since its publication, LSI has been used and is still used by statisticians worldwide. Don Rubin once said, "Bill suggested that I turn to Rao's famous textbook on linear models for its straightforward mathematical clarity, at least relative to some other "math-stat" texts that were in use at the time. Being an official dinosaur, I still use it as a "go to" resource" (Efron *et al.*, 2020). To quote Efron: "When the fat second edition of Rao's magisterial book on linear statistical inference arrived on my desk, it was a big event in the department, not just for me (The book is still in use, though it has gotten a little beat up)" (Efron *et al.*, 2020). One can get almost all the fundamental concepts in probability, linear and abstract algebra, distribution theory, linear models, the theory of least squares and analysis of variance, large sample techniques, and multivariate analysis between the two covers of LSI. In fact, in the preface of LSI, Rao states, "*the aim has been to provide in a single volume a full discussion of the wide range of statistical methods useful for consulting statisticians and, at same time, to present in a rigorous manner the mathematical and logical tools employed in deriving statistical procedures, with which a research worker should be familiar.*"

Personally, this book has served the role of a statistical dictionary throughout our academic journey. This comprehensive treatment of statistical methods, along with all the abstract tools needed to derive them, is also a signature style of our undergraduate and graduate (B.Stat. and M.Stat.) education [https://www.isical.ac.in/∼deanweb/brochure_bstat.pdf] at ISI, a learning experience that shaped both authors' scholarly outlooks. Indeed, the B.Stat. and M.Stat. degrees came out of a number of courses in statistics that were developed by C. R. Rao as the head of the Research and Training School at ISI. The three-year long B. Stat. and two-year long M. Stat. program prepared students in various aspects of statistics over the course of ten semesters. Every semester had five courses, and many of the earlier ones would give student rigorous exposure to skills that Rao thought Statisticians should need in their arsenal. Joining the B. Stat. program straight out of high school, we were introduced to three-semester long sequences of real analysis, probability, linear (including one course on abstract) algebra, computer programming (in lower level languages such as C or Fortran) data structures, two-semester long elective on a domain science of one's choice (economics, physics or biology), and only one sequence on Statistical Methods where key ideas will be introduced in intuitive albeit somewhat informal manner. Only after these introductory courses will come the more formal statistical topics in their full glory: linear models, parametric and nonparametric inference, stochastic processes, sample survey and design of experiments. By then, the students are well-trained to think through abstract concepts and recognize them in action in commonly used statistical methods. This integration of abstract and real enabled a generation of students to comfortably navigate between the two worlds.

Our professors, many of them leading researchers in their fields, would take advantage of this unique curriculum to creatively teach important concepts in a classroom that left long-lasting impressions on us. As a concrete example of this pedagogical style, we recall how we learned about linear regression in our first year B. Stat. Statistical Methods courses. We did not have any textbook. It was typical of our professor (Prof. Probal Chaudhuri) to come to class, pose a statistical question in simple terms, and encourage us to solve them using the tools we learned from our other courses such as analysis, algebra, and computer programming.

In one class, he drew a bivariate scatter plot of X and Y on the blackboard, and asked us to find the formulae for a "reasonable" straight line (*i.e.* two numbers, a slope β_0 and an intercept β_1) that passes through the plot. After a lively discussion in the classroom on how to even define "reasonable", the class settled on two loss functions: the squared error loss $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ and least absolute deviation (LAD) $\sum_{i=1}^{n}|y_i - \beta_0 - \beta_1 x_i|$, of two variables β_0 and β_1. We chose to focus with the first because it is differentiable. Students said they have only learned how to differentiate with respect to one variable in high school. So the professor asked: what if I tell you the value of β_0? Can you then find the best β_1 by taking a derivative? Alternately, if I tell you the value of β_1, can you find the best β_0? After some back-of-the-envelope calculations, the students came up with formulae that only involved some weighted means. Then the professor asked: if you keep computing β_0 and β_1 alternately by plugging in the most recent value of the other, would you eventually find their best values? The class was split: some of us thought it will surely work, while others were more skeptical. At this point, the professor reminded us that we don't need to wonder, we already knew enough programming to implement this strategy and see for ourselves. That programming exercise was our Statistics homework for the day. By the time we solved the homework problem and tried it on multiple synthetic data sets that we created ourselves, we had not only learned the concept of simple linear regression, but also a way to solve it using knowledge from our programming class.

We came back to the same problem later in the course, after learning about partial derivatives in other classes, and solved it analytically. This time we not only learned how the closed form solutions of β_0 and β_1 in simple linear regression looked and why the formulae made intuitive sense, we also recognized that even though this strategy is not applicable to the LAD problem, the alternating minimization algorithm introduced earlier is still a potential path to pursue. We revisited the linear regression problem a third time in our Statistics Methods course after we learned enough about vectors and matrices in our linear algebra class. This time, when solving the problem with our newly acquired skills, we recognized that the complicated formulae we derived earlier using multivariable calculus was a special case of a very simple-looking matrix-valued formulae: $\hat{\beta} = (X^\top X)^{-1} X^\top Y$, which even generalized to regression with more than two predictors. This experience helped us appreciate the power of reformulating a complex statistical problem in the language of matrix and vectors.

This pedagogical theme resonated throughout our entire 3-year undergraduate (B. Stat.) and 2 year Masters (M. Stat.) training in ISI. We received rigorous training in real and complex analysis, probability and measure theory, differential equations, and were able to see their application in designing statistical methods. The upshot of this learning style was that Statistics was never about formulae or recipe, it was the experience of solving a realistic problem by combining our intuition with some incredibly powerful yet seemingly disjoint

abstract techniques. Later in our careers, we benefitted a lot from this outlook while doing both methodological and interdisciplinary research. Whenever we tried to adopt this learning style in our own undergraduate classes, we appreciated Rao's vision of presenting rigorous mathematical and logical tools in tandem with statistical methods. A solid foundation in the abstract tools can help students feel the joy of discovery when learning about statistical methods, and appreciate them through a developer's lens.

Importance of domain knowledge

Fifteen years ago, at the breakfast table in ISI boys' hostel, we asked Professor Rao for his advice to junior statisticians like us. He offered many valuable insights, but one in particular we remember clearly to this date. He advised us to study, along with statistics, another domain science rigorously. He stressed that it does not matter what the subject is: it could be physics, chemistry, biology or economics. But if we don't acquire expertise in another domain, he said, someone else will get the credit for our core innovation. Coming from a legend of mathematical statistics, this seemed quite unusual at the moment. We both took elective biology courses during our B. Stat. years, but never fully grasped their role in an otherwise quantitative curriculum. Years later, while doing our postdocs, each of us would spend a fair part of two years in molecular biology labs (SB in Brown/Celniker lab at LBNL, JD in Dave lab at Duke), learning from the domain experts in an immersive environment. Our postdoc advisors, Bin Yu and David Dunson, stressed the crucial importance of this immersive experience for carrying out good scientific work. The experience fundamentally changed the way we approach and conduct research, and also form new collaborations. In our academic journey through this era of data science and its widespread impact across disciplinary boundaries, Prof. Rao's words remain all the more relevant.

Prof. C. R. Rao's significant legacy can perhaps be best summarized by the popular Sanskrit phrase: *deepena prajjwalito deepah*, meaning '*from one lamp, another is lit.*' Prof. Rao's name will be remembered for a long time to come as one of the 'developers of statistics as an independent discipline,' through his many path-breaking contributions, his role in statistics education, and his influence on the numerous statisticians like us in the present and future generations.

References

Champkin, J. (2011). C. R. Rao. *Significance*, **8**, 175–178.

DasGupta, A. (2024). C. R. Rao: Paramount statistical scientist (1920 to 2023). *Proceedings of the National Academy of Sciences*, **121**, e2321318121.

Efron, B., Amari, S.-i., Rubin, D. B., Rao, A. S. R. S., and Cox, D. R. (2020). C. R. Rao's Century. *Significance*, **17**, 36–38.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, **37**, 81–91.

Rao, C. R. (1965). *Linear Statistical Inference and its Applications*. John Wiley & Sons, Inc., New York-London-Sydney.

Stigler, S. M. (1980). Stigler's law of eponymy. *Transactions of the New York Academy of Sciences*, **39**, 147–157.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 37–43
http://www.ssca.org.in/journal

List of 103 Selected Research Papers of C. R. Rao

Compiled by
V. K. Gupta, Bikas K. Sinha and Bimal K. Sinha
in consultation with T. J. Rao, B. L. S. Prakasa Rao and T. Krishna Kumar

Received: 15 July 2024; Revised: 27 July 2024; Accepted: 28 July 2024

The thought in preparing this list of 103 references of Professor C. R. Rao was to highlight the research papers of his lifetime equal in number of years Professor C. R. Rao lived [although short by a few weeks only]. The choice of these papers was made by the group who complied this list. The idea was to have at least one paper for every year since 1942 and fair distribution over the research interests of Professor C. R. Rao. There is no claim being made that this has been the best choice. Unfortunately, we could not find any journal publications for the years 2011, 2015 and 2019.

REFERENCES

1. Nair, K. R. and Rao, C. R. (1941). Confounded designs for asymmetrical factorial experiments. *Science and Culture*, **6**, 313-314.
2. Nair, K. R. and Rao, C. R. (1942). Confounded designs for k x p^m x q^n x ... type of factorial experiments. *Science and Culture*, **7**, 361.
3. Rao, C. R. (1943). Certain experimental arrangements in quasi-latin square. *Current Science*, **12**, 322.
4. Rao, C. R. (1944). On linear estimation and testing of hypotheses. *Current Science*, **13**, 154-155.
5. Rao, C. R. (1945a). Generalisation of Markoff's theorem and tests of linear hypotheses. *Sankhya*, **7**, 9-16.
6. Rao, C. R. (1945b). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, **37**, 81-89.
7. Rao, C. R. (1946a). Hypercubes of strength d leading to confounded designs in factorial experiments. *Bulletin Calcutta Mathematical Society*, **38**, 67-78.
8. Rao, C. R. (1946b). On the linear combination of observations and the general theory of least squares. *Sankhya*, **7**, 237-256.
9. Rao, C. R. (1947a). General methods of analysis for incomplete block designs. *Journal of the American Statistical Association*, **42**, 541-561.
10. Rao, C. R. (1947b). Factorial experiments derivable from combinatorial arrangements of arrays. *Journal of the Royal Statistical Society (Supplement)*, **9**, 128-139.

Corresponding Author: V. K. Gupta
Email: vkgupta1751@yahoo.co.in

11. Nair, K. R. and Rao, C. R. (1948a). Confounded designs for asymmetrical factorial experiments. *Journal of the Royal Statistical Society - Series B*, **10**, 109-131.
12. Rao C. R. (1948b). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, **44**, 50-57.
13. Rao C. R. (1949a). Sufficient statistics and minimum variance estimates. *Mathematical Proceedings of the Cambridge Philosophical Society*, **45**, 213-218.
14. Mahalanobis, P. C., Majumdar D. N., and Rao, C. R. (1949b). Anthropometric survey of the United Provinces: A statistical study, *Sankhya*, **9**, 90–324.
15. Rao, C. R. (1950a). The theory of fractional replication in factorial experiments. *Sankhya*, **10**, 81-86.
16. Rao, C. R. (1950b). Methods of scoring linkage data given the simultaneous segregation of three factors. *Heredity*, **4**, 37-59.
17. Rao, C. R. (1951). A simplified approach to factorial experiments and the punched card technique in the construction and analysis of designs. *Bulletin of the Institute for International Statistics*, **33**, 1-28.
18. Kishen, K. and Rao, C. R. (1952a). An examination of various inequality relations among parameters of the balanced incomplete block design. *Journal of the Indian Society of Agricultural Statistics*, **4**, 137-144.
19. Rao, C. R. (1952b). Some theorems on minimum variance estimation. *Sankhya*, **12**, 27-42.
20. Rao, C. R. (1953). Discriminant function for genetic differentiation and selection (Part IV of statistical inference applied to classificatory problems). *Sankhya*, **12**, 229-246.
21. Rao, C. R. (1954). A general theory of discrimination when the information about the alternative population distribution is based on samples. *Annals of Mathematical Statistics*, **25**, 651-670.
22. Rao, C. R. (1955a). Analysis of dispersion for multiply classified data with unequal numbers in cells. *Sankhya*, **15**, 253-280.
23. Rao, C. R. (1955b). Estimation and tests of significance in factor analysis. *Psychometrika*, **20**, 93-111.
24. Kallianpur, G. and Rao, C. R. (1955c). On Fisher's lower bound to asymptotic variance of a consistent estimate. *Sankhya*, **15**, 331-342.
25. Rao, C. R. (1956a). On the recovery of inter block information in varietal trials. *Sankhya*, **17**, 105–114.
26. Rao, C. R. (1956b). A general class of quasifactorial and related designs. *Sankhya*, **17**, 165-174.
27. Rao, C. R. (1957). Maximum likelihood estimation for multinomial distribution. *Sankhya*, **18**, 139-148.
28. Rao, C. R. (1958a). Some statistical methods for comparison of growth curves. *Biometrics*, **14**, 1-17.
29. Majumdar, D. N., Rao, C. R., and Mahalanobis, P. C. (1958b). Bengal anthropometric survey, 1945: A statistical study. *Sankhya*, **19**, 201-408.
30. Rao, C. R. (1959). Expected values of mean squares in the analysis of incomplete block experiments and some comments based on them. *Sankhya*, **21**, 327-336.

31. Rao, C. R. (1960a). Experimental designs with restricted randomization. *Bulletin of the Institute for International Statistics*, **37**, 397-404.

32. Rao, C. R. (1960b). Multivariate Analysis: An indispensable statistical aid in applied research. *Sankhya*, **22**, 317-338.

COUNT UPTO 1960 ... 32 REFERENCES LISTED

33. Rao, C. R. (1961). A study of BIB designs with replications 11 to 15. *Sankhya - Series A*, **23**, 117-127.

34. Rao, C. R. (1962). Efficient estimates and optimum inference procedures in large samples (with discussion). *Journal of the Royal Statistical Society - Series B*, **24**, 46-72.

35. Rao, C. R. and Varadarajan, V. S. (1963). Discrimination of Gaussian processes. *Sankhya - Series A*, **25**, 303-330.

36. Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya – Series A*, **26**, 329-358.

37. Rao, C. R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, **52**, 447-458.

38. Rao, C. R. (1966a). Discriminant function between composite hypotheses and related problems. *Biometrika*, **53**, 339-345.

39. Rao, C. R. (1966b). Characterization of the distribution of random variables in linear structural relations. *Sankhya*, **28**, 251-260.

40. Rao, C. R. (1967a). Calculus of generalized inverses of matrices Part I. General theory. *Sankhya - Series A*, **29**, 317–342.

41. Rao, C. R. (1967b). On some characterizations of the normal law. *Sankhya - Series A*, **29**, 1-14.

42. Khatri, C. G. and Rao, C. R. (1968a). Solutions to some functional equations and their applications to characterization of probability distributions. *Sankhya - Series A*, **30**, 167-180.

43. Ramachandran, B. and Rao, C. R. (1968b). Some results on characteristic functions and characterizations of the normal and generalized stable laws. *Sankhya - Series A*, **30**, 125-140.

44. Rao, C. R. (1969). A decomposition theorem for vector variables with a linear structure. *Annals of Mathematical Statistics*, **40**, 1845-1849.

45. Rao, C. R. (1970). Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical Association*, **65**, 161–172.

COUNT UPTO 1970 ... 45 REFERENCES LISTED

46. Rao, C. R. (1971a). Unified theory of linear estimation. *Sankhya - Series A*, **33**, 371–394. [Corrigenda (1972): **34**, p. 194 and p. 477].

47. Rao, C. R. (1971b). Estimation of variance and covariance components—MINQUE theory. *Journal of Multivariate Analysis*, **1**, 257-275.

48. Rao, C. R. (1971c). Some aspects of statistical inference in problems of sampling from finite populations [with discussion and a reply by the author]. *In Foundations of Statistical Inference*, Proceedings of Symposium, University of Waterloo, March 31-April 9, 1970 (V. P. Godambe and D. A. Sprott, Eds.), Holt, Rinehart and Winston of Canada, Toronto, Ontario, 177-202.

49. Rao, C. R. and Mitra, S. K. (1972a). Generalized inverse of a matrix and its applications. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, Lucien M. Le Cam, Jerzy Neyman and Elizabeth L. Scott (Eds.), Berkeley: University of California Press, 601-620.

50. Khatri, C. G. and Rao, C. R. (1972b). Functional equations and characterization of probability laws through linear functions of random variables. *Journal of Multivariate Analysis*, **2**, 162-173.

51. Rao, C. R. (1973a). Unified theory of least squares. *Communications in Statistics*, **1**, 1-8.

52. Baksalary, J. K., Rao, C. R., and Markiewicz, A. (1973b). A study of the influence of the 'natural restrictions' on estimation problems in the singular Gauss-Markov model. *Journal of Statistical Planning and Inference*, **31**, 335-351.

53. Mitra, S. K. and Rao, C. R. (1974). Projections under seminorms and generalized Moore Penrose inverses. *Linear Algebra and its Applications*, **9**, 155-167.

54. Rao, C. R. (1975). Simultaneous estimation of parameters in different linear models and applications to biometric problems. *Biometrics*, **31**, 545-554.

COUNT UPTO 1975 ... 54 REFERENCES LISTED

55. Rao, C. R. (1976). Characterization of prior distributions and solution to a compound decision problem. *Annals of Statistics*, **4**, 823-835.

56. Rao, C. R. (1977). A natural example of weighted distributions: A classroom exercise. *American Statistician*, **31**, 24-26.

57. Patil, G. P. and Rao, C. R. (1978). Weighted distributions and size biased sampling with applications to wildlife populations and human families. *Biometrics*, **34**, 179-189.

58. Rao, C. R. (1979). MINQE theory and its relation to ML and MML estimation of variance components. *Sankhya -Series B*, **41**, 138-153.

59. Rao, C. R. and Kleffe, J. (1980). Estimation of variance components. In P. R. Krishnaiah (Ed.) *Handbook of Statistics*, **1**, 1-40.

COUNT UPTO 1980 ... 59 REFERENCES LISTED

60. Khatri, C. G. and Rao, C. R. (1981). Some extensions of the Kantorovich inequality and statistical applications. *Journal of Multivariate Analysis*, **11**, 498-505.

61. Lau, K. S. and Rao, C. R. (1982). Integrated Cauchy functional equation and characterizations of the exponential law. *Sankhya - Series A*, **44**, 72-90.

62. Rao, C. R. (1983). Likelihood ratio tests for relationships between two covariance matrices. In *Studies in Econometrics, Time Series, and Multivariate Statistics*. 529-544, New York, Academic Press.

63. Miiller, J., Rao, C. R., and Sinha, Bimal K. (1984). Inference on parameters in a linear model: A review of recent results. In K. Hinkelmann (Ed.), *Experimental Design, Statistical Models, and Genetic Studies*, Statistics Textbooks Monographs **50**, Marcel Dekker, New York, 277- 295.

64. Rao, C. R., Reatring, I. P., and Mason, R. D. (1985a). The Pitman nearness criterion and its determination. *Communications in Statistics – Theory and Methods*, **15**, 3173-3191.

65. Rao, C. R. (1985b). The inefficiency of least squares: extensions of Kantorovich inequality. *Linear Algebra and its Applications*, **70**, 249-255.

COUNT UPTO 1985 ... 65 REFERENCES LISTED

66. Patil, G. P., Rao, C. R., and Ratnaparkhi, M. V. (1986a). On discrete weighted distributions and their use in model choice for observed data. *Communications in Statistics*, **15**, 907-918.

67. Rao, C. R. and Shanbhag, D. N. (1986b). Recent results on characterization of probability distributions: a unified approach through extensions of Deny's theorem. *Advances in Applied Probability*, **18**, 660-678.

68. Khatri, C. G. and Rao, C. R. (1987). Effects of estimated noise covariance matrix in optimal signal detection. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **35**, 671-679.

69. Hedayat, A. S., Rao, C. R., and Stufken, J. (1988). Sampling plans excluding contagious units. *Journal of Statistical Planning and Inference*, **19**, 159-170.

70. Bai, Z. D., Rao, C. R., and Zhao, L. C. (1989). Kernel estimators of density function of directional data. *Multivariate Statistics and Probability, Essays in Memory of Paruchuri R. Krishnaiah*, 24-39.

71. Bai, Z. D., Rao, C. R., and Yin, Y. Q. (1990). Least absolute deviations analysis of variance. *Sankhya - Series A*, **52**, 166-177.

COUNT UPTO 1990 ... 71 REFERENCES LISTED

72. Bai, Z. D. and Rao, C. R. (1991). Edgeworth expansion of a function of sample means. *The Annals of Statistics*, **19**, 1295-1315.

73. Babu, G. J., Rao, C. R., and Rao, M. B. (1992). Nonparametric estimation of specific occurrence / exposure rate in risk and survival analysis. *Journal of the American Statistical Association*, **87**, 84-89.

74. Rao, C. R. and Zhao, L. C. (1993a). Asymptotic behavior of maximum likelihood estimates of superimposed exponential signals. *IEEE Transactions on Signal Processing*, **41**, 1461-1464.

75. Rao, C. R. (1993b). Statistics must have a purpose - Mahalanobis dictum. *Sankhya - Series A*, **55**, 331-349.

76. Rao, C. R., Zhao, L. C., and Zhou, B. (1994). Maximum likelihood estimation of 2-D superimposed exponential signals. *IEEE Transactions on Signal Processing*, **42**, 1795-1802.

77. Rao, C. R. and Mukerjee, R. (1995). Comparison of Bartlett-type adjustments for the efficient score statistic. *Journal of Statistical Planning and Inference*, **46**, 137-146.

COUNT UPTO 1995 ... 77 REFERENCES LISTED

78. Rao, C. R. and Suryawanshi, S. (1996). Statistical analysis of shape of objects based on landmark data. *Proceedings of the National Academy of Sciences*, **93**, 12132-12136.
79. Rao, C. R., Pathak, P. K., and Koltchinskii, V. I. (1997). Bootstrap by sequential resampling. *Journal of Statistical Planning and Inference*, **64**, 257-281.
80. Rao, C. R. and Shanbhag, D. N. (1998). Further versions of the convolution equation. *Journal of the Indian Society of Agricultural Statistics*, **51**, 361-378.
81. Bai, Z. D., Rao C. R., and Wu, Y. (1999). Model selection with data-oriented penalty. *Journal of Statistical Planning and Inference*, **77**, 103-118.
82. Rao, C. R. (2000a). Statistical proofs of some matrix inequalities. *Linear Algebra and its Applications*, **321**, 307-320.
83. Szekely, G. J. and Rao, C. R. (2000b). Identifiability of distributions of independent random variables by linear combinations and moments. *Sankhya - Series A*, **62**, 193-202.

COUNT UPTO 2000 ... 83 REFERENCES LISTED

84. Rao, C. R. (2001a). Pre and post least squares: The emergence of robust estimation. *Journal of Statistical Research*, **32**, 1-18.
85. Rao, C. R. (2001b). Statistics: Reflections on the past and visions for the future. *Communications in Statistics: Theory and Methods*, **30**, 2235-2254.
86. Rao, C.R. (2002). Karl Pearson chi-square test the dawn of statistical inference. In: Huber-Carol, C., Balakrishnan, N., Nikulin, M.S., Mesbah, M. (Eds) *Goodness-of-Fit Tests and Model Validity. Statistics for Industry and Technology*, Birkhäuser, Boston, MA.
87. Balakrishnan, N. and Rao, C. R. (2003). Some efficiency properties of best linear unbiased estimators. *Journal of Statistical Planning and Inference*, **113**, 551-555.
88. Shelton, P. and Rao, C. R. (2004). A note on testing for serial correlation in large number of small samples using tail probability approximations. *Communications in Statistics - Theory and Methods*, **33**, 1767-1777.
89. Rao, C. R. and Wu, Y. (2005). Linear model selection by cross-validation. *Journal of Statistical Planning and Inference*, **128**, 231-240.

COUNT UPTO 2005 ... 89 REFERENCES LISTED

90. Rao, C. R. (2006). Statistical proofs of some matrix theorems. *International Statistical Review / Revue Internationale de Statistique*, **74**, 169-185.
91. Rao, C. R., Rao, M. B., and Zhang H. (2007). One bulb? Two bulbs? How many bulbs light up? — A discrete probability problem involving dermal patches. *Sankhya - Series A*, **69**, 137-161.
92. Rao, C. R. (2008). Life and works of Ronald Aylmer Fisher. *Journal of Statistical Theory and Practice*, **2**,131-141.

93. Rao, C. R., Shanbhag, D. N., Theofanis, S., and Rao, M. B. (2009). Some properties of extreme stable laws and related infinitely divisible random variables. *Journal of Statistical Planning and Inference*, **139**, 802-813.

94. Rao, C. R. (2010a). Quadratic entropy and analysis of diversity. *Sankhya - Series A*, **72**, 70–80.

95. Rao, C. R. (2010b). Entropy and cross entropy: Characterizations and applications. In Alladi, K., Klauder, J. R., and Rao, C. R. (Eds.) *The Legacy of Alladi Ramakrishnan in the Mathematical Sciences*. Springer.

96. Rao, C. R., Wu, Y., and Shi, X. (2010c). An M-estimation-based criterion for simultaneous change point analysis and variable selection in a regression problem. *Journal of Statistical Theory and Practice*, **4**, 773-801.

COUNT UPTO 2010 ... 96 REFERENCES LISTED

97. Hyunsook, L., Babu, J. G., and Rao, C. R. (2012). A Jackknife type approach to statistical model selection. *Journal of Statistical Planning and Inference*, **142**, 301-311.

98. Pathak, P. K. and Rao, C. R. (2013). The sequential bootstrap. *Handbook of Statistics - Machine Learning: Theory and Applications*, **31**, 3-18. Elsevier/North-Holland, Amsterdam.

99. Rao, C. R., Shi, X., and Wu, Y. (2014). Approximation of the expected value of the harmonic mean and some applications. *Proceedings of the National Academy of Sciences*, **111**, 15681-15686.

100. Rao, T. J. and Rao, C. R. (2016). Review of certain recent advances in randomized response techniques. *Handbook of Statistics*, **34**, 1-12.

101. Shi, Xiaoping, Wu, Yuehua, and Rao, C. R. (2017). Consistent and powerful graph-based change-point test for high-dimensional data. *Physical Sciences*, **114**, 3873-3878.

102. Shi, X., Wu, Y., and Rao, C. R. (2018). Consistent and powerful non-Euclidean graph-based change-point test with applications to segmenting random interfered video data. *Proceedings of the National Academy of Sciences*, **115**, 5914-5919.

103. Jin, B., Wu, Y., Rao, C. R., and Hou, Li (2020). Estimation and model selection in general spatial dynamic panel data models. *Proceedings of the National Academy of Sciences*, **117**, 5235-5241.

FINAL COUNT ... UPTO 2020 ... 103 REFERENCES LISTED

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 45–66
http://www.ssca.org.in/journal

Some Novel Limiting Distributions Arising in Order Restricted Inference

Sayan Ghosh[1] **and Ori Davidov**[2]
[1]*Department of Mathematics, Birla Institute of Technology and Sciences Pilani, Hyderabad Campus, Hyderabad 500078, India*
[2]*Department of Statistics, University of Haifa, Mount Carmel, Haifa 3498838, Israel*

Received: 21 Dember 2023; Revised: 9 February 2024; Accepted: 23 February 2024

Abstract

In this paper, we develop a methodology for testing the hypothesis that the true value of a parameter lies in the union of multiple cones against the alternative that it does not. We propose a test statistic for such problems and derive its novel asymptotic null distribution. The least favourable asymptotic null value and the corresponding least favourable asymptotic null distribution are obtained. The proposed test is uniformly more powerful than conventional tests discussed in the literature. Some illustrative examples are provided and a simulation study evaluating its performance is presented.

Key words: Hypothesis Testing; Convex Cones; Least Favourable Configuration; Asymptotic Distribution.

AMS Subject Classifications: 62F03, 62F30

1. Introduction

Testing problems are typically formulated as $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ versus $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_1$. Usually, the null $\boldsymbol{\Theta}_0$ as well as the alternative $\boldsymbol{\Theta}_1$ are simple sets such as singletons or linear spaces. There are however various applications in which the null and/or the alternative are more complicated sets. In particular, we consider testing

$$H_0 : \boldsymbol{\theta} \in \bigcup_{i=1}^{K} \mathcal{C}_i \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \notin \bigcup_{i=1}^{K} \mathcal{C}_i, \tag{1}$$

where $\mathcal{C}_1, \ldots, \mathcal{C}_K$ are arbitrary distinct convex cones in $\mathbb{R}^m$ defined by systems of linear inequalities. In this paper we address the case where $m = 2$. Some comments on the corresponding theory for general m are deferred to Section 7. It is further assumed that there exists an unconstrained estimator $\boldsymbol{S}_n$ for $\boldsymbol{\theta} \in \mathbb{R}^2$ such that as $n \to \infty$

$$\sqrt{n}(\boldsymbol{S}_n - \boldsymbol{\theta}) \Rightarrow \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}) \tag{2}$$

Corresponding Author: Ori Davidov
Email: davidov@stat.haifa.ac.il

where $\Rightarrow$ denotes convergence in distribution and $\boldsymbol{\Sigma}$ is a positive definite matrix. As indicated below, there are many problems of interest that can be formulated as in (1).

Robertson and Wegman (1978) were among the first to test hypotheses of the type (1) but with $K = 1$. This setting is known in the literature as *testing against an ordering* and classified by Silvapulle and Sen (2004) as a Type B problem. In general the union of convex cones is neither convex nor a cone. Therefore Type B problems are a simple special case of (1). Other special cases of (1) in $\mathbb{R}^2$ have also been addressed in the literature. Berger and Sinclair (1984) examined the problem of testing a null hypothesis that the parameter of interest belongs to a union of linear subspaces which they applied to the testing of symmetric spacings among ordered normal means. Another paper involving linear spaces is by Berger (1997) who tested $H_0 : \min\{|\theta_1|, |\theta_2|\} = 0$ against $H_1 : \min\{|\theta_1|, |\theta_2|\} > 0$. Thus under the null the pair (θ_1, θ_2) lies on the axes whereas under the alternative it does not. If θ_i measures the effect of treatment i then the null states that at least one treatment has no effect whereas under the alternative both treatments have effects.

This paper is organized as follows. In Section 2, we discuss the preliminaries related to testing (1). We introduce relevant notations and setup for the testing problem. Then we define the proposed test statistic and show that it is identical to the likelihood ratio test statistic and to the intersection union test statistic for (1) in some cases. In Section 3 we consider the problem of testing (1) for two quadrants in $\mathbb{R}^2$. We obtain the least favourable null values and the least favourable null distribution of the proposed test statistic for finite samples. In Section 4, we consider the union of multiple distinct arbitrary convex cones in $\mathbb{R}^2$ and obtain the least favourable null values along with the least favourable null distribution of the proposed test statistic for large samples. In Section 5, some examples and testing problems in $\mathbb{R}^2$ are provided as illustration. A simulation study is performed to evaluate the proposed test in Section 6. Finally, in Section 7 we provide a brief summary of our work and some possible extensions.

2. Preliminaries

We begin with some notations. Let $\Pi_{\boldsymbol{\Sigma}}(\boldsymbol{S} \mid \mathcal{C})$ denote the projection of $\boldsymbol{S}$ onto $\mathcal{C}$ with respect to $\boldsymbol{\Sigma}$ and let $\|\boldsymbol{S}\|^2_{\boldsymbol{\Sigma}}$ be the respective norm. Note that when $\boldsymbol{\Sigma} = \boldsymbol{I}$ the latter reduces to the usual projection and the standard euclidean distance. For the hypotheses in (1), we propose the test statistic

$$T_n = \min\{n\|\boldsymbol{S}_n - \Pi_{\boldsymbol{\Sigma}}(\boldsymbol{S}_n \mid \mathcal{C}_1)\|^2_{\boldsymbol{\Sigma}}, \ldots, n\|\boldsymbol{S}_n - \Pi_{\boldsymbol{\Sigma}}(\boldsymbol{S}_n \mid \mathcal{C}_K)\|^2_{\boldsymbol{\Sigma}}\}, \tag{3}$$

where $\boldsymbol{S}_n$ was described in (2). In general the variance matrix $\boldsymbol{\Sigma}$ is unknown in which case T_n is computed with respect to a consistent estimator $\boldsymbol{\Sigma}_n$ thereof. It is clear that T_n essentially minimizes the squared distance between $\boldsymbol{S}_n$ and $\boldsymbol{\theta}$ over various values of $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}_0$. The following result shows the relationship between the proposed test, the likelihood ratio test (LRT) and the intersection union test (IUT).

Theorem 1: If $\boldsymbol{S}$ follows a $\mathcal{N}_2(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ distribution with known $\boldsymbol{\Sigma}$ then the statistic (3) as a function of $\boldsymbol{S}$ is the LRT statistic for the hypotheses in (1). Moreover, (3) is the IUT statistic if and only if the cones $\mathcal{C}_1, \ldots, \mathcal{C}_K$ are all congruent.

Theorem 1 provides a meaningful motivation for using the statistic (3) when (2) holds

as $n \to \infty$ as well as in situations in which $\boldsymbol{\Sigma}$ is unknown but can be consistently estimated from the data. Although under the stated conditions the LRT and the IUT statistics coincide, their critical values are in general different. Further note that if we set $\boldsymbol{G}_n = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}_n$ then $\sqrt{n}(\boldsymbol{G}_n - \boldsymbol{\eta}) \Rightarrow \mathcal{N}_2(\boldsymbol{0}, \boldsymbol{I})$ where $\boldsymbol{\eta} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta}$. In addition testing the hypotheses (1) using $\boldsymbol{S}_n$ is equivalent to testing

$$H_0 : \boldsymbol{\eta} \in \bigcup_{i=1}^{K} \mathcal{C}_i^* \quad \text{versus} \quad H_1 : \boldsymbol{\eta} \notin \bigcup_{i=1}^{K} \mathcal{C}_i^*$$

using $\boldsymbol{G}_n$ where $\mathcal{C}_i^* = \boldsymbol{\Sigma}^{-1/2}\mathcal{C}_i = \{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta} : \boldsymbol{\theta} \in \mathcal{C}_i\}$ are the transformed cones. Therefore without any loss of generality we will henceforth primarily consider the case where $\boldsymbol{\Sigma} = \boldsymbol{I}$.

3. The case of two cones

We start by investigating the important special case of two quadrants. Let $\mathcal{C}_1 = \{\boldsymbol{\theta} \in \mathbb{R}^2 : \theta_1 \geq 0, \theta_2 \geq 0\}$ denote the positive quadrant and let $\mathcal{C}_2 = \{\boldsymbol{\theta} \in \mathbb{R}^2 : \theta_1 \leq 0, \theta_2 \leq 0\}$ denote the negative quadrant. Consider testing the hypotheses

$$H_0 : \boldsymbol{\theta} \in \mathcal{C}_1 \cup \mathcal{C}_2 \quad \text{against} \quad H_1 : \boldsymbol{\theta} \notin \mathcal{C}_1 \cup \mathcal{C}_2 \tag{4}$$

using a single observation $\boldsymbol{S} = (S_1, S_2)^T$ from $\mathcal{N}_2(\boldsymbol{\theta}, \boldsymbol{I})$. Note that under the null θ_1 and θ_2 are either both non-negative or both non-positive. This problem is of independent interest.

Theorem 2: Suppose that $\boldsymbol{S}$ follows $\mathcal{N}_2(\boldsymbol{\theta}, \boldsymbol{I})$. Then the LRT statistic for (4) is

$$T = \min\{S_1^2, S_2^2\}\mathbb{I}(\boldsymbol{S} \notin \mathcal{C}_1 \cup \mathcal{C}_2). \tag{5}$$

Furthermore for all $c \geq 0$ we have

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \mathbb{P}_{\boldsymbol{\theta}}(T \geq c) = \frac{1}{2}\mathbb{P}(\chi_0^2 \geq c) + \frac{1}{2}\mathbb{P}(\chi_1^2 \geq c). \tag{6}$$

Equation (6), where χ_i^2 is a chi–square RV with i degrees of freedom and $\chi_0^2 \equiv 0$, provides us with a formula with which we can compute the p–values associated with the test statistic (5). The value of $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ for which (6) holds is called the *least favourable configuration* or null value. The distribution of the statistic T when $\boldsymbol{\theta}$ is the least favourable is called the *least favourable null distribution*. The proof of Theorem 2 shows that the least favourable configurations are of the form $(0, \pm\infty)$ and $(\pm\infty, 0)$, *i.e.*, they lie on the axes at an infinite distance from the origin while the least favourable null distribution of T is given by (6). It follows that for any other value of $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ and any c

$$\mathbb{P}_{\boldsymbol{\theta}}(T \geq c) < \frac{1}{2}\mathbb{P}(\chi_0^2 \geq c) + \frac{1}{2}\mathbb{P}(\chi_1^2 \geq c).$$

Letting $T(\theta_1, \theta_2)$ denote the LRT statistic at (θ_1, θ_2) we can restate the conclusion of Theorem 2 in the language of stochastic order relations (Shaked and Shanthikumar (2007)) as $T(\theta_1, \theta_2) \preceq_{st} T(0, \pm\infty)$ and $T(\theta_1, \theta_2) \preceq_{st} T(\pm\infty, 0)$ where $\preceq_{st}$ denotes the usual stochastic order. Both relations hold for all $(\theta_1, \theta_2) \in \boldsymbol{\Theta}_0$. It can also be shown that $T(0, 0) \preceq_{st} T(0, \theta_2)$

and $T(0,0) \preceq_{st} T(\theta_1, 0)$. In particular, $T(0,0)$ is distributed as $(1/2)\chi_0^2 + (1/2)\min\{Q_1, Q_2\}$ where Q_1 and Q_2 are independent χ_1^2 RVs.

In the proof of Theorem 2 a closed form expression for $\mathbb{P}_{\boldsymbol{\theta}}(T \geq c)$ was found facilitating the analysis and enabling one to find the least favourable configuration and null distribution. In general though, $\mathbb{P}_{\boldsymbol{\theta}}(T \geq c)$ is not amenable to a simple analysis nor is it given by a simple formula. Consequently, an asymptotic analysis yielding workable formulas of the type (6) is necessary.

4. The general case

In this section a general asymptotic theory for multiple cones is developed. First note that any convex cone in $\mathbb{R}^2$ is of the form

$$\mathcal{C} = \text{conic}(\boldsymbol{u}, \boldsymbol{v}) = \{\lambda_1 \boldsymbol{u} + \lambda_1 \boldsymbol{v} : \lambda_1 \geq 0, \lambda_2 \geq 0\},$$

where $\boldsymbol{u}$ and $\boldsymbol{v}$ are unit vectors lying on the extreme rays of $\mathcal{C}$. Further note that the angle between $\boldsymbol{u}$ and $\boldsymbol{v}$, *i.e.*, $\measuredangle(\boldsymbol{u}, \boldsymbol{v})$ is smaller than π.

Let $\mathcal{C}_1, \ldots, \mathcal{C}_K$ be K distinct convex cones where $\mathcal{C}_i = \text{conic}(\boldsymbol{u}_i, \boldsymbol{v}_i)$ for $i = 1, \ldots, K$. For convenience it is further assumed that for $\boldsymbol{e}_1 = (1,0)^T$ we have:

$$\measuredangle(\boldsymbol{e}_1, \boldsymbol{u}_i) < \measuredangle(\boldsymbol{e}_1, \boldsymbol{v}_i)$$

for all i and

$$\measuredangle(\boldsymbol{e}_1, \boldsymbol{u}_1) < \measuredangle(\boldsymbol{e}_1, \boldsymbol{u}_2) < \cdots < \measuredangle(\boldsymbol{e}_1, \boldsymbol{u}_K).$$

Thus the cone $\mathcal{C}_1$ is the cone whose rays make the smallest angle with the positive real axis followed by the cone $\mathcal{C}_2$, and so forth. Similarly within each cone the ray associated with $\boldsymbol{u}_i$ has a smaller angle than the ray $\boldsymbol{v}_i$.

We say cones $\mathcal{C}_i$ and $\mathcal{C}_j$ are adjacent if the interior of the cone $\text{conic}(\boldsymbol{v}_i, \boldsymbol{u}_j)$ is a subset of $\boldsymbol{\Theta}_1$. The angle between $\boldsymbol{v}_i$ and $\boldsymbol{u}_j$ may be smaller than $\pi/2$, between $\pi/2$ and π or larger than π. If $\measuredangle(\boldsymbol{v}_i, \boldsymbol{u}_j) \leq \pi/2$, we set

$$\mathcal{R}_{ij} = \text{conic}(\boldsymbol{v}_i, \boldsymbol{u}_j). \tag{7}$$

If $\pi/2 < \measuredangle(\boldsymbol{v}_i, \boldsymbol{u}_j) \leq \pi$, we further divide the cone $\text{conic}(\boldsymbol{v}_i, \boldsymbol{u}_j)$ into three conic regions

$$\mathcal{R}_i(\boldsymbol{v}_i) = \text{conic}(\boldsymbol{v}_i, \boldsymbol{u}_{j*}), \quad \mathcal{R}'_{ij} = \text{conic}(\boldsymbol{u}_{j*}, \boldsymbol{v}_{i*}), \quad \mathcal{R}_j(\boldsymbol{u}_j) = \text{conic}(\boldsymbol{v}_{i*}, \boldsymbol{u}_j) \tag{8}$$

where $\boldsymbol{u}_{j*}, \boldsymbol{v}_{i*} \in \text{conic}(\boldsymbol{v}_i, \boldsymbol{u}_j)$, $\boldsymbol{u}_{j*}$ is orthogonal to $\boldsymbol{u}_j$ and $\boldsymbol{v}_{i*}$ is orthogonal to $\boldsymbol{v}_i$. Finally if $\measuredangle(\boldsymbol{v}_i, \boldsymbol{u}_j) > \pi$, then we divide the region bounded by $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ into three conic regions

$$\mathcal{R}_i(\boldsymbol{u}_i) = \text{conic}(\boldsymbol{u}_i, \boldsymbol{u}_{i*}), \quad \mathcal{R}''_{ij} = \text{conic}(\boldsymbol{u}_{i*}, \boldsymbol{v}_{j*}), \quad \mathcal{R}_j(\boldsymbol{v}_j) = \text{conic}(\boldsymbol{v}_{j*}, \boldsymbol{v}_j) \tag{9}$$

where $\boldsymbol{u}_{i*}, \boldsymbol{v}_{j*} \in \text{conic}(\boldsymbol{u}_i, \boldsymbol{v}_j)$, $\boldsymbol{u}_{i*}$ is orthogonal to $\boldsymbol{u}_i$ and $\boldsymbol{v}_{j*}$ is orthogonal to $\boldsymbol{v}_j$.

Remark 1: If $K = 2$ the cones $\mathcal{C}_1$ and $\mathcal{C}_2$ are doubly adjacent. Moreover, if $\measuredangle(\boldsymbol{v}_1, \boldsymbol{u}_2) \leq \pi/2$ and $\measuredangle(\boldsymbol{v}_2, \boldsymbol{u}_1) \leq \pi/2$ then we label the regions between the cones by $\mathcal{R}_{12}$ and $\mathcal{R}_{21}$. The modification when the above mentioned angles are larger than $\pi/2$ is obvious.

The following result provides the number of possible regions between the cones or between their polar cones for various geometric arrangements of the cones.

Lemma 1: Let N, N' and N'' denote the number of regions of the type $\mathcal{R}_{ij}$, $\mathcal{R}'_{ij}$ and $\mathcal{R}''_{ij}$ respectively. Then $N + N' + N'' = K$ where $N \leq K$, $N' \leq 3$ and $N'' \leq 1$. In particular, if $N'' = 1$, then $N \leq K - 1$ and $N' \leq 1$.

It is well known that $\Pi_{\boldsymbol{I}}(\boldsymbol{S} \mid \mathcal{C}_i) = \mathbf{0}$ if and only if $\boldsymbol{S} \in \mathcal{C}_i^0$ where $\mathcal{C}_i^0$ denotes the polar cone of $\mathcal{C}_i$. Thus, it follows that $\Pi_{\boldsymbol{I}}(\boldsymbol{S} \mid \bigcup_{i=1}^{K} \mathcal{C}_i) = \mathbf{0}$ if and only if $\boldsymbol{S} \in \bigcap_{i=1}^{K} \mathcal{C}_i^0$. This event has a positive probability if the set $\bigcap_{i=1}^{K} \mathcal{C}_i^0 \supset \{\mathbf{0}\}$ and zero probability if $\bigcap_{i=1}^{K} \mathcal{C}_i^0 = \{\mathbf{0}\}$. In the first case we denote the set $\bigcap_{i=1}^{K} \mathcal{C}_i^0$ by $\mathcal{R}''_{pq}$ as defined above whereas in the latter case we set $\mathcal{R}''_{pq} = \emptyset$. In other words:

$$N'' = \begin{cases} 0 & \text{if } \bigcap_{i=1}^{K} \mathcal{C}_i^0 = \{\mathbf{0}\} \\ 1 & \text{if } \bigcap_{i=1}^{K} \mathcal{C}_i^0 \supset \{\mathbf{0}\} \end{cases},$$

We now introduce some useful additional notations.

Definition 1: Let $\mathcal{R} = \text{conic}(\boldsymbol{u}, \boldsymbol{v})$ and denote its interior angle by $\gamma = \measuredangle(\boldsymbol{u}, \boldsymbol{v})$ where $0 < \gamma \leq \pi/2$. Let $\boldsymbol{S}$ be a $\mathcal{N}_2(\mathbf{0}, \boldsymbol{I})$ RV. Then conditional on $\boldsymbol{S} \in \mathcal{R}$ we define

$$\chi^2_{1,1}(\gamma) = \mathrm{d}^2(\boldsymbol{S}, \mathrm{bd}(\mathcal{R})) = \min\{\mathrm{d}^2(\boldsymbol{S}, \mathrm{ray}(\boldsymbol{u})), \mathrm{d}^2(\boldsymbol{S}, \mathrm{ray}(\boldsymbol{v}))\} \tag{10}$$

where $\mathrm{d}(\cdot, \cdot)$ is the euclidean distance and $\mathrm{bd}(\mathcal{R})$ is the boundary of the cone $\mathcal{R}$ defined by the rays $\mathrm{ray}(\boldsymbol{u})$ and $\mathrm{ray}(\boldsymbol{v})$.

For example when $\mathcal{R}$ is any quadrant then $\gamma = \pi/2$ and $\mathbb{P}(\chi^2_{1,1}(\pi/2) \geq c) = [\mathbb{P}(\chi^2_1 \geq c)]^2$. When $0 < \gamma < \pi/2$ we have the following, numerically simple to evaluate formula.

Lemma 2: For $c \geq 0$ and $\mathcal{R}$ as in definition 1

$$\mathbb{P}(\chi^2_{1,1}(\gamma) \geq c, \boldsymbol{S} \in \mathcal{R}) = \frac{\gamma}{2\pi}(P_1 + P_2 + P_3 + P_4) \tag{11}$$

where $P_1 = \mathbb{P}(D_1 \geq \sqrt{c}, D_2 \geq \sqrt{c})$, $P_2 = \mathbb{P}(D_1 \geq \sqrt{c}, D_2 \leq -\sqrt{c})$, $P_3 = \mathbb{P}(D_1 \leq -\sqrt{c}, D_2 \geq \sqrt{c})$, $P_4 = \mathbb{P}(D_1 \leq -\sqrt{c}, D_2 \leq -\sqrt{c})$ and $(D_1, D_2)^T$ has a bivariate normal distribution with mean $\mathbf{0}$, unit variances and correlation $-\cos(\gamma)$.

Now let $\mathcal{C}_i$ and $\mathcal{C}_j$ be adjacent cones with an angle $0 < \delta \leq \pi$ between their boundaries. By definition 1, if $0 < \delta \leq \pi/2$ then $\mathcal{R} = \mathcal{R}_{ij}$ and $\gamma = \delta$. However if $\pi/2 < \delta \leq \pi$ then $\mathcal{R} = \mathcal{R}'_{ij}$ and $\gamma = \pi - \delta$. Conditional on $\boldsymbol{S} \in \mathcal{R}$, in both cases we have

$$\chi^2_{1,1}(\gamma) = \min\{\mathrm{d}^2(\boldsymbol{S}, \mathrm{ray}(\boldsymbol{u}_j)), \mathrm{d}^2(\boldsymbol{S}, \mathrm{ray}(\boldsymbol{v}_i))\} = \min\{(\boldsymbol{u}_{j*}^T \boldsymbol{S})^2, (\boldsymbol{v}_{i*}^T \boldsymbol{S})^2\}.$$

Moreover in Lemma 2, $D_1 = \boldsymbol{u}_{j*}^T \boldsymbol{S}$ and $D_2 = \boldsymbol{v}_{i*}^T \boldsymbol{S}$ so the correlation coefficient between them is $-\cos(\gamma)$ if $0 < \delta \leq \pi/2$ and $\cos(\gamma)$ if $\pi/2 < \delta \leq \pi$.

Next we consider angles associated with the cones in (1) and the regions in (7)-(9). Let $\rho_1, \ldots, \rho_K$ denote the interior angles of the cones $\mathcal{C}_1, \ldots, \mathcal{C}_K$ and set $\rho = \sum_{i=1}^K \rho_i$. By assumption $\rho > 0$. The interior angles of the cones $\mathcal{R}_{ij}$, $\mathcal{R}'_{ij}$ and $\mathcal{R}''_{ij}$, as defined in (7), (8) and (9) respectively, are all denoted by γ_{ij}. Let $\mathcal{P}$ denote the set of indices (i, j) for all pairs of adjacent cones $\mathcal{C}_i$ and $\mathcal{C}_j$ except (p, q). It is clear that for all $(i, j) \in \mathcal{P}$ we have $0 < \gamma_{ij} \leq \pi/2$. Furthermore, if $\mathcal{R}''_{pq} \neq \emptyset$, then $0 < \gamma_{pq} < \pi$ denotes the interior angle of $\mathcal{R}''_{pq}$, otherwise $\gamma_{pq} = 0$. Following (8) let $\tau_i(\boldsymbol{v}_i)$ and $\tau_j(\boldsymbol{u}_j)$ be the interior angles of the cones $\mathcal{R}_i(\boldsymbol{v}_i)$ and $\mathcal{R}_j(\boldsymbol{u}_j)$ respectively. Similarly, if $\gamma_{pq} > 0$, *i.e.*, when (9) holds let $\tau_p(\boldsymbol{u}_p)$ and $\tau_q(\boldsymbol{v}_q)$ be the interior angles of $\mathcal{R}_p(\boldsymbol{u}_p)$ and $\mathcal{R}_q(\boldsymbol{v}_q)$ respectively. Finally set $\tau = \sum_{(i,j)\in\mathcal{P}}(\tau_i(\boldsymbol{v}_i)+\tau_j(\boldsymbol{u}_j))+\tau_p(\boldsymbol{u}_p)+\tau_q(\boldsymbol{v}_q)$. Note that $\tau_i(\boldsymbol{v}_i), \tau_j(\boldsymbol{u}_j) < \pi/2$ for all $(i, j) \in \mathcal{P}$. Moreover $\tau_p(\boldsymbol{u}_p) = \tau_q(\boldsymbol{v}_q) = \pi/2$. Also $\tau = 0$ if and only only if $\gamma_{ij} \leq \pi/2$ for all $(i, j) \in \mathcal{P}$ and $\gamma_{pq} = 0$. We have

$$\rho + \sum_{(i,j)\in\mathcal{P}} \gamma_{ij} + \tau + \gamma_{pq} = 2\pi. \tag{12}$$

We are now ready to state the main results of this section.

Theorem 3: Consider $\boldsymbol{\Theta}_0$ in (1) and the statistic T_n in (3). If $n \to \infty$, then we have

$$T_n \Rightarrow \begin{cases} \chi^2_0 & \text{if } \boldsymbol{\theta} \in \text{int}(\boldsymbol{\Theta}_0) \\ \dfrac{1}{2}\chi^2_0 + \dfrac{1}{2}\chi^2_1 & \text{if } \boldsymbol{\theta} \in \text{ray}(\boldsymbol{\Theta}_0) \\ \dfrac{\rho}{2\pi}\chi^2_0 + \displaystyle\sum_{(i,j)\in\mathcal{P}} \frac{\gamma_{ij}}{2\pi}\chi^2_{1,1}(\gamma_{ij}) + \frac{\tau}{2\pi}\chi^2_1 + \frac{\gamma_{pq}}{2\pi}\chi^2_2 & \text{if } \boldsymbol{\theta} = \mathbf{0} \end{cases} \tag{13}$$

where $\text{ray}(\boldsymbol{\Theta}_0)$ is the collection of all rays generating the cones in $\boldsymbol{\Theta}_0$.

Theorem 3 provides the limiting distribution of the LRT statistic for various values of $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. Let $T_{\boldsymbol{I}}$, $T_{\boldsymbol{R}}$ and $T_{\boldsymbol{O}}$ denote the limits of the LRT statistic when $\boldsymbol{\theta}$ is in the interior of $\boldsymbol{\Theta}_0$, on a ray of $\boldsymbol{\Theta}_0$ and the origin, respectively. For the form of the corresponding limits, see Equation (13) in the statement of Theorem 3. Clearly $T_{\boldsymbol{I}} \equiv 0$ so both $T_{\boldsymbol{O}}$ and $T_{\boldsymbol{R}}$ are stochastically larger than $T_{\boldsymbol{I}}$. It follows that the least favourable configuration and the limiting least favourable distribution are not associated with the interior points of $\boldsymbol{\Theta}_0$.

Remark 2: Note that if $\boldsymbol{S}_n$ is normally distributed then the distribution of T_n at $\boldsymbol{\theta} = \mathbf{0}$, which we denote by $T_{\boldsymbol{O}}$, is exact.

Suppose now that a size α test is desired. Let $c_{\alpha,\boldsymbol{R}}$ and $c_{\alpha,\boldsymbol{O}}$ denote the size α critical values associated with $T_{\boldsymbol{R}}$ and $T_{\boldsymbol{O}}$ respectively. These values solve the equations

$$\mathbb{P}(T_{\boldsymbol{R}} \geq c_{\alpha,\boldsymbol{R}}) = \alpha, \quad \text{and} \quad \mathbb{P}(T_{\boldsymbol{O}} \geq c_{\alpha,\boldsymbol{O}}) = \alpha.$$

Incidentally, it is easy to see that $c_{\alpha,\boldsymbol{R}}$ is equal to $(1 - 2\alpha)$-quantile of the χ^2_1 distribution whereas it may be necessary to compute $c_{\alpha,\boldsymbol{O}}$ numerically. Clearly the overall limiting critical value of the test is

$$c_\alpha = \min\{c_{\alpha,\boldsymbol{R}}, c_{\alpha,\boldsymbol{O}}\}.$$

In principle, finding the appropriate limiting critical value for any α is easy. It is worth noting that there are many cases in which we have either $c_{\alpha,\boldsymbol{R}} > c_{\alpha,\boldsymbol{O}}$ or $c_{\alpha,\boldsymbol{R}} < c_{\alpha,\boldsymbol{O}}$ for all

$0 \leq \alpha \leq 1$. The first situation arises when $T_{\boldsymbol{R}} \succeq_{\text{st}} T_{\boldsymbol{O}}$ whereas the second situation arises when $T_{\boldsymbol{R}} \preceq_{\text{st}} T_{\boldsymbol{O}}$. If either order relation holds then finding the limiting critical value is immediate. However, there are situations where an ordering does not exist, *i.e.*, $c_{\alpha,\boldsymbol{R}} > c_{\alpha,\boldsymbol{O}}$ for some values of α and $c_{\alpha,\boldsymbol{R}} < c_{\alpha,\boldsymbol{O}}$ for others. To summarize, for any $c \geq 0$:

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} \lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}}(T_n \geq c) = \begin{cases} \mathbb{P}(T_{\boldsymbol{R}} \geq c) & \text{if } T_{\boldsymbol{O}} \preceq_{\text{st}} T_{\boldsymbol{R}} \\ \mathbb{P}(T_{\boldsymbol{O}} \geq c) & \text{if } T_{\boldsymbol{R}} \preceq_{\text{st}} T_{\boldsymbol{O}} \\ \max\{\mathbb{P}(T_{\boldsymbol{R}} \geq c), \mathbb{P}(T_{\boldsymbol{O}} \geq c)\} & \text{otherwise} \end{cases}. \tag{14}$$

Equation (14) helps us to compute the limiting p–values associated with the test statistic (5). The value of $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ for which (14) holds is called the *least favourable limiting null value* of T_n. The distribution of the statistic T_n when $\boldsymbol{\theta}$ is the least favourable is called the *least favourable limiting null distribution*. In the first case of (14), any point on a ray in $\text{ray}(\boldsymbol{\Theta}_0)$ is the least favourable limiting null value of T_n and $\mathbb{P}(T_{\boldsymbol{R}} \geq c)$ is the least favourable limiting null distribution. In the second case, the origin is the least favourable limiting null value of T_n and $\mathbb{P}(T_{\boldsymbol{O}} \geq c)$ is the least favourable limiting null distribution. In the third case, their union is the least favourable limiting null value of T_n and $\max\{\mathbb{P}(T_{\boldsymbol{R}} \geq c), \mathbb{P}(T_{\boldsymbol{O}} \geq c)\}$ is the least favourable limiting null distribution. The next result shows that the least favourable limiting null distribution of the LRT statistic T_n is determined by the geometry of the cones.

Theorem 4: The least favourable limiting null distribution of (3) for testing (1) is that of $T_{\boldsymbol{O}}$ if and only if $\tau \geq \pi$ and that of $T_{\boldsymbol{R}}$ if and only if $\rho \geq \pi$.

Next we revisit the LRT and IUT for (1) in $\mathbb{R}^2$. By Theorem 1 the LRT and IUT statistics coincide if and only if all cones are congruent. As discussed earlier, the LRT rejects the null hypothesis if $T_n > c_\alpha$ where $c_\alpha = \min\{c_{\alpha,\boldsymbol{R}}, c_{\alpha,\boldsymbol{O}}\}$. The IUT rejects the null if and only if for all $i \in \{1, \ldots, K\}$ we find that $\Lambda^{(i)} > c_\alpha^{(i)}$ where $\Lambda^{(i)}$ is the LRT and $c_\alpha^{(i)}$ is the critical value for testing $H_0^{(i)} : \boldsymbol{\theta} \in \mathcal{C}_i$ against $H_1^{(i)} : \boldsymbol{\theta} \notin \mathcal{C}_i$. Thus the IUT combines K Type B problems in each of which the least favourable null value is the origin. Hence, we reject $H_0^{(i)}$ if $\Lambda^{(i)}$ is larger than the $1-\alpha$ quantile of the RV

$$\frac{\rho_i}{2\pi}\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{\pi - \rho_i}{2\pi}\chi_2^2.$$

For example, consider testing the hypotheses in (4). By Theorems 3 and 4, the null in (1) is rejected if T_n is larger than the $1-\alpha$ quantile of the RV $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$. Since the quadrants are congruent it is easy to see that the IUT rejects the null only if T_n is larger than $(1-\alpha)$ quantile of the RV $\frac{1}{4}\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{1}{4}\chi_2^2$, which is larger than the critical value of the LRT. The following result compares the LRT and the IUT for cones in two dimensions.

Theorem 5: The LRT for (1) is asymptotically uniformly more powerful than the IUT for cones in $\mathbb{R}^2$.

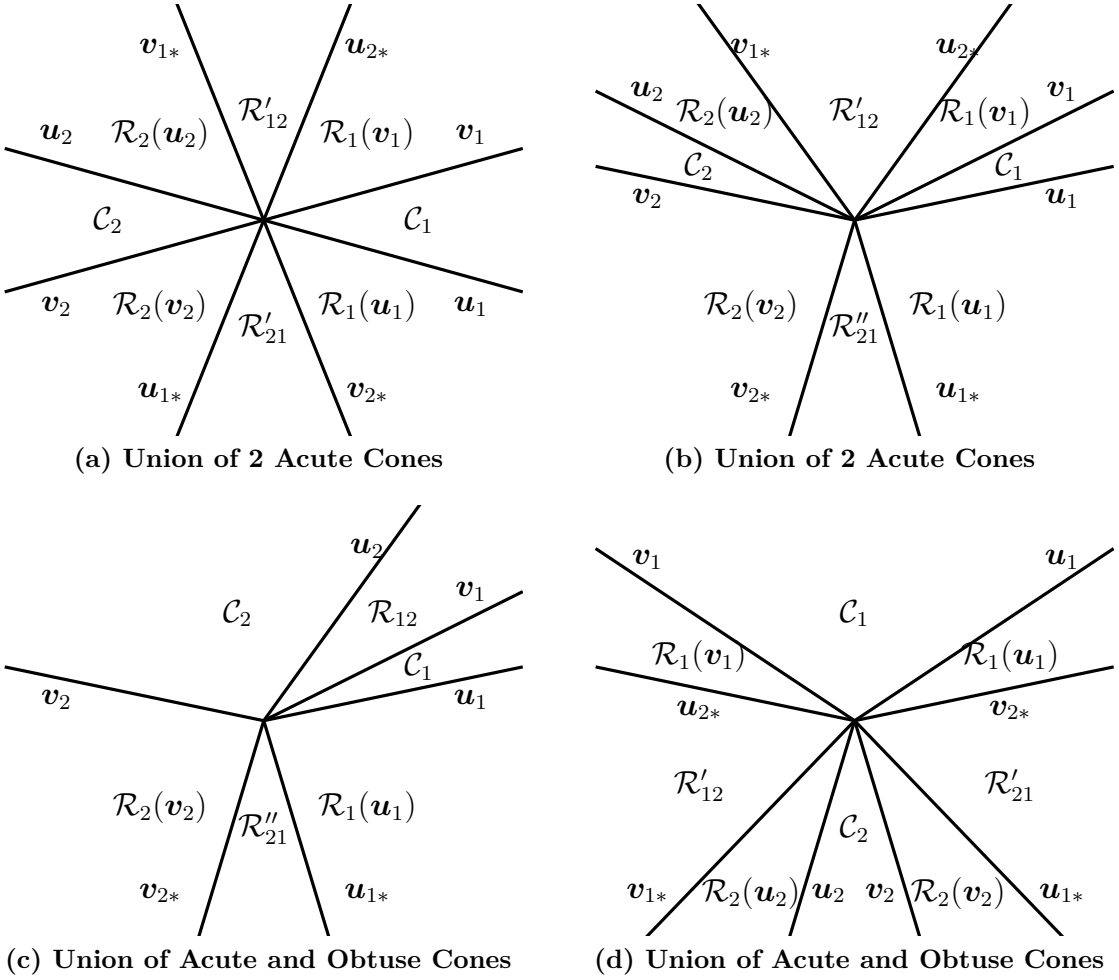
Numerical examples illustrating Theorem 5 are given in Section 6.

5. Some examples and testing problems in $\mathbb{R}^2$

We begin this section by providing some synthetic examples that exemplify our notations and illustrate the applications of Theorems 3 and 4. The synthetic examples are followed by examples of problems analyzed in the literature.

5.1. Synthetic examples

In Figure 1 several examples, depicting various geometric settings, are displayed.



(a) Union of 2 Acute Cones **(b) Union of 2 Acute Cones**

(c) Union of Acute and Obtuse Cones **(d) Union of Acute and Obtuse Cones**

Figure 1: Partition of $\mathbb{R}^2$ by cones $\mathcal{C}_1$ and $\mathcal{C}_2$

Example 1: (**Union of Acute Cones I, Figure 1(A)**): Here $\mathcal{R}''_{21} = \emptyset$ so $\gamma_{21} = 0$. The interior angles ρ_1 and ρ_2 are both smaller than $\pi/2$ so $\rho < \pi$. If $\tau \geq \pi$, then by Theorem 4 the least favourable limiting null value of T_n is the origin and the least favourable limiting null distribution is that of $T_{\boldsymbol{O}}$. However if $\tau < \pi$ then we have an indeterminate case where there is no stochastic ordering between $T_{\boldsymbol{O}}$ and $T_{\boldsymbol{R}}$.

Example 2: (**Union of Acute Cones II, Figure 1(B)**): Here $\mathcal{R}''_{21} \neq \emptyset$ so $\gamma_{21} > 0$ and again $\rho < \pi$. Moreover $\tau_1(\boldsymbol{u}_1) = \tau_2(\boldsymbol{v}_2) = \pi/2$ so it is clear that $\tau > \pi$. Hence by Theorem 4, the least favourable limiting null value of T_n is the origin and the least favourable limiting null distribution is that of $T_{\boldsymbol{O}}$.

Example 3: (**Union of Acute and Obtuse Cones I, Figure 1(C)**): Here $\mathcal{R}''_{21} \neq \emptyset$ so $\gamma_{21} > 0$ and also $\rho < \pi$. Moreover $\tau_1(\boldsymbol{u}_1) = \tau_2(\boldsymbol{v}_2) = \pi/2$ so $\tau = \pi$. Hence by Theorem 4, the least favourable limiting null value of T_n is the origin and the least favourable limiting null distribution is that of $T_{\boldsymbol{O}}$.

Example 4: (**Union of Acute and Obtuse Cones II, Figure 1(D)**): Here $\mathcal{R}''_{21} = \emptyset$ so $\gamma_{21} = 0$. Let $\pi/2 < \rho_1 < \pi$ and $\rho_2 < \pi/2$ denote the interior angles. If $\rho \geq \pi$, then by Theorem 4, the least favourable limiting null value of T_n lies in ray($\boldsymbol{\Theta}_0$) and the least favourable limiting null distribution is that of $T_{\boldsymbol{R}}$. However if $\rho < \pi$, then we have an indeterminate case where there is no stochastic ordering between $T_{\boldsymbol{O}}$ and $T_{\boldsymbol{R}}$.

Next consider the case where $K \geq 2$. In particular, suppose each cone has interior angle η. Suppose further that angles between all adjacent cones are also equal. It follows that $\rho = K\eta$ and the angle between adjacent cones is $(2\pi - K\eta)/K$. Of course, it is assumed that $0 < \rho < 2\pi$. Clearly the angle between the adjacent cones is always smaller than π and therefore $\gamma_{pq} = 0$. Further note that τ is positive if and only if

$$\eta < \frac{2\pi}{K} - \frac{\pi}{2} \quad \text{and} \quad K \in \{2, 3\}. \tag{15}$$

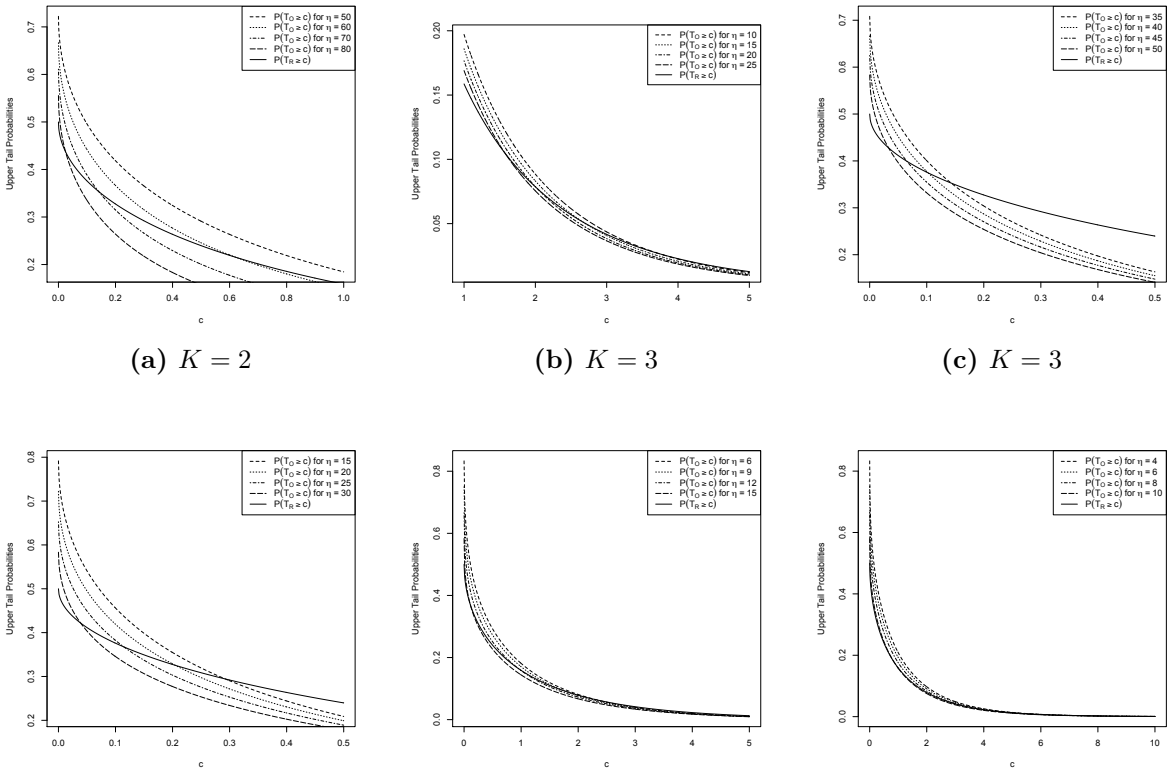
Therefore if τ is positive, then $\tau = 2(\pi - 2\eta)$ when $K = 2$ and $\tau = \pi - 6\eta$ when $K = 3$. By Theorem 3 the limiting null distribution of T_n at the origin, *i.e.*, $T_{\boldsymbol{O}}$ is

$$T_{\boldsymbol{O}} \stackrel{d}{=} \begin{cases} \frac{\eta}{\pi}\chi^2_0 + \frac{\eta}{\pi}\chi^2_{1,1}(\eta) + \frac{\pi - 2\eta}{\pi}\chi^2_1 & \text{if } K = 2 \text{ and } \eta < \frac{\pi}{2} \\ \frac{3\eta}{2\pi}\chi^2_0 + \frac{3\eta + \pi}{2\pi}\chi^2_{1,1}(\eta + \pi/3) + \frac{\pi - 6\eta}{2\pi}\chi^2_1 & \text{if } K = 3 \text{ and } \eta < \frac{\pi}{6}\,. \\ \frac{K\eta}{2\pi}\chi^2_0 + \frac{2\pi - K\eta}{2\pi}\chi^2_{1,1}(2\pi/K - \eta) & \text{otherwise} \end{cases} \tag{16}$$

Now we investigate the relationship between $T_{\boldsymbol{O}}$ given in (16) and $T_{\boldsymbol{R}}$ as given in (13). If $\pi/K \leq \eta < 2\pi/K$ then $\rho \geq \pi$ so by Theorem 4, the least favourable limiting null distribution of T_n is that of $T_{\boldsymbol{R}}$. Further note that $\tau \geq \pi$ if and only if $\eta \leq \pi/4$ and $K = 2$ in which case the least favourable limiting null distribution of T_n is that of $T_{\boldsymbol{O}}$. For any other values of η and K, Theorem 4 can not be used to identify the least favourable null distribution and the corresponding size α critical value so this determination must be made numerically as illustrated in Figure 2.

Figure 2(A)-(F) present plots of the tail probabilities $\mathbb{P}(T_{\boldsymbol{O}} \geq c)$ and $\mathbb{P}(T_{\boldsymbol{R}} \geq c)$ for various values of $c > 0$ and choices of K and η. In particular K and η were chosen so $\rho < \pi$. In addition $\tau < \pi$ in Figures 2(A) and 2(B) whereas $\tau = 0$ in Figures 2(C)-(F). It is clear that for the above choices the tail probabilities cross and consequently the RVs $T_{\boldsymbol{R}}$ and $T_{\boldsymbol{O}}$ are not stochastically ordered. In other words, the least favourable limiting null distribution of T_n is not the same for all size α critical values. It is also clear that whenever the tail probabilities of $T_{\boldsymbol{R}}$ and $T_{\boldsymbol{O}}$ cross then there exists a unique value c^* satisfying $\mathbb{P}(T_{\boldsymbol{O}} \geq c) \geq \mathbb{P}(T_{\boldsymbol{R}} \geq c)$ for all $c \leq c^*$ whereas $\mathbb{P}(T_{\boldsymbol{R}} \geq c) \geq \mathbb{P}(T_{\boldsymbol{O}} \geq c)$ for all $c \geq c^*$. Moreover for a fixed K, c^* is monotonically decreasing in η. Similarly if η is fixed, then c^* is monotonically decreasing in K. In the majority of cases plotted we find that $c_\alpha = c_{\alpha,\boldsymbol{R}}$.

Remark 3: Note that when $\pi/2 < \eta < \pi$, the cones are obtuse and $K \leq 3$. Thus the intersection of their polar cones contains only the origin and hence $\gamma_{pq} = 0$. Since $\rho > \pi$, by Theorem 4 the least favourable limiting null values of T_n lie in ray($\boldsymbol{\Theta}_0$) and the least favourable limiting null distribution is that of $T_{\boldsymbol{R}}$ which is the same as that for testing over union of two quadrants. Note that the null distribution of T_n at the origin is a function of dependent χ^2_1 RVs in the first case but a function of independent χ^2_1 RVs in the latter case.



(a) $K = 2$ **(b)** $K = 3$ **(c)** $K = 3$

(d) $K = 5$ **(e)** $K = 10$ **(f)** $K = 15$

Figure 2: Stochastic ordering between T_O and T_R for various no. of cones (K) and values of interior angles (η)

5.2. Examples from the literature

We conclude this section by providing some examples of relevant testing problems in $\mathbb{R}^2$ which have appeared in the literature. These examples illustrate the application of Theorems 3–5 and show the simplicity and superiority (in power) of our testing procedure as compared to those in the literature.

Example 5: Consider first testing the hypotheses $H_0 : \min\{|\theta_1|, |\theta_2|\} = 0$ against $H_1 : \min\{|\theta_1|, |\theta_2|\} > 0$. Variants of this problem have been studied by Cohen *et al.* (1983) and Berger (1997) where the IUT had been advocated. Note that by setting $\mathcal{C}_1 = \{\boldsymbol{\theta} : \theta_1 \geq 0, \theta_2 = 0\}$, $\mathcal{C}_2 = \{\boldsymbol{\theta} : \theta_1 = 0, \theta_2 \geq 0\}$, $\mathcal{C}_3 = \{\boldsymbol{\theta} : \theta_1 \leq 0, \theta_2 = 0\}$ and $\mathcal{C}_4 = \{\boldsymbol{\theta} : \theta_1 = 0, \theta_2 \leq 0\}$ we can reformulate the problem as in (1). Next, it is clear that the statistic (3) reduces to $T_n = \min\{S^2_{n,1}, S^2_{n,2}\}$. The least favourable null value and distribution are $\boldsymbol{\theta} = \mathbf{0}$ and $\chi^2_{1,1}(\pi/2)$ respectively and the critical value is the $\sqrt{1-\alpha}$ quantile of a χ^2_1 RV. Since the cones are congruent, T_n is the same as the IUT by Theorem 1 but more powerful than the IUT by Theorem 5.

Example 6: Laska and Meisner (1989) tested (1) with $\mathcal{C}_i = \{\boldsymbol{\theta} \in \mathbb{R}^m : \theta_i \leq 0\}$. In their formulation $\theta_i = \mu_0 - \mu_i$ where μ_0 is the mean response under treatment $\mathcal{T}_0$ and μ_i is the mean response under treatment $\mathcal{T}_i$. Thus under the null some treatments are superior to $\mathcal{T}_0$

whereas under the alternative all treatments are inferior to $\mathcal{T}_0$. This problem is known in the literature as the *sign testing problem* and has received considerable attention (e.g., Berger (1982) and Cohen *et al.* (1983)). It is easy to verify that when $m = 2$ this testing problem can be reformulated as $H_0 : \boldsymbol{\theta} \in \mathcal{Q}_2 \cup \mathcal{Q}_3 \cup \mathcal{Q}_4$ and $H_1 : \boldsymbol{\theta} \in \mathcal{Q}_1$ where $\mathcal{Q}_1, \ldots, \mathcal{Q}_4$ denote the quadrants of $\mathbb{R}^2$ in clockwise direction. Interestingly, this problem is the complement of a Type B problem since $\boldsymbol{\Theta}_1$ is a single convex cone. Berger (1982) proposed testing H_0 against H_1 using the IUT. The LRT for this problem is $T_n = \min\{S^2_{n,1}, S^2_{n,2}\}\mathbb{I}(\boldsymbol{S}_n \in \mathcal{Q}_1)$. Since $\rho = 3\pi/2$, by Theorem 4 the least favourable limiting null values of T_n are in ray$(\boldsymbol{\Theta}_0)$, *i.e.*, the rays defining the first quadrant, and the least favourable limiting null distribution is $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$. By Theorem 5, the proposed test is more powerful than the IUT although they are identical by Theorem 1 due to congruence of the cones.

Example 7: Gail and Simon (1985) as well as Silvapulle (2001) tested (1) for $K = 2$ where $\mathcal{C}_1 = \{\boldsymbol{\theta} \in \mathbb{R}^m : \boldsymbol{\theta} \geq \mathbf{0}\}$ and $\mathcal{C}_2 = \{\boldsymbol{\theta} \in \mathbb{R}^m : \boldsymbol{\theta} \leq \mathbf{0}\}$. Here θ_i is the difference between the mean responses to treatments $\mathcal{T}_1$ and $\mathcal{T}_2$, say, in the i^{th} group where $i = 1, \ldots, m$. If $\mathcal{T}_1$ is more beneficial than $\mathcal{T}_2$ ($\theta_i \geq 0$) in some groups but more harmful than $\mathcal{T}_2$ ($\theta_i \leq 0$) in others, it is said that there is *crossover interaction* between treatments and groups. Thus under the null there is no crossover interaction whereas under the alternative, there is such interaction. The hypotheses of interest in $\mathbb{R}^2$ are $H_0 : \boldsymbol{\theta} \in \mathcal{C}_1 \cup \mathcal{C}_2$ and $H_1 : \boldsymbol{\theta} \notin \mathcal{C}_1 \cup \mathcal{C}_2$ where $\mathcal{C}_1$ and $\mathcal{C}_2$ are the non-negative and the non-positive quadrants. By Theorem 2, the test statistic is given by $T_n = \min\{S^2_{n,1}, S^2_{n,2}\}\mathbb{I}(\boldsymbol{S}_n \notin \mathcal{C}_1 \cup \mathcal{C}_2)$. Since $\rho = \pi$, by Theorem 4 the least favourable limiting null values of T_n lie in ray$(\boldsymbol{\Theta}_0)$ and the least favourable limiting null distribution is $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$. Gail and Simon (1985) and Silvapulle (2001) did not assume that the variance is known but their statistic is of the same form as the LRT T_n.

Example 8: Berger (1989) and Liu and Berger (1995) tested (1) with $\mathcal{C}_i = \{\boldsymbol{\theta} \in \mathbb{R}^m : \boldsymbol{b}_i^T\boldsymbol{\theta} \leq 0\}$ where $\boldsymbol{\theta}$ is the mean vector of a multivariate normal distribution and $\boldsymbol{b}_i$s are non-redundant vectors. If $\boldsymbol{\theta}$ denotes a vector of means then under the null some linear combinations, e.g., contrasts, are negative whereas under the alternative, all linear combinations are positive. Consider the above problem in $\mathbb{R}^2$ where $\mathcal{C}_i = \{\boldsymbol{\theta} \in \mathbb{R}^2 : \boldsymbol{b}_i^T\boldsymbol{\theta} \leq 0\}$ (a half-space), $\boldsymbol{\theta}$ is the mean vector of a bivariate normal distribution and $\boldsymbol{b}_i$s are non-redundant vectors. Here $\boldsymbol{\Theta}_0$ is a union of multiple convex cones whereas $\boldsymbol{\Theta}_1$ is a single convex cone, which is the complement of a Type B problem. Berger (1989) and Liu and Berger (1995) applied the IUT to this problem. The LRT statistic T_n is given by (3). Since $\rho > \pi$, by Theorem 4 the least favourable asymptotic null values of T_n lie in ray$(\boldsymbol{\Theta}_0)$ and the least favourable asymptotic null distribution is $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$. By Theorem 5, the proposed test based on T_n is uniformly more powerful than the IUT.

6. Simulation study

We performed a small simulation study comparing the power of the LRT to that of the IUT. We considered $K = 2$ cones and both congruent (C) as well as non-congruent (NC) pairs of cones. See Table 1 for the settings of the study. We fixed $n = 100$ and $\alpha = 0.05$. The critical values were computed by simulation at the least favourable null values. For computing power, the point in the alternative for union of quadrants is of the form (θ_1, θ_2) where θ_1 and θ_2 have different signs; otherwise it is of the form $(0, \theta_2)$ where $\theta_2 > 0$. These points were chosen so that the LRT has a power of around 0.8. From Table 2 it is observed that in each of the settings in Table 1 the LRT is more powerful than the IUT for congruent

Table 1: Congruent and Non-congruent pairs of cones under various settings

Settings	Geometry	Type	Cones
1	$\rho = \pi$	Congruent	$\{\theta_1 \geq 0, \theta_2 \geq 0\}$,$\{\theta_1 \leq 0, \theta_2 \leq 0\}$
2	$\rho = \pi$	Non-congruent	$\{\theta_2 \geq -\sqrt{3}\theta_1, \theta_2 \leq \sqrt{3}\theta_1\}$,$\{\sqrt{3}\theta_2 \geq \theta_1, \sqrt{3}\theta_2 \leq -\theta_1\}$
3	$\rho > \pi$	Congruent	$\{\theta_2 \geq -\theta_1, \theta_2 \leq 2\theta_1\}$,$\{\theta_2 \geq \theta_1, \theta_2 \leq -2\theta_1\}$
4	$\rho > \pi$	Non-congruent	$\{\theta_2 \geq -\theta_1, \theta_2 \leq 3\theta_1\}$,$\{\theta_2 \geq \theta_1, \theta_2 \leq -2\theta_1\}$
5	$\tau = \pi$	Congruent	$\{\theta_2 \geq \theta_1, \theta_2 \leq 2\theta_1\}$,$\{\theta_2 \geq -\theta_1, \theta_2 \leq -2\theta_1\}$
6	$\tau = \pi$	Non-congruent	$\{\theta_2 \geq \theta_1, \theta_2 \leq 2.1\theta_1\}$, $\{\theta_2 \geq -\theta_1, \theta_2 \leq -2\theta_1\}$
7	$\tau > \pi$	Congruent	$\{4\theta_2 \geq \theta_1, 3\theta_2 \leq \theta_1\}$,$\{4\theta_2 \geq -\theta_1, 3\theta_2 \leq -\theta_1\}$
8	$\tau > \pi$	Non-congruent	$\{4\theta_2 \geq \theta_1, 3\theta_2 \leq \theta_1\}$,$\{4\theta_2 \geq -\theta_1, 2\theta_2 \leq -\theta_1\}$

as well as non-congruent pairs of cones. Although not reported in Table 2, it is observed that the powers of the LRT and the IUT decrease or increase as $\boldsymbol{\theta}$ is closer to or further from $\mathbf{0}$. Moreover the ratio of the power of the LRT to that of the IUT increases or decreases as $\boldsymbol{\theta}$ is closer to or further from $\mathbf{0}$, and equals 1 for large $\boldsymbol{\theta}$.

Table 2: Powers of LRT ($\mathbf{P_{LRT}}$) and IUT ($\mathbf{P_{IUT}}$) under settings in Table 1 for $\alpha = 0.05$ and selected $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$

Settings	$\boldsymbol{\theta}$	P_{LRT}	P_{IUT}
1	(-0.29,0.29)	0.8009	0.6814
2	(0,0.5)	0.8014	0.6988
3	(0,0.66)	0.8084	0.6623
4	(0,0.82)	0.8069	0.6863
5	(0,0.75)	0.8042	0.7367
6	(0,0.76)	0.7989	0.7299
7	(0,0.32)	0.8080	0.7023
8	(0,0.33)	0.8021	0.6894

7. Discussion

As noted in the introduction, the existing literature has focused on testing (1) in situations where the null parameter space is either a linear subspace or single convex cone. In this paper, we develop a general framework to address multiple cone problems in two dimensions. We consider situations where the null parameter space can be expressed as the union of multiple closed convex cones in $\mathbb{R}^2$, which encompasses a large class of problems. We propose a test statistic which is equivalent to the LRT under normality and coincides with the the IUT in some special cases. Since the finite sampling distributions of these test statistics usually do not have closed-form expressions, we derive their asymptotic null distributions. We also obtain their least favourable asymptotic null values and the corresponding

least favourable asymptotic null distributions based on the geometry of the cones. These distributions are used to determine the size α critical values, which depend on the stochastic ordering of the test statistics. Finally, we show that our tests are uniformly more powerful than the conventional IUTs discussed in the literature. In future, we hope to address the more challenging problem of testing (1) for arbitrary convex cones in any finite dimension.

In fact the scope of (1) is much broader than the references in Sections 1 and 3. Some important classes of problems which can be formulated as (1) in higher dimensions include problems of model selection arising in order restricted inference (Mack and Wolfe (1981), Pan and Wolfe (1996), Pan (1997), Rueda *et al.* (2016), Wei *et al.* (2019), Panda (2019), Larriba *et al.* (2016), Larriba *et al.* (2020), Peddada *et al.* (2003), Peddada *et al.* (2005)); problems in the theory of ranking and selection; and problems in mathematical psychology which involve the verification of transitivity axioms underlying social choice theory (Oliveira *et al.* (2018), Iverson and Falmagne (1985), Tversky (1969), Regenwetter *et al.* (2011), Davis-Stober (2009), Myung *et al.* (2005), Heck and Davis-Stober (2019)). For example, in the theory of ranking and selection, Nettleton (2009) considered the problem of testing for the supremacy of a multinomial cell probability. The supremacy of the K^{th} cell probability is established by rejecting the null hypothesis $H_0 : \theta_K \leq \max\{\theta_1, \ldots, \theta_{K-1}\}$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^T$ denotes the vector of multinomial cell probabilities. Clearly the null can be rewritten as $H_0 : \boldsymbol{\theta} \in \bigcup_{j=1}^{K-1} \mathcal{C}_j$, where $\mathcal{C}_j = \{\boldsymbol{\theta} \in \mathrm{P} : \theta_K \leq \theta_j\}$ and P is the set of K-dimensional probability vectors whose components sum to 1.

Acknowledgments

The work of Sayan Ghosh was conducted while being a post doctoral fellow at the University of Haifa. The work of Ori Davidov was partially supported by the Israel Science Foundation Grants No. 456/17 and 2200/22 and gratefully acknowledged.

References

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, **24**, 295–300.

Berger, R. L. (1989). Uniformly more powerful tests for hypotheses concerning linear inequalities and normal means. *Journal of the American Statistical Association*, **84**, 192–199.

Berger, R. L. (1997). *Advances in Statistical Decision Theory and Applications*. Birkhäuser, Boston.

Berger, R. L. and Sinclair, D. F. (1984). Testing hypotheses concerning unions of linear subspaces. *Journal of the American Statistical Association*, **79**, 158–163.

Cohen, A., Gatsonis, C., and Marden, J. I. (1983). *Hypothesis Tests and Optimality Properties in Discrete Multivariate Analysis*. Academic Press, New York.

Davis-Stober, C. P. (2009). Analysis of multinomial models under inequality constraints: applications to measurement theory. *Journal of Mathematical Psychology*, **53**, 1–13.

Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, **41**, 361–372.

Heck, D. W. and Davis-Stober, C. P. (2019). Multinomial models with linear inequality constraints: Overview and improvements of computational methods for Bayesian inference. *Journal of Mathematical Psychology*, **91**, 70–87.

Iverson, G. and Falmagne, J. C. (1985). Statistical issues in measurement. *Mathematical Social Sciences*, **10**, 131–153.

Larriba, Y., Rueda, C., Fernández, M., and Peddada, S. (2016). Order Restricted Inference for Oscillatory Systems for Detecting Rhythmic Signals. *Nucleic Acids Research*, **44**, 1–8.

Larriba, Y., Rueda, C., Fernández, M., and Peddada, S. (2020). Order restricted inference in chronobiology. *Statistics in Medicine*, **39**, 265–278.

Laska, E. M. and Meisner, M. J. (1989). Testing whether identified treatment is best. *Biometrics*, **45**, 1139–1151.

Liu, H. and Berger, R. L. (1995). Uniformly more powerful, one-sided tests for hypotheses about linear inequalities. *Annals of Statistics*, **23**, 55–72.

Mack, G. A. and Wolfe, D. A. (1981). K-Sample Rank Tests for Umbrella Alternative. *Journal of the American Statistical Association*, **76**, 175–181.

Myung, J. I., Karabatsos, G., and Iverson, G. J. (2005). A Bayesian approach to testing decision making axioms. *Journal of Mathematical Psychology*, **49**, 205–225.

Nettleton, D. (2009). Testing for the supremacy of a Multinomial cell probability. *Journal of the American Statistical Association*, **104**, 1052–1059.

Oliveira, I. F. D., Ailon, N., and Davidov, O. (2018). A new and flexible approach to the analysis of paired comparison data . *Journal of Machine Learning Research*, **19**, 1–29.

Pan, G. (1997). Confidence subset containing the unknown peaks of an umbrella ordering. *Journal of the American Statistical Association*, **92**, 307–314.

Pan, G. and Wolfe, D. A. (1996). Comparing groups with umbrella orderings. *Journal of the American Statistical Association*, **91**, 311–317.

Panda, S. (2019). The arrival of circadian medicine. *Nature Reviews Endocrinology*, **15**, 67–69.

Peddada, S., Harris, S., Zajd, J., and Harvey, E. (2005). ORIOGEN: order restricted inference for ordered gene expression data. *Bioinformatics*, **21**, 3933–3934.

Peddada, S., Lobenhofer, E. K., Li, L., Afshari, C. A., Weinberg, C. R., and Umbach, D. M. (2003). Gene selection and clustering for time-courseand dose–response microarray experimentsusing order-restricted inference. *Bioinformatics*, **19**, 834–841.

Regenwetter, M., Dana, J., and Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, **118**, 42–56.

Robertson, T. and Wegman, E. J. (1978). Likelihood ratio tests for order restrictions in exponential families. *Annals of Statistics*, **6**, 485–505.

Rueda, C., Ugarte, M. D., and Militino, A. F. (2016). Checking unimodality using isotonic regression: an application to breast cancer mortality rates. *Stochastic Environmental Research and Risk Assessment*, **30**, 1277–1288.

Saumard, A. and Wellner, J. A. (2014). Log-concavity and strong log-concavity: A review. *Statistics Surveys*, **8**, 45–114.

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic Orders*. Springer, New York.

Silvapulle, M. J. (2001). Tests against qualitative interaction: Exact critical values and robust tests. *Biometrics*, **57**, 1157–1165.

Silvapulle, M. J. and Sen, P. K. (2004). *Constrained Statistical Inference*. Wiley, New York.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, **76**, 31–48.

Wei, Y., Wainwright, M. J., and Guntuboyina, A. (2019). The geometry of hypothesis testing over convex cones: generalized likelihood ratio tests and minimax radii. *Annals of Statistics*, **47**, 994–1024.

APPENDIX

PROOFS

Proof of Theorem 1:

Proof: Since $\boldsymbol{S} \sim \mathcal{N}_2(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, the kernel of the log-likelihood is given by

$$l(\boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{S}-\boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{S}-\boldsymbol{\theta}) = -\frac{1}{2}\|\boldsymbol{S}-\boldsymbol{\theta}\|^2_{\boldsymbol{\Sigma}}. \tag{17}$$

It follows that the global, unrestricted MLE of $\boldsymbol{\theta}$ is $\widehat{\boldsymbol{\theta}} = \boldsymbol{S}$ and

$$l(\widehat{\boldsymbol{\theta}}) = 0. \tag{18}$$

The restricted MLE solves

$$\begin{aligned}\widetilde{\boldsymbol{\theta}} &= \arg\max\{l(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \bigcup_{i=1}^{K} \mathcal{C}_i\} = \arg\min\{\|\boldsymbol{S}-\boldsymbol{\theta}\|^2_{\boldsymbol{\Sigma}} : \boldsymbol{\theta} \in \bigcup_{i=1}^{K} \mathcal{C}_i\} \\ &= \arg\min\{\|\boldsymbol{S}-\boldsymbol{\theta}\|^2_{\boldsymbol{\Sigma}} : \boldsymbol{\theta} \in \{\widetilde{\boldsymbol{\theta}}_1, \ldots, \widetilde{\boldsymbol{\theta}}_K\}\}\end{aligned}$$

where for $i = 1, \ldots, K$ we define $\widetilde{\boldsymbol{\theta}}_i = \arg\min\{\|\boldsymbol{S}-\boldsymbol{\theta}\|^2_{\boldsymbol{\Sigma}} : \boldsymbol{\theta} \in \mathcal{C}_i\}$ which is nothing but the projection of $\boldsymbol{S}$ on $\mathcal{C}_i$ with respect to $\boldsymbol{\Sigma}$ denoted by $\Pi_{\boldsymbol{\Sigma}}(\boldsymbol{S}_n \mid \mathcal{C}_i)$. In other words

$$\widetilde{\boldsymbol{\theta}} = \arg\min\{\|\boldsymbol{S} - \Pi_{\boldsymbol{\Sigma}}(\boldsymbol{S} \mid \mathcal{C}_i)\|^2_{\boldsymbol{\Sigma}} : i \in \{1, \ldots, K\}\}, \tag{19}$$

so

$$l(\widetilde{\boldsymbol{\theta}}) = -\frac{1}{2}\min\{\|\boldsymbol{S} - \Pi_{\boldsymbol{\Sigma}}(\boldsymbol{S} \mid \mathcal{C}_i)\|^2_{\boldsymbol{\Sigma}} : i \in \{1, \ldots, K\}\}. \tag{20}$$

Now the LRT statistic is given by

$$\Lambda = 2\{l(\widehat{\boldsymbol{\theta}}) - l(\widetilde{\boldsymbol{\theta}})\}$$

which, using (18) and (20), reduces to

$$\Lambda = \min\{\|\boldsymbol{S} - \Pi_{\boldsymbol{\Sigma}}(\boldsymbol{S} \mid \mathcal{C}_i)\|^2_{\boldsymbol{\Sigma}} : i \in \{1, \ldots, K\}\} \tag{21}$$

as claimed in (21) with $\boldsymbol{S}_n$ replaced by $\boldsymbol{S}$.

Now note that testing (1) is equivalent to testing $\bigcup_{i=1}^{K} H_0^{(i)}$ versus $\bigcap_{i=1}^{K} H_1^{(i)}$ where $H_0^{(i)}$: $\boldsymbol{\theta} \in \mathcal{C}_i$ and $H_1^{(i)} : \boldsymbol{\theta} \notin \mathcal{C}_i$. It is clear that the LRT statistic for individual tests $H_0^{(i)}$ versus $H_1^{(i)}$, each of which is a Type B problem (Silvapulle and Sen (2004)), is $\Lambda^{(i)} = \|\boldsymbol{S} - \tilde{\boldsymbol{\theta}}_i\|^2 = \|\boldsymbol{S} - \Pi_{\boldsymbol{\Sigma}}(\boldsymbol{S} \mid \mathcal{C}_i)\|^2$. Since the least favourable null value for any Type B problem is the origin $H_0^{(i)}$ is rejected if $\Lambda^{(i)}$ is larger than $c_\alpha^{(i)}$, the $1-\alpha$ quantile of the RV

$$\sum_{k=1}^{K} w_k(\mathcal{C}_i, \boldsymbol{\Sigma})\chi_k^2. \tag{22}$$

The IUT rejects the null in (1) if and only if $\Lambda^{(i)} > c_\alpha^{(i)}$ for every i. Note that $c_\alpha^{(i)} = c_\alpha^{(j)}$ if and only if the weights in (22) satisfy $w_k(\mathcal{C}_i, \boldsymbol{\Sigma}) = w_k(\mathcal{C}_j, \boldsymbol{\Sigma})$ for all $k = 1, \ldots, K$ or in other words that all cones are congruent. Since the critical values $c_\alpha^{(i)}$ are equal for each of the K tests, it follows that the IUT statistic is

$$\min\{\Lambda^{(1)}, \ldots, \Lambda^{(K)}\},$$

which is the same as the LRT statistic in (21) as a function of $\boldsymbol{S}$. □

Proof of Theorem 2:

Proof: By Theorem 1 the LRT statistic for (4) is (3) which reduces to

$$T = \min\{\|\boldsymbol{S} - \Pi(\boldsymbol{S} \mid \mathcal{C}_1)\|_2^2, \|\boldsymbol{S} - \Pi(\boldsymbol{S} \mid \mathcal{C}_2)\|_2^2\}. \tag{23}$$

Note that when $\boldsymbol{S} \in \mathcal{C}_1$ then $\Pi(\boldsymbol{S} \mid \mathcal{C}_1) = \boldsymbol{S}$ so $\|\boldsymbol{S} - \Pi(\boldsymbol{S} \mid \mathcal{C}_1)\|_2^2 = 0$ and similarly when $\boldsymbol{S} \in \mathcal{C}_2$. Thus if $\boldsymbol{S} \in \mathcal{C}_1 \cup \mathcal{C}_2$ then $T = 0$. Next, if $\boldsymbol{S} \notin \mathcal{C}_2 \cup \mathcal{C}_2$ then for $i \in \{1, 2\}$, $\Pi(\boldsymbol{S} \mid \mathcal{C}_i) = (S_1, 0)^T$ or $(0, S_2)^T$ so $\|\boldsymbol{S} - \Pi(\boldsymbol{S} \mid \mathcal{C}_i)\|_2^2 = S_1^2$ or S_2^2. Consequently,

$$T = \min\{S_1^2, S_2^2\}\mathbb{I}(\boldsymbol{S} \notin \mathcal{C}_1 \cup \mathcal{C}_2)$$

as claimed. Let $c > 0$. Then for any $\boldsymbol{\theta} \in \mathbb{R}^2$ we have

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\theta}}(T \geq c) &= \mathbb{P}_{\boldsymbol{\theta}}(\min\{S_1^2, S_2^2\} \geq c, \boldsymbol{S} \notin \mathcal{C}_1 \cup \mathcal{C}_2) \\
&= \mathbb{P}_{\boldsymbol{\theta}}(S_1^2 \geq c, S_2^2 \geq c, S_1 \geq 0, S_2 \leq 0) + \mathbb{P}_{\boldsymbol{\theta}}(S_1^2 \geq c, S_2^2 \geq c, S_1 \leq 0, S_2 \geq 0) \\
&= \mathbb{P}_{\boldsymbol{\theta}}(S_1 \geq \sqrt{c})\mathbb{P}_{\boldsymbol{\theta}}(S_2 \leq -\sqrt{c}) + \mathbb{P}_{\boldsymbol{\theta}}(S_1 \leq -\sqrt{c})\mathbb{P}_{\boldsymbol{\theta}}(S_2 \geq \sqrt{c}) \\
&= [1 - \Phi(\sqrt{c} - \theta_1)]\Phi(-\sqrt{c} - \theta_2) + \Phi(-\sqrt{c} - \theta_1)[1 - \Phi(\sqrt{c} - \theta_2)].
\end{aligned} \tag{24}$$

Here, as usual, $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and distribution function of a standard normal RV. Denote $\mathbb{P}_{\boldsymbol{\theta}}(T \geq c)$ by $H(\boldsymbol{\theta}) = H(\theta_1, \theta_2)$. Our goal is to maximize $H(\theta_1, \theta_2)$ over $(\theta_1, \theta_2) \in \boldsymbol{\Theta}_0$. We will first show that if $\boldsymbol{\theta} \in \mathcal{C}_1$ and $\theta_1 > \theta_2 > 0$ then

$$H(\theta_1, \theta_2) < H(\theta_1, 0). \tag{25}$$

Since $\theta_1 > \theta_2 > 0$, it follows that

$$\begin{aligned}
H(\theta_1, 0) - H(\theta_1, \theta_2) &= [1 - \Phi(\sqrt{c} - \theta_1)][\Phi(-\sqrt{c}) - \Phi(-\sqrt{c} - \theta_2)] \\
&\quad - \Phi(-\sqrt{c} - \theta_1)[\Phi(\sqrt{c}) - \Phi(\sqrt{c} - \theta_2)] \\
&> [1 - \Phi(\sqrt{c} - \theta_2)][\Phi(-\sqrt{c}) - \Phi(-\sqrt{c} - \theta_2)] \\
&\quad - \Phi(-\sqrt{c} - \theta_2)[\Phi(\sqrt{c}) - \Phi(\sqrt{c} - \theta_2)].
\end{aligned} \tag{26}$$

Let $p = 1 - \Phi(\sqrt{c} - \theta_2)$, $q = \Phi(-\sqrt{c}) - \Phi(-\sqrt{c} - \theta_2)$, $r = \Phi(-\sqrt{c} - \theta_2)$ and $s = \Phi(\sqrt{c}) - \Phi(\sqrt{c} - \theta_2)$. Thus establishing (25) is equivalent to showing that $pq > rs$ or $p/s > r/q$. Observe that p, q, r and s are all strictly positive. Furthermore $p > q$, $p > r$, $p > s$, $s > q$ and $p > q + r$, from which we deduce that $p/s > (q + r)/s$. It follows that showing $(q + r)/s > r/q$ will complete the proof of (25). Suppose the latter does not hold, *i.e.*, $(q + r)/s \leq r/q$ which in turn implies that $q^2 \leq r(s - q) < 0$. Since $q > 0$, we have a contradiction. Thus $(q + r)/s > r/q$ and consequently $pq > rs$ so (25) holds. A similar argument can be used to show that if $\boldsymbol{\theta} \in \mathcal{C}_1$ which satisfy $\theta_2 > \theta_1 > 0$ then

$$H(\theta_1, \theta_2) < H(0, \theta_2). \tag{27}$$

Next we consider the case where $\theta_1 = \theta_2 = \theta > 0$. Now,

$$H(\theta, \theta) = 2\Phi(-\sqrt{c} - \theta))\Phi(-\sqrt{c} + \theta) < 2[\Phi(-\sqrt{c})]^2 = H(0, 0),$$

where the inequality above is a consequence of the log–concavity of $\Phi(\cdot)$ (see Saumard and Wellner (2014)). Thus,

$$H(\theta, \theta) < H(0, 0). \tag{28}$$

It follows from (25), (27) and (28) that for any $\boldsymbol{\theta}$ in the interior of $\mathcal{C}_1$ there exists a $\boldsymbol{\theta}^*$ on the boundary of $\mathcal{C}_1$ for which

$$H(\boldsymbol{\theta}^*) > H(\boldsymbol{\theta}). \tag{29}$$

Repeating the above arguments we can show that (29) holds also for $\boldsymbol{\theta} \in \mathcal{C}_2$. Thus $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} H(\boldsymbol{\theta})$ is attained on the set $\{(0, x) \cup (x, 0) : x \in \mathbb{R}\}$, *i.e.*, the boundary of $\boldsymbol{\Theta}_0$. Next consider the function $H(\theta_1, 0)$ with $\theta_1 \geq 0$. Clearly,

$$H(\theta_1, 0) = \Phi(-\sqrt{c})[1 - \Phi(\sqrt{c} - \theta_1) + \Phi(-\sqrt{c} - \theta_1)] \tag{30}$$

and therefore

$$\frac{\partial}{\partial \theta_1} H(\theta_1, 0) = \Phi(-\sqrt{c})[\phi(\sqrt{c} - \theta_1) - \phi(-\sqrt{c} - \theta_1)] \geq 0$$

since $\phi(\sqrt{c} - \theta_1) \geq \phi(-\sqrt{c} - \theta_1)$ whenever $\theta_1 \geq 0$. This implies that

$$\sup_{\boldsymbol{\theta}_1 \geq 0} H(\theta_1, 0) = \lim_{\theta_1 \to \infty} H(\theta_1, 0) = \Phi(-\sqrt{c}) = 1 - \Phi(\sqrt{c}) = \mathbb{P}(\mathcal{N}(0, 1) \geq \sqrt{c}) = \frac{1}{2}\mathbb{P}(\chi_1^2 \geq c).$$

Since the function $H(\theta_1, \theta_2)$ is permutation invariant and odd we have

$$\sup_{\boldsymbol{\theta}_1 \geq 0} H(\theta_1, 0) = \sup_{\boldsymbol{\theta}_1 \leq 0} H(\theta_1, 0) = \sup_{\boldsymbol{\theta}_2 \geq 0} H(0, \theta_2) = \sup_{\boldsymbol{\theta}_2 \leq 0} H(0, \theta_2) = \frac{1}{2}\mathbb{P}(\chi_1^2 \geq c). \tag{31}$$

Furthermore it is easy to see that:

$$\begin{aligned} \lim_{\theta_1 \to \infty} \mathbb{P}_{(\theta_1, 0)}(T = 0) = \lim_{\theta_1 \to -\infty} \mathbb{P}_{(\theta_1, 0)}(T = 0) = \lim_{\theta_2 \to \infty} \mathbb{P}_{(0, \theta_2)}(T = 0) = \lim_{\theta_2 \to -\infty} \mathbb{P}_{(0, \theta_2)}(T = 0) \\ = \frac{1}{2}\chi_0^2. \end{aligned} \tag{32}$$

It now follows from (31) and (32) that for $c \geq 0$

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \mathbb{P}_{\boldsymbol{\theta}}(T \geq c) = \frac{1}{2}\mathbb{P}(\chi_0^2 \geq c) + \frac{1}{2}\mathbb{P}(\chi_1^2 \geq c). \tag{33}$$

as claimed. □

Proof of Lemma 1:

Proof: Note that $\mathcal{R}_{ij}$, $\mathcal{R}'_{ij}$ and $\mathcal{R}''_{ij}$ denote regions between the boundaries of adjacent cones or their polar cones. Since there are K such regions, we have $N + N' + N'' = K$. The upper bound for each summand is attained if all the regions between various pairs of adjacent cones are of the same type. If the angle between every pair of adjacent cones is less than $\pi/2$, then there are K regions of the type $\mathcal{R}_{ij}$, *i.e.*, $N \leq K$. If the angle between every pair of adjacent cones is between $\pi/2$ and π, then we have $K \leq 3$ and hence at most 3 regions of the type $\mathcal{R}'_{ij}$, *i.e.*, $N' \leq 3$. If the angle between some pair of adjacent cones is greater than π, then the angles between all other pairs of such cones are each less than π. Hence there is at most one region of the type $\mathcal{R}''_{ij}$, *i.e.*, $N'' \leq 1$. Suppose now that $N'' = 1$. If the angles between all other pairs of adjacent cones are each less than $\pi/2$, then there are $K - 1$ regions of the type $\mathcal{R}_{ij}$, *i.e.*, $N \leq K - 1$. Moreover there can be at most one other pair of adjacent cones with an angle between $\pi/2$ and π between them, and hence at most one region of the type $\mathcal{R}'_{ij}$, *i.e.*, $N' \leq 1$. □

Proof of Lemma 2:

Proof: First note that $\mathrm{d}^2(\boldsymbol{S}, \mathrm{ray}(\boldsymbol{u})) = (\boldsymbol{u}_*^T \boldsymbol{S})^2$ and similarly $\mathrm{d}^2(\boldsymbol{S}, \mathrm{ray}(\boldsymbol{v})) = (\boldsymbol{v}_*^T \boldsymbol{S})^2$. Since $\boldsymbol{S} \sim \mathcal{N}_2(\boldsymbol{0}, \boldsymbol{I}$, both $(\boldsymbol{u}_*^T \boldsymbol{S})^2$ and $(\boldsymbol{v}_*^T \boldsymbol{S})^2$ are distributed as χ^2_1 RVs. Moreover the correlation coefficient between $\boldsymbol{u}_*^T \boldsymbol{S}$ and $\boldsymbol{v}_*^T \boldsymbol{S}$ is $\boldsymbol{u}_*^T \boldsymbol{v}_*$. Since $\measuredangle(\boldsymbol{u}_*, \boldsymbol{v}_*) = \pi - \measuredangle(\boldsymbol{u}, \boldsymbol{v}) = \pi - \gamma$, we have $\boldsymbol{u}_*^T \boldsymbol{v}_* = \cos(\measuredangle(\boldsymbol{u}_*, \boldsymbol{v}_*)) = \cos(\pi - \gamma) = -\cos(\gamma)$. Let $D_1 = \boldsymbol{u}_*^T \boldsymbol{S}$ and $D_2 = \boldsymbol{v}_*^T \boldsymbol{S}$ so $\mathrm{Var}(D_1) = \mathrm{Var}(D_2) = 1$ and the correlation coefficient between D_1 and D_2 is $\boldsymbol{u}_*^T \boldsymbol{v}_* = -\cos(\gamma)$. Further we have $\chi^2_{1,1}(\gamma) = \min\{D_1^2, D_2^2\}$ which is a function of the length of $\boldsymbol{S}$. Since $\boldsymbol{S} \sim \mathcal{N}_2(\boldsymbol{0}, \boldsymbol{I})$, the length and direction of $\boldsymbol{S}$ are independently distributed. Thus

$$
\begin{aligned}
\mathbb{P}(\chi^2_{1,1}(\gamma) \geq c, \boldsymbol{S} \in \mathcal{R}) &= \mathbb{P}(\boldsymbol{S} \in \mathcal{R}) \mathbb{P}(\chi^2_{1,1}(\gamma) \geq c) \\
&= \frac{\gamma}{2\pi} \mathbb{P}(\min\{D_1^2, D_2^2\} \geq c) \\
&= \frac{\gamma}{2\pi} \mathbb{P}(D_1^2 \geq c, D_2^2 \geq c) \\
&= \frac{\gamma}{2\pi} [1 - \mathbb{P}(-\sqrt{c} \leq D_1 \leq \sqrt{c}) - \mathbb{P}(D_1 \leq -\sqrt{c}, -\sqrt{c} \leq D_2 \leq \sqrt{c}) \\
&\quad - \mathbb{P}(D_1 \geq \sqrt{c}, -\sqrt{c} \leq D_2 \leq \sqrt{c})] \\
&= \frac{\gamma}{2\pi} [\mathbb{P}(D_1 \geq \sqrt{c}, D_2 \geq \sqrt{c}) + \mathbb{P}(D_1 \geq \sqrt{c}, D_2 \leq -\sqrt{c}) \\
&\quad + \mathbb{P}(D_1 \leq -\sqrt{c}, D_2 \geq \sqrt{c}) + \mathbb{P}(D_1 \leq -\sqrt{c}, D_2 \leq -\sqrt{c})]
\end{aligned}
$$

where $(D_1, D_2)^T$ has a bivariate normal distribution with mean $\boldsymbol{0}$, unit variances and correlation $-\cos(\gamma)$. □

Proof of Theorem 3:

Proof: Recall that the LRT statistic T_n for (1) is

$$
T_n = \min\{n\|\boldsymbol{S}_n - \Pi(\boldsymbol{S}_n \mid \mathcal{C}_1)\|^2, \ldots, n\|\boldsymbol{S}_n - \Pi(\boldsymbol{S}_n \mid \mathcal{C}_K)\|^2\}, \tag{34}
$$

which minimizes the squared distance between $\boldsymbol{S}_n$ and each of the K cones. If $\boldsymbol{S}_n \in \mathcal{C}_i$ for any i, then $\Pi(\boldsymbol{S}_n \mid \mathcal{C}_i) = \boldsymbol{S}_n$, thus $T_n = 0$. If $\boldsymbol{S}_n$ lies in $\mathcal{R}_i(\boldsymbol{u}_i)$ for some i, then $\Pi(\boldsymbol{S}_n \mid \mathcal{C}_i) = (\boldsymbol{u}_i^T \boldsymbol{S}_n)\boldsymbol{u}_i$ and $\|\boldsymbol{S}_n - \Pi(\boldsymbol{S}_n \mid \mathcal{C}_r)\|^2 > \|\boldsymbol{S}_n - \Pi(\boldsymbol{S}_n \mid \mathcal{C}_i)\|^2$ where $r \neq i$, thus $T_n = n(\boldsymbol{u}_{i*}^T \boldsymbol{S}_n)^2$. Similarly T_n can be obtained for $\boldsymbol{S}_n$ lying in $\mathcal{R}_i(\boldsymbol{v}_i)$ or $\mathcal{R}_j(\boldsymbol{u}_j)$ or $\mathcal{R}_j(\boldsymbol{v}_j)$. If $\boldsymbol{S}_n$ lies in $\mathcal{R}_{ij}$ or $\mathcal{R}'_{ij}$ for some $(i,j) \in \mathcal{P}$, then $\Pi(\boldsymbol{S}_n \mid \mathcal{C}_i) = (\boldsymbol{v}_i^T \boldsymbol{S}_n)\boldsymbol{v}_i$ and $\Pi(\boldsymbol{S}_n \mid \mathcal{C}_j) = (\boldsymbol{u}_j^T \boldsymbol{S}_n)\boldsymbol{u}_j$. Also $\|\boldsymbol{S}_n - \Pi(\boldsymbol{S}_n \mid \mathcal{C}_r)\|^2 > \max\{\|\boldsymbol{S}_n - \Pi(\boldsymbol{S}_n \mid \mathcal{C}_i)\|^2, \|\boldsymbol{S}_n - \Pi(\boldsymbol{S}_n \mid \mathcal{C}_j)\|^2\}$ where $r \notin \{i,j\}$, thus $T_n = \min\{n(\boldsymbol{u}_{j*}^T \boldsymbol{S}_n)^2, n(\boldsymbol{v}_{i*}^T \boldsymbol{S}_n)^2\}$ Finally if $\boldsymbol{S}_n \in \mathcal{R}''_{pq}$, then $\boldsymbol{S}_n \in \mathcal{C}_i^0$ and $\Pi(\boldsymbol{S}_n \mid \mathcal{C}_i) = \boldsymbol{0}$ for every i, thus $T_n = n\|\boldsymbol{S}_n\|^2$. To summarize,

$$T_n = \begin{cases} 0 & \text{if } \boldsymbol{S}_n \in \mathcal{C}_i \\ \min\{n(\boldsymbol{u}_{j*}^T \boldsymbol{S}_n)^2, n(\boldsymbol{v}_{i*}^T \boldsymbol{S}_n)^2\} & \text{if } \boldsymbol{S}_n \in \mathcal{R}_{ij} \text{ or } \mathcal{R}'_{ij} \\ n(\boldsymbol{u}_{i*}^T \boldsymbol{S}_n)^2 \text{ or } n(\boldsymbol{v}_{i*}^T \boldsymbol{S}_n)^2 & \text{if } \boldsymbol{S}_n \in \mathcal{R}_i(\boldsymbol{u}_i) \text{ or } \mathcal{R}_i(\boldsymbol{v}_i) \\ n\|\boldsymbol{S}_n\|^2 & \text{if } \boldsymbol{S}_n \in \mathcal{R}''_{pq} \end{cases} \tag{35}$$

for all $(i,j) \in \mathcal{P}$. Next, we evaluate the limiting distribution of T_n for various values of $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. Suppose first that $\boldsymbol{\theta} \in \text{int}(\boldsymbol{\Theta}_0)$, *i.e.*, $\boldsymbol{\theta}$ lies in the interior of $\boldsymbol{\Theta}_0$. If so, the ball $\mathcal{B}(\boldsymbol{\theta}, \delta)$ is a subset of $\text{int}(\boldsymbol{\Theta}_0)$ for some $\delta > 0$. Since $\boldsymbol{S}_n$ is consistent for $\boldsymbol{\theta}$ it follows that $\mathbb{P}(\boldsymbol{S}_n \in \mathcal{B}(\boldsymbol{\theta}, \epsilon)) \to 1$ as $n \to \infty$ for all $\epsilon < \delta$. Therefore $T_n \xrightarrow{p} 0$ and consequently $\boldsymbol{\theta} \in \text{int}(\boldsymbol{\Theta}_0)$ implies that

$$T_n \Rightarrow \chi_0^2, \tag{36}$$

as $n \to \infty$. Next, consider the situation when $\boldsymbol{\theta} \in \text{ray}(\boldsymbol{\Theta}_0)$. Without loss of generality let $\boldsymbol{\theta} = \lambda \boldsymbol{u}_1$ for some fixed $\lambda > 0$, *i.e.*, $\boldsymbol{\theta}$ lies on one of the rays generating the cone $\mathcal{C}_1$. The ray through $\boldsymbol{\theta}$ partitions $\mathcal{B}(\boldsymbol{\theta}, \epsilon)$ into two half circles $\mathcal{B}_1 \subset \mathcal{C}_1$ and $\mathcal{B}_2 \subset \mathcal{R}_1(\boldsymbol{u}_1)$. Observe that $\mathcal{B}_1 - \boldsymbol{\theta} = \boldsymbol{\theta} - \mathcal{B}_2$ where $\mathcal{B}_1 - \boldsymbol{\theta} = \{\boldsymbol{S} - \boldsymbol{\theta} \mid \boldsymbol{S} \in \mathcal{B}_1\}$ and $\boldsymbol{\theta} - \mathcal{B}_2 = \{\boldsymbol{\theta} - \boldsymbol{S} \mid \boldsymbol{S} \in \mathcal{B}_2\}$. The distribution of $\boldsymbol{S}_n$ is spherically symmetric around $\boldsymbol{\theta}$ so $\boldsymbol{S}_n - \boldsymbol{\theta} \stackrel{d}{=} \boldsymbol{\theta} - \boldsymbol{S}_n$. Consequently,

$$\begin{aligned} \mathbb{P}(\boldsymbol{S}_n \in \mathcal{B}_1) &= \mathbb{P}(\boldsymbol{S}_n - \boldsymbol{\theta} \in \mathcal{B}_1 - \boldsymbol{\theta}) = \mathbb{P}(\boldsymbol{S}_n - \boldsymbol{\theta} \in \boldsymbol{\theta} - \mathcal{B}_2) \\ &= \mathbb{P}(\boldsymbol{\theta} - \boldsymbol{S}_n \in \boldsymbol{\theta} - \mathcal{B}_2) = \mathbb{P}(\boldsymbol{S}_n - \boldsymbol{\theta} \in \mathcal{B}_2 - \boldsymbol{\theta}) = \mathbb{P}(\boldsymbol{S}_n \in \mathcal{B}_2). \end{aligned}$$

Moreover, since $\boldsymbol{S}_n$ is consistent for $\boldsymbol{\theta}$ we have $\mathbb{P}(\boldsymbol{S}_n \in \mathcal{B}(\boldsymbol{\theta}, \epsilon)) \to 1$ so $\mathbb{P}(\boldsymbol{S}_n \in \mathcal{B}_1) = \mathbb{P}(\boldsymbol{S}_n \in \mathcal{B}_2) \to 1/2$ as $n \to \infty$. Thus, $\mathbb{P}(\boldsymbol{S}_n \in \mathcal{C}_1) = \mathbb{P}(\boldsymbol{S}_n \in \mathcal{R}_1(\boldsymbol{u}_1)) = 1/2 + o_P(1)$ and the LRT statistic equals

$$T_n = 0 \times \mathbb{I}(\boldsymbol{S}_n \in \mathcal{C}_1) + n(\boldsymbol{u}_{1*}^T \boldsymbol{S}_n)^2 \times \mathbb{I}(\boldsymbol{S}_n \in \mathcal{R}_1(\boldsymbol{u}_1)) + o_P(1).$$

Observe that $\boldsymbol{S}_n = \boldsymbol{\theta} + \bar{\boldsymbol{Z}}_n$ where $\bar{\boldsymbol{Z}}_n$ is the average of n IID $\mathcal{N}_2(\boldsymbol{0}, \boldsymbol{I})$ RVs as $n \to \infty$. It follows that

$$n(\boldsymbol{u}_{1*}^{T/} \boldsymbol{S}_n)^2 = n(\boldsymbol{u}_{1*}^T(\boldsymbol{\theta} + \bar{\boldsymbol{Z}}_n))^2 = n(\boldsymbol{u}_{1*}^T(\lambda \boldsymbol{u} + \bar{\boldsymbol{Z}}_n))^2 = (\boldsymbol{u}_{1*}^T(\sqrt{n}\bar{\boldsymbol{Z}}_n))^2 \Rightarrow \chi_1^2$$

and therefore, if $\boldsymbol{\theta} = \lambda \boldsymbol{u}_1$

$$T_n \Rightarrow \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2. \tag{37}$$

Obviously (37) remains unchanged if $\boldsymbol{\theta}$ lies on any other ray of $\boldsymbol{\Theta}_0 \setminus \{\mathbf{0}\}$. Finally, suppose $\boldsymbol{\theta} = \mathbf{0}$. Here $\sqrt{n}\boldsymbol{S}_n \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{I})$ as $n \to \infty$ which is spherically symmetric so the direction and length of $\boldsymbol{S}_n$ are statistically independent. We have already seen that $n(\boldsymbol{u}_{j*}^T \boldsymbol{S}_n)^2$ and $n(\boldsymbol{v}_{i*}^T \boldsymbol{S}_n)^2$ are each distributed as a χ_1^2 RV whereas $n\|\boldsymbol{S}_n\|^2$ is distributed as a χ_2^2 RV as $n \to \infty$. It follows from (35) that: (i) if

$$\boldsymbol{S}_n \in \bigcup_{i=1}^{K} \mathcal{C}_i,$$

an event that has probability $\rho/2\pi$, then $T_n \Rightarrow \chi_0^2$; (ii) if $(i,j) \in \mathcal{P}$ and $\boldsymbol{S}_n \in \mathcal{R}_{ij}$ or $\mathcal{R}'_{ij}$, an event that has probability $\gamma_{ij}/2\pi$, then $T_n \Rightarrow \chi_{1,1}^2(\gamma_{ij})$; (iii) if $(i,j) \in \mathcal{P}$ and

$$\boldsymbol{S}_n \in \bigcup_{(i,j)\in\mathcal{P}} (\mathcal{R}_i(\boldsymbol{u}_i) \cup \mathcal{R}_i(\boldsymbol{v}_i) \cup \mathcal{R}_j(\boldsymbol{u}_j) \cup \mathcal{R}_j(\boldsymbol{v}_j)),$$

an event that has probability $\tau/2\pi$, then $T_n \Rightarrow \chi_1^2$; and (iv) if $\boldsymbol{S}_n \in \mathcal{R}''_{pq}$, an event that has probability $\gamma_{pq}/2\pi$, then $T_n \Rightarrow \chi_2^2$. Putting it all together we find that when $\boldsymbol{\theta} = \mathbf{0}$, we have

$$T_n \Rightarrow \frac{\rho}{2\pi}\chi_0^2 + \sum_{(i,j)\in\mathcal{P}} \frac{\gamma_{ij}}{2\pi}\chi_{1,1}^2(\gamma_{ij}) + \frac{\tau}{2\pi}\chi_1^2 + \frac{\gamma_{pq}}{2\pi}\chi_2^2. \tag{38}$$

Equations (36), (37) and (38) establish the result. □

Proof of Theorem 4:

Proof: First we prove that the given conditions are sufficient. Recall that for any $c \geq 0$

$$\mathbb{P}(T_{\boldsymbol{R}} \geq c) = \frac{1}{2}\mathbb{P}(\chi_0^2 \geq c) + \frac{1}{2}\mathbb{P}(\chi_1^2 \geq c) \tag{39}$$

and

$$\mathbb{P}(T_{\boldsymbol{O}} \geq c) = \frac{\rho}{2\pi}\mathbb{P}(\chi_0^2 \geq c) + \sum_{(i,j)\in\mathcal{P}} \frac{\gamma_{ij}}{2\pi}\mathbb{P}(\chi_{1,1}^2(\gamma_{ij}) \geq c) + \frac{\tau}{2\pi}\mathbb{P}(\chi_1^2 \geq c) + \frac{\gamma_{pq}}{2\pi}\mathbb{P}(\chi_2^2 \geq c), \tag{40}$$

where τ or γ_{pq} may be equal to 0. If $\tau \geq \pi$ then both

$$(A_{\boldsymbol{O}}) \quad \frac{\tau}{2\pi} \geq \frac{1}{2} \quad \text{and} \quad (B_{\boldsymbol{O}}) \quad \frac{\rho}{2\pi} < \frac{1}{2}$$

hold. By $(B_{\boldsymbol{O}})$ the first two terms on the right hand side of (40) are larger than the first term on the right hand side of (39). By $(A_{\boldsymbol{O}})$ the same is true when comparing the last two terms in (40) to the second term of (39). Therefore $\mathbb{P}(T_{\boldsymbol{O}} \geq c) > \mathbb{P}(T_{\boldsymbol{R}} \geq c)$ so

$$T_{\boldsymbol{O}} \succeq_{\text{st}} T_{\boldsymbol{R}}.$$

Hence, we conclude that the least favourable limiting null distribution for T_n is that of $T_{\boldsymbol{O}}$ when $\tau \geq \pi$.

Now suppose that $\rho \geq \pi$ then $\gamma_{pq} = 0$. This is because if $\gamma_{pq} > 0$, then $\tau \geq \pi$ so $\rho < \pi$. We have

$$(A_{\boldsymbol{R}}) \quad \frac{\rho}{2\pi} \geq \frac{1}{2} \quad \text{and} \quad (B_{\boldsymbol{R}}) \quad \sum_{(i,j)\in\mathcal{P}} \gamma_{ij} + \tau \leq \pi.$$

Hence,

$$\begin{aligned}
\mathbb{P}(T_{\boldsymbol{O}} \geq c) &= \frac{\rho}{2\pi}\mathbb{P}(\chi_0^2 \geq c) + \sum_{(i,j)\in\mathcal{P}} \frac{\gamma_{ij}}{2\pi}\mathbb{P}(\chi_{1,1}^2(\gamma_{ij}) \geq c) + \frac{\tau}{2\pi}\mathbb{P}(\chi_1^2 \geq c) \\
&< \frac{1}{2}\mathbb{P}(\chi_0^2 \geq c) + \frac{\sum_{(i,j)\in\mathcal{P}} \gamma_{ij} + \tau}{2\pi}\mathbb{P}(\chi_1^2 \geq c) \\
&\leq \frac{1}{2}\mathbb{P}(\chi_0^2 \geq c) + \frac{1}{2}\mathbb{P}(\chi_1^2 \geq c) = \mathbb{P}(T_{\boldsymbol{R}} \geq c)
\end{aligned}$$

where the first inequality is a consequence of $(A_{\boldsymbol{R}})$ and the second of $(B_{\boldsymbol{R}})$. Thus

$$T_{\boldsymbol{R}} \succeq_{\text{st}} T_{\boldsymbol{O}},$$

i.e., the least favourable limiting null distribution for T_n is that of $T_{\boldsymbol{R}}$ when $\rho \geq \pi$.

Next we prove that the above conditions are necessary. Suppose first that $\tau < \pi$. Then $\gamma_{pq} = 0$ since $\gamma_{pq} > 0$ implies $\tau \geq \pi$. For $c > 0$, we have

$$\begin{aligned}
\mathbb{P}(T_{\boldsymbol{O}} \geq c) - \mathbb{P}(T_{\boldsymbol{R}} \geq c) &= \sum_{(i,j)\in\mathcal{P}} \frac{\gamma_{ij}}{2\pi}\mathbb{P}(\chi_{1,1}^2(\gamma_{ij} \geq c) + (\frac{\tau}{2\pi} - \frac{1}{2})\mathbb{P}(\chi_1^2 \geq c) \\
&< (\sum_{(i,j)\in\mathcal{P}} \frac{\gamma_{ij}}{2\pi} + \frac{\tau}{2\pi} - \frac{1}{2})\mathbb{P}(\chi_1^2 \geq c) \\
&= \frac{\sum_{(i,j)\in\mathcal{P}} \gamma_{ij} + \tau - \pi}{2\pi}\mathbb{P}(\chi_1^2 \geq c) \\
&= \frac{\pi - \rho}{2\pi}\mathbb{P}(\chi_1^2 \geq c). \qquad (41)
\end{aligned}$$

If $\rho \geq \pi$ then the RHS of (41) is negative so $\mathbb{P}(T_{\boldsymbol{O}} \geq c) < \mathbb{P}(T_{\boldsymbol{R}} \geq c)$ for all $c > 0$. However, if $\rho < \pi$ then the RHS of (41) is positive. So there is at least one $c > 0$ satisfying $\mathbb{P}(T_{\boldsymbol{O}} \geq c) < \mathbb{P}(T_{\boldsymbol{R}} \geq c)$. Thus the condition $\tau \geq \pi$ is necessary for the limiting null distribution of T_n to be that of $T_{\boldsymbol{O}}$.

Finally suppose that $\rho < \pi$. If $\gamma_{pq} > 0$, then $\tau \geq \pi$ so $\mathbb{P}(T_{\boldsymbol{R}} \geq c) < \mathbb{P}(T_{\boldsymbol{O}} \geq c)$ for all $c > 0$ as shown earlier. If $\gamma_{pq} = 0$ then using (41), we have for $c > 0$

$$\begin{aligned}
\mathbb{P}(T_{\boldsymbol{O}} \geq c) - \mathbb{P}(T_{\boldsymbol{R}} \geq c) &= \sum_{(i,j)\in\mathcal{P}} \frac{\gamma_{ij}}{2\pi}\mathbb{P}(\chi_{1,1}^2(\gamma_{ij} \geq c) + (\frac{\tau}{2\pi} - \frac{1}{2})\mathbb{P}(\chi_1^2 \geq c) \\
&> (\frac{\tau}{2\pi} - \frac{1}{2})\mathbb{P}(\chi_1^2 \geq c). \qquad (42)
\end{aligned}$$

If $\tau \geq \pi$ then the RHS of (42) is positive so $\mathbb{P}(T_{\boldsymbol{R}} \geq c) < \mathbb{P}(T_{\boldsymbol{O}} \geq c)$ for all $c > 0$. However if $\tau < \pi$ then the RHS of (42) is negative. So there is at least one $c > 0$ satisfying $\mathbb{P}(T_{\boldsymbol{R}} \geq c) < \mathbb{P}(T_{\boldsymbol{O}} \geq c)$. Thus the condition $\rho \geq \pi$ is necessary for the limiting null distribution of T_n to be that of $T_{\boldsymbol{R}}$. □

Proof of Theorem 5:

Proof: Let $\Lambda^{(i)}$ denote the LRT statistic for testing $H_0^{(i)} : \boldsymbol{\theta} \in \mathcal{C}_i$ against $H_1^{(i)} : \boldsymbol{\theta} \notin \mathcal{C}_i$ and let $c_\alpha^{(i)}$ denote its critical value. Clearly, (cf. Silvapulle and Sen (2004)) $c_\alpha^{(i)}$ is the $1-\alpha$ quantile of the RV

$$\frac{\rho_i}{2\pi}\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{\pi-\rho_i}{2\pi}\chi_2^2. \tag{43}$$

The size α IUT rejects the null $\bigcup_{i=1}^{K} H_0^{(i)}$ if and only if $\Lambda^{(i)} > c_\alpha^{(i)}$ for every i. Recall that the size α LRT for (1) rejects the null if $T_n > c_\alpha$ where c_α is its critical value as discussed in Section 3. The cones $\mathcal{C}_1, \ldots, \mathcal{C}_K$ satisfy one of three possibilities: (I) $\rho \geq \pi$; (II) $\tau \geq \pi$; or (III) $\rho < \pi, \tau < \pi$. If (I) holds then by Theorems 3 and 4 the asymptotic critical value for T_n is the $1-\alpha$ quantile of the RV

$$\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2. \tag{44}$$

It follows that $c_\alpha < \min\{c_\alpha^{(1)}, \ldots, c_\alpha^{(K)}\}$ so the LRT has higher power than the IUT. If (II) holds then the asymptotic critical value for T_n is the $1-\alpha$ quantile of the RV

$$\frac{\rho}{2\pi}\chi_0^2 + \sum_{(i,j)\in\mathcal{P}} \frac{\gamma_{ij}}{2\pi}\chi_{1,1}^2(\gamma_{ij}) + \frac{\tau}{2\pi}\chi_1^2 + \frac{\gamma_{pq}}{2\pi}\chi_2^2. \tag{45}$$

Observe that

$$\sum_{(i,j)\in\mathcal{P}} \frac{\gamma_{ij}}{2\pi}\chi_{1,1}^2(\gamma_{ij}) + \frac{\tau}{2\pi}\chi_1^2 + \frac{\gamma_{pq}}{2\pi}\chi_2^2 \preceq_{st} \frac{\sum_{(i,j)\in\mathcal{P}}\gamma_{ij} + \tau}{2\pi}\chi_1^2 + \frac{\gamma_{pq}}{2\pi}\chi_2^2 \tag{46}$$

and since (II) holds we have

$$\frac{\sum_{(i,j)\in\mathcal{P}}\gamma_{ij} + \tau}{2\pi}\chi_1^2 + \frac{\gamma_{pq}}{2\pi}\chi_2^2 \preceq_{st} \frac{1}{2}\chi_1^2 + \frac{\pi-\rho}{2\pi}\chi_2^2 \tag{47}$$

where the upper bound on the left hand side of (47) is attained when τ along with all $\gamma_{ij} \in \mathcal{P}$ are minimized and γ_{pq} is maximized. Combining (46) and (47), we conclude that the RV in (45) is stochastically smaller than the RV

$$\frac{\rho}{2\pi}\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{\pi-\rho}{2\pi}\chi_2^2 \tag{48}$$

which is itself stochastically smaller than the RV in (43). Thus $c_\alpha < \min\{c_\alpha^{(1)}, \ldots, c_\alpha^{(K)}\}$ so the LRT is more powerful than the IUT. Finally if (III) holds then c_α is the $1-\alpha$ quantile of the RV in (44) for some values of α and of the RV in (45) for others so the LRT is more powerful. Note that in each of the three cases the rejection probability of the null for the LRT is greater than that for the IUT for all values of $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$ so the LRT is asymptotically uniformly more powerful than the IUT for (1). □

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 67–70
http://www.ssca.org.in/journal

Cramer-Rao Posterior Bounds in the Spirit of van Trees

Malay Ghosh
University of Florida

Received: 07 January 2024; Revised: 06 May 2024; Accepted: 06 May 2024

Abstract

The paper obtains posterior Cramer-Rao bounds for arbitrary parameter vector in the spirit of van Trees. The relationship with the classical Cramer-Rao bound is also discussed.

Key words: Cramer-Rao; Parametric functions; Posterior variances.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

The present short article is a humble tribute to the late Professor C. R. Rao, one of the legendary heroes in the world of statistics. In his professional life, which spanned for nearly eight decades, Professor Rao made a number of pioneering contributions virtually in every single area of statistics, impacting the academic life of several generations of statisticians. It is needless to say that the legacy left behind by Professor Rao will continue its impact on future generations of statisticians as well.

Two major contributions of Professor Rao for which he is most well known in the general scientific community, crossing the boundaries of statisics, are the Cramer-Rao inequality and the Rao-Blackwell Theorem. What is indeed more remarkable is that these results constituted part of the Masters Thesis of Professor Rao. These two fundamental results, accessible even to beginning undrgraduate students in statistics, have far reaching implications, well beyond what possibly was envisaged by their authors. For example, the Rao-Blackwell Theorem, discovered independently by Rao and Blackwell, involves an implicit idea of projection from a certain space of random variables to a second space, spanned by sufficient statistics, resulting thereby in loss reduction under convexity.

The other work, namely the Cramer-Rao inequality, discovered independently by Cramer and Rao, is being used repeatedly by scientists even outside statistics, notably by those working in Quantum Physics, Electrical Engineering and Computer Science.

Early extensions of the Cramer-Rao inequality appear in the articles of Bhattacharyya (1946), Hammersley (1950) and Chapman and Robbins (1951). More recently, there has been a surge of extensions of this inequality, primarily for solving problems in science and engineering, as mentioned in the preceding paragraph.

Corresponding Author: Malay Ghosh
Email: ghoshm@ufl.edu

One very useful extension of the Cramer-Rao inequality appears in the book of Van Trees (2004) who provided a lower bound for the Bayes risk of estimators of one dimensional parameters of interest. This was followed later in a series of articles of Bobrovsky *et al.* (1987), Borovkov and Sakhanenko (1980), Brown and Gajek (1990) and many others. Very important consequences of these results leading to local asymptotic minimaxity in the spirit of Hajek and LeCam are proved in Gill and Levit (1995) and Gassiat *et al.* (2013).

In contrast to the above, Ghosh (1993), obtained Cramer-Rao type bounds for posterior variances. Whereas the original Cramer-Rao inequality is based on the Fisher Information number based on the likelihood and the Bayes risk results involve Fisher information number of both the likelihood and the prior, the lower bound obtained by Ghosh involves the posterior analog of the classical Fisher information number.

The present work extends the work of Ghosh (1993) to the multiparameter case, quite in the spirit of Van Trees (2004) as well as Gill and Levit (1995). In particular, for a vector valued parameter, a lower bound is provided for posterior expected weighted squared norms of the difference of parameter vectors and their posterior means. The technical details are given in the following section.

2. The main results

We begin with the posterior lower bound for the variance-covariance matrix of a vector-valued parameter. Indeed, the same lower bound can be provided for the posterior mean squared error matrix for an arbitrary estimator of a parameter of interest. But the sharpest bound is one for the posterior variance-covariance matrix, since for an arbitrary estimator $\boldsymbol{e}(\boldsymbol{X})$ of a parameter vector $\boldsymbol{\psi}(\boldsymbol{\theta})$,

$$\begin{aligned} & E[(\boldsymbol{\psi}(\boldsymbol{\theta}) - \boldsymbol{e}(\boldsymbol{X}))(\boldsymbol{\psi}(\boldsymbol{\theta}) - \boldsymbol{e}(\boldsymbol{X}))^T|\boldsymbol{X}] \\ &= V[\boldsymbol{\psi}(\boldsymbol{\theta})|\boldsymbol{X}] + [E(\boldsymbol{\psi}(\boldsymbol{\theta})|\boldsymbol{X}) - \boldsymbol{e}(\boldsymbol{X}))(E(\boldsymbol{\psi}(\boldsymbol{\theta})|\boldsymbol{X}) - \boldsymbol{e}(\boldsymbol{X}))^T] \\ &\geq V[\boldsymbol{\psi}(\boldsymbol{\theta})|\boldsymbol{X}]. \end{aligned}$$

Throughout this section, we will consider the following set up. Let X be a real or vector-valued random variable with pdf $f(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$. We denote by $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ the lower and upper end points of $\boldsymbol{\theta}$. Consider an arbitrary prior $\pi(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$. We will denote the posterior of $\boldsymbol{\theta}$ given X by $\pi(\boldsymbol{\theta}|x)$. Consider an $s(\leq p)$-dimensional function $\psi(\boldsymbol{\theta}) = (\psi_1(\boldsymbol{\theta}), \ldots, \psi_s(\boldsymbol{\theta}))^T$ of $\boldsymbol{\theta}$. We begin with the following lemma.

Lemma 1: Let (i) $\pi(\boldsymbol{\theta}|x) \to 0$ as $\boldsymbol{\theta} \to \boldsymbol{r}_1$ or $\boldsymbol{r}_2$ and (ii) $\psi_i(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|x) \to 0$ as $\boldsymbol{\theta} \to \boldsymbol{r}_1$ or $\boldsymbol{r}_2$ for all $i = 1, \ldots, s$. Then

$$E[\psi(\boldsymbol{\theta})\{\nabla \log \pi(\boldsymbol{\theta}|x)\}^T|x] = -E(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}}|x),$$

where ∇ denotes the gradient operator.

Proof: In view of assumptions (i) and (ii), for all $1 \leq i \leq s$ and $1 \leq j \leq p$, integration by

parts yields

$$\begin{aligned} E[\psi_i(\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}\nabla\log\pi(\boldsymbol{\theta}|x)|x] &= \int_{\boldsymbol{r}_1}^{\boldsymbol{r}_2}\psi_i(\boldsymbol{\theta})[(\frac{\partial}{\partial\theta_j}\pi(\boldsymbol{\theta}|x))/\pi(\boldsymbol{\theta}|x)]\pi(\boldsymbol{\theta}|x)d\boldsymbol{\theta} \\ &= -\int_{\boldsymbol{r}_1}^{\boldsymbol{r}_2}\frac{\partial\psi_i(\boldsymbol{\theta})}{\partial\theta_j}\pi(\boldsymbol{\theta}|x)d\boldsymbol{\theta} = -E[\frac{\partial\psi_i(\boldsymbol{\theta})}{\partial\theta_j}|x]. \end{aligned}$$

This proves the result.

We now prove the first main result of this section. The result provides a multiparameter posterior Cramer-Rao type lower bound for a vector-valued function of parameters.

Theorem 1: Assume the conditions of Lemma 1, and assume in addition that $V[\nabla\log\pi(\boldsymbol{\theta}|x)]$ is positive definite. Then

$$V[\boldsymbol{\psi}(\boldsymbol{\theta})|x] \geq E(\frac{\partial\boldsymbol{\psi}}{\partial\boldsymbol{\theta}}|x)[V(\nabla\log\pi(\boldsymbol{\theta}|x))]^{-1}E[(\frac{\partial\boldsymbol{\psi}}{\partial\boldsymbol{\theta}})^T|x].$$

Proof: Let $\boldsymbol{u} = \boldsymbol{\psi}(\boldsymbol{\theta}) - E(\boldsymbol{\psi}(\boldsymbol{\theta})|x)$ and $\boldsymbol{v} = \nabla\log\pi(\boldsymbol{\theta}|x)$. Consider the matrix

$$E[\begin{pmatrix}\boldsymbol{u}\\ \boldsymbol{v}\end{pmatrix}(\boldsymbol{u}^T\boldsymbol{v}^T)|x] = E\left[\begin{pmatrix}\boldsymbol{u}\boldsymbol{u}^T\boldsymbol{u}\boldsymbol{v}^T\\ \boldsymbol{v}\boldsymbol{u}^T\boldsymbol{v}\boldsymbol{v}^T\end{pmatrix}|x\right],$$

which by construction is non negative definite. This immediately leads to

$$E(\boldsymbol{u}\boldsymbol{u}^T|x) \geq [E(\boldsymbol{u}\boldsymbol{v}^T|x)][E(\boldsymbol{v}\boldsymbol{v}^T|x)]^{-1}[E(\boldsymbol{v}\boldsymbol{u}^T|x)].$$

In view of Assumption (i), $E(\boldsymbol{v}^T|x) = \boldsymbol{0}$. Also, $E(\boldsymbol{v}\boldsymbol{v}^T|x) = V[\nabla\log\pi(\boldsymbol{\theta}|x))|x]$. The conclusion follows now by applying Lemma 1.

Remark 1: In the particular case when $\boldsymbol{\psi}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ so that $\frac{\partial\boldsymbol{\psi}}{\partial\boldsymbol{\theta}} = \boldsymbol{I}_p$, one gets the inequality $V(\boldsymbol{\theta}|x) \geq [V(\nabla\log\pi(\boldsymbol{\theta}|x))|x]^{-1}$. The classical Cramer-Rao inequality says that for unbiased estimators $\boldsymbol{T}(X)$ of a parameter vector $\boldsymbol{\theta}$, $V[\boldsymbol{T}(X)|\boldsymbol{\theta}] \geq \boldsymbol{I}^{-1}(\boldsymbol{\theta})$, where $\boldsymbol{I}(\boldsymbol{\theta})$ denotes the Fisher Information matrix. $V(\nabla\log\pi(\boldsymbol{\theta}|x))$ is the posterior analog of the classical Fisher Information matrix. It may be noted that while $\boldsymbol{I}(\boldsymbol{\theta}) = V[(\frac{\partial\log f(X|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}})|\boldsymbol{\theta}]$, $V[\nabla\log\pi(\boldsymbol{\theta}|x)|x] = V[(\frac{\partial\log\pi(\boldsymbol{\theta}|x}{\partial\boldsymbol{\theta}})|x]$.

Remark 2: Equality holds in Theorem 1 when $\boldsymbol{\psi}(\boldsymbol{\theta}) - E[\boldsymbol{\psi}(\boldsymbol{\theta})|x]$ and $\nabla\log\pi(\boldsymbol{\theta}|x)$ are linearly related. As a simple example, consider $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n|\boldsymbol{\theta}$ are iid $\mathrm{N}(\boldsymbol{\theta},\boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ known, and the prior $\pi(\boldsymbol{\theta})$ is $\mathrm{N}(\boldsymbol{\mu},\boldsymbol{\Lambda})$. Then the posterior $\pi(\boldsymbol{\theta}|x)$ is $\mathrm{N}((\boldsymbol{I}-\boldsymbol{B})\boldsymbol{x}+\boldsymbol{B}\boldsymbol{\mu},(\boldsymbol{I}-\boldsymbol{B})\boldsymbol{\Sigma}/n)$, where $\boldsymbol{B} = n^{-1}\boldsymbol{\Sigma}(n^{-1}\boldsymbol{\Sigma}+\boldsymbol{\Lambda})^{-1}$. Then

$$\nabla\log\pi(\boldsymbol{\theta}|x) = [(\boldsymbol{I}-\boldsymbol{B})\boldsymbol{\Sigma}/n)^{-1}(\boldsymbol{\theta}-((\boldsymbol{I}-\boldsymbol{B})\boldsymbol{x}+\boldsymbol{B}\boldsymbol{\mu})).$$

Then $\nabla\log\pi(\boldsymbol{\theta}|x)$ is linearly related to $\boldsymbol{\theta}$ and accordingly $V(\boldsymbol{\theta}|x) = [V(\nabla\log\pi(\boldsymbol{\theta}|x))|x]^{-1}$.

Remark 3: It is important to point out that while the classical Cramer-Rao inequality is based on $\nabla\log f(x|\boldsymbol{\theta})$, the van Tress inequality is based on $\nabla\log(f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta}))$. Ours is based on $\nabla\log\pi(\boldsymbol{\theta}|x)$ instead.

Our next result provides a lower bound for $E[||\boldsymbol{\psi}(\boldsymbol{\theta}) - E(\boldsymbol{\psi}(\boldsymbol{\theta})|x)||^2|x] = E[\text{tr}(V(\boldsymbol{\psi}(\boldsymbol{\theta})|x))]$.

Theorem 2: $E[\text{tr}(V(\boldsymbol{\psi}(\boldsymbol{\theta})|x)] \geq E[\{\text{tr}(\frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}})\}^2|x]/E[||\nabla \log \pi(\boldsymbol{\theta})|x||^2|x]$.
Proof: In view of Lemma 1, it follows that

$$\text{tr}E[\{\boldsymbol{\psi}(\boldsymbol{\theta}) - E(\boldsymbol{\psi}(\boldsymbol{\theta})|x)\}\{\nabla \log \pi(\boldsymbol{\theta}|x)\}^T|x] = -\text{tr}E[(\frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}})|x] = -E[\text{tr}(\frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}})|x].$$

The above is equivalent to

$$-E[\text{tr}(\frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}})|x] = E[\{(\nabla \log \pi(\boldsymbol{\theta}|x)^T(\boldsymbol{\psi}(\boldsymbol{\theta}) - E(\boldsymbol{\psi}(\boldsymbol{\theta})|x))\}|x].$$

Now an application of the Cauchy-Schwarz inequality yields

$$[E\text{tr}(\frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}})|x]^2 \leq E[||\nabla \log \pi(\boldsymbol{\theta}|x)||^2|x]E\{(||\boldsymbol{\psi}(\boldsymbol{\theta}) - E(\boldsymbol{\psi}(\boldsymbol{\theta})|x)||^2)|x\}.$$

This yields the result. Remark 4. The above result can be generalized easily. Note that for an arbitrary positive definite matrix $\boldsymbol{B}$, and two random vectors $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$,

$$E(\boldsymbol{Z}_1^T\boldsymbol{Z}_2) = E(\boldsymbol{Z}_1^T\boldsymbol{B}^{-1/2}\boldsymbol{B}^{1/2}\boldsymbol{Z}_2) \leq E(\boldsymbol{Z}_1^T\boldsymbol{B}^{-1}\boldsymbol{Z}_1)E(\boldsymbol{Z}_2^T\boldsymbol{B}\boldsymbol{Z}_2).$$

Writing $\boldsymbol{Z}_1 = \boldsymbol{\psi}(\boldsymbol{\theta}) - E(\boldsymbol{\psi}(\boldsymbol{\theta})|x)$ and $\boldsymbol{Z}_2 = \nabla \log \pi(\boldsymbol{\theta}|x)$, one gets the weighted squared error posterior risk

$$E[(\boldsymbol{\psi}(\boldsymbol{\theta}) - E(\boldsymbol{\psi}(\boldsymbol{\theta})|x))^T\boldsymbol{B}^{-1}(\boldsymbol{\psi}(\boldsymbol{\theta}) - E(\boldsymbol{\psi}(\boldsymbol{\theta})|xa))]$$
$$\geq [E\text{tr}(\frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}})|x]^2/E[(\nabla \log \pi(\boldsymbol{\theta}|x))^T\boldsymbol{B}\nabla \log \pi(\boldsymbol{\theta}|x)|x].$$

References

Bhattacharyya, A. (1946). On some analogues of the amount of information and their use in statistical estimation. *Sankhyā: The Indian Journal of Statistics*, **1**, 1–14.

Bobrovsky, B.-Z., Mayer-Wolf, E., and Zakai, M. (1987). Some classes of global Cramér-Rao bounds. *The Annals of Statistics*, **1**, 1421–1438.

Borovkov, A. and Sakhanenko, A. (1980). On estimates of the expected quadratic risk (in russian). *Probability and Mathematical Statistics*, **1**, 185–95.

Brown, L. D. and Gajek, L. (1990). Information inequalities for the Bayes risk. *The Annals of Statistics*, **18**, 1578–1594.

Chapman, D. G. and Robbins, H. (1951). Minimum variance estimation without regularity assumptions. *The Annals of Mathematical Statistics*, **1**, 581–586.

Gassiat, E., Pollard, D., and Stoltz, G. (2013). Revisiting the van Trees inequality in the spirit of Hajek and Le Cam. *Unpublished Manuscript*, .

Ghosh, M. (1993). Cramer-Rao bounds for posterior variances. *Statistics & Probability Letters*, **17**, 173–178.

Gill, R. D. and Levit, B. Y. (1995). Applications of the van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli*, **1**, 59–79.

Hammersley, J. M. (1950). On estimating restricted parameters. *Journal of the Royal Statistical Society. Series B (Methodological)*, **12**, 192–240.

Van Trees, H. L. (2004). *Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory*. Wiley, New York.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 71–94
http://www.ssca.org.in/journal

Hierarchical Bayesian Probit Models for Sub-Areas and Ordinal Data

Lu Chen[1] **and Balgobin Nandram**[2]
[1]*National Institute of Statistical Sciences, Washington D.C., US*
[2]*Department of Mathematical Sciences*
Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA, US

Received: 12 April 2024; Revised: 16 May 2024; Accepted: 28 May 2024

Abstract

Many population-based surveys have polychotomous responses from a number of individuals in each household within small areas. An example is the second Nepal Living Standards Survey (NLSS II), in which health categorical data for each individual from the sampled households (sub-areas) are available in sampled wards (small areas). When the survey responses are ordinal, the sub-area hierarchical Bayesian probit models are considered to make inference about finite population proportions of individuals with different health statuses within the small areas. A standard assumption is that the ordered categorical responses are determined by an unobservable continuous variable. We discuss how to fit the model to avoid poor mixing problems in Markov chain Monte Carlo methods when simulating samples from the joint posterior distribution. The application is on health status data in the NLSS II, and the sub-area and the small area models are compared. The results show that the sub-area models are preferred over the small area models that ignore households (sub-areas) within the wards (areas). Our theoretical and methodological work can help provide small area official statistics for numerous surveys worldwide.

Key words: Bayesian Inference; Hierarchical Bayesian model; Metropolis-Hastings algorithm; Ordinal Variables; Small Area Estimation

1. Introduction

Most sample surveys are designed to provide reliable estimates of totals, means and other parameters of interest for large areas or domains (e.g., state level, national level). Such estimates are usually called "direct" estimates if they are only based on the domain-specific sample data. However, direct estimates are not reliable for the areas or domains for which only small samples or no samples are available. In recent years, more and more policymakers demand small area estimates. In fact, many new programs, such as fund allocation for needed areas, new educational or health programs, rely heavily on these estimates. Taking the cost and operation issues into consideration, it is not practical to conduct surveys with large

Corresponding Author: Lu Chen
Email: lchen@niss.org

enough sample sizes within the areas. In particular, small area estimation (SAE) deals with the problem of how to produce reliable "indirect" estimates of characteristics of interest for the small areas or domains.

Small area models are generally classified into two broad types. The basic area level model was introduced originally for SAE by Fay and Herriot (1979). The area level model is applied when individual auxiliary information is not available. Unit level model was first proposed for SAE by Battese, Harter and Fuller (1988). The generalized linear mixed model (GLMM) is one extension of the basic unit-level models. It was considered for SAE by MacGibbon and Tomberlin (1989). GLMM is useful in the case that the small area quantities of interest are finite population proportions.

In this paper, we are particularly interested in the small area models that can capture hierarchical structures, such as the Nepal Living Standard Survey II (NLSS II) data. The sampling scheme of NLSS II is a two-stage stratified sampling design. Nepal is stratified into primary sample units (wards) and within each ward, twelve households (sub-area) are systematically selected and all individuals from the selected households are interviewed. Although the above basic models are very popular and in common use in producing reliable estimates, the hierarchical structure of the data and the consistency between the estimates for different levels may not hold. Therefore, we focus on two-fold models, an important extension of basic small area models.

Hierarchical Bayesian methods are very popular in the two-fold models. Yan and Sedransk (2007) studied the case that the data follow a normal model with a two-stage (three-stage) hierarchical structure while the fitted model has a one-stage (two-stage) hierarchical structure by using posterior predictive p-values. Yan and Sedransk (2010) discussed the ability to detect a three-stage model when a two-stage model is actually fitted. Nandram (2016) and Chen and Nandram (2022) showed that it is important to consider the sample design within each area and proposed a two-fold small-area Beta-Binomial model. Lee *et al.* (2017) use a Bayesian method to infer about a finite population proportion when binary data are collected using a two-fold sample design from small areas. Erciulescu *et al.* (2018), Chen *et al.* (2022), and Nandram *et al.* (2023) illustrated hierarchical Bayesian approaches to provide estimates for the sub-area models with and without constraints. Chen and Nandram (2023) proposed a hierarchical Bayesian logistic regression model for binary data in small area estimation. This model is a unit level model with the sub-area effect. The results show that two-fold models can capture the heterogeneity between samples within not only small areas but also sub-areas.

Many population-based surveys have polychotomous responses from a number of individuals in each household within small areas, and many responses are ordered. For example, in the NLSS II, the answers to the question on health status range from 1 to 4, four options (excellent, good, fair, poor). There are few studies for ordinal response variables in SAE. Early papers on regression models for ordinal data include McKelvey and Zavoina (1975), McCullagh (1980), and Winship and Mare (1984). Nandram (1989) discussed the discrimination between the log-log link and logit link models for ordinal data. The textbook of Agresti (2010) gives a thorough treatment of ordinal data, while O'Connel (2006) provides applied researchers in the social sciences with accessible and comprehensive coverage of analysis for ordinal outcomes.

Albert and Chib (1993) discussed the algorithm to fit the Bayesian ordinal regression model with probit link. They introduced an underlying continuous variable, Z with a standard normal cumulative distribution function Φ. The ordinal response variable, Y_i is then observed in category t if Z_i comes from Normal$(\underset{\sim}{x}_i^T\underset{\sim}{\beta}, 1)$ between the cutpoints $\theta_{t-1} < Z_i \leq \theta_t$ and $\underset{\sim}{x}$ are the covariates. To capture the ordinal nature of the observed data, the cut-points are constrained to be monotonically increasing, $-\infty = \theta_0 < \theta_1 < \ldots < \theta_{T-1} < \theta_T = +\infty$. In addition, they assume that Z_i follows scale mixture of normal distributions, that is, Normal$(\underset{\sim}{x}_i^T\underset{\sim}{\beta}, \lambda_i^{-1})$. They assume that the underlying continuous variables follow the normal distribution without subgroup random effects. In this paper, we focus on the heterogeneous variances among the small areas and the subareas and conduct the subgroup analysis. We start with their models and build additional models with the small area and sub-area random effects.

For the probit analysis, Holmes and Held (2006) discussed Albert and Chib (1993) algorithm and showed that it gives a poorly mixing Gibbs sampler. They showed how to solve this mixing problem by adding latent variables and using the block Gibbs sampler (*i.e.*, some variables are drawn simultaneously). In this paper, we discuss how to fit the heterogeneous model to avoid poor mixing problems in Markov chain Monte Carlo methods when simulating samples from the joint posterior distribution.

In Section 2, a full description of the area and sub-area hierarchical Bayesian ordered probit models is given. In Section 3, we apply the models to the NLSS II data to predict the four health conditions of the household proportions of members for both sampled and nonsampled households. The comparisons between the small area models and the subarea models are presented. Finally, in Section 4, we make concluding remarks and discuss the future work. Technical details are given in the appendices.

2. Bayesian ordered probit models with covariates

In this section, we discuss two hierarchical Bayesian ordered probit models with covariates: the heterogeneous small area model and the heterogeneous sub-area model. We explain in detail about how to draw samples from the joint posterior distributions of heterogeneous models to avoid poor mixing problems in MCMC algorithm.

Suppose that the Y_i are categorical responses, falling in $t = 1, \ldots, T$, categories. Then Y_i follows a multinomial distribution with parameter $\underset{\sim}{p}$ where p_{it} denotes the probability that the i^{th} observation falls in the response category t. The cumulative probabilities are

$$\gamma_{it} = P(Y_i \leq t) = p_{i1} + \cdots + p_{it}.$$

Let $g(\cdot)$ denote a link function mapping probabilities to the real line, $g(\gamma_{it}) = \theta_t + \underset{\sim}{x}_i^T\underset{\sim}{\beta}$, where $\underset{\sim}{x}_i^T$ is a vector of explanatory variables for the i^{th} observation and $\underset{\sim}{\beta}$ is the corresponding set of regression parameters. The θ_t parameters are constant representing the baseline value for category t. Notice that the predictors do not include a column of ones for the intercept term since the constants are written explicitly. In this paper, we primary discuss the model with probit link function within the Bayesian paradigm, that is, $g(\cdot) = \Phi^{-1}(\cdot)$.

2.1. Heterogeneous small area model

In this section, we focus on the model with heterogeneous variances among small areas. In the small area models, we ignore the differences among households. Assume that there are ℓ areas, within the i^{th} area there are M_i individuals. For sampling, m_i individuals are selected from the M_i units available. Suppose that the independent response $y_{ij},\ i=1,\ldots,\ell,\ j=1,\ldots,m_i$ are observed and y_{ij} takes one of T ordered categories, *i.e.*, $y_{ij}\in\{1,2,...,T\}$.

The interest is to provide indirect estimates of the finite population proportions of the small areas in each category which are

$$\bar{P}^a_{it}=\frac{1}{M_i}\sum_{j=1}^{M_i} I(y_{ij}=t)=f^a_i\bar{I}^a_{sti}+(1-f^a_i)\bar{I}^a_{nsti}\ ,\ i=1,\ldots,\ell,\ t=1,\ldots,T,$$

where a denotes the small area estimates, and $\bar{I}^a_{sti}=\sum_{j=1}^{m_i} I(y_{ij}=t)/m_i$, $\bar{I}^a_{nsti}=\sum_{j=m_i+1}^{M_i} I(y_{ij}=t)/(M_i-m_i)$, and $f^a_i=m_i/M_i$ are sampled proportions, non-sampled proportions and sample fraction respectively in the small area model. Bayesian predictive inference is required for non-sample proportions.

Define the underlying continuous variable z_{ij}, where the z_{ij} follows Normal$(\underset{\sim}{x}^T_{ij}\underset{\sim}{\beta}+\nu_i,\lambda_i^{-1})$ with small area random effect ν_i. If $\theta_{t-1}<z_{ij}\le\theta_t$, then $y_{ij}=t$. Define $\theta_0=-\infty$ and $\theta_T=\infty$. θ_t is a constant representing the baseline value for category t. Since the variances of latent variable z_{ij} vary within areas, we called this ordered probit model heterogeneous small area model. Therefore, the small area Bayesian ordinal probit model with heterogeneous variances is

$$z_{ij}|\underset{\sim}{\nu},\underset{\sim}{\beta},\lambda_i,\underset{\sim}{x},\underset{\sim}{y}\overset{ind}{\sim}\text{Normal}(\underset{\sim}{x}^T_{ij}\underset{\sim}{\beta}+\nu_i,\lambda_i^{-1}),\tag{1}$$

where $\theta_{t-1}<z_{ij}\le\theta_t$ if $y_{ij}=t$ and the priors are

$$\nu_i|\delta^2\overset{iid}{\sim}\text{Normal}(0,\delta^2),\ i=1,\ldots,\ell,$$

$$\underset{\sim}{\beta}\sim\text{MN}\left(\underset{\sim}{\beta}_0,1000\Sigma_0\right),$$

$$\lambda_i|a\overset{iid}{\sim}\text{Gamma}(a,a),\ i=1,\ldots,\ell,$$

$$\pi(a,\delta^2)=\frac{1}{(1+a)^2}\frac{1}{(1+\delta^2)^2},$$

$$\pi(\theta_t)=(n-1)!I(\theta_1<\ldots<\theta_{T-1}),\ t=1,\ldots,T-1.$$

A diffuse prior is placed on the coefficient $\underset{\sim}{\beta}$. The prior of λ_i is gamma distribution, which makes the latent variable z_{ij} follows a student's t distribution. We placed the shrinkage priors on both a and δ^2 so that they are proper but with heavy tail. The detail of how to obtain a sample from joint posterior density is shown in Appendix A.

If the variances among the small areas are the same, no λ_i but 1, we call that model

a homogeneous small area model. That is,

$$z_{ij}|\underset{\sim}{\nu},\underset{\sim}{\beta},\underset{\sim}{x},\underset{\sim}{y} \overset{ind}{\sim} \text{Normal}(\underset{\sim}{x}_{ij}^T\underset{\sim}{\beta}+\nu_i,1), \tag{2}$$

where $\theta_{t-1} < z_{ij} \leq \theta_t$ if $y_{ij} = t$ and the priors are

$$\nu_i|\delta^2 \overset{iid}{\sim} \text{Normal}(0,\delta^2),\ i=1,\ldots,\ell,$$

$$\underset{\sim}{\beta} \sim \text{MN}\left(\underset{\sim}{\beta}_0, 1000\Sigma_0\right),$$

$$\pi(a,\delta^2) = \frac{1}{(1+a)^2}\frac{1}{(1+\delta^2)^2},$$

$$\pi(\theta_t) = (n-1)!I(\theta_1 < \ldots < \theta_{T-1}),\ t=1,\ldots,T-1.$$

Since the heterogeneous model is more general than the homogeneous model, the methods can be easily applied to the homogeneous model and the computations are simpler.

2.2. Heterogeneous sub-area model

We focus on the model with heterogeneous variances among sub-areas and discuss how to fit it. In sub-area models, we assume that there are ℓ areas, within the i^{th} area there are N_i sub-areas (households) and within the j^{th} sub-areas, there are M_{ij} individuals. For sampling, n_i sub-areas are sampled from the N_i sub-areas and all individuals are selected in sampled sub-areas, that is, $m_{ij} = M_{ij}$. Let $y_{ijk},\ k=1,\ldots,m_{ij}, j=1,\ldots,n_i, i=1,\ldots\ldots,\ell$, denote the categorical response and y_{ijk} takes one of T ordered categories, *i.e.*, $y_{ijk} \in \{1,2,...,T\}$.

The interest is also to provide estimates of the finite population proportions of small areas in each category which are

$$\bar{P}_{it}^s = \frac{1}{\sum_{j=1}^{N_i} M_{ij}} \sum_{j=1}^{N_i}\sum_{k=1}^{M_{ij}} I(y_{ijk}=t) = f_i^s \bar{I}_{sti}^s + (1-f_i^s)\bar{I}_{nsti}^s\ ,\ i=1,\ldots,\ell,\ t=1,\ldots,T,$$

where s denotes the estimates considering sub areas, and $\bar{I}_{sti}^s = \sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} I(y_{ijk}=t)/\sum_{j=1}^{n_i} m_{ij}$, $\bar{I}_{nsti}^s = \sum_{j=n_i+1}^{N_i}\sum_{k=1}^{M_{ij}} I(y_{ijk}=t)/\sum_{j=n_i+1}^{N_i} M_{ij}$, and $f_i^s = \sum_{j=1}^{n_i} M_{ij}/\sum_{j=1}^{N_i} M_{ij}$ are sampled proportions, non-sampled proportions and sample fraction in sub-area models respectively. Bayesian predictive inference is required for non-sample proportions.

Let the z_{ijk} follow Normal$(\underset{\sim}{x}_{ijk}^T\beta+\nu_i+\mu_{ij},\lambda_i^{-1})$ distribution with the small area random effects ν_i and sub-area random effects μ_{ij}. If $\theta_{t-1} < z_{ijk} \leq \theta_t$, then $y_{ijk} = t$. Since the variance of latent variable z_{ijk} are different among small areas, we call this ordered probit model a heterogeneous sub-area model.

Our sub-area Bayesian ordered probit model as

$$z_{ijk}|\underset{\sim}{\nu},\underset{\sim}{\beta},\lambda_i,\underset{\sim}{x},\underset{\sim}{y} \overset{ind}{\sim} \text{Normal}(\underset{\sim}{x}_{ijk}^T\underset{\sim}{\beta}+\nu_i+\mu_{ij},\lambda_i^{-1}), \tag{3}$$

where $\theta_{t-1} < z_{ijk} \leq \theta_t$ if $y_{ijk} = t$ and the priors are

$$\mu_{ij}|\sigma^2 \overset{iid}{\sim} \text{Normal}(0, \sigma^2),\ j = 1, \ldots, n_i,$$

$$\nu_i|\delta^2 \overset{iid}{\sim} \text{Normal}(0, \delta^2),\ i = 1, \ldots, \ell,$$

$$\underset{\sim}{\beta} \sim \text{MN}\left(\underset{\sim}{\beta}_0, 1000\Sigma_0\right),$$

$$\lambda_i|a \overset{iid}{\sim} \text{Gamma}(a, a),\ i = 1, \ldots, \ell,$$

$$\pi(a, \sigma^2, \delta^2) = \frac{1}{(1+a)^2}\frac{1}{(1+\delta^2)^2}\frac{1}{(1+\sigma^2)^2},$$

$$\pi(\theta_t) = (T-1)!I(\theta_1 < \ldots < \theta_{T-1}),\ t = 1, \ldots, T-1.$$

Similarly, the detail of how to obtain a sample from this joint posterior density is shown in Appendix B.

If the variances among areas are the same, no λ_i but 1, we call that model as homogeneous sub-area model, that is

$$z_{ijk}|\underset{\sim}{\nu}, \underset{\sim}{\beta}, \underset{\sim}{x}, \underset{\sim}{y} \overset{ind}{\sim} \text{Normal}(\underset{\sim}{x}_{ijk}^T\underset{\sim}{\beta} + \nu_i + \mu_{ij}, 1), \tag{4}$$

where $\theta_{t-1} < z_{ijk} \leq \theta_t$ if $y_{ijk} = t$ and the priors are

$$\mu_{ij}|\sigma^2 \overset{iid}{\sim} \text{Normal}(0, \sigma^2),\ j = 1, \ldots, n_i,$$

$$\nu_i|\delta^2 \overset{iid}{\sim} \text{Normal}(0, \delta^2),\ i = 1, \ldots, \ell,$$

$$\underset{\sim}{\beta} \sim \text{MN}\left(\underset{\sim}{\beta}_0, 1000\Sigma_0\right),$$

$$\pi(a, \sigma^2, \delta^2) = \frac{1}{(1+a)^2}\frac{1}{(1+\delta^2)^2}\frac{1}{(1+\sigma^2)^2},$$

$$\pi(\theta_t) = (T-1)!I(\theta_1 < \ldots < \theta_{T-1}),\ t = 1, \ldots, T-1.$$

Since the heterogeneous model is more general than the homogeneous model, the methods can be easily applied to the homogeneous model and the computations are simpler.

2.3. Prediction

In this paper, our interest is to predict the finite population proportions of the 102 sampled wards in both sampled and non-sampled households. The covariates of individuals in non-sampled households and the size of non-sampled households are unknown. Bayesian bootstrap (Rubin 1981) is used to draw them. The bootstrapping is done within sampled wards. The detail of the Bayesian bootstrap procedure is shown in Appendix C. Bayesian predictive inference for the individuals in the non-sampled sub-areas within the sampled small areas can be made once the set of samples are obtained from the posterior distribution.

For the small area models, we can draw samples of the non-sampled underlying vari-

able, $z_{ij}^{(h)},\ h=1,\ldots,M,\ j=m_i+1,\ldots,M_i,\ i=1,\ldots,\ell$, based on the likelihood functions in the models, where h denote the h^{th} samples drawn from the predictive distribution and we draw M samples in total. Then given the set of samples of $\underset{\sim}{\theta}$, the non-sampled responses, y_{ij}, can be predicted based on the criteria:

$$\theta_{t-1}^{(h)} < z_{ij} \leq \theta_t^{(h)},\ \text{then} \quad y_{ij}=t,\ t=1,\ldots,T.$$

For the sub-area models, we can draw samples of the non-sampled underlying variable, $z_{ijk}^{(h)},\ \ h=1,\ldots,M,\ k=1,\ldots,M_{ij},\ j=n_i+1,\ldots,N_i,\ i=1,\ldots,\ell$, based on the likelihood functions in the models. Then given the set of samples of $\underset{\sim}{\theta}$, the non-sampled responses, y_{ijk}, can be predicted based on the criteria:

$$\theta_{t-1}^{(h)} < z_{ijk} \leq \theta_t^{(h)},\ \text{then} \quad y_{ijk}=t,\ t=1,\ldots,T.$$

3. Application

3.1. Nepal living standards survey II

In this section, we describe the second Nepal Living Standards Survey (NLSS II) and the responses and the covariates. The performance of our method is studied using NLSS II, conducted in the years 2003-2004. NLSS is a national household survey in Nepal, actually population based (*i.e.*, interviews are done for all individual household members). Sometimes the head of the household answers the questions. NLSS follows the World Bank Living Standards Measurement Survey methodology with a two-stage stratified sampling scheme. It is an integrated survey which covers samples from the whole country. The main objective of the NLSS is to collect data from Nepalese households and provide information to monitor progress in national living standards. We study the polychotomous variable, health status, from the health section of the questionnaire.

The sampling design of NLSS II is two-stage stratified sampling. One selects the primary units (small areas) in the first stage and then some of the units (sub-areas) are selected from the secondary stage. Figure 1 shows that the area level of NLSS II is wards (circle) and the sub-area level is all selected households (house shape). That is, Nepal is stratified into primary sample units (wards) and within each ward, twelve households (sub-areas) are systematically selected. All household members in the sample were interviewed. Note that any analysis is done for each stratum.

According to the 2001 census data, only about 0.091% of households and only 0.904% of wards were sampled. NLSS II was designed to provide reliable estimates only at stratum level or even larger areas than stratum. It cannot give reliable estimates in small areas (ward or household level) since the sample sizes are too small. Therefore, we need to use statistical models to fit the available data and find reliable estimates in small areas.

3.2. Response variables and covariates

NLSS II has sparse counts of household members within the wards for four health status groups: excellent, good, fair and poor, denoted by 1 to 4. The distribution of all

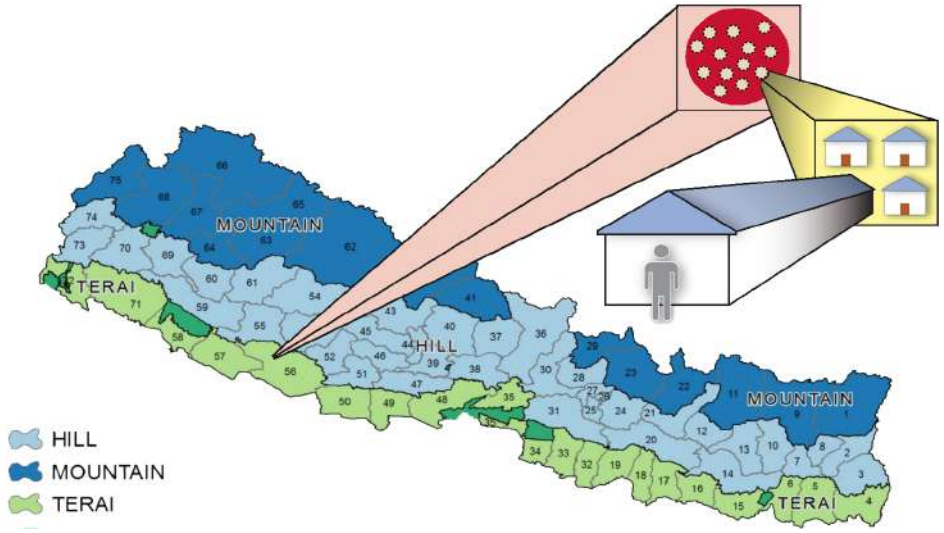


Figure 1: Illastration of NLSS II two-stage sampling design

responses of the health conditions in each stratum is shown in Table 1. Notice that the counts in the fair and poor cells are very sparse. There are six strata in the NLSS II. In this paper, we study the Rural Terai, the largest stratum in Nepal. It has 102 wards with 7,034 individuals in the sample of 12,239 wards in the population with 9,744,810 people. The number of people in the sample is 7,034 with 3,950 in the "excellent" cell, 2,926 in the "good" cell, 153 in the "fair" cell and 5 in the "poor" cell with percentages 56.1%, 41.6%, 2.1% and 0.02%. Notice that the counts in the last cell are mostly zeros.

Table 1: Distributions of wards and households in the sample and the distributions of the responses in each health cell

Stratum	Ward	Household	Individual	Excellent	Good	Fair	Poor
Mountains	32	384	1,949	1,262	658	24	4
Kathemandu	34	408	1,954	1,494	459	1	0
Urban Hills	28	336	1,467	820	626	20	1
Rural Hills	96	1,152	5,755	3,028	2,613	110	4
Urban Terai	34	408	2,104	1,239	811	52	2
Rural Terai	102	1,224	7,034	3,950	2,926	153	5

We choose four relevant covariates which can influence health status from the NLSS II survey for our sub-area logistic model and ordered probit models. They are age, nativity, sex and religion. We created binary variables: nativity (Indigenous = 1, Non-indigenous = 0), religion ((Hindu = 1, Non-Hindu = 0), and sex (Male = 1, Female = 0).

Table 2 shows some details of these 4 covariates. In the model fitting, we standardize age covariate. Elder age and children's age are more vulnerable than younger age. Indigenous people can have different health status from migrated people.

In NLSS II, the ordinal response variable of health status has 4 categories, from 1 to 4, where 1 means excellent health condition and 4 means poor health condition, respectively. When respondents answer this question, there is an underlying order among 1 to 4. Then the baseline values are $-\infty = \theta_0^* < \theta_1^* < \theta_2^* < \theta_3^* < \theta_4^* = +\infty$. In order to make the computation simpler, we subtract θ_2^* in each side and then $-\infty < \theta_1 < 0 < \theta_2 < +\infty$, where $\theta_1 = \theta_1^* - \theta_2^*$ and $\theta_2 = \theta_3^* - \theta_2^*$.

Table 2: Summaries of the four covariates: age, gender, nativity, and religion

Covariates		Frequency	Percentage
Age	0-14	7,765	38.32
	15-59	10,951	54.04
	60+	1,547	7.64
Gender	Male	9,763	48.18
	Female	10,500	51.82
Nativity	Indigenous	11,903	41.25
	Non-Indigenous	8,360	58.75
Religion	Hindu	16,378	80.83
	Non-Hindu	3,385	19.17

3.3. Numerical results

In this section, we show the numerical results and comparisons among the four models: homogeneous and heterogeneous wards models (small area models); homogeneous and heterogeneous household models (sub-area models).

3.3.1. MCMC diagnostics

For each of four models, we run 12,000 MCMC iterations, burn in 2,000 and thin every 10th to obtain 1,000 converged posterior samples. Table 3 and Table 5 give the p-values of the Geweke test and the effective sample sizes for the parameters $\underset{\sim}{\beta}$, θ_1, θ_2 and δ^2 of the homogeneous models. Table 4 gives the p-values of the Geweke test and the effective sample sizes for the parameters $\underset{\sim}{\beta}$, θ_1, θ_2, a and δ^2 of the heterogeneous area model. Table 6 gives the p-values of the Geweke test and the effective sample sizes for the parameters $\underset{\sim}{\beta}$, θ_1, θ_2, a, δ^2 and σ^2 for the heterogeneous household model. The p-values are all large, so we do not reject the null hypothesis test which is that the Markov chain is in the stationary distribution. The effective sample sizes are not too far away from 1,000. These model diagnostic summaries indicate that the MCMC chains converge and strongly mixing.

3.3.2. Model comparisons

For evaluating and comparing these models, the Bayesian posterior predictive p-value (Meng,1994), the deviance information criterion (DIC) and the logarithm of the pseudo marginal likelihood (LPML) are computed.

In the ordered probit models, denote $\Omega = (\underset{\sim}{\nu}, \underset{\sim}{\mu}, \underset{\sim}{\theta}, \underset{\sim}{\beta}, \underset{\sim}{\lambda})$. Since the responses y_{ijk} follow

Table 3: Summary of MCMC diagnostics: posterior mean, posterior standard deviation, the p-values of the Geweke test and the effective sample sizes for the homogeneous wards model

Model	Homogeneous Wards Model			
	Mean	SD	Geweke pval	Effective Size
β_1	0.08817	0.02238	0.34	1000
β_2	0.00038	0.02818	0.65	900
β_3	-0.02079	0.02555	0.57	1000
β_4	-0.37098	0.02379	0.71	1123
θ_1	-0.50001	0.00021	0.11	1000
θ_2	0.59635	0.60952	0.54	1000
δ^2	0.59320	0.11752	0.35	1092

Table 4: Summary of MCMC diagnostics: posterior mean, posterior standard deviation, the p-values of the Geweke test and the effective sample sizes for the heterogeneous wards model

Model	Heterogeneous Wards Model			
	Mean	SD	Geweke pval	Effective Size
β_1	0.1711	0.0148	0.78	1000
β_2	-0.0404	0.0164	0.48	910
β_3	-0.0074	0.0140	0.46	1000
β_4	-0.3467	0.0103	0.17	1000
θ_1	-0.5028	0.0044	0.28	908
θ_2	0.5963	0.6706	0.69	1000
a	0.8645	0.9327	0.51	1000
δ^2	1.4927	0.0396	0.12	875

Table 5: Summary of MCMC diagnostics: posterior mean, posterior standard deviation, the p-values of the Geweke test and the effective sample sizes for the homogeneous household model

Model	Homogeneous Household Model			
	Mean	SD	Geweke pval	Effective Size
β_1	0.10234	0.02344	0.88	1000
β_2	-0.00755	0.02477	0.71	1006
β_3	-0.02113	0.02562	0.93	888
β_4	-0.05413	0.02141	0.93	598
θ_1	-0.50000	0.00020	0.67	1000
θ_2	0.58684	0.10882	0.66	1000
δ^2	0.55568	0.13674	0.78	901
σ^2	0.03291	0.01905	0.30	855

Table 6: Summary of MCMC diagnostics: posterior mean, posterior standard deviation, the p-values of the Geweke test and the effective sample sizes for the heterogeneous household model

Model	Heterogeneous Household Model			
	Mean	SD	Geweke pval	Effective Size
β_1	0.1752	0.0149	0.78	1000
β_2	-0.0132	0.0177	0.39	1000
β_3	-0.0403	0.0159	0.69	1000
β_4	0.0089	0.0192	0.24	1000
θ_1	-0.4843	0.0044	0.51	888
θ_2	0.5584	0.0622	0.65	1000
a	0.8498	0.0997	0.52	1000
δ^2	1.6472	0.9333	0.16	1000
σ^2	0.3367	0.1957	0.68	1000

multinomial distributions, we consider a measure of form

$$T(\underset{\sim}{y},\Omega)=\sum_{t=1}^{T}\sum_{i=1}^{\ell}\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}}\frac{(I(y_{ijk}=t)-p_{ijkt})^2}{n_t p_{ijkt}(1-p_{ijkt})},$$

where $n_t=\sum_{i=1}^{\ell}\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}}I(y_{ijk}=t)$ is the total number of y_{ijk} in t category and $p_{ijkt}=\Phi(\theta_t-\underset{\sim}{x}_{ijk}^T\underset{\sim}{\beta}-\nu_i-\mu_{ij})$. We calculate $T(\underset{\sim}{y}^{rep},\Omega)$ for each of 1,000 samples, and then seeing what percent are above single calculated $T(\underset{\sim}{y}^{obs},\Omega)$. The Bayesian posterior predictive p-value (BPP) is used in order to check the discrepancy between data and the posited model. The BPPs of all models shown in Table 7 are not in the extreme range (close to 0 or 1). Therefore, they are appropriate and adequate to make inference for the finite population proportions of interest. Note that the BPP cannot be used for ranking the models, but for checking if the model is good or not.

In addition, we calculated their DICs and LPMLs. The deviance information criterion (DIC) (Spiegelhalter *et al.* 2002) is a Bayesian measure of goodness-of-fit,

$$\text{DIC}=2\left\{\frac{1}{M}\sum_{h=1}^{M}D(\underset{\sim}{y},\Omega^{(h)})\right\}-D(\underset{\sim}{y},\hat{\Omega}),$$

where $\hat{\Omega}$ is a point estimate for Ω such as the mean of the posterior simulations, $\Omega^{(h)}$ are posterior simulations and $D(\underset{\sim}{y},\hat{\Omega})=-\log f(y|\Omega)$. DIC has been suggested as a criterion of model fit when the goal is to pick a model with best out-of-sample predictive power. A smaller value of DIC indicates a better fit and it provides reasonable assessments of model fit while considering the model complexity.

Similar to the DIC, LPML is also based on the same cross-validation (leave-one-out) procedure. A summary statistic of the conditional predictive ordinate (CPO) values is LPML. CPO is defined as the predictive density of observation i given all the other observa-

tions, that is, $\text{CPO}_i = p(y_i|y_{(i)}) = \int p(y_i|\Omega)p(\Omega|y_{(i)})d\Omega$, where $y_{(i)}$ is the data y without i^{th} observation. If observations are conditionally independent, a harmonic mean approximation of CPO is $\widehat{\text{CPO}_i} = \left\{\frac{1}{M}\sum_{h=1}^{M}\frac{1}{p(y_i|\Omega^{(h)})}\right\}^{-1}$, where $\Omega^{(h)}, h = 1, \ldots, M$ are samples from the posterior distribution. Then,

$$\text{LPML} = \sum_i \log(\widehat{\text{CPO}_i}).$$

Larger values of LPML indicate better fitting models (Geisser and Eddy 1979).

The DICs of the heterogeneous and homogeneous wards models are 1,852.58 and 4,039.93 respectively. The LPML of heterogeneous and homogeneous wards models are -1,096.38 and -1,838.45 respectively. So the heterogeneous ward model is better than the homogeneous one. The DICs of heterogeneous and homogeneous household model are 1,329.35 and 1,927.68 respectively. The LPML of the heterogeneous and homogeneous area models are -1,056.01 and -1,272.85 respectively. So the heterogeneous household model is better than the homogeneous one. Overall, based on the DIC and LPML, the heterogeneous household model has the smallest DIC and the largest LPML. The household models have relatively small DIC and large LPML. The household models are better when fitting the NLSS II health data.

Table 7: Comparison of BPP and DIC among four models: heterogeneous household model (HES), heterogeneous wards model (HEA), homogeneous household model (HOS), homogeneous wards model (HOA) for NLSS II data

Model	BPP	DIC	LPML
HEA	0.415	1852.58	-1096.38
HES	0.475	1329.35	-1056.01
HOA	0.155	4039.93	-1838.45
HOS	0.280	1927.68	-1272.85

We are interested in the finite population proportions of four health conditions in the small areas. We use all four ordered probit models to predict the nonsampled households in the 102 sampled wards. Bayesian bootstraps are used to generate unknown household sizes and nonsampled covariates within sampled wards and the bootstrapping is done within wards. The 2001 Census could potentially provide these two pieces of information, but there is a mismatch between the households in the census and the NLSS II (a record linkage can be performed). We note, however, that there is linkage between the wards, but this information is not useful to household estimates. In this application, we know the total number of households and individuals in each sampled wards, and we have sampled household information. Therefore, we decide to use these information in the Bayesian bootstrap approach to generate nonsampled household sizes and corresponding nonsampled covariates within sampled wards.

Based on 1,000 samples of parameters from the joint posterior distribution, we get 1,000 values of $\bar{P}_{it}$; order these values and pick the 95% prediction interval to be $(\bar{P}_{it}^{(25)}, \bar{P}_{it}^{(975)})$, $t = 1, 2, 3, 4$, where the values are arranged in increasing order.

The health status proportions of the 102 sampled wards based on both sampled and non-sampled households ($\bar{P}_t,\ t = 1, 2, 3, 4$) under all four models are shown in Figure 2. The proportions in excellent health condition are similar among all four models. The estimates from the heterogeneous household model are slightly less than those from the other models. The proportions in good health condition from both household models are more than those from the small area models. The proportions of fair condition and poor condition from the area models are relatively similar. The proportions of fair condition from the household models are larger than the proportions of poor conditions, which is consistent with the observed data. The error bars are the 95% credible intervals of $\bar{P}_t$. We can notice that the 95% credible intervals of the estimates in the homogeneous wards model are widest among all four models. The 95% credible intervals in the heterogeneous model have relatively the narrowest among all four models.

We examine plots to further compare the predictive inference of the finite population proportions of the four health conditions between the heterogeneous ward model and the heterogeneous household model. Figure 3 shows the comparison of the finite population proportions of four health conditions in each household within the sampled wards respectively between two models. One of our interest is to provide estimates for sampled wards. We can get the finite population proportions of health status in each sampled wards by taking the average on those estimates for households in each ward. Figure 4 shows the comparison of the finite population proportions of the four health conditions in sampled wards respectively between the two models. We can see that the points do not fall reasonably well on the 45^o line, which indicates that everything being equal, the model with sub-area random effects can capture more information, the heterogeneity of different households (sub-areas).

4. Concluding remarks and future works

In this paper, we study several hierarchical Bayesian ordered probit models for polychotomous responses. The sub-area models can capture the heterogeneity among the sub-areas (households) within the small areas (wards) and borrow strength from the sub-areas to obtain more efficient estimators. A full Bayesian analysis is provided for each model and predictive inference of the finite population proportions of the small areas is conducted. We have demonstrated our application to health status data from NLSS II.

We discussed one posterior computation algorithm to avoid poor mixing problems that the Gibbs sampler may cause. NLSS II health data were used in order to examine the performance of two models. We have performed a Bayesian predictive inference for the finite population proportion of each health status in the sampled wards based on the sampled and non-sampled households. BPP and DIC are used to assess and compare our ordinal probit models. The sub-area models perform better than the small area models.

In the paper, we assume the samples are self-weighted. However, if the sample unit cannot represent the target population, survey weights should be used to adjust selection bias. In the future, incorporating survey weights into the models can be explored. The observed biased samples actually followed a weighted distribution instead of the original distribution that the random samples follow. In order to predict and make inference about the finite population, the surrogate sampling approach by Nandram (2007) can be used to predict the finite population proportions.

We focus on parametric statistical models in this paper. Nonparametric Bayesian models using the stick-breaking priors can be considered to robustify the inference by embedding parametric models in nonparametric models. Ishwaran and James (2001) discussed the Gibbs samplers that can be used to fit posteriors of Bayesian hierarchical models based on stick-breaking priors. They are more flexible and better than the stick breaking prior of the Dirichlet process.

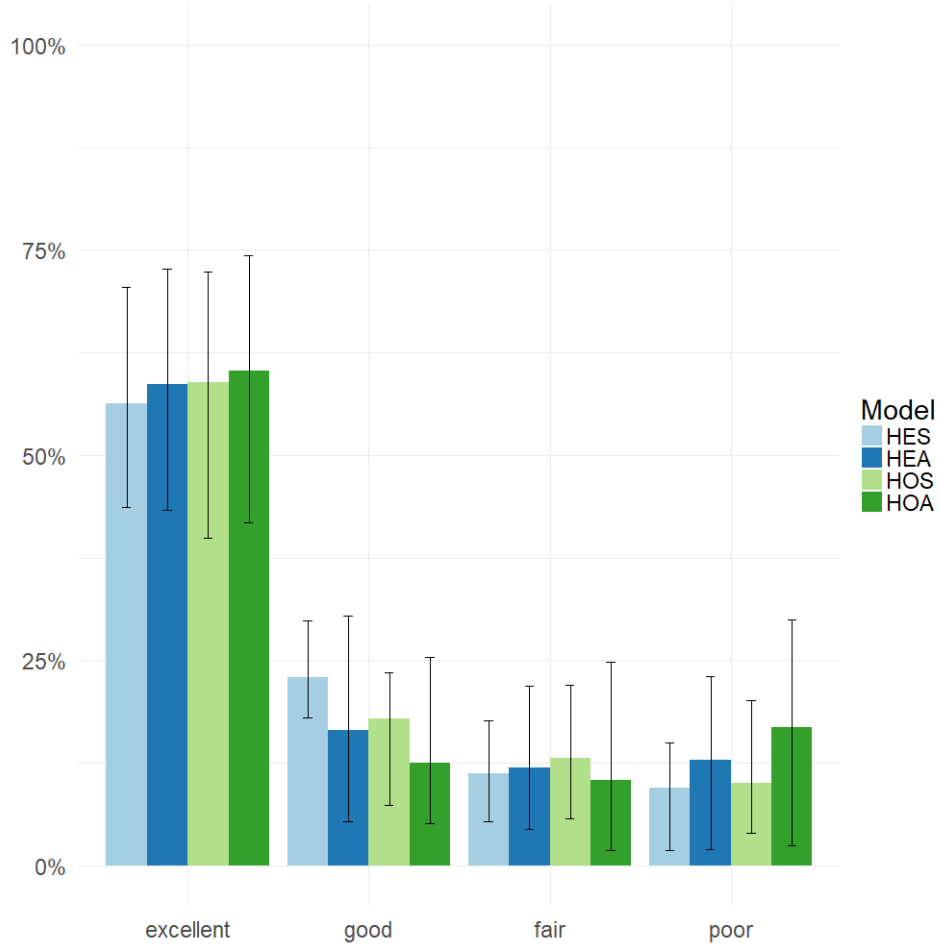


Figure 2: Comparison of finite population proportions of each health condition cell among four models

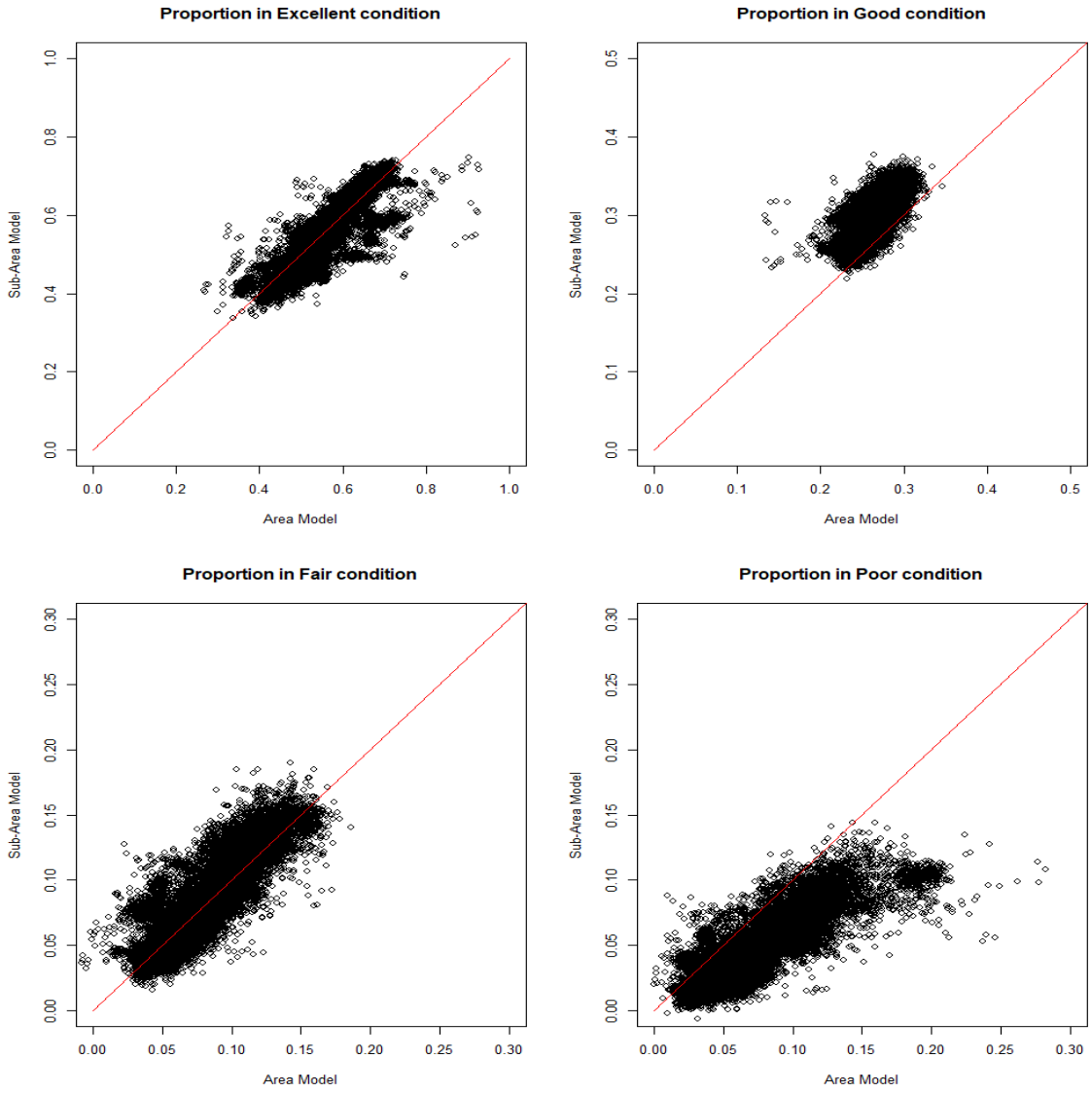


Figure 3: Comparison of the wards model and household models for prediction of the finite population proportions of 4 different health conditions of household

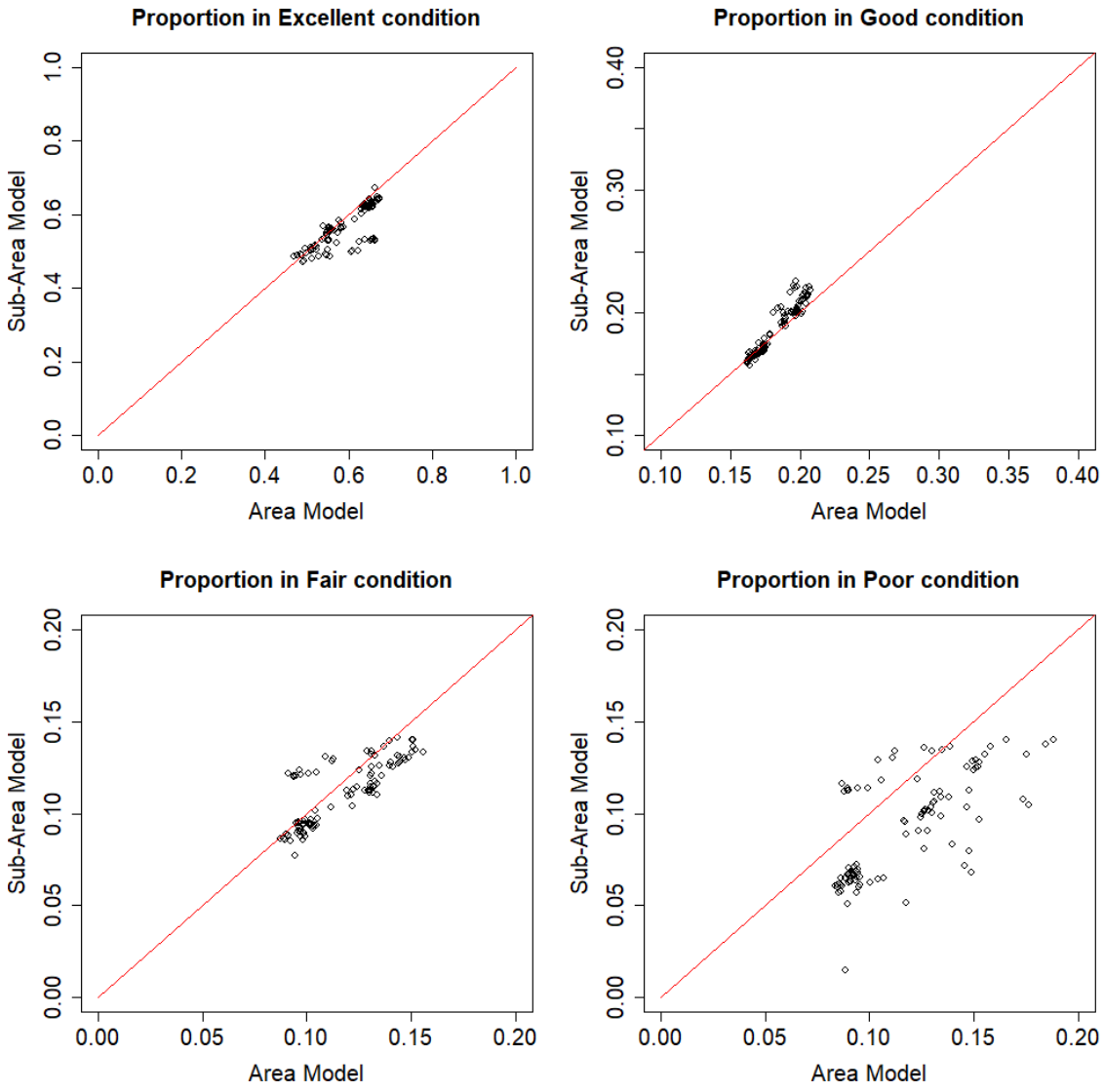


Figure 4: Comparison of the wards model and household models for prediction of the finite population proportions of 4 different health conditions of wards

Acknowledgements

We are very pleased and honored to be invited to write a paper for the special issue in memory of Professor C.R. Rao. We thank the reviewer for these informative comments.

References

Agresti, A. (2010). *Analysis of Ordinal Categorical Data.* John Wiley & Sons, Vol. **656**. DOI:10.1002/9780470594001.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669-679. DOI: 10.2307/2290350.

Battese, G. E., Harter, R., and Fuller, W. A. (1988). An error components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28-36.

Central Bureau of Statistics Thapathali. Nepal Living Standards Survey 2003/04. *Statistical Report*, Kathmandu, Nepal.

Chen, L. and Nandram, B. (2023). Bayesian logistic regression model for sub-areas. *Stats*, **6**, 209-231. DOI: 10.3390/stats6010013.

Chen, L. and Nandram, B. (2020). A hierarchical Bayesian Beta-Binomial model for sub-areas. In *Annual Conference of the Society of Statistics, Computer and Applications*, 23-40.

Chen, L., Nandram, B., and Cruze, N. B. (2022). Hierarchical Bayesian model with inequality constraints for US county estimates. *Journal of Official Statistics*, **38**, 709-732. DOI: 10.2478/jos-2022-0032.

Erciulescu, A. L., Cruze, N. B., and Nandram, B. (2019). Model-based county level crop estimates incorporating auxiliary sources of information. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, **182**, 283-303. DOI: 10.1111/rssa.12390.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269-277. DOI: 10.1080/01621459.1979.10482505.

Fuller, W. A. and Goyeneche, J. J. (1998). Estimation of the state variance component. *unpublished*.

Geisser, S. (1980). Discussion on sampling and Bayes' inference in scientific modeling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, 383-430.

Geisser, S. and Eddy, W. F. (1977). Predictive sample reuse approaches to model selection. *University of Minnesota*.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161-173. DOI: 10.1198/01621450 1750332758.

Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, **1**, 145-168. DOI: 10.1214/06-BA105.

Lee, D., Nandram, B., and Kim, D. (2017). Bayesian predictive inference of a proportion under a two-fold small area model with heterogeneous correlations. *Survey Methodology*, **43**, 69-93.

Malec, D., Davis, W., and Cao, X. (1999). Model-based small area estimates of overweight prevalence using sample selection adjustment. *Statistics in Medicine*, **18**, 3189-3200. DOI: 10.1002/(sici)1097-0258(19991215)18:23<3189::aid-sim309>3.0.co;2-c.

MacGibbon, B. and Tomberlin, T. J. (1989). Small area estimation of proportions via empirical Bayes techniques. *Survey Methodology*, **15**, 237-252.

McKelvey R.D. and Zavoina W. (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology*, **4**, 103-120. DOI: 10.1080/0022250X.1975.9989847.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, **42**, 109-127. DOI: 10.1111/j.2517-6161.1980.tb01109.x.

Meng, X.L. (1994). Posterior predictive P-values. *The Annals of Statistics*, **22**, 1142-1160. DOI:10.1214/AOS/1176325622.

Nandram, B. (1989). Discrimination between complementary log-log and logistic model for ordinal data. *Communications in Statistics, Theory and Methods*, **18**, 2155-2164.

Nandram, B. (1998). A Bayesian analysis of the three-stage hierarchical multinomial model. *Journal of Statistical Computation and Simulation*, **61**, 97-126.

Nandram, B. (2000). Bayesian generalized linear models for inference about small areas. In *Generalized linear models: A Bayesian Perspective, Eds. D. K. Dey, S. K. Ghosh and B. K. Mallick, Marcel Dekker, Chapter 6*, 91-114.

Nandram, B. (2007). Bayesian predictive inference under information sampling via surrogate samples. *Bayesian Statistics and Its Applications*, 357-374.

Nandram, B. (2016). Bayesian predictive inference of a proportion under a two-fold small area model. *Journal of Official Statistics*, **32**, 187-207. DOI: 10.1515/jos-2016-0009.

Nandram, B., Chen, L., Fu, S-T., and Manandhar, B. (2018). Bayesian logistic regression for small areas with numerous households. *Statistics and Application*, **1**, 171-205. arXiv:1806.00446 [stat.ME].

Nandram, B., Chen, L., and Manandhar, B. (2018). Bayesian analysis of multinomial counts from small areas and sub-areas. *In JSM proceedings: Section on Survey Research Methods*, 1140–1162. Vancouver.

Nandram, B. and Erhardt, E. (2005). Fitting Bayesian two-stage generalized linear models using random samples via the SIR algorithm. *Sankhya*, **66**, 733-755. DOI: 10.2307/2505 3398.

Nandram, B., Cruze, N. B., and Erciulescu, A. L. (2023). Bayesian small area models under inequality constraints with benchmarking and double shrinkage. *Survey Methodology*, **49**, 517-546. http://www.statcan.gc.ca/pub/12-001-x/2023002/article/00004-eng.htm.

Nandram, B. and Sedransk, J. (1993). Bayesian predictive inference for a finite population proportion: two-stage cluster sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 399-408. DOI:10.1111/J.2517-6161.1993.TB01910.X.

O'Connell, A. A. (2006). Logistic regression models for ordinal response variables. *Sage*, **146**, 140-153. DOI:10.4135/9781412984812.

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, **9**, 130-134. DOI:10.1214/AOS/1176345338.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 583–639. DOI:10.1111/1467-9868.00353.

Winship, C. and Mare, R. D. (1984). Regression models with ordinal variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 109-142. DOI:10.2307/2095465.

Yan, G. and Sedransk, J. (2007). Bayesian diagnostic techniques for detecting hierarchical structure. *Bayesian Analysis*, **2**, 735-760. DOI:10.1214/07-BA230.

Yan, G. and Sedransk, J. (2010). A note on Bayesian residuals as a hierarchical model diagnostic technique. *Statistical Papers*, **51**, 1-10. DOI:10.1007/S00362-007-0111-2.

APPENDIX

A. Computation method for the heterogeneous small area model

Using Bayes' theorem, the joint posterior distribution of the heterogeneous small area model in Section 2.1 is

$$\begin{aligned}\pi(\underset{\sim}{z},\nu,\underset{\sim}{\beta},\underset{\sim}{\lambda},\underset{\sim}{\theta},a,\delta^2|\underset{\sim}{y}) \propto & \prod_{i=1}^{\ell}\prod_{j=1}^{m_i}\left\{\sqrt{\lambda_i}e^{-\frac{\lambda_i}{2}(z_{ij}-\underset{\sim}{x}_{ij}^T\underset{\sim}{\beta}-\nu_i)^2}\sum_{t=1}^{T}[I(y_{ij}=t,\theta_{t-1}<z_{ij}\le\theta_t)]\right\}\\ & \times\left(\frac{1}{\delta^2}\right)^{\frac{\ell}{2}}\prod_{i=1}^{\ell}\left\{e^{-\frac{1}{2\delta^2}\nu_i^2}\right\}\times\exp\left\{-(\underset{\sim}{\beta}-\beta_0)^T(1000\Sigma_0)^{-1}(\underset{\sim}{\beta}-\beta_0)\right\}\\ & \times\left\{\prod_{i=1}^{\ell}\frac{a^a\lambda_i^{a-1}e^{-a\lambda_i}}{\Gamma(a)}\right\}\frac{1}{(1+a)^2}\frac{1}{(1+\delta^2)^2}.\end{aligned}$$

In order to jointly draw samples of $\underset{\sim}{z}$ and $\underset{\sim}{\lambda}$, we integrate out $\underset{\sim}{z}$ from the joint posterior distribution $\pi(\underset{\sim}{z},\underset{\sim}{\lambda}|\underset{\sim}{\nu},\underset{\sim}{x},\underset{\sim}{\beta},\underset{\sim}{\theta},a,\underset{\sim}{y})$. That is,

$$\begin{aligned}\pi(\lambda_i|\underset{\sim}{\nu},\underset{\sim}{x},\underset{\sim}{\beta},\underset{\sim}{\theta},a,\underset{\sim}{y}) &= \int\pi(z,\lambda_i|\underset{\sim}{\nu},\underset{\sim}{x},\underset{\sim}{\beta},\underset{\sim}{\theta},a,\underset{\sim}{y})d\underset{\sim}{z}\\ &\propto\prod_{j=1}^{m_i}\left\{\int\sqrt{\lambda_i}e^{-\frac{\lambda_i}{2}(z_{ij}-\underset{\sim}{x}_{ij}^T\underset{\sim}{\beta}-\nu_i)^2}\sum_{t=1}^{T}[I(y_{ij}=t,\theta_{t-1}<z_{ij}\le\theta_t)]\,d\underset{\sim}{z}\right\}\frac{a^a\lambda_i^{a-1}e^{-a\lambda_i}}{\Gamma(a)}\\ &=\prod_{j=1}^{m_i}\left\{\sum_{t=1}^{T}\int_{\theta_{t-1}}^{\theta_t}\left[\sqrt{\lambda_i}e^{-\frac{\lambda_i}{2}(z_{ij}-\underset{\sim}{x}_{ij}^T\underset{\sim}{\beta}-\nu_i)^2}\right]I(y_{ij}=t)dz_{ij}\right\}\frac{a^a\lambda_i^{a-1}e^{-a\lambda_i}}{\Gamma(a)}\\ &=\prod_{j=1}^{m_i}\left\{\sum_{t=1}^{T}\left[\Phi\left(\sqrt{\lambda_i}(\theta_t-\underset{\sim}{x}_{ij}^T\underset{\sim}{\beta}-\nu_i)\right)-\Phi\left(\sqrt{\lambda_i}(\theta_{t-1}-\underset{\sim}{x}_{ij}^T\underset{\sim}{\beta}-\nu_i)\right)\right]I(y_{ij}=t)\right\}\\ &\quad\times\frac{a^a\lambda_i^{a-1}e^{-a\lambda_i}}{\Gamma(a)}.\end{aligned}$$

Then Metropolis-Hastings algorithm is used to draw samples of λ_i from the marginal conditional distribution. Given λ_i and samples of $\underset{\sim}{\beta},\nu_i$ and $\underset{\sim}{\theta}$, we draw z_{ij} from truncated normal $\mathrm{N}(\underset{\sim}{x}_{ij}^T\underset{\sim}{\beta}+\nu_i,\lambda_i^{-1})$, where $y_{ij}=t$ if $\theta_{t-1}<z_{ij}\le\theta_t$.

To implement the Gibbs sampler once we get a sample of $\underset{\sim}{\lambda}$ and $\underset{\sim}{z}$, we need to draw samples from the full conditional posterior distributions of $\underset{\sim}{\nu}$, $\underset{\sim}{\beta}$, a, δ^2 and $\underset{\sim}{\theta}$.

First, the conditional distribution of $\underset{\sim}{\nu}$ is

$$\nu_i|\lambda_i,\underset{\sim}{z},\underset{\sim}{\beta},\delta^2,\underset{\sim}{y}\overset{ind}{\sim}\mathrm{Normal}\left(\frac{\lambda_i\sum_{j=1}^{n_i}(z_{ij}-\underset{\sim}{x}_{ij}^T\underset{\sim}{\beta})}{\frac{1}{\delta^2}+n_i\lambda_i},(\frac{1}{\delta^2}+n_i\lambda_i)^{-1}\right).$$

Second, the conditional distribution of $\underset{\sim}{\beta}$ is $\underset{\sim}{\beta}|\underset{\sim}{\nu},\underset{\sim}{\lambda},\underset{\sim}{z},\underset{\sim}{x},\underset{\sim}{y} \sim \text{MN}\left(\underset{\sim}{\hat{\beta}},\Sigma_{\underset{\sim}{\hat{\beta}}}\right)$, where

$$\underset{\sim}{\hat{\beta}} = \Sigma_{\underset{\sim}{\hat{\beta}}}\left(\sum_{i=1}^{\ell}\sum_{j=1}^{m_j}\lambda_i(z_{ij}-\nu_i)\underset{\sim}{x}_{ij} + (1000\Sigma_0)^{-1}\underset{\sim}{\beta}_0\right),$$

$$\Sigma_{\underset{\sim}{\hat{\beta}}} = \left(\sum_{i=1}^{\ell}\sum_{j=1}^{n_i}\lambda_i z_{ij}\underset{\sim}{x}_{ij}\underset{\sim}{x}_{ij}^T + (1000\Sigma_0)^{-1}\right)^{-1}.$$

Third, the fully conditional distribution of θ_t, given $\underset{\sim}{z}$, $\underset{\sim}{\theta}_{(t)} = \{\theta_s, s \neq t\}$ and data, is given by

$$\pi(\theta_t|\underset{\sim}{z},\underset{\sim}{\theta}_{(t)},\underset{\sim}{y}) \propto \prod_{i=1}^{\ell}\prod_{j=1}^{m_i}\left[I(y_{ij}=t,\theta_{t-1}<z_{ij}\leq\theta_t) + I(y_{ij}=t+1,\theta_t<z_{ij}\leq\theta_{t+1})\right].$$

Notice that this conditional density is uniform density on the interval. That is

$$\theta_t|\underset{\sim}{z},\underset{\sim}{\theta}_{(t)},\underset{\sim}{y} \sim \text{Uniform}\Big(\max\{\max\{z_{ij},y_{ij}=t\},\theta_{t-1}\},\min\{\min\{z_{ij},y_{ij}=t+1\},\theta_{t+1}\}\Big).$$

Fourth, given the sample of $\underset{\sim}{\lambda}$, we can use grid method to draw a. Transform a to $\phi_1 = \frac{a}{1+a}$, which is in $(0,1)$. The conditional posterior distribution of ϕ_1 is

$$\pi(\phi_1|\underset{\sim}{\lambda}) \propto \left(\prod_{i=1}^{\ell}\frac{a^a\lambda_i^{a-1}e^{-a\lambda_i}}{\Gamma(a)}\right)_{\phi_1=\frac{a}{1+a}}.$$

Fifth, to draw δ^2 we also use the grid method. Transform δ^2 to $\phi_2 = \frac{\delta^2}{1+\delta^2}$, which is in $(0,1)$. The conditional posterior distribution of ϕ_2 is

$$\pi(\phi_2|\underset{\sim}{\nu}) \propto \left\{\left(\frac{1}{\delta^2}\right)^{\frac{\ell}{2}}\exp\left(-\frac{1}{2\delta^2}\sum_{i=1}^{\ell}\nu_i^2\right)\right\}|_{\phi_2=\frac{\delta^2}{1+\delta^2}}.$$

To implement the algorithm, we chose starting points $\underset{\sim}{\beta}^{(0)},\underset{\sim}{\theta}^{(0)}$ equal to the maximum likelihood estimators (MLE) based on the previous paper by Chen and Nandram (2023), $\lambda_i^{(0)} = 1$ and $\nu_i^{(0)} = 1$. We first draw a and δ^2 using grid method, and then jointly draw a sample of $\underset{\sim}{\lambda}$ and $\underset{\sim}{z}$, and simulate from the conditional distribution of ν_i, $\underset{\sim}{\beta}$ and θ_t.

B. Computation method for the heterogeneous sub-area model

Using Bayes' theorem, the joint posterior distribution of the Heterogeneous Sub-Area Model in Section 2.2 is

$$\begin{aligned}
&\pi(\underset{\sim}{z},\underset{\sim}{\nu},\underset{\sim}{\mu},\underset{\sim}{\beta},\underset{\sim}{\lambda},a,\sigma^2,\delta^2|\underset{\sim}{y}) \\
&\quad\propto \prod_{i=1}^{\ell}\prod_{j=1}^{n_i}\prod_{k=1}^{m_{ij}}\left\{\sqrt{\lambda_i}e^{-\frac{\lambda_i}{2}(z_{ijk}-\underset{\sim}{x}_{ijk}^T\underset{\sim}{\beta}-\nu_i-\mu_{ij})^2}\sum_{t=1}^{T}\left[I(y_{ijk}=t,\theta_{t-1}<z_{ijk}\le\theta_t)\right]\right\} \\
&\quad\times\left(\frac{1}{\sigma^2}\right)^{\frac{\ell}{2}}\prod_{i=1}^{\ell}\prod_{j=1}^{n_i}\left\{e^{-\frac{1}{2\sigma^2}\mu_{ij}^2}\right\}\left(\frac{1}{\delta^2}\right)^{\frac{\ell}{2}}\prod_{i=1}^{\ell}\left\{e^{-\frac{1}{2\delta^2}\nu_i^2}\right\} \\
&\quad\times\exp\left\{-(\underset{\sim}{\beta}-\beta_0)^T(1000\Sigma_0)^{-1}(\underset{\sim}{\beta}-\beta_0)\right\}\left\{\prod_{i=1}^{\ell}\frac{a^a\lambda_i^{a-1}e^{-a\lambda_i}}{\Gamma(a)}\right\} \\
&\quad\times\frac{1}{(1+a)^2}\frac{1}{(1+\delta^2)^2}\frac{1}{(1+\sigma^2)^2}.
\end{aligned}$$

The method to fit the sub-area probit model is discussed in the following steps. In order to jointly draw samples of $\underset{\sim}{z}$ and $\underset{\sim}{\lambda}$, we integrate out $\underset{\sim}{z}$ from the joint posterior distribution $\pi(\underset{\sim}{z},\underset{\sim}{\lambda}|\underset{\sim}{\mu},\underset{\sim}{\nu},\underset{\sim}{x}.\underset{\sim}{\beta},a,\underset{\sim}{y})$. That is,

$$\begin{aligned}
&\pi(\lambda_i|\underset{\sim}{\nu},\underset{\sim}{x}.\underset{\sim}{\beta},a,\underset{\sim}{y})=\int\pi(\lambda_i|\underset{\sim}{\mu},\underset{\sim}{\nu},\underset{\sim}{x},\underset{\sim}{\beta},a,\underset{\sim}{y})d\underset{\sim}{z} \\
&\propto\prod_{j=1}^{n_i}\prod_{k=1}^{m_{ij}}\left\{\int\sqrt{\lambda_i}e^{-\frac{\lambda_i}{2}(z_{ijk}-\underset{\sim}{x}_{ijk}^T\underset{\sim}{\beta}-\nu_i-\mu_{ij})^2}\sum_{t=1}^{T}[I(y_{ijk}=t,\theta_{t-1}<z_{ijk}\le\theta_t)]\,d\underset{\sim}{z}\right\}\frac{a^a\lambda_i^{a-1}e^{-a\lambda_i}}{\Gamma(a)} \\
&=\prod_{j=1}^{n_i}\prod_{k=1}^{m_{ij}}\left\{\sum_{t=1}^{T}\int_{\theta_{t-1}}^{\theta_t}\left[\sqrt{\lambda_i}e^{-\frac{\lambda_i}{2}(z_{ijk}-\underset{\sim}{x}_{ijk}^T\underset{\sim}{\beta}-\nu_i-\mu_{ij})^2}\right]I(y_{ijk}=t)dz_{ijk}\right\}\frac{a^a\lambda_i^{a-1}e^{-a\lambda_i}}{\Gamma(a)} \\
&=\prod_{j=1}^{n_i}\prod_{k=1}^{m_{ij}}\left\{\sum_{t=1}^{T}\left[\Phi\left(\sqrt{\lambda_i}(\theta_t-\underset{\sim}{x}_{ijk}^T\underset{\sim}{\beta}-\nu_i-\mu_{ij})\right)-\Phi\left(\sqrt{\lambda_i}(\theta_{t-1}-\underset{\sim}{x}_{ijk}^T\underset{\sim}{\beta}-\nu_i-\mu_{ij})\right)\right]I(y_{ijk}=t)\right\} \\
&\qquad\times\frac{a^a\lambda_i^{a-1}e^{-a\lambda_i}}{\Gamma(a)}.
\end{aligned}$$

Then we can use accept-reject algorithm to draw samples of $\lambda_i,\ i=1,\ldots,\ell$. Once we get the sample, we can draw z_{ijk}. Similarly, we first draw a sample $\underset{\sim}{\beta}^*$ from prior MN$\left(\underset{\sim}{\beta}_0,1000\Sigma_0\right)$, draw a sample ν_i^* from prior Normal$(0,\delta^2)$ and draw a sample μ_{ij}^* from prior Normal$(0,\sigma^2)$ and given $\underset{\sim}{\beta}^*,\underset{\sim}{\mu}^*,\underset{\sim}{\nu}^*,\underset{\sim}{\lambda}$ and data, we can draw sample z_{ijk} from truncated Normal$(\underset{\sim}{x}_{ijk}^T\underset{\sim}{\beta}^*+\nu_i^*+\mu_{ij}^*,\lambda_i^{-1})$, where $\theta_{t-1}<z_{ijk}\le\theta_t$ if $y_{ijk}=t,\ t=1,\ldots,T,\ i=1,\ldots,\ell,\ j=1,\ldots,n_i,\ k=1,\ldots,m_{ij}$.

To implement the Gibbs sampler once we get a sample of $\underset{\sim}{\lambda}$ and $\underset{\sim}{z}$, we need to draw samples from the full conditional posterior distributions of $\underset{\sim}{\mu}$, $\underset{\sim}{\nu}$, $\underset{\sim}{\beta}$, a, σ^2, δ^2 and $\underset{\sim}{\theta}$.

First, the conditional distribution of $\underset{\sim}{\nu}$ is

$$\mu_{ij}|\nu_i,\lambda_i,\underset{\sim}{z},\underset{\sim}{\beta},\sigma^2,\underset{\sim}{y}\overset{ind}{\sim}\text{Normal}\left(\frac{\lambda_i\sum_{k=1}^{m_{ij}}(z_{ijk}-\underset{\sim}{x}_{ijk}^T\underset{\sim}{\beta}-\nu_i)}{\frac{1}{\sigma^2}+m_{ij}\lambda_i},(\frac{1}{\sigma^2}+m_{ij}\lambda_i)^{-1}\right).$$

Second, the conditional distribution of $\underset{\sim}{\nu}$ is

$$\nu_i|\underset{\sim}{\mu},\lambda_i,\underset{\sim}{z},\underset{\sim}{\beta},\delta^2,\underset{\sim}{y} \overset{ind}{\sim} \text{Normal}\left(\frac{\lambda_i\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}}(z_{ijk}-\underset{\sim}{x}_{ijk}^T\underset{\sim}{\beta}-\mu_{ij})}{\frac{1}{\delta^2}+\sum_{j=1}^{n_i}m_{ij}\lambda_i},(\frac{1}{\delta^2}+\sum_{j=1}^{n_i}m_{ij}\lambda_i)^{-1}\right).$$

Third, the conditional distribution of $\underset{\sim}{\beta}$ is $\underset{\sim}{\beta}|\underset{\sim}{\mu},\underset{\sim}{\nu},\underset{\sim}{\lambda},\underset{\sim}{z},\underset{\sim}{x},\underset{\sim}{y} \sim \text{MN}\left(\underset{\sim}{\hat{\beta}},\Sigma_{\underset{\sim}{\hat{\beta}}}\right)$, where

$$\underset{\sim}{\hat{\beta}} = \Sigma_{\underset{\sim}{\hat{\beta}}}\left(\sum_{i=1}^{\ell}\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}}\lambda_i(z_{ijk}-\nu_i-\mu_{ij})\underset{\sim}{x}_{ijk}+(1000\Sigma_0)^{-1}\underset{\sim}{\beta}_0\right),$$

$$\Sigma_{\underset{\sim}{\hat{\beta}}} = \left(\sum_{i=1}^{\ell}\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}}\lambda_i\underset{\sim}{x}_{ijk}\underset{\sim}{x}_{ijk}^T+(1000\Sigma_0)^{-1}\right)^{-1}.$$

Fourth, the fully conditional distribution of θ_t given $\underset{\sim}{z}$, $\underset{\sim}{\theta}_{(t)}=\{\theta_s, s\neq t\}$ and data is given by

$$\pi(\theta_t|\underset{\sim}{z},\underset{\sim}{\theta}_{(t)},\underset{\sim}{y}) \propto \prod_{i=1}^{\ell}\prod_{j=1}^{m_i}\left[I(y_{ij}=t,\theta_{t-1}<z_{ij}\leq\theta_t)+I(y_{ij}=t+1,\theta_t<z_{ij}\leq\theta_{t+1})\right].$$

Notice that this conditional density is uniform density on the interval. That is

$$\theta_t|\underset{\sim}{z},\underset{\sim}{\theta}_{(t)},\underset{\sim}{y} \sim \text{Uniform}\left[\max\left\{\max\left\{z_{ij},y_{ij}=t\right\},\theta_{t-1}\right\},\min\left\{\min\left\{z_{ij},y_{ij}=t+1\right\},\theta_{t+1}\right\}\right].$$

Fifth, given the sample of $\underset{\sim}{\lambda}$, we can use grid method to draw a. Transform a to $\phi_1=\frac{a}{1+a}$, which is in $(0,1)$. The conditional posterior distribution of ϕ_1 is

$$\pi(\phi_1|\underset{\sim}{\lambda}) \propto \left(\prod_{i=1}^{\ell}\frac{a^a\lambda_i^{a-1}e^{-a\lambda_i}}{\Gamma(a)}\right)_{\phi_1=\frac{a}{1+a}}.$$

Sixth, to draw δ^2 we also use grid method. Transform δ^2 to $\phi_2=\frac{\delta^2}{1+\delta^2}$, which is in $(0,1)$. The conditional posterior distribution of ϕ_2 is

$$\pi(\phi_2|\underset{\sim}{\nu}) \propto \left\{\left(\frac{1}{\delta^2}\right)^{\frac{\ell}{2}}\exp\left(-\frac{1}{2\delta^2}\sum_{i=1}^{\ell}\nu_i^2\right)\right\}_{\phi_2=\frac{\delta^2}{1+\delta^2}}.$$

Seventh, to draw σ^2 we also use grid method. Transform σ^2 to $\phi_3=\frac{\sigma^2}{1+\sigma^2}$, which is in $(0,1)$. The conditional posterior distribution of ϕ_2 is

$$\pi(\phi_2|\underset{\sim}{\mu}) \propto \left\{\left(\frac{1}{\sigma^2}\right)^{\frac{\sum_{i=1}^{\ell}n_i}{2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{\ell}\sum_{j=1}^{n_i}\mu_{ij}^2\right)\right\}_{\phi_3=\frac{\sigma^2}{1+\sigma^2}}.$$

To implement the algorithm, we chose start points $\underset{\sim}{\beta}^{(0)},\underset{\sim}{\theta}^{(0)}$ equal to the MLE based on the

previous paper by Chen and Nandram (2023), $\lambda_i^{(0)} = 1,\ \mu^0(0)_{ij} = 1,\ \nu^0(0)_i = 1$. We first draw a, δ^2, and σ^2 using the grid method, and then jointly draw a sample of $\underset{\sim}{\lambda}$ and $\underset{\sim}{z}$, and simulate from the conditional distribution of $\mu_{ij},\ \nu_i\ ,\underset{\sim}{\beta}$ and θ_t.

C. Bayesian bootstrap

Our interest is to predict the finite population proportions of 102 sampled wards for all households. The covariates of individuals in non-sampled households and the size of non-sampled households are unknown. We know the total number of households and individuals in each sampled ward and we have all information about the sampled households. Therefore, we decide to use these information in the Bayesian bootstrap approach to generate the non-sampled household sizes and corresponding non-sampled covariates within sampled wards. The Bayesian bootstrap (Rubin 1981) method is used sample the sampled households to impute the non-sampled households. There are $n = 12$ sampled households in the sampled wards and everyone is sampled from the sampled households. We know the sizes and covariates of all sampled households, and we simple need to have the sample sizes and the covariates for all the non-sampled households in any sampled ward to do Bayesian predictive inference in each sampled ward; the procedure is done independently for each sampled ward.

Let N denote the number of households in one of the sampled wards. We simply need to fill in the sizes of the households and their covariates. This procedure is equivalent to simply sampling the households. Denote the labels of the sampled households by $1,\ldots,n$ to provide the information (sizes and covariates) of the non-sampled households with labels, $n+1,\ldots,N$. Denote the sampled indicators of each household by $I_i,\ i = 1,\ldots,n$. After the bootstrap is executed, because it is based on sampling with replacement, there will be N_i^* non-sampled households corresponding to the i^{th} sampled household, and $\sum_{i=1}^n N_i^* = N - n$.

The Bayesian bootstrap assumes that

$$\underset{\sim}{I} \mid \underset{\sim}{p} \sim \text{Multinomial}(n, \underset{\sim}{p}),$$

where we actually observed $I_i = 1, i = 1,\ldots,n$,

$$\underset{\sim}{p} \sim \text{Dirichlet}(Q),$$

Haldane's improper prior, where Q is a vector of zeros. Then, a posterior

$$\underset{\sim}{p} \mid \underset{\sim}{I} \sim \text{Dirichlet}(\underset{\sim}{j}), \tag{5}$$

where $\underset{\sim}{j}$ is a vector of ones. Therefore,

$$\underset{\sim}{N}^* \mid \underset{\sim}{p}, \underset{\sim}{I} \sim \text{Multinomial}(N^*, \underset{\sim}{p}). \tag{6}$$

To execute the bootstrap, simply draw $\underset{\sim}{p}$ from (5) and $\underset{\sim}{N}^*$ from (6).

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 95–118
http://www.ssca.org.in/journal

The Fundamental BLUE Equation in Linear Models Revisited

Stephen J. Haslett[1], Jarkko Isotalo[2], Augustyn Markiewicz[3] and Simo Puntanen[2]

[1]*School of Mathematical and Computational Sciences and Environmental Health Intelligence NZ, Massey University, Palmerston North, New Zealand;*
Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, Australia;
Faculty of Engineering and Information Sciences, University of Wollongong, Australia
[2]*Faculty of Information Technology and Communication Sciences, Tampere University, FI-33014 Tampere, Finland*
[3]*Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, Wojska Polskiego 28, PL-60637 Poznań, Poland*

Received: 22 April 2024; Revised: 07 June 2024; Accepted: 09 June 2024

Abstract

In the world of linear statistical models there is a particular matrix equation, $\mathbf{G}(\mathbf{X} : \mathbf{V}\mathbf{X}^{\perp}) = (\mathbf{X} : \mathbf{0})$, which is sufficiently important that it is sometimes called the fundamental BLUE equation. In this equation, $\mathbf{X}$ is a model matrix, $\mathbf{V}$ is the covariance matrix of $\mathbf{y}$ in the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, and we are interested in finding the best linear estimator, BLUE, of $\mathbf{X}\boldsymbol{\beta}$. Any solution $\mathbf{G}$ for this equation has the property that $\mathbf{G}\mathbf{y}$ provides a representation for the BLUE of $\mathbf{X}\boldsymbol{\beta}$: this is the message of the the fundamental BLUE equation, whose main developer was the late Professor C. R. Rao in early 1970s. In this article we revisit some interesting features and consequences of this equation. We do not provide essentially new results – the aim is to offer a compact easy-to-follow review including also some recent related results by the authors.

Key words: BLUE; BLUP; Covariance matrix; Equality of the BLUEs; Linear sufficiency; Misspecified model.

AMS Subject Classifications: 62J05, 62J10

1. Introduction

Our main focus in this paper is the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, denoted as triplet

$$\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}.$$

Here $\mathbf{y}$ is an n-dimensional observable random vector, and $\boldsymbol{\varepsilon}$ is an unobservable random error vector with a known (possibly singular) covariance matrix $\mathrm{cov}(\boldsymbol{\varepsilon}) = \mathbf{V} = \mathrm{cov}(\mathbf{y})$ and

Corresponding Author: Simo Puntanen
Email: simo.puntanen@tuni.fi

expectation $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$. The matrix $\mathbf{X}$ is a known $n \times p$ matrix, *i.e.*, $\mathbf{X} \in \mathbb{R}^{n\times p}$. Vector $\boldsymbol{\beta}$ is a vector of fixed (but unknown) parameters; here symbol $'$ stands for the transpose. We will also denote $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. If we want to emphasize what is the covariance matrix, we may use notation $\mathscr{M}(\mathbf{V})$.

As for notations, the symbols $\mathrm{r}(\mathbf{A})$, $\mathbf{A}^-$, $\mathbf{A}^+$, $\mathscr{C}(\mathbf{A})$, and $\mathscr{C}(\mathbf{A})^\perp$, denote, respectively, the rank, a generalized inverse, the Moore–Penrose inverse, the column space, and the orthogonal complement of the column space of the matrix $\mathbf{A}$. By $(\mathbf{A} : \mathbf{B})$ we denote the columnwise partitioned matrix with $\mathbf{A}_{a\times b}$ and $\mathbf{B}_{a\times c}$ as submatrices. By $\mathbf{A}^\perp$ we denote any matrix satisfying $\mathscr{C}(\mathbf{A}^\perp) = \mathscr{C}(\mathbf{A})^\perp$. We will write $\mathbf{P}_\mathbf{A} = \mathbf{A}\mathbf{A}^+ = \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$ to denote the orthogonal projector onto $\mathscr{C}(\mathbf{A})$ and $\mathbf{Q}_\mathbf{A} = \mathbf{I}_a - \mathbf{P}_\mathbf{A}$, where $\mathbf{I}_a$ is the identity matrix of order a with a being the number of rows in $\mathbf{A}$. In particular, we denote

$$\mathbf{H} = \mathbf{P}_\mathbf{X}\,, \quad \mathbf{M} = \mathbf{I}_n - \mathbf{P}_\mathbf{X}\,, \quad \mathbf{M}_i = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}_i}\,,\ i = 1, 2\,.$$

The following special cases or extensions of $\mathscr{M}$ will be considered in this paper:

(a) The partitioned linear model is denoted as

$$\mathscr{M}_{12} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}\} = \{\mathbf{y},\, \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,\, \mathbf{V}\} = \{\mathbf{y},\, \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2,\, \mathbf{V}\}\,.$$

(b) In addition to the *full* model $\mathscr{M}_{12}$, we will consider the *small* models $\mathscr{M}_i = \{\mathbf{y},\, \mathbf{X}_i\boldsymbol{\beta}_i,\, \mathbf{V}\}$, $i = 1, 2$, and the *reduced* model

$$\mathscr{M}_{12\cdot 2} = \{\mathbf{M}_2\mathbf{y},\, \mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1,\, \mathbf{M}_2\mathbf{V}\mathbf{M}_2\},$$

which is obtained by premultiplying the model $\mathscr{M}_{12}$ by $\mathbf{M}_2 = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}_2}$.

(c) Let $\mathbf{y}_*$ denote a $q \times 1$ unobservable random vector containing new observations. The new observations are assumed to be generated from

$$\mathbf{y}_* = \mathbf{X}_*\boldsymbol{\beta} + \boldsymbol{\varepsilon}_*\,,$$

where $\mathbf{X}_*$ is a known $q\times p$ matrix, $\boldsymbol{\beta}$ is the same vector of fixed but unknown parameters as in $\mathscr{M}$, and $\boldsymbol{\varepsilon}_*$ is a q-dimensional random error vector. We further assume that

$$\mathrm{E}\begin{pmatrix}\mathbf{y}\\ \mathbf{y}_*\end{pmatrix} = \begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{X}_*\boldsymbol{\beta}\end{pmatrix} = \begin{pmatrix}\mathbf{X}\\ \mathbf{X}_*\end{pmatrix}\boldsymbol{\beta}\,, \quad \mathrm{cov}\begin{pmatrix}\mathbf{y}\\ \mathbf{y}_*\end{pmatrix} = \begin{pmatrix}\mathbf{V} & \mathbf{V}_{12}\\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{pmatrix} = \boldsymbol{\Psi},$$

where $\boldsymbol{\Psi}$ is known. We denote this setup shortly as

$$\mathscr{M}_* = \left\{\begin{pmatrix}\mathbf{y}\\ \mathbf{y}_*\end{pmatrix},\, \begin{pmatrix}\mathbf{X}\\ \mathbf{X}_*\end{pmatrix}\boldsymbol{\beta},\, \begin{pmatrix}\mathbf{V} & \mathbf{V}_{12}\\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{pmatrix}\right\}. \tag{1}$$

We call $\mathscr{M}_*$ "the linear model with new observations". Our main interest in $\mathscr{M}_*$ lies in predicting $\mathbf{y}_*$ on the basis of observable $\mathbf{y}$. Notice the crucial role of the cross-covariance matrix $\mathrm{cov}(\mathbf{y}, \mathbf{y}_*) = \mathbf{V}_{12} \in \mathbb{R}^{n\times q}$. The mixed linear model can be interpreted as a special case of $\mathscr{M}_*$; see Sec. 4.

Under the model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$, the statistic $\mathbf{Gy}$, where $\mathbf{G}$ is an $n \times n$ matrix, is the best linear unbiased estimator, BLUE, of $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ if $\mathbf{Gy}$ is unbiased, *i.e.*, $\mathbf{GX} = \mathbf{X}$, and it has the smallest covariance matrix in the Löwner sense among all unbiased linear estimators of $\mathbf{X}\boldsymbol{\beta}$; shortly denoted

$$\operatorname{cov}(\mathbf{Gy}) \leq_{\mathrm{L}} \operatorname{cov}(\mathbf{Cy}) \quad \text{for all } \mathbf{C} \in \mathbb{R}^{n\times n}\colon\ \mathbf{CX} = \mathbf{X}\,.$$

The BLUE of an estimable parametric function $\boldsymbol{\mu}_* = \mathbf{X}_*\boldsymbol{\beta}$, where $\mathbf{X}_* \in \mathbb{R}^{q\times p}$, is defined in the corresponding way. Estimability of $\mathbf{X}_*\boldsymbol{\beta}$ means that it has a linear unbiased estimator which happens if and only if $\mathscr{C}(\mathbf{X}_*') \subseteq \mathscr{C}(\mathbf{X}')$. In particular, $\boldsymbol{\mu}_1 = \mathbf{X}_1\boldsymbol{\beta}_1$ is estimable in the partitioned model if and only if

$$\mathscr{C}(\mathbf{X}_1) \cap \mathscr{C}(\mathbf{X}_2) = \{\mathbf{0}\}\,.$$

The random vector $\mathbf{Ay}$ is a linear unbiased predictor (LUP) of $\mathbf{y}_*$ if $\mathrm{E}(\mathbf{y}_* - \mathbf{Ay}) = \mathbf{0}$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$. Such a matrix $\mathbf{A} \in \mathbb{R}^{q\times n}$ exists if and only if $\mathscr{C}(\mathbf{X}_*') \subseteq \mathscr{C}(\mathbf{X}')$, *i.e.*, $\mathbf{X}_*\boldsymbol{\beta}$ is estimable under $\mathscr{M}$ and then we say that $\mathbf{y}_*$ is predictable under $\mathscr{M}_*$. Now a LUP $\mathbf{Ay}$ is the best linear unbiased predictor, BLUP, for $\mathbf{y}_*$, if the covariance matrix of the prediction error, subject to the unbiasedness of the prediction, is minimized:

$$\operatorname{cov}(\mathbf{y}_* - \mathbf{Ay}) \leq_{\mathrm{L}} \operatorname{cov}(\mathbf{y}_* - \mathbf{A}_\#\mathbf{y}) \quad \text{for all } \mathbf{A}_\# : \mathbf{A}_\#\mathbf{X} = \mathbf{X}_*\,.$$

Our matrix expressions will use generalized inverses heavily and in this context it is essential to know whether the expressions are independent of the choice of the generalized inverses involved. Lemma 2.2.4 of Rao and Mitra (1971) gives the condition under which the matrix product $\mathbf{AB}^-\mathbf{C}$ is invariant with respect to the choice of $\mathbf{B}^-$.

Proposition 1: For nonnull matrices $\mathbf{A}$ and $\mathbf{C}$ the following holds:

(a) $\mathbf{AB}^-\mathbf{C} = \mathbf{AB}^+\mathbf{C}$ for all $\mathbf{B}^- \iff \mathscr{C}(\mathbf{C}) \subseteq \mathscr{C}(\mathbf{B})$ & $\mathscr{C}(\mathbf{A}') \subseteq \mathscr{C}(\mathbf{B}')$.

(b) $\mathbf{AA}^-\mathbf{C} = \mathbf{C}$ for some (and hence for all) $\mathbf{A}^- \iff \mathscr{C}(\mathbf{C}) \subseteq \mathscr{C}(\mathbf{A})$.

(c) $\mathbf{C}'\mathbf{A}^-\mathbf{A} = \mathbf{C}'$ for some (and hence for all) $\mathbf{A}^- \iff \mathscr{C}(\mathbf{C}) \subseteq \mathscr{C}(\mathbf{A}')$.

Suppose that the matrix equation

$$\mathbf{YB} = \mathbf{A} \tag{2}$$

is solvable for $\mathbf{Y}$, *i.e.*, $\mathscr{C}(\mathbf{A}') \subseteq \mathscr{C}(\mathbf{B}')$. Then it is well known, see, *e.g.*, Rao and Mitra (1971, p. 24), that the general solution $\mathbf{Y}_0$ to (2) can be written, for example, as

$$\mathbf{Y}_0 = \mathbf{AB}^+ + \mathbf{E}(\mathbf{I} - \mathbf{P}_\mathbf{B}) = \mathbf{AB}^+ + \mathbf{EQ}_\mathbf{B}\,, \quad \text{where } \mathbf{E} \text{ is free to vary,} \tag{3a}$$

$$\mathbf{Y}_0 = \{\text{one solution to } \mathbf{YB} = \mathbf{A}\} + \{\text{general solution to } \mathbf{YB} = \mathbf{0}\}. \tag{3b}$$

For later considerations, we collect some useful results into the following proposition.

Proposition 2: Consider the partitioned model $\mathscr{M}_{12}(\mathbf{V})$, and let "$\oplus$" refer to the direct sum and "$\boxplus$" to the direct sum of orthogonal subspaces. Then

(a) $\mathscr{C}(\mathbf{X}_1 : \mathbf{X}_2) = \mathscr{C}(\mathbf{X}_1 : \mathbf{M}_1\mathbf{X}_2)$, *i.e.*, $\mathscr{C}(\mathbf{X}) = \mathscr{C}(\mathbf{X}_1) \boxplus \mathscr{C}(\mathbf{M}_1\mathbf{X}_2)$.

(b) $\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{X} : \mathbf{VM}) = \mathscr{C}(\mathbf{X}) \oplus \mathscr{C}(\mathbf{VM}) = \mathscr{C}(\mathbf{X}) \boxplus \mathscr{C}(\mathbf{MV})$.

(c) $\mathbf{M} = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}} = \mathbf{I}_n - (\mathbf{P}_{\mathbf{X}_1} + \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}) = \mathbf{M}_1\mathbf{Q}_{\mathbf{M}_1\mathbf{X}_2} = \mathbf{M}_1\mathbf{M}$.

(d) $\mathbf{Q}_{(\mathbf{X}:\mathbf{V})} = \mathbf{I}_n - (\mathbf{P}_{\mathbf{X}} + \mathbf{P}_{\mathbf{MV}}) = \mathbf{M} - \mathbf{P}_{\mathbf{MV}} = \mathbf{M}\mathbf{Q}_{\mathbf{MV}} = \mathbf{M}\mathbf{Q}_{(\mathbf{X}:\mathbf{V})}$.

(e) $\mathrm{r}(\mathbf{AB}) = \mathrm{r}(\mathbf{A}) - \dim \mathscr{C}(\mathbf{A}') \cap \mathscr{C}(\mathbf{B}^{\perp})$ for conformable $\mathbf{A}$ and $\mathbf{B}$.

We assume the model $\mathscr{M}(\mathbf{V})$ to be consistent in the sense that $\mathbf{y}$ lies in $\mathscr{C}(\mathbf{X} : \mathbf{V})$ with probability 1, *i.e.*, the observed numerical value of $\mathbf{y}$ satisfies

$$\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{X}) \oplus \mathscr{C}(\mathbf{VM}) = \mathscr{C}(\mathbf{X}) \boxplus \mathscr{C}(\mathbf{MV}),$$

so that

$$\mathbf{y} = \mathbf{Xa} + \mathbf{VMb} \quad \text{for some vectors } \mathbf{a} \in \mathbb{R}^p \text{ and } \mathbf{b} \in \mathbb{R}^n. \tag{4}$$

There is one special class of matrices worth particular attention and that is the set $\mathscr{W}_{\geq}$ of nonnegative definite matrices defined as

$$\mathscr{W}_{\geq} = \left\{ \mathbf{W} \in \mathbb{R}^{n\times n} : \mathbf{W} = \mathbf{V} + \mathbf{XUU}'\mathbf{X}',\ \mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V}) \right\}. \tag{5}$$

In (5) $\mathbf{U}$ can be any matrix comprising p rows as long as $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$ is satisfied. One obvious choice is $\mathbf{U} = \mathbf{I}_p$. In particular, if $\mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{V})$, we can choose $\mathbf{U} = \mathbf{0}$. The extended version of $\mathscr{W}_{\geq}$ is

$$\mathscr{W} = \left\{ \mathbf{W} \in \mathbb{R}^{n\times n} : \mathbf{W} = \mathbf{V} + \mathbf{XTX}',\ \mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V}) \right\}. \tag{6}$$

Above $\mathbf{T} \in \mathbb{R}^{p\times p}$ is free to vary subject to condition $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$. Notice that $\mathbf{W}$ belonging to $\mathscr{W}$ is not necessarily nonnegative definite and it can be nonsymmetric. We may use the notations $\mathscr{W}(\mathscr{A})$ or $\mathscr{W}(\mathbf{V})$ to indicate that the model $\mathscr{A}$ or the covariance matrix $\mathbf{V}$ is under consideration. Proposition 3 collects together some properties of the class $\mathscr{W}$.

Proposition 3: Let $\mathbf{V} \in \mathbb{R}^{n\times n}$ be nonnegative definite and let $\mathbf{X} \in \mathbb{R}^{n\times p}$ and define $\mathbf{W}$ as $\mathbf{W} = \mathbf{V} + \mathbf{XTX}'$, where $\mathbf{T} \in \mathbb{R}^{p\times p}$. Then the following statements are equivalent:

(a) $\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{W})$,

(b) $\mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{W})$,

(c) $\mathbf{X}'\mathbf{W}^{-}\mathbf{X}$ is invariant for any choice of $\mathbf{W}^{-}$,

(d) $\mathscr{C}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X}) = \mathscr{C}(\mathbf{X}')$ for any choice of $\mathbf{W}^{-}$,

(e) $\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-}\mathbf{X} = \mathbf{X}$ for any choices of $\mathbf{W}^{-}$ and $(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}$.

Observe that obviously $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{W}')$ since

$$\mathscr{C}(\mathbf{W}') = \mathscr{C}(\mathbf{V} + \mathbf{X}\mathbf{T}'\mathbf{X}') \subseteq \mathscr{C}(\mathbf{W})\,, \quad \mathrm{r}(\mathbf{W}') = \mathrm{r}(\mathbf{W})\,,$$

and hence in statements (a)–(e) $\mathbf{W}$ can be replaced with with $\mathbf{W}'$. For further properties of $\mathscr{W}$, see, *e.g.*, Baksalary and Mathew (1990, Th. 2), and Puntanen *et al.* (2011, Sec. 12.3). Haslett *et al.* (2022a) provide an extensive review of the class $\mathscr{W}$.

Let us cite Puntanen *et al.* (2011, Sec. 5.13):

Proposition 4: Consider the model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ and let $\mathbf{W} \in \mathscr{W}(\mathscr{M})$. Then

(a) $\mathscr{C}(\mathbf{V}\mathbf{M})^\perp = \mathscr{C}(\mathbf{W}^+\mathbf{X} : \mathbf{Q}_\mathbf{W}) = \mathscr{C}[(\mathbf{W}^+)'\mathbf{X} : \mathbf{Q}_\mathbf{W}]\,,$

(b) $\mathscr{C}(\mathbf{W}^+\mathbf{X})^\perp = \mathscr{C}(\mathbf{W}\mathbf{M} : \mathbf{Q}_\mathbf{W}) = \mathscr{C}(\mathbf{V}\mathbf{M} : \mathbf{Q}_\mathbf{W})\,.$

It appears to be useful to denote

$$\dot{\mathbf{M}} = \mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M}\,.$$

The matrix $\dot{\mathbf{M}}$ is unique with respect to the choice of the generalized inverse $(\mathbf{M}\mathbf{V}\mathbf{M})^-$ if and only if $\mathbb{R}^n = \mathscr{C}(\mathbf{X} : \mathbf{V})$. However, for example, $\mathbf{V}\dot{\mathbf{M}}\mathbf{P}_\mathbf{W}$ is always unique. It is noteworthy that using the Moore–Penrose inverse the following holds:

$$\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^+\mathbf{M} = (\mathbf{M}\mathbf{V}\mathbf{M})^+\mathbf{M} = \mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^+ = (\mathbf{M}\mathbf{V}\mathbf{M})^+. \tag{7}$$

In particular, for a positive definite $\mathbf{V}$ we have, for any $(\mathbf{M}\mathbf{V}\mathbf{M})^-$,

$$\begin{aligned}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M} &= \mathbf{V}^{-1/2}\mathbf{P}_{\mathbf{V}^{1/2}\mathbf{M}}\mathbf{V}^{-1/2} = \mathbf{V}^{-1/2}(\mathbf{I}_n - \mathbf{P}_{(\mathbf{V}^{1/2}\mathbf{M})^\perp})\mathbf{V}^{-1/2} \\ &= \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1} =: \mathbf{V}^{-1}(\mathbf{I}_n - \mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}}),\end{aligned}$$

where we have used the obvious fact $\mathscr{C}(\mathbf{V}^{1/2}\mathbf{M})^\perp = \mathscr{C}(\mathbf{V}^{-1/2}\mathbf{X})$.

We will use the following notation:

$$\mathbf{P}_{\mathbf{X};\mathbf{W}^+} = \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^+, \quad \mathbf{P}_{\mathbf{X}_*;\mathbf{W}^+} = \mathbf{X}_*(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^+.$$

Notice that $\mathbf{P}_{\mathbf{X};\mathbf{W}^+}$ and $\mathbf{P}_{\mathbf{X}_*;\mathbf{W}^+}$ are invariant for any choice of the generalized inverses $\mathbf{W}^-$ and $(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-$ but this invariance does not concern the matrix

$$\mathbf{P}_{\mathbf{X};\mathbf{W}^-} = \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^-.$$

Proposition 5: Consider the linear model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$. Let $\mathbf{T}$ be any $p \times p$ matrix such that the matrix $\mathbf{W} = \mathbf{V} + \mathbf{X}\mathbf{T}\mathbf{X}'$ satisfies the condition $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$, *i.e.*, $\mathbf{W} \in \mathscr{W}(\mathscr{M})$, and denote $\dot{\mathbf{M}} = \mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M}$. Then

(a) $\mathbf{P}_\mathbf{W}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M}\mathbf{P}_\mathbf{W} = \mathbf{W}^+ - \mathbf{W}^+\mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^+.$

(b) $\mathbf{P_W M(MVM)^- MP_W} = \mathbf{(MVM)^+} = \mathbf{P_W \dot{M} P_W}$.

(c) $\mathbf{P_{X;W^+}} = \mathbf{X(X'W^- X)^- X'W^+} = \mathbf{P_W - VM(MVM)^- MP_W}$.

(d) $\mathbf{P_{X;W^+}} = \mathbf{H - HVM(MVM)^- MP_W} = \mathbf{H - HVM(MVM)^+ M}$, where $\mathbf{H} = \mathbf{P_X}$.

The result (a) is the the most crucial one in Proposition 5. For the proof of (a), see Puntanen *et al.* (2011, Prop. 15.2) and Isotalo *et al.* (2008, Cor. 2.2). Notice that in light of Proposition 2, we have

$$\mathbf{P_W} = \mathbf{P_X} + \mathbf{P_{MV}} = \mathbf{H} + \mathbf{P_{MVM}}\,, \quad \mathbf{MP_W} = \mathbf{MP_{MV}} = \mathbf{P_{MV}} = \mathbf{P_{MVM}}\,,$$

which implies (b) of Proposition 5. Premultiplying (a) by $\mathbf{W}$ and using $\mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{W})$ gives (c), *i.e.*,

$$\begin{aligned}\mathbf{X(X'W^- X)^- X'W^+} &= \mathbf{P_W - VM(MVM)^- MP_W} = \mathbf{P_W - VM(MVM)^+ MP_W} \\ &= \mathbf{P_W - V(MVM)^+ P_W} = \mathbf{P_W - V(MVM)^+} \\ &= \mathbf{P_W - VM(MVM)^+ M}\,, \end{aligned} \tag{8}$$

where we have used (7) and the fact that $\mathscr{C}[\mathbf{(MVM)^+}] = \mathscr{C}\mathbf{(MVM)} \subseteq \mathscr{C}(\mathbf{W})$. From (8) we immediately confirm that $\mathbf{X(X'W^- X)^- X'W^+}$ is invariant with respect to the choice of $\mathbf{W} \in \mathscr{W}$ supposing that $\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{W})$ is holding. Premultiplying (8) by $\mathbf{H} = \mathbf{P_X}$ gives

$$\begin{aligned}\mathbf{X(X'W^- X)^- X'W^+} &= \mathbf{H - HVM(MVM)^- MP_W} = \mathbf{H - HVM(MVM)^+ MP_W} \\ &= \mathbf{H - HV(MVM)^+ P_W} = \mathbf{H - HV(MVM)^+} \\ &= \mathbf{H - HVM(MVM)^+ M}\,. \end{aligned} \tag{9}$$

Remark 1: The equality (9) follows from (a) of Proposition 5. However, it is interesting to prove (9) directly. This is done by first observing that

$$\mathbf{X(X'W^- X)^- X'W^+(X : VM)} = \mathbf{[H - HVM(MVM)^+ MP_W](X : VM)}\,, \tag{10}$$

and then confirming that

$$\mathbf{X(X'W^- X)^- X'W^+ Q_W} = \mathbf{[H - HVM(MVM)^+ MP_W]Q_W}\,. \tag{11}$$

Now (10) and (11) together imply (9). □

2. The fundamental BLUE equation

Theorem 1 below provides so-called fundamental BLUE equations.

Theorem 1: [BLUE] Consider the linear model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$.

(a) Then the linear estimator $\mathbf{Gy}$ is the BLUE for $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ if and only if $\mathbf{G} \in \mathbb{R}^{n\times n}$ satisfies the equation

$$\mathbf{G(X : VX^\perp)} = \mathbf{(X : 0)}\,. \tag{12}$$

(b) Let $\boldsymbol{\mu}_* = \mathbf{X}_*\boldsymbol{\beta}$, where $\mathbf{X}_* \in \mathbb{R}^{q\times p}$, be estimable so that $\mathscr{C}(\mathbf{X}_*') \subseteq \mathscr{C}(\mathbf{X}')$. Then $\mathbf{By}$ is the BLUE of $\boldsymbol{\mu}_*$ if and only if $\mathbf{B} \in \mathbb{R}^{q\times n}$ satisfies the equation

$$\mathbf{B}(\mathbf{X} : \mathbf{V}\mathbf{X}^\perp) = (\mathbf{X}_* : \mathbf{0})\,. \tag{13}$$

(c) Let $\boldsymbol{\mu}_1 = \mathbf{X}_1\boldsymbol{\beta}_1$ be estimable in the partitioned model $\mathscr{M}_{12}$. Then $\mathbf{Cy}$ is the BLUE of $\boldsymbol{\mu}_1$ if and only if

$$\mathbf{C}(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{V}\mathbf{X}^\perp) = (\mathbf{X}_1 : \mathbf{0} : \mathbf{0})\,. \tag{14}$$

For the proofs, see, *e.g.*, Rao (1973, p. 282) and for coordinate-free approach Drygas (1970, p. 55) and Zmyślony (1980). For further proofs see, for example, Groß (2004), Kala (1981, Th. 3.1), Puntanen *et al.* (2000), Puntanen *et al.* (2011, Th. 10), and Baksalary (2004).

For Theorem 2, characterizing the BLUP, see, *e.g.*, Christensen (2020, Th. 6.6.3), and Isotalo and Puntanen (2006, p. 1015).

Theorem 2: [BLUP] Consider the linear model with new observations defined as $\mathscr{M}_*$ in (1), where $\mathscr{C}(\mathbf{X}_*') \subseteq \mathscr{C}(\mathbf{X}')$, *i.e.*, $\mathbf{y}_*$ is predictable. Then:

(a) $\mathbf{Ay}$ is the BLUP for $\mathbf{y}_*$ if and only if $\mathbf{A}(\mathbf{X} : \mathbf{V}\mathbf{X}^\perp) = (\mathbf{X}_* : \mathbf{V}_{21}\mathbf{X}^\perp)$.

(b) $\mathbf{By}$ is the BLUE of $\boldsymbol{\mu}_* = \mathbf{X}_*\boldsymbol{\beta}$ if and only if $\mathbf{B}(\mathbf{X} : \mathbf{V}\mathbf{X}^\perp) = (\mathbf{X}_* : \mathbf{0})$.

(c) $\mathbf{Dy}$ is the BLUP for $\boldsymbol{\varepsilon}_*$ if and only if $\mathbf{D}(\mathbf{X} : \mathbf{V}\mathbf{X}^\perp) = (\mathbf{0} : \mathbf{V}_{21}\mathbf{X}^\perp)$.

Theorems 1 and 2 offer extremely handy tools to check whether a given estimator/predictor is a BLUE/BLUP. Moreover, they provide convenient ways to introduce various representations for the BLUE/BLUP. The solutions are unique if and only if $\mathscr{C}(\mathbf{X} : \mathbf{V}\mathbf{X}^\perp) = \mathbb{R}^n$. Trivially, one choice for $\mathbf{X}^\perp$ is $\mathbf{M} = \mathbf{I}_n - \mathbf{P}_\mathbf{X}$. In view of (3b), the *general* solution for $\mathbf{G}$ in (12) can be expressed as

$$\{\text{one solution to } \mathbf{G}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{X} : \mathbf{0})\} + \{\text{general sol. to } \mathbf{G}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{0} : \mathbf{0})\}. \tag{15}$$

Suppose that $\mathbf{W} \in \mathscr{W}(\mathscr{M})$ where $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$. We observe immediately that

$$\mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^+(\mathbf{X} : \mathbf{V}\mathbf{M}) = \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^+(\mathbf{X} : \mathbf{W}\mathbf{M}) = (\mathbf{X} : \mathbf{0})\,,$$

and so

$$\begin{aligned}\mathbf{P}_{\mathbf{X};\mathbf{W}^+}\mathbf{y} &= \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^+\mathbf{y} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^-\mathbf{y} = \mathbf{P}_{\mathbf{X}_*;\mathbf{W}^-}\mathbf{y} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \tilde{\boldsymbol{\mu}}\,,\end{aligned}$$

where we have used the consistency condition (4) to replace $\mathbf{W}^+$ with $\mathbf{W}^-$. Correspondingly, for an estimable $\boldsymbol{\mu}_* = \mathbf{X}_*\boldsymbol{\beta}$ we have

$$\mathbf{P}_{\mathbf{X}_*;\mathbf{W}^+}\mathbf{y} = \mathbf{X}_*(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^+\mathbf{y} = \mathrm{BLUE}(\mathbf{X}_*\boldsymbol{\beta}) = \tilde{\boldsymbol{\mu}}_*\,. \tag{16}$$

Moreover, in view of (8) and (9) and the consistency of the model $\mathscr{M}$, we have

$$\tilde{\boldsymbol{\mu}} = \mathbf{P}_{\mathbf{W}}\mathbf{y} - \mathbf{VM}(\mathbf{MVM})^{-}\mathbf{MP}_{\mathbf{W}}\mathbf{y} = \mathbf{y} - \mathbf{VM}(\mathbf{MVM})^{-}\mathbf{My}, \tag{17a}$$

$$\begin{aligned}\tilde{\boldsymbol{\mu}} &= \mathbf{HP}_{\mathbf{W}}\mathbf{y} - \mathbf{HVM}(\mathbf{MVM})^{-}\mathbf{MP}_{\mathbf{W}}\mathbf{y} = \mathbf{Hy} - \mathbf{HVM}(\mathbf{MVM})^{-}\mathbf{My} \\ &= \mathrm{OLSE}(\boldsymbol{\mu}) - \mathbf{HVM}(\mathbf{MVM})^{-}\mathbf{My},\end{aligned} \tag{17b}$$

which hold for all $\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V})$ and $\mathrm{OLSE}(\boldsymbol{\mu})$ refers to the ordinary least squares estimator of $\boldsymbol{\mu}$. One of the first references to (17b) is Albert (1973). Notice that in light of (17a) the BLUE's residual can be expressed as

$$\tilde{\boldsymbol{\varepsilon}} := \mathbf{y} - \tilde{\boldsymbol{\mu}} = \mathbf{VM}(\mathbf{MVM})^{-}\mathbf{My}\,.$$

If $\mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{V})$, then $\mathscr{M}$ is said to be a weakly singular linear model. In this situation we can choose $\mathbf{T} = \mathbf{0}$ in (6) and thereby replace $\mathbf{W}$ with $\mathbf{V}$ so that

$$\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-}\mathbf{y}\,. \tag{18}$$

3. How to solve the fundamental BLUE equation?

In the previous section we have shown that certain expressions satisfy the fundamental BLUE equation. It is another question how to end up into these solutions, *i.e.*, how to *introduce* these solutions. And this is just what we aim to do in this section. We believe that our approaches – some not much used in literature as straightforward as they are – may increase the insight of the meaning of the fundamental BLUE equations.

To begin, notice that part (b) of Theorem 1 can be expressed so that $\mathbf{By} = \mathrm{BLUE}(\boldsymbol{\mu}_*)$ if and only if the following two conditions are satisfied:

$$\text{(i) } \mathbf{By} \text{ is unbiased for } \boldsymbol{\mu}_*\,, \quad \text{(ii) } \mathbf{By} \text{ is uncorrelated with } \mathbf{My}. \tag{19}$$

How to solve (19)? As said, by simple substitution we can check that $\mathbf{P}_{\mathbf{X}_*;\mathbf{W}^{+}}\mathbf{y}$ is indeed the BLUE for $\boldsymbol{\mu}_* = \mathbf{X}_*\boldsymbol{\beta}$ under $\mathscr{M}$. We may now raise the question how to *introduce* a solution $\mathbf{B}$ for fundamental BLUE equation

$$\mathbf{B}(\mathbf{X} : \mathbf{VM}) = (\mathbf{X}_* : \mathbf{0})\,, \tag{20}$$

where $\mathbf{X}_* = \mathbf{LX}$ for some $\mathbf{L}$ so that $\boldsymbol{\mu}_* = \mathbf{X}_*\boldsymbol{\beta}$ is estimable. Notice that then

$$\mathbf{X}_*\mathbf{X}^{+} = \mathbf{LXX}^{+} = \mathbf{LP}_{\mathbf{X}} = \mathbf{LH}\,.$$

■ **Solution 1:** The general solution to the unbiasedness condition (i) in (19), *i.e.*, to $\mathbf{BX} = \mathbf{X}_*$, can be expressed, *e.g.*, as

$$\mathbf{B}_0 = \mathbf{X}_*\mathbf{X}^{+} + \mathbf{E}(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}}) = \mathbf{LH} + \mathbf{EM}\,, \quad \text{where } \mathbf{E} \text{ is free to vary.}$$

Hence the requirement (ii) in (19), *i.e.*, $\mathbf{B}_0\mathbf{VM} = \mathbf{0}$, is satisfied if and only if

$$\mathbf{LHVM} + \mathbf{EMVM} = \mathbf{0}\,, \quad \text{i.e.,} \quad \mathbf{EMVM} = -\mathbf{LHVM}\,,$$

from which we get the general expression for $\mathbf{E}$:

$$\mathbf{E}_0 = -\mathbf{LHVM}(\mathbf{MVM})^+ + \mathbf{E}_1\mathbf{Q}_{\mathbf{MV}}\,, \quad \text{where } \mathbf{E}_1 \text{ is free to vary.} \tag{21}$$

In view of the decomposition

$$\mathbf{Q}_{(\mathbf{X}:\mathbf{V})} = \mathbf{I}_n - (\mathbf{P}_{\mathbf{X}} + \mathbf{P}_{\mathbf{MV}}) = -\mathbf{H} + \mathbf{Q}_{\mathbf{MV}}\,,$$

we have $\mathbf{Q}_{\mathbf{MV}} = \mathbf{H} + \mathbf{Q}_{(\mathbf{X}:\mathbf{V})}$, and thereby by (21) we have

$$\mathbf{E}_0 = -\mathbf{LHVM}(\mathbf{MVM})^+ + \mathbf{E}_1(\mathbf{H} + \mathbf{Q}_{(\mathbf{X}:\mathbf{V})})\,,$$

and hence the expression for the general solution to $\mathbf{B}$ in (20) can be written as

$$\begin{aligned}
\mathbf{B}_0 &= \mathbf{LH} + \mathbf{E}_0\mathbf{M} = \mathbf{LH} - \mathbf{LHVM}(\mathbf{MVM})^+\mathbf{M} + \mathbf{E}_1\mathbf{Q}_{(\mathbf{X}:\mathbf{V})}\mathbf{M} \\
&= \mathbf{L}[\mathbf{H} - \mathbf{HVM}(\mathbf{MVM})^+\mathbf{M}] + \mathbf{E}_1\mathbf{Q}_{(\mathbf{X}:\mathbf{V})} \\
&= \mathbf{LX}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^+ + \mathbf{E}_1\mathbf{Q}_{(\mathbf{X}:\mathbf{V})} \\
&= \mathbf{X}_*(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^+ + \mathbf{E}_1\mathbf{Q}_{(\mathbf{X}:\mathbf{V})}\,,
\end{aligned} \tag{22}$$

where $\mathbf{E}_1$ is free to vary. In (22) we have used (9).

■ **Solution 2:** An alternative way to introduce a representation for $\mathbf{B}$ is to start from $\mathbf{BVM} = \mathbf{0}$, which by Proposition 4 is equivalent to

$$\mathscr{C}(\mathbf{B}') \subseteq \mathscr{C}(\mathbf{VM})^\perp = \mathscr{C}\big[(\mathbf{W}^+)'\mathbf{X} : \mathbf{Q}_{\mathbf{W}}\big]\,,$$

where $\mathbf{W} \in \mathscr{W}(\mathscr{M})$, so that

$$\mathbf{B}' = (\mathbf{W}^+)'\mathbf{XR} + \mathbf{Q}_{\mathbf{W}}\mathbf{S}\,, \quad \text{for some } \mathbf{S} \text{ and } \mathbf{R}.$$

Now the unbiasedness condition $\mathbf{X}'\mathbf{B}' = \mathbf{X}'_*$ holds if and only if

$$\mathbf{X}'\mathbf{W}^+\mathbf{XR} + \mathbf{X}'\mathbf{Q}_{\mathbf{W}}\mathbf{S} = \mathbf{X}'\mathbf{W}^+\mathbf{XR} = \mathbf{X}'_*\,,$$

from which it follows that the general expression for $\mathbf{R}$ can be expressed as

$$\mathbf{R} = (\mathbf{X}'\mathbf{W}^+\mathbf{X})^-\mathbf{X}'_* + \mathbf{Q}_{\mathbf{X}'\mathbf{W}^+}\mathbf{E}_3 = (\mathbf{X}'\mathbf{W}^+\mathbf{X})^-\mathbf{X}'_* + \mathbf{Q}_{\mathbf{X}'}\mathbf{E}_3\,,$$

where $\mathbf{E}_3$ is free to vary. Hence the general expression for $\mathbf{B}'$ satisfying $\mathbf{B}(\mathbf{X} : \mathbf{VM}) = (\mathbf{X}_* : \mathbf{0})$ can be written as as

$$\begin{aligned}
\mathbf{B}'_0 &= (\mathbf{W}^+)'\mathbf{X}(\mathbf{X}'\mathbf{W}^+\mathbf{X})^-\mathbf{X}'_* + (\mathbf{W}^+)'\mathbf{X}\mathbf{Q}_{\mathbf{X}'}\mathbf{E}_3 + \mathbf{Q}_{\mathbf{W}}\mathbf{S} \\
&= (\mathbf{W}^+)'\mathbf{X}(\mathbf{X}'\mathbf{W}^+\mathbf{X})^-\mathbf{X}'_* + \mathbf{Q}_{\mathbf{W}}\mathbf{S}\,,
\end{aligned}$$

so that

$$\mathbf{B}_0 = \mathbf{X}_*(\mathbf{X}'\mathbf{W}^+\mathbf{X})^-\mathbf{X}'\mathbf{W}^+ + \mathbf{S}'\mathbf{Q}_{\mathbf{W}} = \mathbf{P}_{\mathbf{X}_*;\mathbf{W}^+} + \mathbf{S}'\mathbf{Q}_{\mathbf{W}}\,,$$

where $\mathbf{S}$ is free to vary. Thus we have obtained the same presentation as in (22).

■ **Solution 3:** It is clear that there exists a matrix $\mathbf{X}^{\sim}$ such that $\mathbf{X}\mathbf{X}^{\sim}\mathbf{y}$ is the BLUE for $\mathbf{X}\boldsymbol{\beta}$, *i.e.*,

$$\mathbf{X}\mathbf{X}^{\sim}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{X} : \mathbf{0}), \tag{23}$$

so that $\mathbf{X}^{\sim} \in \{\mathbf{X}^{-}\}$. According to Rao and Mitra (1971, Th. 2.4.1), the general expression for a generalized inverse $\mathbf{X}^{\sim} \in \{\mathbf{X}^{-}\}$ can be written as

$$\mathbf{X}^{\sim} = \mathbf{X}^{+} + \mathbf{E}_3(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}}) + (\mathbf{I}_p - \mathbf{P}_{\mathbf{X}'})\mathbf{E}_4,$$

where $\mathbf{E}_3$ and $\mathbf{E}_4$ are free to vary. Now

$$\mathbf{X}\mathbf{X}^{\sim} = \mathbf{H} + \mathbf{X}\mathbf{E}_3\mathbf{M}, \tag{24}$$

and hence (23) holds if and only if

$$\mathbf{X}\mathbf{X}^{\sim}\mathbf{V}\mathbf{M} = \mathbf{H}\mathbf{V}\mathbf{M} + \mathbf{X}\mathbf{E}_3\mathbf{M}\mathbf{V}\mathbf{M} = \mathbf{0},$$

i.e.,

$$\mathbf{X}\mathbf{E}_3\mathbf{M}\mathbf{V}\mathbf{M} = -\mathbf{H}\mathbf{V}\mathbf{M}. \tag{25}$$

One solution for $\mathbf{X}\mathbf{E}_3$ in (25) is $\mathbf{X}\mathbf{E}_3 = -\mathbf{H}\mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{+}$, and thus $\mathbf{X}\mathbf{X}^{\sim}$ in (24) can be written as

$$\mathbf{X}\mathbf{X}^{\sim} = \mathbf{H} - \mathbf{H}\mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{+}\mathbf{M}. \tag{26}$$

Notice that $\mathbf{X}^{\sim}$ satisfying (23) can be written as

$$\mathbf{X}^{=} = \mathbf{X}^{+} - \mathbf{X}^{+}\mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{+}\mathbf{M}.$$

Another choice for $\mathbf{X}^{\sim}$ satisfying (23) is obviously

$$\mathbf{X}^{\#} = (\mathbf{X}'\mathbf{W}^{+}\mathbf{X})^{+}\mathbf{X}'\mathbf{W}^{+}.$$

It is easy to confirm that actually $\mathbf{X}^{=} = \mathbf{X}^{\#}$.

■ **Solution 4:** A very straightforward way to find a general solution to $\mathbf{B}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{X}_* : \mathbf{0})$ is to write

$$\mathbf{B}_0 = (\mathbf{X}_* : \mathbf{0})(\mathbf{X} : \mathbf{V}\mathbf{M})^{+} + \mathbf{E}\mathbf{Q}_{(\mathbf{X}:\mathbf{V})} =: \mathbf{B}_1 + \mathbf{E}\mathbf{Q}_{(\mathbf{X}:\mathbf{V})},$$

where the matrix $\mathbf{E}$ is free to vary. It is easy to confirm that $(\mathbf{X} : \mathbf{V}\mathbf{M})^{+}$ can be written as

$$(\mathbf{X} : \mathbf{V}\mathbf{M})^{+} = \begin{pmatrix} \mathbf{X}^{+}[\mathbf{I}_n - \mathbf{V}(\mathbf{M}\mathbf{V}\mathbf{M})^{+}] \\ (\mathbf{M}\mathbf{V}\mathbf{M})^{+} \end{pmatrix}. \tag{27}$$

Therefore, when $\mathbf{X}_* = \mathbf{L}\mathbf{X}$ for some $\mathbf{L} \in \mathbb{R}^{q\times n}$ so that $\mathbf{X}_*\mathbf{X}^{+} = \mathbf{L}\mathbf{H}$,

$$\begin{aligned} \mathbf{B}_1 &= (\mathbf{X}_* : \mathbf{0})(\mathbf{X} : \mathbf{V}\mathbf{M})^{+} = \mathbf{X}_*\mathbf{X}^{+}[\mathbf{I}_n - \mathbf{V}(\mathbf{M}\mathbf{V}\mathbf{M})^{+}] \\ &= \mathbf{L}[\mathbf{H} - \mathbf{H}\mathbf{V}(\mathbf{M}\mathbf{V}\mathbf{M})^{+}] = \mathbf{L}\mathbf{X}(\mathbf{X}'\mathbf{W}^{+}\mathbf{X})^{+}\mathbf{X}'\mathbf{W}^{+}. \end{aligned} \tag{28}$$

■ **Solution 5:** In the partitioned model $\mathscr{M}_{12} = \{\mathbf{y}, (\mathbf{X}_1 : \mathbf{X}_2)\boldsymbol{\beta}, \mathbf{V}\}$, one expression for the BLUE of $\boldsymbol{\mu}_1$ can be obtained from (16) yielding

$$\mathrm{BLUE}(\boldsymbol{\mu}_1 \mid \mathscr{M}_{12}) = \tilde{\boldsymbol{\mu}}_1(\mathscr{M}_{12}) = (\mathbf{X}_1 : \mathbf{0})(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{+}\mathbf{y} =: \mathbf{P}_{\#1}\mathbf{y}.$$

Premultiplying the model $\mathscr{M}_{12}$ by $\mathbf{M}_2$ yields the reduced model

$$\mathscr{M}_{12\cdot 2} = \{\mathbf{M}_2\mathbf{y},\, \mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1,\, \mathbf{M}_2\mathbf{V}\mathbf{M}_2\}\,.$$

The fundamental BLUE equation for estimating $\boldsymbol{\theta}_1 := \mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1$ under $\mathscr{M}_{12\cdot 2}$ is now

$$\mathbf{L}(\mathbf{M}_2\mathbf{X}_1 : \mathbf{M}_2\mathbf{V}\mathbf{M}_2 \cdot \mathbf{Q}_{\mathbf{M}_2\mathbf{X}_1}) = (\mathbf{M}_2\mathbf{X}_1 : \mathbf{0})\,. \tag{29}$$

To find a solution for $\mathbf{L}$ in (29), we observe that choosing $\mathbf{W} = \mathbf{V} + \mathbf{X}\mathbf{X}' \in \mathscr{W}(\mathscr{M}_{12})$ we have $\mathbf{M}_2\mathbf{W}\mathbf{M}_2 \in \mathscr{W}(\mathscr{M}_{12\cdot 2})$. Hence one solution for $\mathbf{L}$ in (29) is

$$\mathbf{L} = \mathbf{M}_2\mathbf{X}_1(\mathbf{X}_1'\dot{\mathbf{M}}_2\mathbf{X}_1)^{-}\mathbf{X}_1'\dot{\mathbf{M}}_2\mathbf{y} =: \mathbf{M}_2 \cdot \mathbf{P}_{\mathbf{X}_1;\dot{\mathbf{M}}_2}\,,$$

where

$$\dot{\mathbf{M}}_2 = \mathbf{M}_2(\mathbf{M}_2\mathbf{W}\mathbf{M}_2)^{+}\mathbf{M}_2\,,$$

and so $\mathbf{L}\mathbf{y} = \mathrm{BLUE}(\boldsymbol{\theta}_1 \mid \mathscr{M}_{12\cdot 2})$ and $\mathbf{P}_{\mathbf{X}_1;\dot{\mathbf{M}}_2}\mathbf{y} = \mathrm{BLUE}(\boldsymbol{\mu}_1 \mid \mathscr{M}_{12\cdot 2})$, *i.e.*,

$$\mathbf{X}_1(\mathbf{X}_1'\dot{\mathbf{M}}_2\mathbf{X}_1)^{-}\mathbf{X}_1'\dot{\mathbf{M}}_2\mathbf{y} = \mathbf{P}_{\mathbf{X}_1;\dot{\mathbf{M}}_2}\mathbf{y} = \tilde{\boldsymbol{\mu}}_1(\mathscr{M}_{12\cdot 2})\,.$$

It is easy to confirm that

$$\mathbf{P}_{\mathbf{X}_1;\dot{\mathbf{M}}_2}(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{V}\mathbf{M}) = (\mathbf{X}_1 : \mathbf{0} : \mathbf{0})\,,$$

so that the BLUEs of $\boldsymbol{\mu}_1$ under $\mathscr{M}_{12}$ and $\mathscr{M}_{12\cdot 2}$ coincide, which is the message of the Frisch–Waugh–Lovell theorem, see, *e.g.*, Groß and Puntanen (2000, Sec. 6).

Actually the following holds, see Haslett *et al.* (2023, Prop. 3.1),

$$\mathbf{P}_{1\#} = (\mathbf{X}_1 : \mathbf{0})(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{+} = \mathbf{X}_1(\mathbf{X}_1'\dot{\mathbf{M}}_2\mathbf{X}_1)^{-}\mathbf{X}_1'\dot{\mathbf{M}}_2 = \mathbf{P}_{\mathbf{X}_1;\dot{\mathbf{M}}_2}\,,$$

and hence

$$\begin{aligned}\mathbf{P}_{1\#} &= (\mathbf{X}_1 : \mathbf{0})\mathbf{X}^{\sim}, \text{ where } \mathbf{X}^{\sim} = (\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{+} \in \{\mathbf{X}^{-}\}\,,\\ \mathbf{P}_{\mathbf{X}_1;\dot{\mathbf{M}}_2} &= \mathbf{X}_1\ \mathbf{X}_1^{\sim}, \qquad \text{where } \mathbf{X}_1^{\sim} = (\mathbf{X}_1'\dot{\mathbf{M}}_2\mathbf{X}_1)^{-}\mathbf{X}_1'\dot{\mathbf{M}}_2 \in \{\mathbf{X}_1^{-}\}\,.\end{aligned}$$

■ **Solution 6:** Let $\mathbf{W} \in \mathscr{W}_{\geq}(\mathscr{M})$. Then it is clear that

$$\mathbf{G}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{X} : \mathbf{0}) \iff \mathbf{G}(\mathbf{X} : \mathbf{W}\mathbf{M}) = (\mathbf{X} : \mathbf{0})\,.$$

Observing that $\mathscr{M}_{\mathbf{W}} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{W}\}$ is a weakly singular linear model we can conclude, parallel to (18), that

$$\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-}\mathbf{y} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_{\mathbf{W}}) = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M})\,. \tag{30}$$

For (30) see also Christensen (2020, Th. 10.1.3).

■ **Solution 7:** (Pandora's Box.) Rao (1971, Th. 3.1) proved that the matrix $\mathbf{G}$ is a solution to the fundamental equation $\mathbf{G}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{X} : \mathbf{0})$ if and only if there exists a matrix $\mathbf{L}$ such that $\mathbf{G}$ is a solution to

$$\begin{pmatrix}\mathbf{V} & \mathbf{X}\\ \mathbf{X}' & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{G}'\\ \mathbf{L}\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ \mathbf{X}'\end{pmatrix}.$$

Let us denote

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{V} & \mathbf{X} \\ \mathbf{X}' & \mathbf{0} \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & -\mathbf{C}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{V} & \mathbf{X} \\ \mathbf{X}' & \mathbf{0} \end{pmatrix}^{-} \in \{\mathbf{\Gamma}^{-}\},$$

so that $\mathbf{C}$ is a generalized inverse of $\mathbf{\Gamma}$. Rao (1971) showed that the matrix $\mathbf{C}$ is like a Pandora's Box, providing surprisingly many useful results concerning the model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$. For example, denoting $\tilde{\boldsymbol{\mu}} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M})$, the following holds:

$$\tilde{\boldsymbol{\mu}} = \mathbf{X}\mathbf{C}_{12}'\mathbf{y}, \quad \mathrm{cov}(\tilde{\boldsymbol{\mu}}) = \mathbf{X}\mathbf{C}_{22}\mathbf{X}', \quad \tilde{\boldsymbol{\varepsilon}} = \mathbf{y} - \tilde{\boldsymbol{\mu}} = \mathbf{V}\mathbf{C}_{11}\mathbf{y}.$$

4. Solutions for BLUPs

Let us define the sets $\{\mathbf{P}_{\mathbf{y}_*|\mathscr{M}_*}\}$, $\{\mathbf{P}_{\mathbf{X}_*|\mathscr{M}_*}\}$, and $\{\mathbf{P}_{\boldsymbol{\varepsilon}_*|\mathscr{M}_*}\}$ as follows:

$$\mathbf{A} \in \{\mathbf{P}_{\mathbf{y}_*|\mathscr{M}_*}\} \iff \mathbf{A}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{X}_* : \mathbf{V}_{21}\mathbf{M}), \tag{31a}$$

$$\mathbf{B} \in \{\mathbf{P}_{\mathbf{X}_*|\mathscr{M}_*}\} \iff \mathbf{B}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{X}_* : \mathbf{0}), \tag{31b}$$

$$\mathbf{D} \in \{\mathbf{P}_{\boldsymbol{\varepsilon}_*|\mathscr{M}_*}\} \iff \mathbf{D}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{0} : \mathbf{V}_{21}\mathbf{M}). \tag{31c}$$

Using (27), one solution to $\mathbf{A}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{X}_* : \mathbf{V}_{21}\mathbf{M})$ can be written as

$$\mathbf{A}_1 = (\mathbf{X}_* : \mathbf{V}_{21}\mathbf{M})(\mathbf{X} : \mathbf{V}\mathbf{M})^{+} = \mathbf{B}_1 + \mathbf{V}_{21}(\mathbf{M}\mathbf{V}\mathbf{M})^{+},$$

where by (28), $\mathbf{B}_1 = \mathbf{X}_*(\mathbf{X}'\mathbf{W}^{+}\mathbf{X})^{+}\mathbf{X}'\mathbf{W}^{+}$. Putting (31b) and (31c) together yields

$$\begin{pmatrix} \mathbf{B} \\ \mathbf{D} \end{pmatrix} (\mathbf{X} : \mathbf{V}\mathbf{M}) = \begin{pmatrix} \mathbf{X}_* & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{21}\mathbf{M} \end{pmatrix},$$

which implies that

$$(\mathbf{B} + \mathbf{D})(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{X}_* : \mathbf{V}_{21}\mathbf{M}),$$

and thereby $(\mathbf{B} + \mathbf{D})\mathbf{y}$ is a BLUP for $\mathbf{y}_*$ and we have the following result:

$$\mathrm{BLUP}(\mathbf{y}_*) = \mathrm{BLUE}(\mathbf{X}_*\boldsymbol{\beta}) + \mathrm{BLUP}(\boldsymbol{\varepsilon}_*).$$

From part (c) of Theorem 2 we observe that $\mathbf{D}\mathbf{y}$ is the BLUP for $\boldsymbol{\varepsilon}_*$ if $\mathbf{D} = \mathbf{K}\mathbf{M}$ for some matrix $\mathbf{K} \in \mathbb{R}^{q \times n}$ such that $\mathbf{K}\mathbf{M}\mathbf{V}\mathbf{M} = \mathbf{V}_{21}\mathbf{M}$, from which one solution to $\mathbf{K}$ is $\mathbf{K} = \mathbf{V}_{21}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}$ yielding the following expression:

$$\mathrm{BLUP}(\boldsymbol{\varepsilon}_*) = \mathbf{D}\mathbf{y} = \mathbf{V}_{21}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{y} = \mathbf{V}_{21}\dot{\mathbf{M}}\mathbf{y}.$$

Further representations, see Haslett *et al.* (2014, Th. 2), are

$$\begin{aligned} \mathrm{BLUP}(\boldsymbol{\varepsilon}_*) &= \mathbf{V}_{21}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{y} = \mathbf{V}_{21}\mathbf{V}^{-}\mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{y} \\ &= \mathbf{V}_{21}\mathbf{W}^{-}\mathbf{W}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{y} = \mathbf{V}_{21}\mathbf{V}^{-}(\mathbf{y} - \tilde{\boldsymbol{\mu}}) \\ &= \mathbf{V}_{21}\mathbf{V}^{-}(\mathbf{I}_n - \mathbf{G})\mathbf{y} = \mathbf{V}_{21}\mathbf{W}^{-}(\mathbf{I}_n - \mathbf{G})\mathbf{y}, \end{aligned}$$

where $\mathbf{G} = \mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-}$. If $\mathbf{V}$ is positive definite and $\mathrm{r}(\mathbf{X}) = p$, we obtain

$$\mathrm{BLUP}(\mathbf{y}_*) = \mathrm{BLUE}(\mathbf{X}_*\boldsymbol{\beta}) + \mathrm{BLUP}(\boldsymbol{\varepsilon}_*) = \mathbf{X}_*\tilde{\boldsymbol{\beta}} + \mathbf{V}_{21}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})\,,$$

where $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$.

One application of the model $\mathscr{M}_*$ is the *linear mixed model*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad \text{or shortly,} \quad \mathscr{L} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{D}, \mathbf{R}\}\,,$$

where $\mathbf{X}_{n\times p}$ and $\mathbf{Z}_{n\times q}$ are known matrices, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of unknown fixed effects, $\mathbf{u}$ is an unobservable vector (q elements) of random effects with $\mathrm{E}(\mathbf{u}) = \mathbf{0}$, $\mathrm{cov}(\mathbf{u}) = \mathbf{D}$, $\mathrm{cov}(\mathbf{e}, \mathbf{u}) = \mathbf{0}$, and $\mathrm{E}(\mathbf{e}) = \mathbf{0}$, $\mathrm{cov}(\mathbf{e}) = \mathbf{R}$. In this situation we have

$$\mathrm{cov}\begin{pmatrix}\mathbf{e}\\ \mathbf{u}\end{pmatrix} = \begin{pmatrix}\mathbf{R} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}\end{pmatrix} =: \boldsymbol{\Lambda}\,, \quad \mathrm{cov}\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma} & \mathbf{ZD}\\ (\mathbf{ZD})' & \mathbf{D}\end{pmatrix} =: \boldsymbol{\Omega}\,.$$

The mixed model can be expressed as a version of the model with "new future observations", the new (unobservable) observations being, for example, in $\mathbf{u} = \mathbf{0}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_*$:

$$\mathscr{L}_* := \left\{\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\end{pmatrix}, \begin{pmatrix}\mathbf{X}\\ \mathbf{0}\end{pmatrix}\boldsymbol{\beta}, \begin{pmatrix}\boldsymbol{\Sigma} & \mathbf{ZD}\\ (\mathbf{ZD})' & \mathbf{D}\end{pmatrix}\right\}. \tag{32}$$

Corresponding to (1) we have

$$\begin{aligned}
&\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \quad \boldsymbol{\varepsilon} = \mathbf{Z}\mathbf{u} + \mathbf{e}\,, \quad \mathrm{cov}(\boldsymbol{\varepsilon}) = \mathrm{cov}(\mathbf{y}) = \mathbf{ZDZ}' + \mathbf{R} =: \boldsymbol{\Sigma}\,,\\
&\mathbf{y}_* = \mathbf{u}\,, \quad \mathbf{X}_* = \mathbf{0}\,,\\
&\boldsymbol{\varepsilon}_* = \mathbf{u}\,, \quad \mathrm{cov}(\boldsymbol{\varepsilon}_*) = \mathbf{D}, \quad \mathrm{cov}(\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}_*) = \mathbf{ZD}.
\end{aligned}$$

Now under the mixed model $\mathscr{L}$, $\mathbf{B}_1\mathbf{y}$ is the BLUE for $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{B}_2\mathbf{y}$ is the BLUP for $\mathbf{u}$ if and only if

$$\begin{pmatrix}\mathbf{B}_1\\ \mathbf{B}_2\end{pmatrix}(\mathbf{X} : \boldsymbol{\Sigma}\mathbf{M}) = \begin{pmatrix}\mathbf{X} & \mathbf{0}\\ \mathbf{0} & (\mathbf{ZD})'\mathbf{M}\end{pmatrix}. \tag{34}$$

Thus the $\mathrm{BLUP}(\mathbf{u})$ can be written as

$$\mathrm{BLUP}(\mathbf{u}) = \mathbf{DZ}'\mathbf{W}^{-}(\mathbf{y} - \tilde{\boldsymbol{\mu}}) = \mathbf{DZ}'\mathbf{M}(\mathbf{M}\boldsymbol{\Sigma}\mathbf{M})^{-}\mathbf{M}\mathbf{y},$$

where $\mathbf{W} = \boldsymbol{\Sigma} + \mathbf{X}\mathbf{X}'$. For example, in the simple situation when $\mathbf{X}$ has full column rank and $\boldsymbol{\Sigma} = \mathbf{ZDZ}' + \mathbf{R}$ is positive definite, we have

$$\mathrm{BLUP}(\mathbf{u}) = \mathbf{DZ}'\boldsymbol{\Sigma}^{-1}(\mathbf{I}_n - \mathbf{X}\tilde{\boldsymbol{\beta}})\,, \quad \tilde{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}.$$

Remark 2: We can write up the mixed model (32) as

$$\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\end{pmatrix} = \begin{pmatrix}\mathbf{X}\\ \mathbf{0}\end{pmatrix}\boldsymbol{\beta} + \begin{pmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_*\end{pmatrix}, \quad \text{where } \mathrm{cov}\begin{pmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_*\end{pmatrix} = \boldsymbol{\Omega}\,. \tag{35}$$

It noteworthy that even as (35) looks like a standard linear model it is not quite correct: the random vector $\mathbf{u}$ is *unobservable*. On the other hand, keeping $\mathbf{u}$ fixed (but unknown)

and denoting $\mathbf{y}_0 = \mathbf{u} + \boldsymbol{\varepsilon}_*$ we get a fixed partitioned model with supplemented stochastic restrictions on $\mathbf{u}$:

$$\mathscr{F} := \left\{ \begin{pmatrix} \mathbf{y} \\ \mathbf{y}_0 \end{pmatrix}, \begin{pmatrix} \mathbf{X} & \mathbf{Z} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{pmatrix}, \begin{pmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix} \right\}.$$

We get an interesting version of $\mathscr{F}$ by putting $\mathbf{y}_0 = \mathbf{0}$:

$$\mathscr{F}_\# := \left\{ \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{X} & \mathbf{Z} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{pmatrix}, \begin{pmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix} \right\} = \left\{ \mathbf{y}_\#,\, \mathbf{X}_\# \boldsymbol{\pi},\, \boldsymbol{\Lambda} \right\}.$$

Of course $\mathscr{F}_\#$ is not a proper model since $\mathbf{y}_0 = \mathbf{0}$. In the full rank case fitting the model $\mathscr{F}_\#$ yields to so-called Henderson equations and the BLUE of $\mathbf{X}\boldsymbol{\beta}$ and BLUP of $\mathbf{u}$ are obtained by minimizing the following quadratic form $f(\boldsymbol{\beta}, \mathbf{u})$ (keeping $\mathbf{u}$ as a non-random vector):

$$f(\boldsymbol{\beta}, \mathbf{u}) = (\mathbf{y}_\# - \mathbf{X}_\# \boldsymbol{\pi})' \boldsymbol{\Lambda}^{-1} (\mathbf{y}_\# - \mathbf{X}_\# \boldsymbol{\pi}).$$

For further references, see, *e.g.*, Henderson (1950, 1963) and Haslett *et al.* (2015). □

5. Two models with different covariance matrices

Suppose that we have two models $\mathscr{M}(\mathbf{V}_0) = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}_0\}$ and $\mathscr{M}(\mathbf{V}) = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$, which have different covariance matrices. Then we can ask, for example, what is needed that every representation of the BLUE of $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}(\mathbf{V}_0)$ remains BLUE under $\mathscr{M}(\mathbf{V})$. Mitra and Moore (1973, p. 139) give a very clear description of the different problems occurring:

(a) Problem MM-1: When is specific linear representation of the BLUE of $\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}(\mathbf{V}_0)$ also BLUE under $\mathscr{M}(\mathbf{V})$?

(b) Problem MM-2: When does $\mathbf{X}\boldsymbol{\beta}$ have a common representation for the BLUE under $\mathscr{M}(\mathbf{V}_0)$ and $\mathscr{M}(\mathbf{V})$?

(c) Problem MM-3: When does every linear representation of the BLUE of $\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}(\mathbf{V}_0)$ remain BLUE also under $\mathscr{M}(\mathbf{V})$?

As for MM-1, we may mention that Hauke *et al.* (2013) consider conditions under which

$$\mathbf{P}_{\mathbf{X};\mathbf{W}_0^+} \mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{W}_0^- \mathbf{X})^- \mathbf{X}'\mathbf{W}_0^+ \mathbf{y} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}(\mathbf{V})). \tag{36}$$

This happens if and only if $\mathbf{P}_{\mathbf{X};\mathbf{W}_0^+} \mathbf{V}\mathbf{M} = \mathbf{0}$, which further is equivalent to

$$\mathbf{X}'\mathbf{W}_0^+ \mathbf{V}\mathbf{M} = \mathbf{0}, \quad \text{i.e.,} \quad \mathscr{C}(\mathbf{V}\mathbf{M}) \subseteq \mathscr{C}(\mathbf{W}_0^+ \mathbf{X})^\perp = \mathscr{C}(\mathbf{V}_0 \mathbf{M} : \mathbf{Q}_{\mathbf{W}_0}),$$

where we have used Proposition 4. Denoting $\mathbf{Z} = (\mathbf{V}_0 \mathbf{M} : \mathbf{Q}_{\mathbf{W}_0})$, Hauke *et al.* (2013) showed that (36) holds if and only if $\mathbf{V}$ belongs to the class $\mathscr{V}_{mm1}$, say, defined as

$$\mathbf{V} \in \mathscr{V}_{mm1} \iff \mathbf{V} = \mathbf{X}\mathbf{A}\mathbf{A}'\mathbf{X}' + \mathbf{Z}\mathbf{B}\mathbf{B}'\mathbf{Z}' \quad \text{for some matrices } \mathbf{A} \text{ and } \mathbf{B}. \tag{37}$$

Let us take a closer look at MMM-3 in the spirit of Puntanen *et al.* (2011, Sec. 11.1). First, let us denote

$$\mathbf{G} \in \{\mathbf{P}_{\boldsymbol{\mu}\,|\,\mathbf{V}_0}\} \iff \mathbf{G}(\mathbf{X} : \mathbf{V}_0\mathbf{M}) = (\mathbf{X} : \mathbf{0})\,.$$

Let $\mathbf{G}$ be such a matrix that $\mathbf{Gy}$ is the BLUE for $\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}(\mathbf{V}_0)$, Then we say that $\mathbf{Gy}$ remains BLUE under $\mathscr{M}(\mathbf{V})$ if the following implication holds:

$$\mathbf{G}(\mathbf{X} : \mathbf{V}_0\mathbf{M}) = (\mathbf{X} : \mathbf{0}) \implies \mathbf{G}(\mathbf{X} : \mathbf{VM}) = (\mathbf{X} : \mathbf{0})\,.$$

Moreover, let the set of all representations of BLUE of $\boldsymbol{\mu}$ under $\mathscr{M}(\mathbf{V}_0)$ be denoted as

$$\begin{aligned}\mathscr{B}(\boldsymbol{\mu}\,|\,\mathbf{V}_0) &= \{\mathrm{BLUE}(\boldsymbol{\mu}\,|\,\mathbf{V}_0)\} = \{\mathbf{Gy} : \mathbf{G}(\mathbf{X} : \mathbf{V}_0\mathbf{M}) = (\mathbf{X} : \mathbf{0})\} \\ &= \{\mathbf{Gy} : \mathbf{G} \in \{\mathbf{P}_{\boldsymbol{\mu}\,|\,\mathbf{V}_0}\}\}\,. \end{aligned} \tag{38}$$

It is important to understand that the notation of the above type (38) is merely symbolic. Our main interest lies in the *multipliers*, like the members of $\{\mathbf{P}_{\boldsymbol{\mu}\,|\,\mathbf{V}_0}\}$, of the response vector $\mathbf{y}$ which have specific properties. For the property that *every* representation of the BLUE of $\boldsymbol{\mu}$ under $\mathscr{M}(\mathbf{V}_0)$ remains BLUE of $\boldsymbol{\mu}$ under $\mathscr{M}(\mathbf{V})$ we will use the notation

$$\mathscr{B}(\boldsymbol{\mu}\,|\,\mathbf{V}_0) \subseteq \mathscr{B}(\boldsymbol{\mu}\,|\,\mathbf{V})\,, \quad \text{i.e.,} \quad \{\mathbf{P}_{\boldsymbol{\mu}\,|\,\mathbf{V}_0}\} \subseteq \{\mathbf{P}_{\boldsymbol{\mu}\,|\,\mathbf{V}}\}\,. \tag{39}$$

We may consider $\mathscr{M}(\mathbf{V}_0)$ as the original model and $\mathscr{M}(\mathbf{V})$ as the misspecified model; misspecification concerning only the covariance matrix.

Let us next show that (39) is equivalent to

$$\mathscr{C}(\mathbf{VM}) \subseteq \mathscr{C}(\mathbf{V}_0\mathbf{M})\,, \tag{40}$$

which is essentially Rao's result in Theorem 5.3 of his paper in 1971. This is a well-known old but yet a fundamental result whose proof is worth going through. Proceeding as Puntanen *et al.* (2011, p. 270), we observe that a general representation of a member in $\{\mathbf{P}_{\boldsymbol{\mu}\,|\,\mathbf{V}_0}\}$ can be expressed as

$$\mathbf{G}_0 = \mathbf{P}_{\mathbf{X};\mathbf{W}_0^+} + \mathbf{EQ}_{\mathbf{W}_0} = \mathbf{X}(\mathbf{X}'\mathbf{W}_0^-\mathbf{X})^-\mathbf{X}'\mathbf{W}_0^+ + \mathbf{E}(\mathbf{I}_n - \mathbf{P}_{\mathbf{W}_0})\,,$$

where $\mathbf{E}$ is free to vary and $\mathbf{W}_0 \in \mathscr{W}(\mathbf{V}_0)$. Now (39) holds if and only if

$$\mathbf{G}_0(\mathbf{X} : \mathbf{VM}) = (\mathbf{P}_{\mathbf{X};\mathbf{W}_0^+} + \mathbf{EQ}_{\mathbf{W}_0})(\mathbf{X} : \mathbf{VM}) = (\mathbf{X} : \mathbf{0})\,,$$

i.e.,

$$\mathbf{P}_{\mathbf{X};\mathbf{W}_0^+}\mathbf{VM} + \mathbf{EQ}_{\mathbf{W}_0}\mathbf{VM} = \mathbf{0} \quad \text{for all } \mathbf{E}, \tag{41}$$

which implies that $\mathbf{Q}_{\mathbf{W}_0}\mathbf{VM} = \mathbf{0}$, *i.e.*, $\mathscr{C}(\mathbf{VM}) \subseteq \mathscr{C}(\mathbf{W}_0) = \mathscr{C}(\mathbf{X} : \mathbf{V}_0\mathbf{M})$, which further means that

$$\mathbf{VM} = \mathbf{XR} + \mathbf{V}_0\mathbf{MS} \quad \text{for some } \mathbf{R} \text{ and } \mathbf{S}. \tag{42}$$

Substituting (42) into (41) shows that $\mathbf{XR} = \mathbf{0}$ and thereby (39) implies (40). The reverse relation is easy to check. It is worth noting that

$$\mathscr{C}(\mathbf{VM}) \subseteq \mathscr{C}(\mathbf{V}_0\mathbf{M}) \implies \mathscr{C}(\mathbf{X} : \mathbf{VM}) \subseteq \mathscr{C}(\mathbf{X} : \mathbf{V}_0\mathbf{M})$$

but the reverse implication does not hold.

Remark 3: Let us consider conditions under which

$$\mathbf{P}_{\mathbf{X};\mathbf{W}_0^-}\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{W}_0^-\mathbf{X})^-\mathbf{X}'\mathbf{W}_0^-\mathbf{y} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}(\mathbf{V})) \ \text{ for all } \mathbf{W}_0^- \,, \tag{43}$$

i.e.,

$$\mathbf{X}'\mathbf{W}_0^-\mathbf{V}\mathbf{M} = \mathbf{0} \ \text{ for all } \mathbf{W}_0^- \,. \tag{44}$$

Now in view of Proposition 1, (44) holds if and only if

$$\mathbf{X}'\mathbf{W}_0^+\mathbf{V}\mathbf{M} = \mathbf{0} \ \text{ and } \ \mathscr{C}(\mathbf{V}\mathbf{M}) \subseteq \mathscr{C}(\mathbf{W}_0) \,, \tag{45}$$

i.e.,

$$\mathscr{C}(\mathbf{V}\mathbf{M}) \subseteq \mathscr{C}(\mathbf{W}_0^+\mathbf{X})^\perp = \mathscr{C}(\mathbf{V}_0\mathbf{M} : \mathbf{Q}_{\mathbf{W}_0}) \ \text{ and } \ \mathscr{C}(\mathbf{V}\mathbf{M}) \subseteq \mathscr{C}(\mathbf{W}_0) \,, \tag{46}$$

which together imply (40). □

It is clear that $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ has a common representation for the BLUE under $\mathscr{M}(\mathbf{V}_0)$ and $\mathscr{M}(\mathbf{V})$, *i.e.*, $\{\mathbf{P}_{\boldsymbol{\mu}\,|\,\mathbf{V}_0}\} \cap \{\mathbf{P}_{\boldsymbol{\mu}\,|\,\mathbf{V}}\} \neq \{\emptyset\}$, if and only if the equation

$$\mathbf{G}(\mathbf{X} : \mathbf{V}_0\mathbf{M} : \mathbf{V}\mathbf{M}) = (\mathbf{X} : \mathbf{0} : \mathbf{0})$$

has a solution for $\mathbf{G}$, *i.e.*,

$$\mathscr{C}[(\mathbf{X} : \mathbf{0} : \mathbf{0})'] \subseteq \mathscr{C}[(\mathbf{X} : \mathbf{V}_0\mathbf{M} : \mathbf{V}\mathbf{M})'] \,, \tag{47}$$

for which, according to Mitra and Moore (1973, Sec. 3), it is necessary and sufficient that

$$\mathscr{C}(\mathbf{V}_0\mathbf{M} : \mathbf{V}\mathbf{M}) \cap \mathscr{C}(\mathbf{X}) = \{\mathbf{0}\} \,.$$

Suppose that (47) holds. Given $\mathbf{V}_0$, how can we then characterize the class $\mathscr{V}_{mm2}$, say, of matrices $\mathbf{V}$ such that $\{\mathbf{P}_{\boldsymbol{\mu}\,|\,\mathbf{V}_0}\}$ and $\{\mathbf{P}_{\boldsymbol{\mu}\,|\,\mathbf{V}}\}$ are not disjoint? Mitra and Moore (1973, Sec. 3) showed that $\mathscr{V}_{mm2} = \mathscr{V}_{mm1}$ so that

$$\mathbf{V} \in \mathscr{V}_{mm2} \iff \mathbf{V} = \mathbf{X}\mathbf{A}\mathbf{A}'\mathbf{X}' + (\mathbf{V}_0\mathbf{M} : \mathbf{Q}_{\mathbf{W}_0}) \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix} (\mathbf{B}_1' : \mathbf{B}_2') \begin{pmatrix} \mathbf{M}\mathbf{V}_0 \\ \mathbf{Q}_{\mathbf{W}_0} \end{pmatrix} , \tag{48}$$

for some matrices $\mathbf{A}, \mathbf{B}_1$ and $\mathbf{B}_2$.

Let us next consider the following task: Given a covariance matrix $\mathbf{V}_0$, characterize the set $\mathscr{V}$ of covariance matrices such that every representation of the BLUE of $\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}(\mathbf{V}_0)$ remains BLUE under $\mathscr{M}(\mathbf{V})$, *i.e.*,

$$\mathbf{V} \in \mathscr{V} \iff \mathscr{B}(\boldsymbol{\mu} \,|\, \mathbf{V}_0) \subseteq \mathscr{B}(\boldsymbol{\mu} \,|\, \mathbf{V}) \,.$$

We will next show that a necessary condition for $\mathbf{V} \in \mathscr{V}$ is the following:

$$\mathbf{V} = \mathbf{X}\mathbf{A}\mathbf{A}'\mathbf{X}' + \mathbf{V}_0\mathbf{M}\mathbf{B}\mathbf{B}'\mathbf{M}\mathbf{V}_0 \ \text{ for some matrices } \mathbf{A} \text{ and } \mathbf{B}. \tag{49}$$

This is also given by Rao (1971, Th. 5.3) but we will give a slightly different proof. Notice that class $\mathscr{V}_{mm2}$ in (48) is wider than class $\mathscr{V}$ defined in (49).

Since $\mathscr{C}(\mathbf{X} : \mathbf{V}_0\mathbf{M} : \mathbf{Q}_{\mathbf{W}_0}) = \mathbb{R}^n$, where $\mathbf{W}_0 \in \mathscr{W}(\mathbf{V}_0)$, an arbitrary nonnegative definite matrix $\mathbf{V}$ can be expressed as $\mathbf{V} = \mathbf{U}\mathbf{U}'$ where

$$\mathbf{U} = \mathbf{X}\mathbf{L}_1 + \mathbf{V}_0\mathbf{M}\mathbf{L}_2 + \mathbf{Q}_{\mathbf{W}_0}\mathbf{L}_3\,,$$

for some matrices $\mathbf{L}_1, \mathbf{L}_2, \mathbf{L}_3$, so that

$$\mathbf{V} = \mathbf{X}\mathbf{L}_{11}\mathbf{X}' + \mathbf{V}_0\mathbf{M}\mathbf{L}_{22}\mathbf{M}\mathbf{V}_0 + \mathbf{Q}_{\mathbf{W}_0}\mathbf{L}_{33}\mathbf{Q}_{\mathbf{W}_0} + \mathbf{N} + \mathbf{N}', \tag{50}$$

where $\mathbf{L}_{ij} = \mathbf{L}_i\mathbf{L}_j', j = 1, 2, 3$, and

$$\mathbf{N} = \mathbf{X}\mathbf{L}_{12}\mathbf{M}\mathbf{V}_0 + \mathbf{X}\mathbf{L}_{13}\mathbf{Q}_{\mathbf{W}_0} + \mathbf{V}_0\mathbf{M}\mathbf{L}_{23}\mathbf{Q}_{\mathbf{W}_0}\,.$$

Now

$$\mathbf{U}'\mathbf{M} = \mathbf{L}_2'\mathbf{M}\mathbf{V}_0\mathbf{M} + \mathbf{L}_3'\mathbf{Q}_{\mathbf{W}_0}\mathbf{M} = \mathbf{L}_2'\mathbf{M}\mathbf{V}_0\mathbf{M} + \mathbf{L}_3'\mathbf{Q}_{\mathbf{W}_0} =: \mathbf{S}\,,$$

where $\mathbf{Q}_{\mathbf{W}_0}\mathbf{M} = \mathbf{Q}_{\mathbf{W}_0}$ follows from part (d) of Proposition 2. Moreover,

$$\mathscr{C}(\mathbf{U}\mathbf{U}'\mathbf{M}) = \mathscr{C}(\mathbf{X}\mathbf{L}_1\mathbf{S} + \mathbf{V}_0\mathbf{M}\mathbf{L}_2\mathbf{S} + \mathbf{Q}_{\mathbf{W}_0}\mathbf{L}_3\mathbf{S}) \subseteq \mathscr{C}(\mathbf{V}_0\mathbf{M})$$

holds if and only if

$$\mathscr{C}(\mathbf{X}\mathbf{L}_1\mathbf{S} + \mathbf{Q}_{\mathbf{W}_0}\mathbf{L}_3\mathbf{S}) \subseteq \mathscr{C}(\mathbf{V}_0\mathbf{M})\,. \tag{51}$$

Premultiplying (51) by $\mathbf{Q}_{\mathbf{W}_0}$ shows that $\mathbf{Q}_{\mathbf{W}_0}\mathbf{L}_3\mathbf{S} = \mathbf{0}$, *i.e.*,

$$\mathbf{Q}_{\mathbf{W}_0}\mathbf{L}_{32}\mathbf{M}\mathbf{V}_0\mathbf{M} + \mathbf{Q}_{\mathbf{W}_0}\mathbf{L}_{33}\mathbf{Q}_{\mathbf{W}_0} = \mathbf{0}\,. \tag{52}$$

Postmultiplying (52) by $\mathbf{Q}_{\mathbf{W}_0}$ implies that $\mathbf{Q}_{\mathbf{W}_0}\mathbf{L}_{33}\mathbf{Q}_{\mathbf{W}_0} = \mathbf{0}$, *i.e.*,

$$\mathbf{L}_3\mathbf{Q}_{\mathbf{W}_0} = \mathbf{0}\,. \tag{53}$$

Substituting (53) into (51) yields

$$\mathscr{C}(\mathbf{X}\mathbf{L}_1\mathbf{S}) \subseteq \mathscr{C}(\mathbf{V}_0\mathbf{M})\,. \tag{54}$$

The disjointness of $\mathscr{C}(\mathbf{X})$ and $\mathscr{C}(\mathbf{V}_0\mathbf{M})$ implies that (54) holds if and only if

$$\mathbf{X}\mathbf{L}_1\mathbf{S} = \mathbf{X}\mathbf{L}_1(\mathbf{L}_2'\mathbf{M}\mathbf{V}_0\mathbf{M} + \mathbf{L}_3'\mathbf{Q}_{\mathbf{W}_0}) = \mathbf{0}\,,$$

which further is equivalent to

$$\mathbf{X}\mathbf{L}_{12}\mathbf{M}\mathbf{V}_0 = \mathbf{0}\,. \tag{55}$$

Substituting (53) and (55) into (50) proves that (49) is a necessary condition for $\mathbf{V} \in \mathscr{V}$. Its sufficiency is obvious.

Some equivalent statements to (39) are given as follows.

Proposition 6: Consider the linear models $\mathscr{M}(\mathbf{V}_0) = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}_0\}$ and $\mathscr{M}(\mathbf{V}) = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$. Then the following statements are equivalent:

(a) $\mathscr{B}(\boldsymbol{\mu} \,|\, \mathbf{V}_0) \subseteq \mathscr{B}(\boldsymbol{\mu} \,|\, \mathbf{V})$, *i.e.*, $\mathbf{V} \in \mathscr{V}$.

(b) $\mathscr{C}(\mathbf{V}\mathbf{M}) \subseteq \mathscr{C}(\mathbf{V}_0\mathbf{M})$.

(c) $\mathbf{V} = \mathbf{X}\mathbf{A}\mathbf{A}'\mathbf{X}' + \mathbf{V}_0\mathbf{M}\mathbf{B}\mathbf{B}'\mathbf{M}\mathbf{V}_0$, for some matrices $\mathbf{A}$ and $\mathbf{B}$.

(d) $\mathbf{V} = \mathbf{V}_0 + \mathbf{X}\mathbf{C}\mathbf{C}'\mathbf{X}' + \mathbf{V}_0\mathbf{M}\mathbf{D}\mathbf{D}'\mathbf{M}\mathbf{V}_0$, for some matrices $\mathbf{C}$ and $\mathbf{D}$.

For the proof of Proposition 6 and related discussion, see, *e.g.*, Mitra and Moore (1973, Th. 4.1–4.2), Rao (1968, Lemma 5), Rao (1971, Th. 5.2, Th. 5.5), Rao (1973, p. 289), and Baksalary and Mathew (1986, Th. 3).

Consider then the special case when we have models $\mathscr{M}(\mathbf{I}) = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{I}\}$ and $\mathscr{M}(\mathbf{V}) = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$. Then the BLUE of $\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}(\mathbf{I})$ is $\mathbf{P}_\mathbf{X}\mathbf{y} = \mathbf{H}\mathbf{y}$ since the unique solution for $\mathbf{G}$ in $\mathbf{G}(\mathbf{X} : \mathbf{M}) = (\mathbf{X} : \mathbf{0})$ is $\mathbf{H}$. When is $\mathbf{H}\mathbf{y}$, *i.e.*, the ordinary least squares estimator (OLSE) BLUE for $\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}(\mathbf{V})$? The answer is by part (b) of Proposition (6) the inclusion $\mathscr{C}(\mathbf{V}\mathbf{M}) \subseteq \mathscr{C}(\mathbf{M})$, which can be equivalently expressed as any of the following conditions:

$$\mathscr{C}(\mathbf{V}\mathbf{H}) \subseteq \mathscr{C}(\mathbf{H}), \quad \mathbf{H}\mathbf{V} = \mathbf{V}\mathbf{H}, \quad \mathbf{H}\mathbf{V}\mathbf{M} = \mathbf{0}\,.$$

For further references regarding the equality of OLSE and BLUE, see, *e.g.*, Rao (1967), Zyskind (1967), and Markiewicz *et al.* (2010, 2021).

Let $\mathscr{V}_{1/12}$ denote the set of nonnegative definite matrices $\mathbf{V}$ such that every representation of the BLUE of $\boldsymbol{\mu}_1$ under $\mathscr{M}(\mathbf{V}_0)$ remains BLUE under $\mathscr{M}(\mathbf{V})$, *i.e.*,

$$\mathbf{V} \in \mathscr{V}_{1/12} \iff \mathscr{B}(\boldsymbol{\mu}_1 \,|\, \mathbf{V}_0) \subseteq \mathscr{B}(\boldsymbol{\mu}_1 \,|\, \mathbf{V})\,.$$

In view of Haslett and Puntanen (2010a, Th. 2.1, 2023b, Th. 11.4), see also Mathew and Bhimasankaram (1983, Th. 2.1, Th. 2.4), the following holds:

Proposition 7: Consider the partitioned linear models $\mathscr{M}(\mathbf{V}_0)$ and $\mathscr{M}(\mathbf{V})$, where $\boldsymbol{\mu}_1 = \mathbf{X}_1\boldsymbol{\beta}_1$ is estimable. Then the following statements are equivalent:

(a) $\mathscr{B}(\boldsymbol{\mu}_1 \,|\, \mathbf{V}_0) \subseteq \mathscr{B}(\boldsymbol{\mu}_1 \,|\, \mathbf{V})$, *i.e.*, $\mathbf{V} \in \mathscr{V}_{1/12}$.

(b) $\mathscr{C}(\mathbf{M}_2\mathbf{V}\mathbf{M}) \subseteq \mathscr{C}(\mathbf{M}_2\mathbf{V}_0\mathbf{M})$.

(c) $\mathscr{C}(\mathbf{V}\mathbf{M}) \subseteq \mathscr{C}(\mathbf{X}_2 : \mathbf{V}_0\mathbf{M})$.

(d) The matrix $\mathbf{V}$ can be expressed, for some $\mathbf{L}_i$, $\mathbf{L}_{ij} = \mathbf{L}_i\mathbf{L}_j'$, as

$$\mathbf{V} = \mathbf{X}_1\mathbf{L}_{11}\mathbf{X}_1' + \mathbf{X}_2\mathbf{L}_{22}\mathbf{X}_2' + \mathbf{V}_0\mathbf{M}\mathbf{L}_{33}\mathbf{M}\mathbf{V}_0 + \mathbf{Z} + \mathbf{Z}',$$

where $\mathbf{Z} = \mathbf{X}_1\mathbf{L}_{12}\mathbf{X}_2' + \mathbf{X}_2\mathbf{L}_{23}\mathbf{M}\mathbf{V}_0$.

So far in this section we have been dealing with linear models $\mathscr{M}(\mathbf{V}_0) = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}_0\}$ and $\mathscr{M}(\mathbf{V}) = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$. The corresponding considerations can be done for the two models with new future observations. For this purpose, denote

$$\mathscr{A}_1 = \left\{ \begin{pmatrix} \mathbf{y} \\ \mathbf{y}_* \end{pmatrix}, \begin{pmatrix} \mathbf{X} \\ \mathbf{X}_* \end{pmatrix} \boldsymbol{\beta}, \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix} \right\},$$

where $\mathscr{C}(\mathbf{X}_*') \subseteq \mathscr{C}(\mathbf{X}')$. Consider now another model $\mathscr{A}_2$, which may differ from $\mathscr{A}_1$ through its covariance matrix, *i.e.*,

$$\mathscr{A}_2 = \left\{ \begin{pmatrix} \mathbf{y} \\ \mathbf{y}_* \end{pmatrix}, \begin{pmatrix} \mathbf{X} \\ \mathbf{X}_* \end{pmatrix} \boldsymbol{\beta}, \begin{pmatrix} \underline{\mathbf{V}}_{11} & \underline{\mathbf{V}}_{12} \\ \underline{\mathbf{V}}_{21} & \underline{\mathbf{V}}_{22} \end{pmatrix} \right\}.$$

For the proof of the following result see Haslett and Puntanen (2010b).

Proposition 8: Consider the models $\mathscr{A}_1$ and $\mathscr{A}_2$ (with new unobserved future observations), where $\mathscr{C}(\mathbf{X}_*') \subseteq \mathscr{C}(\mathbf{X}')$. Then every representation of the BLUP for $\mathbf{y}_*$ under the model $\mathscr{A}_1$ is also a BLUP for $\mathbf{y}_*$ under the model $\mathscr{A}_2$ if and only if

$$\mathscr{C} \begin{pmatrix} \underline{\mathbf{V}}_{11}\mathbf{M} \\ \underline{\mathbf{V}}_{21}\mathbf{M} \end{pmatrix} \subseteq \mathscr{C} \begin{pmatrix} \mathbf{X} & \mathbf{V}_{11}\mathbf{M} \\ \mathbf{X}_* & \mathbf{V}_{21}\mathbf{M} \end{pmatrix}.$$

Consider then two mixed models:

$$\mathscr{B}_1 = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u},\, \mathbf{D}_1\,, \mathbf{R}_1\}\,, \quad \mathscr{B}_2 = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u},\, \mathbf{D}_2,\, \mathbf{R}_2\}\,.$$

The only difference above concerns the covariance matrices. We may denote $\boldsymbol{\Sigma}_i = \mathbf{Z}\mathbf{D}_i\mathbf{Z}' + \mathbf{R}_i\,,\ i = 1, 2\,.$ For the next proposition, see Haslett and Puntanen (2011).

Proposition 9: Consider the mixed models $\mathscr{B}_1$ and $\mathscr{B}_2$. Then every representation of the BLUP for $\mathbf{u}$ under the model $\mathscr{B}_1$ is also the BLUP for $\mathbf{u}$ under the model $\mathscr{B}_2$ if and only if

$$\mathscr{C} \begin{pmatrix} \boldsymbol{\Sigma}_2\mathbf{M} \\ \mathbf{D}_2\mathbf{Z}'\mathbf{M} \end{pmatrix} \subseteq \mathscr{C} \begin{pmatrix} \mathbf{X} & \boldsymbol{\Sigma}_1\mathbf{M} \\ \mathbf{0} & \mathbf{D}_1\mathbf{Z}'\mathbf{M} \end{pmatrix}.$$

In particular, both the BLUE($\mathbf{X}\boldsymbol{\beta}$) under $\mathscr{B}_1$ continues to be BLUE($\mathbf{X}\boldsymbol{\beta}$) under $\mathscr{B}_2$ and BLUP($\mathbf{u}$) under $\mathscr{B}_1$ continues to be BLUP($\mathbf{u}$) under $\mathscr{B}_2$ if and only if

$$\mathscr{C} \begin{pmatrix} \boldsymbol{\Sigma}_2\mathbf{M} \\ \mathbf{D}_2\mathbf{Z}'\mathbf{M} \end{pmatrix} \subseteq \mathscr{C} \begin{pmatrix} \boldsymbol{\Sigma}_1\mathbf{M} \\ \mathbf{D}_1\mathbf{Z}'\mathbf{M} \end{pmatrix}.$$

6. Further remarks

In this section we very briefly review some recent articles by the authors. Fundamental BLUE/BLUP equations have instrumental role in these papers.

[**A**] Haslett *et al.* (2023), [**B**] Haslett *et al.* (2020).

In these articles we consider the partitioned linear model $\mathscr{M}_{12}$, and the corresponding small model $\mathscr{M}_1$. We focus on comparing the BLUEs of $\boldsymbol{\mu}_1$ under $\mathscr{M}_{12}$ and $\mathscr{M}_1$. Particular attention is paid on the consistency of the model, *i.e.*, whether the realized value of the response vector $\mathbf{y}$ belongs to the column space of $(\mathbf{X}_1 : \mathbf{V})$ or $(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{V})$. In [A] these models are supplemented with the new unobservable random vector $\mathbf{y}_*$, coming from $\mathbf{y}_* = \mathbf{X}_*\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_*$. We will concentrate on comparing the BLUEs of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}$, and BLUPs of $\mathbf{y}_*$ and $\boldsymbol{\varepsilon}_*$ under $\mathscr{M}_{12}$ and $\mathscr{M}_1$.

Let us shortly consider paper [A] to get an idea what kinds of problems we are dealing with here. Denote

$$\mathbf{G}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{W}_1^+\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{W}_1^+, \quad \mathbf{P}_{\mathbf{X}_1;\dot{\mathbf{M}}_2} = \mathbf{X}_1(\mathbf{X}_1'\dot{\mathbf{M}}_2\mathbf{X}_1)^-\mathbf{X}_1'\dot{\mathbf{M}}_2\,,$$

where $\mathbf{W}_1 = \mathbf{V} + \mathbf{X}_1\mathbf{X}_1'$ so that $\mathbf{G}_1\mathbf{y} = \tilde{\boldsymbol{\mu}}_1(\mathscr{M}_1)$ and $\mathbf{P}_{\mathbf{X}_1;\dot{\mathbf{M}}_2}\mathbf{y} = \tilde{\boldsymbol{\mu}}_1(\mathscr{M}_{12})$. We might now be tempted to express the equality $\mathbf{G}_1\mathbf{y} = \mathbf{P}_{\mathbf{X}_1;\dot{\mathbf{M}}_2}\mathbf{y}$ as

$$\tilde{\boldsymbol{\mu}}_1(\mathscr{M}_1) = \tilde{\boldsymbol{\mu}}_1(\mathscr{M}_{12})\,, \quad \text{i.e.,} \quad \mathrm{BLUE}(\boldsymbol{\mu}_1 \mid \mathscr{M}_1) = \mathrm{BLUE}(\boldsymbol{\mu}_1 \mid \mathscr{M}_{12})\,. \tag{56}$$

However, the notation used in (56) can be problematic when the possible values of the response vector $\mathbf{y}$ are taken into account. Doing that, we can consider for example statements like

$$\mathbf{G}_1\mathbf{y} = \mathbf{P}_{\mathbf{X}_1;\dot{\mathbf{M}}_2}\mathbf{y} \quad \text{for all } \mathbf{y} \in \mathscr{C}(\mathbf{X}_1 : \mathbf{V})\,, \tag{57a}$$

$$\mathbf{G}_1\mathbf{y} = \mathbf{P}_{\mathbf{X}_1;\dot{\mathbf{M}}_2}\mathbf{y} \quad \text{for all } \mathbf{y} \in \mathscr{C}(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{V})\,. \tag{57b}$$

The claim (57a) appears to be equivalent to $\{\mathbf{P}_{\boldsymbol{\mu}_1 \mid \mathscr{M}_{12}}\} \subseteq \{\mathbf{P}_{\boldsymbol{\mu}_1 \mid \mathscr{M}_1}\}$.

[**C**] Haslett *et al.* (2021), [**D**] Haslett *et al.* (2023a).

In these articles we consider the partitioned fixed linear model $\mathscr{F}\colon \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$ and the corresponding mixed model $\mathscr{M}\colon \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\mathbf{u} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is random error vector and $\mathbf{u}$ is a random effect vector. Isotalo *et al.* (2006) found conditions under which an arbitrary representation of the BLUE of an estimable parametric function of $\boldsymbol{\beta}_1$ in the fixed model $\mathscr{F}$ remains BLUE in the mixed model $\mathscr{M}$. In paper [C] we extend the results concerning further equalities arising from models $\mathscr{F}$ and $\mathscr{M}$. In paper [D] we establish upper bounds for the Euclidean norm of the difference between the BLUEs of an estimable parametric function of $\boldsymbol{\beta}_1$ under models $\mathscr{F}$ and $\mathscr{M}$.

[**E**] Haslett *et al.* (2023c), [**F**] Haslett *et al.* (2023b), [**G**] Haslett and Puntanen (2023).

We consider the partitioned linear model $\mathscr{M}_{12}(\mathbf{V}_0)$ and the corresponding small model $\mathscr{M}_1(\mathbf{V}_0)$. We define the set $\mathscr{V}_{1/12}$ of nonnegative definite matrices $\mathbf{V}$ such that every representation of the BLUE of $\boldsymbol{\mu}_1$ under $\mathscr{M}_{12}(\mathbf{V}_0)$ remains BLUE under $\mathscr{M}_{12}(\mathbf{V})$. Correspondingly, we can characterize the set $\mathscr{V}_1$ of matrices $\mathbf{V}$ such that every BLUE of $\boldsymbol{\mu}_1$ under $\mathscr{M}_1(\mathbf{V}_0)$ remains BLUE under $\mathscr{M}_1(\mathbf{V})$. In paper E we focus on the mutual relations between the sets $\mathscr{V}_1$ and $\mathscr{V}_{1/12}$.

In article [F] we focus on the mutual relations between the sets $\mathscr{V}_1$ and $\mathscr{V}_{12}$, where $\mathscr{V}_1$ is defined as in [E] and $\mathscr{V}_{12}$ is the set of nonnegative definite matrices $\mathbf{V}$ such that every representation of the BLUE of $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}_{12}(\mathbf{V}_0)$ remains BLUE under $\mathscr{M}_{12}(\mathbf{V})$.

Structural insight into Rao's condition of 1971 can be gained by writing the quadratic form that is permitted to be added to the original covariance in block diagonal form. When the original full linear model is made smaller by reducing the number of regressors, block diagonal or diagonal matrices also provide insight into conditions for the entire set of full, small, and intermediate models each to retain their own BLUEs. The paper [G] outlines the role that such changes in error covariance structure can play in data confidentiality and data encryption, especially when the covariance of the BLUEs is also retained.

[**H**] Haslett *et al.* (2021)

A linear statistic **Fy** is called linearly sufficient for $\mathbf{X}_*\boldsymbol{\beta}$ under $\mathscr{M}(\mathbf{V})$ if there exists a matrix **A** such that **AFy** is the BLUE for $\mathbf{X}_*\boldsymbol{\beta}$, *i.e.*, there exists a matrix **A** such that

$$\mathbf{AF}(\mathbf{X} : \mathbf{VM}) = (\mathbf{X}_* : \mathbf{0}) .$$

Thus we can immediately recognize the crucial role of the fundamental BLUE equation in definition of the linear sufficiency. Originally the concept of linear sufficiency as done by Baksalary and Kala (1981, 1986). The article [H] provides an extensive review of this concept.

Acknowledgements

The authors are grateful to Professor Bikas Kumar Sinha for the invitation to prepare this article for possible publication in the SSCA journal *Statistics and Applications*. Thanks go also to Professor Vinod Kumar Gupta for editorial cooperation.

Part of this research has been done during the meetings of an International Research Group on Multivariate and Mixed Linear Models in the Mathematical Research and Conference Center, Będlewo, Poland, in October 2023 and April 2024, supported by the Stefan Banach International Mathematical Center. Thanks for helpful discussions go to Katarzyna Filipiak.

ORCID

Stephen J. Haslett `https://orcid.org/0000-0002-2775-5468`
Jarkko Isotalo `https://orcid.org/0000-0002-8068-6262`
Augustyn Markiewicz `https://orcid.org/0000-0001-5473-3419`
Simo Puntanen `https://orcid.org/0000-0002-6776-0173`

References

Albert, A. (1973). The Gauss-Markov Theorem for regression models with possibly singular covariances. *SIAM Journal on Applied Mathematics*, **24**, 182–187.

Baksalary, J. K. (2004). An elementary development of the equation characterizing best linear unbiased estimators. *Linear Algebra and its Applications*, **388**, 3–6.

Baksalary, J. K. and Kala, R. (1981). Linear transformations preserving best linear unbiased estimators in a general Gauss–Markoff model. *Annals of Statistics*, **9**, 913–916.

Baksalary, J. K. and Kala, R. (1986). Linear sufficiency with respect to a given vector of parametric functions. *Journal of Statistical Planning and Inference*, **14**, 331–338.

Baksalary, J. K. and Mathew, T. (1986). Linear sufficiency and completeness in an incorrectly specified general Gauss–Markov model. *Sankhyā Series A*, **48**, 169–180.

Baksalary, J. K. and Mathew, T. (1990). Rank invariance criterion and its application to the unified theory of least squares. *Linear Algebra and its Applications*, **127**, 393–401.

Christensen, R. (2020). *Plane Answers to Complex Questions: The Theory of Linear Models*. Fifth Edition, Springer, New York.

Drygas, H. (1970). *The Coordinate-Free Approach to Gauss–Markov Estimation*. Springer, Berlin.

Groß, J. (2004). The general Gauss–Markov model with possibly singular dispersion matrix. *Statistical Papers*, **45**, 311–336.

Groß, J. and Puntanen, S. (2000). Estimation under a general partitioned linear model. *Linear Algebra and its Applications*, **321**, 131–144.

Haslett, S. J., Isotalo, J., Liu, Y., and Puntanen, S. (2014). Equalities between OLSE, BLUE and BLUP in the linear model. *Statistical Papers*, **55**, 543–561.

Haslett, S. J., Isotalo, J., Kala, R., Markiewicz, A., and Puntanen, S. (2021). A review of the linear sufficiency and linear prediction sufficiency in the linear model with new observations. *Multivariate, Multilinear and Mixed Linear Models*. (K. Filipiak, A. Markiewicz, D. von Rosen, eds.) Springer, Cham, pp. 265–318.

Haslett, S. J., Isotalo, J., Markiewicz, A., and Puntanen, S. (2023a). Upper bounds for the Euclidean distances between the BLUEs under the partitioned linear fixed model and the corresponding mixed model. *Applied Linear Algebra, Probability and Statistics: A Volume in Honour of C.R. Rao and Arbind K. Lal*. (R.B. Bapat, K.M. Prasad, S.J. Kirkland, S.K. Neogy, S. Pati, S. Puntanen, eds.) Springer Singapore, Indian Statistical Institute Series. Chapter 3, pp. 27–43.

Haslett, S. J., Isotalo, J., Markiewicz, A., and Puntanen, S. (2023b). Permissible covariance structures for simultaneous retention of BLUEs in small and big linear models. *Applied Linear Algebra, Probability and Statistics: A Volume in Honour of C.R. Rao and Arbind K. Lal*. (R.B. Bapat, K.M. Prasad, S.J. Kirkland, S.K. Neogy, S. Pati, S. Puntanen, eds.) Springer Singapore, Indian Statistical Institute Series. Chapter 11, pp. 197–213.

Haslett, S. J., Isotalo, J., Markiewicz, A., and Puntanen, S. (2023c). Further remarks on permissible covariance structures for simultaneous retention of BLUEs in linear models. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, **27**, 101–112.

Haslett, S. J., Isotalo, J., and Puntanen, S. (2021). Equalities between the BLUEs and BLUPs under the partitioned linear fixed model and the corresponding mixed model. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, **25**, 239–257.

Haslett, S. J., Markiewicz, A., and Puntanen, S. (2020). Properties of BLUEs and BLUPs in full vs. small linear models with new observations. *Recent Developments in Multivariate and Random Matrix Analysis: Festschrift in Honour of Dietrich von Rosen*. (T. Holgersson, M. Singull, eds.) Springer, Cham, 123–146.

Haslett, S. J., Markiewicz, A., and Puntanen, S. (2022a). Properties of the matrix V + XTX' in linear statistical models. *Gujarat Journal of Statistical and Data Science* (formerly *Gujarat Statistical Review*), **38**, 107–131.

Haslett, S. J., Markiewicz, A., and Puntanen, S. (2023). Properties of BLUEs in full versus small linear models. *Communications in Statistics: Theory and Methods*, **52**, 7684–7698.

Haslett, S. J. and Puntanen, S. (2010a). Effect of adding regressors on the equality of the Best Linear Unbiased Estimators (BLUEs) under two linear models. *Journal of Statistical Planning and Inference*, **140**, 104–110.

Haslett, S. J. and Puntanen, S. (2010b). A note on the equality of the BLUPs for new observations under two linear models. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, **14**, 27–33.

Haslett, S. J. and Puntanen, S. (2011). On the equality of the BLUPs under two linear mixed models. *Metrika*, **74**, 381–395.

Haslett, S. J. and Puntanen, S. (2023). Equality of BLUEs for full, small, and intermediate linear models under covariance change, with links to data confidentiality and encryption. *Applied Linear Algebra, Probability and Statistics: A Volume in Honour of C.R. Rao and Arbind K. Lal.* (R.B. Bapat, K.M. Prasad, S.J. Kirkland, S.K. Neogy, S. Pati, S. Puntanen, eds.) Springer Singapore, Indian Statistical Institute Series. Chapter 14, pp. 237–291.

Haslett, S. J., Puntanen, S., and Arendacká, B. (2015). The link between the mixed and fixed linear models revisited. *Statistical Papers*, **56**, 849–861.

Hauke, J., Markiewicz, A., and Puntanen, S. (2013). Revisiting the BLUE in a linear model via proper eigenvectors. *Combinatorial Matrix Theory and Generalized Inverses of Matrices*. (R.B. Bapat, S.J. Kirkland, K.M. Prasad, S. Puntanen, eds.) Springer, India, 73–83.

Henderson, C. R. (1950). Estimation of genetic parameters. *The Annals of Mathematical Statistics*, **21**, 309–310.

Henderson, C. R. (1963). Selection index and expected genetic advance. *Statistical Genetics and Plant Breeding*, National Academy of Sciences – National Research Council Publication No. 982, 141–163.

Isotalo, J., Möls, M., and Puntanen, S. (2006). Invariance of the BLUE under the linear fixed and mixed effects models. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, **10**, 69–76.

Isotalo, J. and Puntanen, S. (2006). Linear prediction sufficiency for new observations in the general Gauss–Markov model. *Communications in Statistics: Theory and Methods*, **35**, 1011–1023.

Isotalo, J., Puntanen, S., and Styan, G. P. H. (2008). A useful matrix decomposition and its statistical applications in linear regression. *Communications in Statistics: Theory and Methods*, **37**, 1436–1457.

Kala, R. (1981). Projectors and linear estimation in general linear models. *Communications in Statistics: Theory and Methods*, **10**, 849–873.

Markiewicz, A., Puntanen, S., and Styan, G. P. H. (2010). A note on the interpretation of the equality of OLSE and BLUE. *Pakistan Journal of Statistics*, **26**, 127–134.

Markiewicz, A., Puntanen, S., and Styan, G. P. H. (2021). The legend of the equality of OLSE and BLUE: highlighted by C.R. Rao in 1967. *Methodology and Applications of Statistics: A Volume in Honor of C.R. Rao on the Occasion of his 100th Birthday.* (B.C. Arnold, N. Balakrishnan, C.A. Coelho, eds.) Springer, Cham. Pages 51–76.

Mathew, T. and Bhimasankaram, P. (1983). On the robustness of LRT in singular linear models. *Sankhyā Series A*, **45**, 301–312.

Mitra, S .K. and Moore, B. J. (1973). Gauss–Markov estimation with an incorrect dispersion matrix. *Sankhyā Series A*, **35**, 139–152.

Puntanen, S., Styan, G. P. H., and Werner, H. J. (2000). Two matrix-based proofs that the linear estimator Gy is the best linear unbiased estimator. *Journal of Statistical Planning and Inference*, **88**, 173–179.

Puntanen, S., Styan, G. P. H., and Isotalo, J. (2011). *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty.* Springer, Heidelberg.

Rao, C. R. (1967). Least squares theory using an estimated dispersion matrix and its application to measurement of signals. *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1.* (L.M. Le Cam, J. Neyman, eds.) Univ. of Calif. Press, Berkeley, pp. 355–372.

Rao, C. R. (1968). A note on a previous lemma in the theory of least squares and some further results. *Sankhyā Series A*, **30**, 259–266.

Rao, C. R. (1971). Unified theory of linear estimation. *Sankhyā Series A*, **33**, 371–394. [Corrigenda (1972): **34**, p. 194 and p. 477]

Rao, C. R. (1973). Representations of best linear estimators in the Gauss–Markoff model with a singular dispersion matrix. *Journal of Multivariate Analysis*, **3**, 276–292.

Rao, C. R. and Mitra, S. K. (1971). *Generalized Inverse of Matrices and Its Applications.* Wiley, New York.

Zmyślony, R. (1980). A characterization of best linear unbiased estimators in the general linear model. *Mathematical Statistics and Probability Theory: In Proceedings Sixth International Conference (Wisła, Poland, 1978).* (W. Klonecki, A. Kozek, J. Rosiński, eds.) Springer, New York, pp. 365–373.

Zyskind, G. (1967). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. *The Annals of Mathematical Statistics*, **38**, 1092–1109.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 119–152
http://www.ssca.org.in/journal

Confidence Ellipsoids of a Multivariate Normal Mean Vector Based on Noise Perturbed and Synthetic Data with Applications

Biswajit Basak[1], Yehenew G. Kifle[2] and Bimal K. Sinha[2,3]
[1]*Department of Statistics, Sister Nivedita University, Kolkata 700156, India*
[2]*Department of Mathematics and Statistics*
University of Maryland Baltimore County, Maryland 21250, USA
[3]*Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Rd, Suitland-Silver Hill, MD 20746, USA*

Received: 19 April 2024; Revised: 09 June 2024; Accepted: 11 June 2024

Abstract

In this paper we address the problem of constructing a confidence ellipsoid of a multivariate normal mean vector based on a random sample from it. The central issue at hand is the sensitivity of the original data and hence the data cannot be directly used/analyzed. We consider a few perturbations of the original data, namely, noise addition and creation of synthetic data based on the plug-in sampling (PIS) method and the posterior predictive sampling (PPS) method. We review some theoretical results under PIS and PPS which are already available based on both frequentist and Bayesian analysis (Klein and Sinha, 2015, 2016; Guin *et al.*, 2023) and derive the necessary results under noise addition. A theoretical comparison of all the methods based on expected volumes of the confidence ellipsoids is provided. A measure of privacy protection (PP) is discussed and its formulas under PIS, PPS and noise addition are derived and the different methods are compared based on PP. Applications include analysis of two multivariate datasets. The first dataset, with $p = 2$, is obtained from the latest Annual Social and Economic Supplement (ASEC) conducted by the US Census Bureau in 2023. The second dataset, with $p = 3$, pertains to renal variables obtained from the book by Harris and Boyd (1995). Using a synthetic version of the original data generated through PIS and PPS methods and also the noise added data, we produce and display the confidence ellipsoids for the unknown mean vector under various scenarios. Finally, the privacy protection measure is evaluated for various methods and different features.

Key words: Bayesian credible Set; Confidence ellipsoid; Noise addition; Plug-in sampling; Posterior predictive sampling; Privacy protection.

AMS Subject Classifications: 62K05, 05B05

Corresponding Author: Bimal K. Sinha
Email: sinha@umbc.edu

1. Introduction

Statistical data analysis under privacy protection has been the focus of statistical research at many government agencies where the charge is to collect public data or information on many aspects of their lives and then analyze and disseminate the information for public use, policy decisions, and further research by other interested parties. Often the information provided by the public as in the decennial census in the USA contain some sensitive features and it is the responsibility of the data collecting agency to ensure that information related to these features are not compromised, properly hidden, and hard to retrieve from subsequent data analysis and released tables. Statistical research dealing with this task falls into the category of Statistical Disclosure Control (SDC) Methods. Fortunately, many novel methods of SDC have been developed and used over the years, notably noise addition/multiplication, swapping, and synthetic data creation (Drechsler and Reiter (2010); Drechsler (2011); Kinney *et al.* (2014); Kinney *et al.* (2011); Kinney *et al.* (2014); Lin and Wise (2012); Little *et al.* (1993); Meng (1994); Klein *et al.* (2014); Klein and Sinha (2013a); Klein and Sinha (2015); Klein and Sinha (2016); Raghunathan *et al.* (2003); Reiter (2003); Reiter (2004); Reiter (2005a); Reiter (2005b); Reiter (2005c); Reiter and Kinney (2012); Reiter and Mitra (2009); Reiter and Raghunathan (2007); Rubin (1987); Rubin (1993); Rubin (1996); Nayak *et al.* (2011); Sinha *et al.* (2011);Klein and Sinha (2013b)). There are three distinct parts in this process: how to perturb or distort the sensitive parts of the information collected, how to carry out proper statistical analysis based on the perturbed data so as to draw valid inference about some population features (like proportions, means, variances, correlation) and a study of the extent to which privacy has been preserved!

The focus of this paper is on multivariate data analysis in the context of sensitive data collected on p continuous features from a random sample of n units of a population. We assume that data follows a multivariate normal (MVN) model with the mean vector $\boldsymbol{\mu}$ and dispersion matrix $\boldsymbol{\Sigma}$, both unknown, and primarily address the problem of constructing confidence sets (CS) for $\boldsymbol{\mu}$ based on suitable perturbations of the original data. Three methods of SDC are considered: noise addition, synthetic data analysis based on Plug-in Sampling (PIS) scheme and synthetic data analysis based on Posterior Predictive Sampling (PPS) scheme. In each case we clearly spell out 1) how to create artificial data, 2) how to analyze it so as to produce a valid CS for $\boldsymbol{\mu}$, and 3) to what extent privacy is protected based on a suitable privacy protection (PP) measure. We should point out that the above methods are widely used in the literature and we have freely used some results which are already available and derived necessary additional results for a complete analysis of the MVN data.

The organization of the paper is as follows. In Section 2 we discuss valid inference based on noise added data, including proper analysis leading to a CS for $\boldsymbol{\mu}$. Section 3 is devoted to valid analysis of synthetic data under PIS and Section 4 to valid analysis under PPS. Both Sections 3 and 4 reside in the frequentist paradigm. We consider Bayesian analysis of PIS and PPS data in Section 5. A comparison of the suggested methods based on the expected volumes is done in Section 6. In Section 7 a measure of privacy protection (PP) suitable for multivariate data is given and explicit expressions of this measure for all the methods are derived. A comparison of the suggested artificial data generation methods based on PP is also given. It should be noted that evaluation of PP depends only on the way the original data are perturbed and not on subsequent data analysis methods. Finally, in Section 8, we apply all the proposed methods in the analysis of two multivariate datasets:

the first, with $p = 2$, is obtained from the US Census Bureau, and the second dataset, encompassing renal variables with $p = 3$, is obtained from the book by Harris and Boyd (1995), providing a comprehensive analysis of both datasets.

Throughout this paper we assume the original data $\boldsymbol{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ are *iid* as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $n > p$ and $\boldsymbol{\Sigma}$ is a positive definite matrix. Define $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ (sample mean), $\boldsymbol{W}_{\mathbf{x}} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ (sample Wishart matrix) and $\hat{\boldsymbol{\Sigma}} = \frac{\boldsymbol{W}_{\mathbf{x}}}{n-1}$. Based on the original data, $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ are jointly sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Define $T^2_{\mathbf{x}} = n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{W}^{-1}_{\mathbf{x}}(\bar{\mathbf{x}} - \boldsymbol{\mu})$, then $\left(\frac{n-p}{p}\right) T^2_{\mathbf{x}} \sim F_{p,n-p}$. A $(1-\gamma)$ level confidence ellipsoid (CE) for $\boldsymbol{\mu}$ based on the original data $\boldsymbol{X}$ will be

$$\Delta(\boldsymbol{\mu}) = \left\{ \boldsymbol{\mu} : T^2_{\mathbf{x}} \leq \left(\frac{p}{n-p}\right) F_{p,n-p;\gamma} \right\}, \tag{1}$$

where $F_{p,n-p;\gamma}$ is the $100(1-\gamma)^{\text{th}}$ percentile of an F-distribution with $(p, n-p)$ degrees of freedoms. The *observed volume* and the *expected volume* of the above CE will be

$$V_{\boldsymbol{\mu}}(\boldsymbol{X}) = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(\frac{p}{n-p} F_{p,n-p;\gamma}\right)^{p/2} |\boldsymbol{W}_{\mathbf{x}}|^{\frac{1}{2}} \tag{2}$$

$$E\left[V_{\boldsymbol{\mu}}(\boldsymbol{X})\right] = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(\frac{p}{n-p} F_{p,n-p;\gamma}\right)^{p/2} \mathscr{C}_{n,p} |\boldsymbol{\Sigma}|^{\frac{1}{2}}, \tag{3}$$

where $E[|\boldsymbol{W}_{\mathbf{x}}|^{\frac{1}{2}}] = \mathscr{C}_{n,p}|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ and $\mathscr{C}_{n,p} = \prod_{i=1}^p \left[2^{\frac{1}{2}} \frac{\Gamma\left(\frac{n-i+1}{2}\right)}{\Gamma\left(\frac{n-i}{2}\right)}\right]$.

2. Inference based on noise added data

In this section our objective is to propose an inferential method of finding a suitable confidence set for the unknown $\boldsymbol{\mu}$ based on the noise added data. The original data $\boldsymbol{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ are assumed to be independent and identically distributed (*iid*) as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $n > p$. Based on these data, one can define the summary statistics $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ (sample mean) and $\boldsymbol{W}_{\mathbf{x}} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ (sample Wishart matrix). Sometimes the unit level/micro data are available and sometimes they are not. We have encountered these two cases in the following subsections.

2.1. Case 1: Unit level data available

When unit level data are available, they can be perturbed by adding some random noise $\mathbf{e}_i \sim N_p(\mathbf{0}, \boldsymbol{R})$, *iid* for $i = 1, \cdots, n$, to the i^{th} level - resulting in $\mathbf{u}_i = \mathbf{x}_i + \mathbf{e}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{R}), i = 1, 2, \ldots, n$, where $\boldsymbol{R}$ is a known positive definite noise dispersion matrix. Our objective is to propose an inferential method of finding a suitable confidence set for the unknown $\boldsymbol{\mu}$ based on the noise added data $\boldsymbol{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$. Define $\bar{\mathbf{u}} = \frac{1}{n}\sum_{i=1}^n \mathbf{u}_i$ and $\boldsymbol{W}_{\mathbf{u}} = \sum_{i=1}^n (\mathbf{u}_i - \bar{\mathbf{u}})(\mathbf{u}_i - \bar{\mathbf{u}})'$. It is very easy to verify that, based on the noise added data $\boldsymbol{U}$, $(\bar{\mathbf{u}}, \boldsymbol{W}_{\mathbf{u}})$ are jointly sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Clearly $\bar{\mathbf{u}} \sim N_p\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}+\boldsymbol{R}}{n}\right)$, independently of $\boldsymbol{W}_{\mathbf{u}} \sim \text{Wishart}_p(\boldsymbol{\Sigma} + \boldsymbol{R}, n-1)$. We define $T^2_{\mathbf{u}} = n(\bar{\mathbf{u}} - \boldsymbol{\mu})'\boldsymbol{W}^{-1}_{\mathbf{u}}(\bar{\mathbf{u}} - \boldsymbol{\mu})$ which follows $\frac{p}{n-p}$ times an F-distribution with degrees of freedoms $(p, n-p)$. Clearly, $T^2_{\mathbf{u}}$ can be looked upon

as a pivot and can be used to find a $(1-\gamma)$ ellipsoid for $\boldsymbol{\mu}$ as given by

$$\Delta^1_{NA}(\boldsymbol{\mu}) = \left\{ \boldsymbol{\mu} : n(\boldsymbol{\mu} - \bar{\mathbf{u}})' \, \boldsymbol{W}_{\mathbf{u}}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{u}}) \leq \frac{p}{n-p} F_{p,n-p;\gamma} \right\}, \tag{4}$$

where $F_{p,n-p;\gamma}$ is the $100(1-\gamma)^{\text{th}}$ percentile of an $F_{p,n-p}$ distribution. The *volume* of the confidence ellipsoid $\Delta^1_{NA}(\boldsymbol{\mu})$ based on the noise added data $\boldsymbol{U}$ is given by

$$V_{\boldsymbol{\mu}}(\boldsymbol{U}) = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(\frac{p}{n-p} F_{p,n-p;\gamma} \right)^{p/2} |\, \boldsymbol{W}_{\mathbf{u}}|^{\frac{1}{2}}. \tag{5}$$

Note that $E(|\, \boldsymbol{W}_{\mathbf{u}}|^{\frac{1}{2}}) = \mathscr{C}_{n,p} |\boldsymbol{\Sigma} + \boldsymbol{R}|^{1/2}$ with $\mathscr{C}_{n,p} = \prod_{i=1}^{p} \left[2^{\frac{1}{2}} \frac{\Gamma\left(\frac{n-i+1}{2}\right)}{\Gamma\left(\frac{n-i}{2}\right)} \right]$, the *expected volume* is obtained as

$$E[V_{\boldsymbol{\mu}}(\boldsymbol{U})] = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(\frac{p}{n-p} F_{p,n-p;\gamma} \right)^{p/2} \mathscr{C}_{n,p} |\boldsymbol{\Sigma} + \boldsymbol{R}|^{\frac{1}{2}}. \tag{6}$$

2.2. Case 2: Unit level data not available

If unit level/micro data is not available on $\boldsymbol{X}$, but only summary statistics $\bar{\mathbf{x}}$ and $\boldsymbol{W}_{\mathbf{x}}$ are available, we define $\bar{\mathbf{u}} = \bar{\mathbf{x}} + \bar{\mathbf{e}}$, where $\bar{\mathbf{e}} \sim N_p(\mathbf{0}, \frac{\boldsymbol{R}}{n})$, independent of $\bar{\mathbf{x}}$, and $\boldsymbol{W}_{\mathbf{u}} = \boldsymbol{W}_{\mathbf{x}} + \boldsymbol{W}_r$, where $\boldsymbol{W}_r \sim \text{Wishart}_p(r, \boldsymbol{R})$ with $r \geq p$, independent of $\boldsymbol{W}$. Consequently we have $\bar{\mathbf{u}} \sim N_p(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}+\boldsymbol{R}}{n})$ and $\boldsymbol{W}_{\mathbf{u}}$ follows a distribution which is the sum of two independent Wishart distributions: $\text{Wishart}_p(n-1, \boldsymbol{\Sigma})$ and $\text{Wishart}_p(r, \boldsymbol{R})$. For the sake of simplicity, we write it as $\boldsymbol{W}_{\mathbf{u}} \sim \boldsymbol{W}_p(n-1, \boldsymbol{\Sigma}) + \boldsymbol{W}_p(r, \boldsymbol{R})$. Define $F_{\mathbf{u}} = n(\bar{\mathbf{u}} - \boldsymbol{\mu})' \boldsymbol{W}_{\mathbf{u}}^{-1} (\bar{\mathbf{u}} - \boldsymbol{\mu})$. Here, it should be noted that the distribution of $F_{\mathbf{u}}$ is not independent of the parameter $\boldsymbol{\Sigma}$ and hence can not be used as a pivot. Our goal is to find F^* which is stochastically larger than $F_{\mathbf{u}}$ and which has a distribution free from the parameter.

Consider $\mathbf{v} = \sqrt{n}(\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}}(\bar{\mathbf{u}}-\boldsymbol{\mu}) \sim N_p(\mathbf{0}, \boldsymbol{I}_p)$, that is $\sqrt{n}(\bar{\mathbf{u}}-\boldsymbol{\mu}) = (\boldsymbol{\Sigma}+\boldsymbol{R})^{\frac{1}{2}}\mathbf{v}$, and $\boldsymbol{W}^*_{\mathbf{u}} = (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}} \boldsymbol{W}_{\mathbf{u}} (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}}$, we can rewrite $F_{\mathbf{u}} = \mathbf{v}'(\boldsymbol{W}^*_{\mathbf{u}})^{-1}\mathbf{v}$. Note that, $\boldsymbol{W}^*_{\mathbf{u}} \sim \boldsymbol{W}_p(n-1, \boldsymbol{A}_1) + \boldsymbol{W}_p(r, \boldsymbol{A}_2)$, where $\boldsymbol{A}_1 = (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}} \boldsymbol{\Sigma} (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}}$ and $\boldsymbol{A}_2 = (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}} \boldsymbol{R} (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}}$ with $\boldsymbol{A}_1 + \boldsymbol{A}_2 = \boldsymbol{I}_p$.

Theorem 1: If $\mathbf{w}_1 \sim \text{Wishart}_p(n-1, \boldsymbol{I}_p)$, independently of $\mathbf{w}_2 \sim \text{Wishart}_p(r, \boldsymbol{I}_p)$, the distribution of $F^* = Max \left\{ \mathbf{v}'\mathbf{w}_1^{-1}\mathbf{v}, \mathbf{v}'\mathbf{w}_2^{-1}\mathbf{v} \right\}$ is stochastically larger than $F_{\mathbf{u}}$ and also free from the parameter.

Proof: Suppose $\boldsymbol{S}_1 = (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}} \boldsymbol{W} (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}} \sim \text{Wishart}_p(n-1, \boldsymbol{A}_1)$, independently of $\boldsymbol{S}_2 = (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}} \boldsymbol{W}_r (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}} \sim \text{Wishart}_p(r, \boldsymbol{A}_2)$, and $\boldsymbol{S} = \boldsymbol{S}_1 + \boldsymbol{S}_2$, then $F_{\mathbf{u}}$ can be written as $F_{\mathbf{u}} = \mathbf{v}' \boldsymbol{S}^{-1} \mathbf{v}$. We first note that,

$$F_{\mathbf{u}} = \mathbf{v}' \boldsymbol{S}^{-1} \mathbf{v} = \underset{\mathbf{l}:\mathbf{l}'\mathbf{l}=1}{Max} \left\{ \frac{(\mathbf{l}'\mathbf{v})^2}{\mathbf{l}'\boldsymbol{S}\mathbf{l}} \right\}. \tag{*}$$

Let $\mathbf{w}_1 \sim \text{Wishart}_p(n-1, \boldsymbol{I}_p)$, independently of $\mathbf{w}_2 \sim \text{Wishart}_p(r, \boldsymbol{I}_p)$. For any $\mathbf{l}$ such that $\mathbf{l}'\mathbf{l} = 1$, $\mathbf{l}'\mathbf{w}_1\mathbf{l} \sim \chi^2_{n-1}$, and $\mathbf{l}'\mathbf{w}_2\mathbf{l} \sim \chi^2_r$. Again $\frac{\mathbf{l}'\boldsymbol{S}_1\mathbf{l}}{\mathbf{l}'\boldsymbol{A}_1\mathbf{l}} \sim \chi^2_{n-1}$ and $\frac{\mathbf{l}'\boldsymbol{S}_1\mathbf{l}}{\mathbf{l}'\boldsymbol{A}_2\mathbf{l}} \sim \chi^2_r$, which implies

$$\begin{aligned} \mathbf{l}'\boldsymbol{S}_1\mathbf{l} &\stackrel{d}{=} (\mathbf{l}'\boldsymbol{A}_1\mathbf{l})(\mathbf{l}'\mathbf{w}_1\mathbf{l}), \\ \text{and } \mathbf{l}'\boldsymbol{S}_2\mathbf{l} &\stackrel{d}{=} (\mathbf{l}'\boldsymbol{A}_2\mathbf{l})(\mathbf{l}'\mathbf{w}_2\mathbf{l}). \end{aligned}$$

Hence

$$\begin{aligned} \mathbf{l}'\boldsymbol{S}\mathbf{l} \;= \mathbf{l}'(\boldsymbol{S}_1+\boldsymbol{S}_2)\mathbf{l} &\stackrel{d}{=} (\mathbf{l}'\boldsymbol{A}_1\mathbf{l})(\mathbf{l}'\mathbf{w}_1\mathbf{l}) + (\mathbf{l}'\boldsymbol{A}_2\mathbf{l})(\mathbf{l}'\mathbf{w}_2\mathbf{l}) \\ &\underset{st}{\geq} (\mathbf{l}'\boldsymbol{A}_1\mathbf{l} + \mathbf{l}'\boldsymbol{A}_2\mathbf{l}) Min\left\{\mathbf{l}'\mathbf{w}_1\mathbf{l}, \mathbf{l}'\mathbf{w}_2\mathbf{l}\right\} \\ &= (\mathbf{l}'\mathbf{l}) Min\left\{\mathbf{l}'\mathbf{w}_1\mathbf{l}, \mathbf{l}'\mathbf{w}_2\mathbf{l}\right\}, \quad [\text{Since, } \boldsymbol{A}_1 + \boldsymbol{A}_2 = \boldsymbol{I}_p] \\ &= Min\left\{\mathbf{l}'\mathbf{w}_1\mathbf{l}, \mathbf{l}'\mathbf{w}_2\mathbf{l}\right\}, \qquad [\text{Since, } \mathbf{l}'\mathbf{l} = 1] \end{aligned}$$

Thus we have

$$\begin{aligned} \frac{(\mathbf{l}'\mathbf{v})^2}{\mathbf{l}'\boldsymbol{S}\mathbf{l}} &\underset{st}{\leq} \frac{(\mathbf{l}'\mathbf{v})^2}{Min\left\{\mathbf{l}'\mathbf{w}_1\mathbf{l}, \mathbf{l}'\mathbf{w}_2\mathbf{l}\right\}} \\ &\stackrel{d}{=} Max\left\{\frac{(\mathbf{l}'\mathbf{v})^2}{\mathbf{l}'\mathbf{w}_1\mathbf{l}}, \frac{(\mathbf{l}'\mathbf{v})^2}{\mathbf{l}'\mathbf{w}_2\mathbf{l}}\right\}. \end{aligned}$$

From (*),

$$\begin{aligned} F_u &= \underset{\mathbf{l}:\mathbf{l}'\mathbf{l}=1}{Max}\left\{\frac{(\mathbf{l}'\mathbf{v})^2}{\mathbf{l}'\boldsymbol{S}\mathbf{l}}\right\} \\ &\underset{st}{\leq} \underset{\mathbf{l}:\mathbf{l}'\mathbf{l}=1}{Max}\left\{Max\left\{\frac{(\mathbf{l}'\mathbf{v})^2}{\mathbf{l}'\mathbf{w}_1\mathbf{l}}, \frac{(\mathbf{l}'\mathbf{v})^2}{\mathbf{l}'\mathbf{w}_2\mathbf{l}}\right\}\right\} \\ &\stackrel{d}{=} Max\left\{\mathbf{v}'\mathbf{w}_1^{-1}\mathbf{v}, \mathbf{v}'\mathbf{w}_2^{-1}\mathbf{v}\right\} \\ &= F^* \end{aligned}$$

Clearly the distribution of F^* is free from $\boldsymbol{\Sigma}$, as all of $\mathbf{v}$, $\mathbf{w}_1$ and $\mathbf{w}_2$ are having distributions free from $\boldsymbol{\Sigma}$.

[Note: Here we have used the symbols $\stackrel{d}{=}$, $\underset{st}{\leq}$ and $\underset{st}{\geq}$, which stands for identically distributed, stochastically smaller and stochastically larger respectively]. □

We determine $F^*_{n,p,r,\gamma}$ such that $P[F^* \leq F^*_{n,p,r,\gamma}] = 1-\gamma$, which implies $P[F_{\mathbf{u}} \leq F^*_{n,p,r,\gamma}] \geq P[F^* \leq F^*_{n,p,r,\gamma}] = 1-\gamma$. Therefore a confidence ellipsoid for $\boldsymbol{\mu}$ with confidence level at least $(1-\gamma)$ is given by

$$\Delta^2_{NA}(\boldsymbol{\mu}) = \left\{\boldsymbol{\mu} : n(\boldsymbol{\mu} - \bar{\mathbf{u}})' \boldsymbol{W}_{\mathbf{u}}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{u}}) \leq F^*_{n,p,r,\gamma}\right\}. \tag{7}$$

The *volume* of the confidence ellipsoid $\Delta^2_{NA}(\boldsymbol{\mu})$ is given by

$$V^*_{\boldsymbol{\mu}} = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}\left(F^*_{n,p,r,\gamma}\right)^{p/2} |\,\boldsymbol{W}_{\mathbf{u}}|^{\frac{1}{2}}. \tag{8}$$

The *expected volume* can be computed by evaluating $E\left[|\boldsymbol{W}_{\mathbf{u}}|^{\frac{1}{2}}\right]$. Recall $\boldsymbol{W}_{\mathbf{u}} = \boldsymbol{W}_{\mathbf{x}} + \boldsymbol{W}_r$, we can use the result, $|\boldsymbol{W}_{\mathbf{x}} + \boldsymbol{W}_r|^{\frac{1}{2}} > Max\left\{|\boldsymbol{W}_{\mathbf{x}}|^{\frac{1}{2}}, |\boldsymbol{W}_r|^{\frac{1}{2}}\right\}$ with probability 1, resulting in $E\left[|\boldsymbol{W}_{\mathbf{x}} + \boldsymbol{W}_r|^{\frac{1}{2}}\right] > E\left[Max\left\{|\boldsymbol{W}_{\mathbf{x}}|^{\frac{1}{2}}, |\boldsymbol{W}_r|^{\frac{1}{2}}\right\}\right] \geq Max\left\{E\left[|\boldsymbol{W}_{\mathbf{x}}|^{\frac{1}{2}}\right], E\left[|\boldsymbol{W}_r|^{\frac{1}{2}}\right]\right\}$. Therefore a lower bound to the *expected volume* will be

$$\begin{aligned} E[V^*_{\boldsymbol{\mu}}] &\geq \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}\left(F^*_{n,p,r,\gamma}\right)^{p/2} Max\left\{\mathscr{C}_{n,p}|\boldsymbol{\Sigma}|^{\frac{1}{2}}, \mathscr{C}_{r+1,p}|\boldsymbol{R}|^{\frac{1}{2}}\right\} \\ &\approx \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}\left(F^*_{n,p,r,\gamma}\right)^{p/2} \mathscr{C}_{n,p}|\boldsymbol{\Sigma}|^{\frac{1}{2}}. \qquad (9) \\ &\quad \text{[Assuming } |\boldsymbol{R}| \text{ to be significantly small.]} \end{aligned}$$

Remark 1: We can do a direct comparison of the expected volume in (6) when unit level data are available and the lower bound of the expected volume in (9) when unit level data are not available in situations when R is *small*. This essentially boils down to a comparison of $[p/(n-p)]F_{p,n-p;\gamma}$ and $F^*_{n,p,r,\gamma}$. However, from the definition of F^* it follows that any percentile of F^* is larger than the corresponding percentile of $\mathbf{v}'\mathbf{w}_1^{-1}\mathbf{v}$. Since the latter percentile is $[p/(n-p)]F_{p,n-p;\gamma}$, it readily follows that $F^*_{n,p,r,\gamma}$ is larger than $[p/(n-p)]F_{p,n-p;\gamma}$, regardless of r. In other words, even the lower bound for the expected volume given in (9) is larger than the exact expected volume in (6), whatever be the df r. Table 1 shows a direct comparison of these two cut-off points.

Table 1: The first table presents $F^*_{n,p,r,\gamma}$ cut-off points for various combinations of n, p and r, while the second table displays the $[p/(n-p)]F_{p,n-p;\gamma}$ cut-off points across different values of n and p, with $\gamma = 0.05$. significance level.

n	r=10			r=15			r=20			r=100		
	p=2	p=3	p=4	p=2	p=3	p=4	p=2	p=3	p=4	p=2	p=3	p=4
25	0.921	1.504	2.379	0.534	0.797	1.070	0.394	0.568	0.739	0.307	0.419	0.539
50	0.969	1.518	2.402	0.532	0.795	1.092	0.382	0.523	0.693	0.134	0.179	0.219
100	0.976	1.531	2.437	0.544	0.820	1.097	0.363	0.522	0.679	0.069	0.090	0.109

$[p/(n-p)]F_{p,n-p;\gamma}$ cut-off points			
n	p=2	p=3	p=4
25	0.298	0.416	0.541
50	0.133	0.179	0.224
100	0.063	0.083	0.103

3. Analysis of synthetic data under plug-in sampling

In this section we briefly review the method of analysing synthetic data obtained under plug-in sampling method, which are derived by (Klein and Sinha, 2016). The main objective here is to obtain a confidence ellipsoid for $\boldsymbol{\mu}$, based on the synthetic data, for a given confidence level.

Under plug-in sampling method, singly imputed synthetic data, denoted by $\boldsymbol{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$, are obtained by drawing *iid* observations from $N_p(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$. Based on these

synthetic data, $\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i$ and $\boldsymbol{W}_{\mathbf{y}} = \sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'$ are jointly sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (See (Klein and Sinha, 2016)). Clearly, given the original data $\boldsymbol{X}$, $\bar{\mathbf{y}} \sim N_p(\bar{\mathbf{x}}, n^{-1}\hat{\boldsymbol{\Sigma}})$ independently of $\boldsymbol{W}_{\mathbf{y}} \sim \text{Wishart}_p(\hat{\boldsymbol{\Sigma}}, n-1)$. The joint pdf (unconditional) of $(\bar{\mathbf{y}}, \boldsymbol{W}_{\mathbf{y}})$ is given by

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\bar{\mathbf{y}}, \boldsymbol{W}_{\mathbf{y}}) \propto \int_{\hat{\boldsymbol{\Sigma}} \in S_n^{++}} \frac{|\boldsymbol{W}_{\mathbf{y}}|^{\frac{n-p-2}{2}} |\boldsymbol{\Sigma} + \hat{\boldsymbol{\Sigma}}|^{-\frac{1}{2}}}{|\boldsymbol{\Sigma}|^{\frac{n-1}{2}} |\hat{\boldsymbol{\Sigma}}|^{\frac{p+1}{2}}} e^{-\frac{1}{2}\left[n(\bar{\mathbf{y}}-\boldsymbol{\mu})'\left(\boldsymbol{\Sigma}+\hat{\boldsymbol{\Sigma}}\right)^{-1}(\bar{\mathbf{y}}-\boldsymbol{\mu}) + Tr(\boldsymbol{W}_{\mathbf{y}}\hat{\boldsymbol{\Sigma}}^{-1}) + (n-1)Tr(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1})\right]} d\hat{\boldsymbol{\Sigma}},$$

where S_n^{++} stands for the set of $p \times p$ positive definite matrices. For the derivation of the above expression we refer to (Klein and Sinha, 2016).

Based on the synthetic data $\boldsymbol{Y}$, consider $T^2_{\mathbf{y}} = n(\bar{\mathbf{y}} - \boldsymbol{\mu})' \boldsymbol{W}_{\mathbf{y}}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu})$, which has a mixture-type distribution mentioned in the following theorem which is derived by (Klein and Sinha, 2016). The theorem also shows that $T^2_{\mathbf{y}}$ is a pivotal quantity and can be used to find a confidence ellipsoid for $\boldsymbol{\mu}$.

Theorem 2: The distribution of $T^2_{\mathbf{y}} = n(\bar{\mathbf{y}} - \boldsymbol{\mu})' \boldsymbol{W}_{\mathbf{y}}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu})$ has the representation: $T^2_{\mathbf{y}} = T_{\mathbf{y}1} \times T_{\mathbf{y}2}$ where $T_{\mathbf{y}1} \sim \frac{1}{\chi^2_{n-p}}$, independent of $T_{\mathbf{y}2}$, and the conditional distribution of $T_{\mathbf{y}2}$, given a Wishart matrix $\boldsymbol{W}^*$, is $\sum_{i=1}^{p}\lambda_i\chi^2_{1i}$ where χ^2_{1i} are independent χ^2 variables each with 1 degree of freedom and $\lambda_1, \ldots, \lambda_p$ are the roots of $|(n-1)\boldsymbol{I}_p + (1-\lambda)\boldsymbol{W}^*| = 0$ where $\boldsymbol{W}^* \sim \text{Wishart}_p(\boldsymbol{I}_p, n-1)$.

Theorem 2 shows that $T^2_{\mathbf{y}}$ can be used as a pivotal quantity, and hence we can construct a $(1-\gamma)$ ellipsoid for $\boldsymbol{\mu}$ based on $T^2_{\mathbf{y}}$ as given by

$$\Delta_1(\boldsymbol{\mu}) = \{\boldsymbol{\mu} : n(\boldsymbol{\mu} - \bar{\mathbf{y}})' \boldsymbol{W}_{\mathbf{y}}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{y}}) \leq a_{n,p,\gamma}\} \tag{10}$$

where $a_{n,p,\gamma}$ is the $(1-\gamma)$ percentile from the distribution of $T^2_{\mathbf{y}}$. The cut-off point $a_{n,p,\gamma}$ can be obtained by simulating from the distribution of $T^2_{\mathbf{y}}$ as given below:

1. Generate $\lambda_1, \lambda_2, \ldots, \lambda_p$, the roots of $|(n-1)\boldsymbol{I}_p + (1-\lambda)\boldsymbol{W}^*| = 0$ where $\boldsymbol{W}^* \sim \text{Wishart}_p(\boldsymbol{I}_p, n-1)$.
2. Generate $T_{\mathbf{y}2} = \sum_{i=1}^{p}\lambda_i\chi^2_{1i}$ where χ^2_{1i} are independent χ^2 variables each with 1 degree of freedom.
3. Generate $T_{\mathbf{y}1} \sim \frac{1}{\chi^2_{n-p}}$, independent of $T_{\mathbf{y}2}$.
4. Finally compute $T^2_{\mathbf{y}} = T_{\mathbf{y}1} \times T_{\mathbf{y}2}$.

The *volume* of the confidence ellipsoid $\Delta_1(\boldsymbol{\mu})$ based on the synthetic data $\boldsymbol{Y}$ is given by

$$V_{\boldsymbol{\mu}}(\boldsymbol{Y}) = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}(a_{n,p,\gamma})^{p/2}|\boldsymbol{W}_{\mathbf{y}}|^{\frac{1}{2}}. \tag{11}$$

Since $E\left(|\boldsymbol{W}_{\mathbf{y}}|^{\frac{1}{2}}\right) = \frac{\mathscr{C}_{n,p}^2}{(n-1)^{p/2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ with $\mathscr{C}_{n,p} = \prod_{i=1}^{p}\left[2^{\frac{1}{2}}\frac{\Gamma\left(\frac{n-i+1}{2}\right)}{\Gamma\left(\frac{n-i}{2}\right)}\right]$, the *expected volume* is obtained as

$$E[V_{\boldsymbol{\mu}}(\boldsymbol{Y})] = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}(a_{n,p,\gamma})^{p/2}\frac{\mathscr{C}_{n,p}^2}{(n-1)^{p/2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}. \tag{12}$$

4. Analysis of synthetic data under posterior predictive sampling

Likewise in the previous section, the original data $\boldsymbol{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ is assumed to be *iid* as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $n > p$. In this section we briefly discuss a method, illustrated in ((Klein and Sinha, 2015)), to obtain confidence ellipsoid for $\boldsymbol{\mu}$ based on a synthetically generated data under posterior predictive sampling. Consider $\bar{\mathbf{x}}$ and $\boldsymbol{W}_{\mathbf{x}}$, as mentioned in the section (1), which are jointly sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Under the posterior predictive sampling method, a vague prior for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is set as $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{\alpha}{2}}$, where $n+\alpha > 2p+3$. The joint posterior distribution of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ given $\boldsymbol{X}$, can be represented as

$$\begin{aligned} \boldsymbol{\Sigma}^{-1}|\boldsymbol{X} &\sim \text{Wishart}_p(\boldsymbol{W}_{\mathbf{x}}^{-1}, n+\alpha-p-2) \\ \boldsymbol{\mu}|(\boldsymbol{\Sigma}, \boldsymbol{X}) &\sim N_p\left(\bar{\mathbf{x}}, \frac{\boldsymbol{\Sigma}}{n}\right). \end{aligned} \tag{13}$$

We draw $(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ from the above posterior and finally a random sample $\boldsymbol{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$ is drawn from $N_p(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, which constitutes the synthetic data. Based on these synthetic data $\boldsymbol{Z}$, one can easily verify that $\bar{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i$ and $\boldsymbol{W}_{\mathbf{z}} = \sum_{i=1}^{n}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})'$ are jointly sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The joint pdf of $\bar{\mathbf{z}}$ and $\boldsymbol{W}_{\mathbf{z}}$ is obtained by integrating out $\boldsymbol{\Sigma}^*$ from the joint pdf of $(\bar{\mathbf{z}}, \boldsymbol{W}_{\mathbf{z}}, \boldsymbol{\Sigma}^*)$ given by

$$\begin{aligned} f(\bar{\mathbf{z}}, \boldsymbol{W}_{\mathbf{z}}, \boldsymbol{\Sigma}^*) \propto e^{-\frac{1}{2}\left[n(\bar{\mathbf{z}}-\boldsymbol{\mu})'(\boldsymbol{\Sigma}+2\boldsymbol{\Sigma}^*)^{-1}(\bar{\mathbf{z}}-\boldsymbol{\mu})+Tr\left(\boldsymbol{W}_{\mathbf{z}}\boldsymbol{\Sigma}^{*-1}\right)\right]}|\boldsymbol{\Sigma}+2\boldsymbol{\Sigma}^*|^{-\frac{1}{2}}|\boldsymbol{\Sigma}|^{\frac{n-p+\alpha-2}{2}} \\ |\boldsymbol{\Sigma}+\boldsymbol{\Sigma}^*|^{-\frac{2n-p+\alpha-3}{2}}|\boldsymbol{\Sigma}^*|^{-\left(\frac{p+1}{2}+\alpha\right)}|\boldsymbol{W}_{\mathbf{z}}|^{\frac{n-p-2}{2}}. \end{aligned}$$

Define $T_{\mathbf{z}}^2 = n(\bar{\mathbf{z}}-\boldsymbol{\mu})'\boldsymbol{W}_{\mathbf{z}}^{-1}(\bar{\mathbf{z}}-\boldsymbol{\mu})$, then the distribution of $T_{\mathbf{z}}^2$, as mentioned in (Klein and Sinha, 2015), given in Theorem (3) below.

Theorem 3: $T_{\mathbf{z}}^2$ has the representation: $T_{\mathbf{z}}^2 = T_{\mathbf{z}1} \times T_{\mathbf{z}2}$ with $T_{\mathbf{z}1} \sim \frac{1}{\chi_{n-p}^2}$, independent of $T_{\mathbf{z}2} = \sum_{i=1}^{n}\lambda_i\chi_{1i}^2$ where χ_{1i}^2 are independent χ^2 random variables each with 1 degree of freedom and $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the roots of $|\boldsymbol{I}_p + (2-\lambda)\tilde{\boldsymbol{\Sigma}}| = 0$, and the distribution of $\tilde{\boldsymbol{\Sigma}}$ is given by

$$f(\tilde{\boldsymbol{\Sigma}}) \propto |\tilde{\boldsymbol{\Sigma}}|^{\frac{n-p-2}{2}} \times |I+\tilde{\boldsymbol{\Sigma}}|^{-\frac{2n+\alpha-p-3}{2}}.$$

From the above theorem it is clear that $T_{\mathbf{z}}^2$ can be used as a pivot and hence a $(1-\gamma)$ level confidence ellipsoid for $\boldsymbol{\mu}$ based on $T_{\mathbf{z}}^2$ is given by

$$\Delta_2(\boldsymbol{\mu}) = \{\boldsymbol{\mu} : n(\boldsymbol{\mu}-\bar{\mathbf{z}})'\boldsymbol{W}_{\mathbf{z}}^{-1}(\boldsymbol{\mu}-\bar{\mathbf{z}}) \leq b_{n,p,\alpha,\gamma}\}, \tag{14}$$

where $b_{n,p,\alpha,\gamma}$ is the $(1-\gamma)$ level cut-off point from the distribution of $T_{\mathbf{z}}^2$ and it can be obtained by simulating from the distribution of $T_{\mathbf{z}}^2$ as discussed below.

1. To generate $\tilde{\boldsymbol{\Sigma}}$ having the density $f(\tilde{\boldsymbol{\Sigma}})$ as defined in Theorem (3), one can generate $A_1 \sim \text{Wishart}_p(\boldsymbol{I}_p, n-1)$ independent of $A_2 \sim \text{Wishart}_p(\boldsymbol{I}_p, n+\alpha-p-2)$, and set $\tilde{\boldsymbol{\Sigma}} = A_1^{\frac{1}{2}} A_2^{-1} A_1^{\frac{1}{2}}$. The proof of this representation of $\tilde{\boldsymbol{\Sigma}}$ appears in the proof of Theorem 8.2.8 of (Muirhead, 1982).
2. Obtain the eigenvalues of $\tilde{\boldsymbol{\Sigma}}$ as $\delta_1, \delta_2, \ldots, \delta_p$ and take $\lambda_i = 2 + \frac{1}{\delta_i}$, $i = 1, \ldots, p$.
3. Generate $T_{\mathbf{z}2} = \sum_{i=1}^{p} \lambda_i \chi_{1i}^2$ where χ_{1i}^2 are independent χ^2 variables each with 1 degree of freedom.
4. Generate $T_{\mathbf{z}1} \sim \frac{1}{\chi_{n-p}^2}$, independent of $T_{\mathbf{z}2}$.
5. Finally compute $T_{\mathbf{z}}^2 = T_{\mathbf{z}1} \times T_{\mathbf{z}2}$.

The *volume* of the confidence ellipsoid $\Delta_2(\boldsymbol{\mu})$ based on the synthetic data $\boldsymbol{Z}$ is given by

$$V_{\boldsymbol{\mu}}(\boldsymbol{Z}) = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} (b_{n,p,\alpha,\gamma})^{p/2} |\boldsymbol{W}_{\mathbf{z}}|^{\frac{1}{2}}, \tag{15}$$

therefore the *expected volume* is

$$E[V_{\boldsymbol{\mu}}(\boldsymbol{Z})] = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} (b_{n,p,\alpha,\gamma})^{p/2} \mathscr{D}_{n,p}^2 \mathscr{E}_{n,p,\alpha} |\boldsymbol{\Sigma}|^{\frac{1}{2}}, \tag{16}$$

where $\mathscr{D}_{n,p} = \prod_{i=1}^{p} \left[\frac{\sqrt{2}\Gamma(\frac{n-i+1}{2})}{\Gamma(\frac{n-i}{2})}\right]$ and $\mathscr{E}_{n,p,\alpha} = \prod_{i=1}^{p} \left[\frac{\Gamma(\frac{n+\alpha-p-i-2}{2})}{\sqrt{2}\Gamma(\frac{n+\alpha-p-i-1}{2})}\right]$.

5. Bayesian analysis of PIS and PPS data

In this section, which is essentially based on Guin *et al.* (2023), we discuss the Bayesian credible confidence ellipsoids (BCCE) for the mean vector $\boldsymbol{\mu}$ and their (frequentist) expected volumes under PIS and PPS.

5.1. BCCE under PIS

Referring to the likelihood function of the released data $\mathbf{y}_1, \cdots, \mathbf{y}_n$ under PIS mentioned in Section 3, we now apply a diffuse prior $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{\delta}{2}}$. This results in the posterior joint distribution of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, which can be represented in the following manner:

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}} | \boldsymbol{W}_{\mathbf{y}}, \bar{\mathbf{y}} &\sim \text{Wishart}_p^{-1}(\boldsymbol{W}_{\mathbf{y}}, n-p+\delta-2) \\
\boldsymbol{\Sigma} | \hat{\boldsymbol{\Sigma}}, \bar{\mathbf{y}}, \boldsymbol{W}_{\mathbf{y}} &\sim \text{Wishart}_p^{-1}\left((n-1)\hat{\boldsymbol{\Sigma}}, n-p+\delta-2\right) \\
\boldsymbol{\mu} | \boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}, \bar{\mathbf{y}}, \boldsymbol{W}_{\mathbf{y}} &\sim N_p\left(\bar{\mathbf{y}}, \frac{1}{n}\left(\boldsymbol{\Sigma}+\hat{\boldsymbol{\Sigma}}\right)\right)
\end{aligned} \tag{17}$$

The above can be further reformulated as:

$$\begin{aligned}
\boldsymbol{W}_{\mathbf{y}}^{-1/2} \hat{\boldsymbol{\Sigma}} \boldsymbol{W}_{\mathbf{y}}^{-1/2} &\sim \text{Wishart}_p^{-1}(\boldsymbol{I}_p, n-p+\delta-2) \\
\hat{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1/2} &\sim \text{Wishart}_p^{-1}((n-1)\boldsymbol{I}_p, n-p+\delta-2) \\
\boldsymbol{\mu} \,|\, \boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}, \bar{\mathbf{y}} &\sim N_p\left(\bar{\mathbf{y}}, \frac{1}{n}\left(\boldsymbol{\Sigma}+\hat{\boldsymbol{\Sigma}}\right)\right)
\end{aligned} \tag{18}$$

which has the benefit that $\boldsymbol{W_y}^{-1/2}\hat{\boldsymbol{\Sigma}}\,\boldsymbol{W_y}^{-1/2}$ is independent of $\hat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1/2}$ and their posterior distributions are unconditional. The posterior distributions are proper as long as $n > \max\{p, 2p-\delta+1\}$. A $(1-\gamma)$ BCCE for $\boldsymbol{\mu}$ can be taken as [(Guin *et al.*, 2023)]

$$\Delta_3(\boldsymbol{\mu}) = \left\{\boldsymbol{\mu} \;:\; T^2_{\mathbf{y}} \le c_{n,p,\delta;\gamma}\right\}, \tag{19}$$

where $T^2_{\mathbf{y}} = n(\bar{\mathbf{y}}-\boldsymbol{\mu})'\,\boldsymbol{W}^{-1}_{\mathbf{y}}(\bar{\mathbf{y}}-\boldsymbol{\mu})$ and the cut-off point $c_{n,p,\delta;\gamma}$ is obtained by simulation through the following steps:

1. Generate $\boldsymbol{B} \sim \text{Wishart}_p^{-1}\left(\boldsymbol{I}_p, n-p+\delta-2\right)$.
2. Generate $\boldsymbol{A}\,|\,\boldsymbol{B} \sim \text{Wishart}_p^{-1}\left((n-1)\boldsymbol{B}, n-p+\delta-2\right)+\boldsymbol{B}$.
3. Generate $\lambda_1, \ldots, \lambda_p$, the roots of $|\boldsymbol{A}-\lambda\boldsymbol{I}_p| = 0$.
4. Generate $T^2_{\mathbf{y}}{=}\sum_{i=1}^p \lambda_i\chi^2_{1i}$ where χ^2_{1i} are independent χ^2_1 variables.

The observed and expected volumes of the above BCCE under PIS are readily obtained as

$$V^B_{\boldsymbol{\mu}}(\boldsymbol{Y}) \;=\; \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)}\left(c_{n,p,\delta;\gamma}/n\right)^{p/2}\,|\,\boldsymbol{W}_{\mathbf{y}}|^{1/2} \tag{20}$$

$$E[V^B_{\boldsymbol{\mu}}(\boldsymbol{Y})] \;=\; \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}\left(c_{n,p,\delta;\gamma}\right)^{p/2}\frac{\mathscr{C}^2_{n,p}}{(n-1)^{p/2}}\,|\boldsymbol{\Sigma}|^{1/2}, \tag{21}$$

where $\mathscr{C}_{n,p} = \prod_{i=1}^p\left[2^{1/2}\Gamma(\frac{n-i+1}{2})/\Gamma(\frac{n-i}{2})\right]$.

5.2. BCCE under PPS

Referring to the likelihood function of the released data $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ under PPS mentioned in Section 4, we now apply a diffuse prior $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{\delta}{2}}$. This results in the posterior joint distribution of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ which can be represented in the following form:

$$\begin{aligned}
\boldsymbol{\Sigma}^*\,|\,\boldsymbol{W}_{\mathbf{z}} &\sim \text{Wishart}_p^{-1}\left(\boldsymbol{W}_{\mathbf{z}}, n-2p+\delta-1+2\alpha\right)\\
\boldsymbol{\Sigma}^{*-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{*-1/2} &\sim \text{B}_{\text{p}}^{\text{II}}\left(\frac{n+\alpha-\delta-1}{2}, \frac{n-p+\delta-2}{2}\right)\\
\boldsymbol{\mu}\,|\,\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*, \bar{\mathbf{z}} &\sim N_p\left(\bar{\mathbf{z}}, \frac{1}{n}\left(\boldsymbol{\Sigma}+2\boldsymbol{\Sigma}^*\right)\right)
\end{aligned} \tag{22}$$

where $\text{B}_{\text{p}}^{\text{II}}(a,b)$ denotes the matrix variate beta type II distribution as described in (Muirhead, 1982). We can reformulate the above posterior distributions as:

$$\begin{aligned}
\boldsymbol{W}_{\mathbf{z}}^{-1/2}\boldsymbol{\Sigma}^*\boldsymbol{W}_{\mathbf{z}}^{-1/2} &\sim \text{Wishart}_p^{-1}\left(\boldsymbol{I}_p, n-2p+\delta-1+2\alpha\right)\\
\boldsymbol{\Sigma}^{*-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{*-1/2} &\sim \text{B}_{\text{p}}^{\text{II}}\left(\frac{n+\alpha-\delta-1}{2}, \frac{n-p+\delta-2}{2}\right)\\
\boldsymbol{\mu}\,|\,\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*, \bar{\mathbf{z}} &\sim N_p\left(\bar{\mathbf{z}}, \frac{1}{n}\left(\boldsymbol{\Sigma}+2\boldsymbol{\Sigma}^*\right)\right)
\end{aligned} \tag{23}$$

which has the benefit that $\boldsymbol{W}_{\mathbf{z}}^{-1/2}\boldsymbol{\Sigma}^{*}\boldsymbol{W}_{\mathbf{z}}^{-1/2}$ is independent of $\boldsymbol{\Sigma}^{*-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{*-1/2}$ and its posterior distribution is unconditional. The posterior distributions are proper as long as $n > \max\{p, 2p-\alpha+1, 3p-\delta, p-\alpha+\delta, 2p-\delta+1-2\alpha\}$.

A BCCE for $\boldsymbol{\mu}$ can be taken as [(Guin *et al.*, 2023)]

$$\Delta_4(\boldsymbol{\mu}) = \{\boldsymbol{\mu} : T_{\mathbf{z}}^2 \leq d_{n,p,\alpha,\delta,\gamma}\}, \tag{24}$$

where $T_{\mathbf{z}}^2 = n(\boldsymbol{\mu}-\bar{\mathbf{z}})'\boldsymbol{W}_{\mathbf{z}}^{-1}(\boldsymbol{\mu}-\bar{\mathbf{z}})$ and the cut-off point $d_{n,p,\alpha,\delta,\gamma}$ is obtained by simulation through the following steps:

1. Generate $\boldsymbol{B} \sim \text{Wishart}_p^{-1}(\boldsymbol{I}_p, n-2p+\delta+2\alpha-1)$ and decompose as $\boldsymbol{B} = \boldsymbol{D}\boldsymbol{D}'$.
2. Generate $\boldsymbol{V}_0 \sim \text{Wishart}_p(\boldsymbol{I}_p, n-p+\delta-2)$, $\boldsymbol{V}_1 \sim \text{Wishart}_p(\boldsymbol{I}_p, n+\alpha-\delta-1)$, $\boldsymbol{C} = \boldsymbol{V}_0^{-1}\boldsymbol{V}_1\boldsymbol{V}_0^{-1}$ and $\boldsymbol{A} = \boldsymbol{D}\boldsymbol{C}\boldsymbol{D}' + 2\boldsymbol{B}$ (Gupta and Nagar, 1999).
3. Generate $\lambda_1, \ldots, \lambda_p$, the roots of $|\boldsymbol{A} - \lambda\boldsymbol{I}_p| = 0$.
4. Generate $T_{\mathbf{z}}^2 = \sum_{i=1}^{p} \lambda_i \chi_{1i}^2$ where χ_{1i}^2 are independent χ_1^2 variables.

The observed and expected volumes of the above BCCE under PPS are readily obtained as

$$V_{\boldsymbol{\mu}}^{B}(\boldsymbol{Z}) = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(d_{n,p,\alpha,\delta;\gamma}\right)^{p/2} |\boldsymbol{W}_{\mathbf{z}}|^{1/2}, \tag{25}$$

$$E[V_{\boldsymbol{\mu}}^{B}(\boldsymbol{Z})] = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(d_{n,p,\alpha,\delta,\gamma}\right)^{p/2} \mathscr{D}_{n,p}^{2}\mathscr{E}_{n,p,\alpha} \times |\boldsymbol{\Sigma}|^{1/2}, \tag{26}$$

where $\mathscr{D}_{n,p} = \prod_{i=1}^{p}\left[\frac{\sqrt{2}\Gamma(\frac{n-i+1}{2})}{\Gamma(\frac{n-i}{2})}\right]$ and $\mathscr{E}_{n,p,\alpha} = \prod_{i=1}^{p}\left[\frac{\Gamma(\frac{n+\alpha-p-i-2}{2})}{\sqrt{2}\Gamma(\frac{n+\alpha-p-i-1}{2})}\right]$.

6. Comparison of the suggested methods based on the expected volumes

6.1. Expressions of observed and expected volumes

In this subsection, we provide a brief overview of various expressions for *observed* and *expected* volumes for $\boldsymbol{\mu}$ within the noise-added data context and also both frequentist and Bayesian frameworks under PIS and PPS methods.

The observed and expected volumes of the confidence ellipsoid for $\boldsymbol{\mu}$ (see 4), derived from noise added data $\boldsymbol{U}$, when unit level data are available, are given below.

$$V_{\boldsymbol{\mu}}(\boldsymbol{U}) = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(\frac{p}{n-p}F_{p,n-p;\gamma}\right)^{p/2} |\boldsymbol{W}_{\mathbf{u}}|^{\frac{1}{2}},$$

$$E[V_{\boldsymbol{\mu}}(\boldsymbol{U})] = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(\frac{p}{n-p}F_{p,n-p;\gamma}\right)^{p/2} \mathscr{C}_{n,p}|\boldsymbol{\Sigma}+\boldsymbol{R}|^{\frac{1}{2}}. \tag{27}$$

where $\mathscr{C}_{n,p} = \prod_{i=1}^{p}\left[2^{\frac{1}{2}}\frac{\Gamma\left(\frac{n-i+1}{2}\right)}{\Gamma\left(\frac{n-i}{2}\right)}\right]$.

If unit level data are not available, the observed volume and a lower bound to the expected volume of the confidence ellipsoid for $\boldsymbol{\mu}$ (see 7) are given by,

$$\begin{aligned} V_{\boldsymbol{\mu}}^{*} &= \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}\left(F_{n,p,r,\gamma}^{*}\right)^{p/2}|\,\boldsymbol{W}_{\mathbf{u}}|^{\frac{1}{2}}, \\ E[V_{\boldsymbol{\mu}}^{*}] &\geq \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}\left(F_{n,p,r,\gamma}^{*}\right)^{p/2} Max\left\{\mathscr{C}_{n,p}|\boldsymbol{\Sigma}|^{\frac{1}{2}}, \mathscr{C}_{r+1,p}|\boldsymbol{R}|^{\frac{1}{2}}\right\} \\ &\geq \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}\left(F_{n,p,r,\gamma}^{*}\right)^{p/2}\mathscr{C}_{n,p}|\boldsymbol{\Sigma}|^{\frac{1}{2}}. \qquad (28) \\ &\quad [\text{Assuming } |\boldsymbol{R}| \text{ to be significantly small}] \end{aligned}$$

Below are the observed and expected volumes of the confidence ellipsoid for $\boldsymbol{\mu}$ (see 10), derived from synthetic data $\boldsymbol{Y}$ using the PIS method.

$$\begin{aligned} V_{\boldsymbol{\mu}}(\boldsymbol{Y})_{PIS} &= \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}(a_{n,p,\gamma})^{p/2}|\,\boldsymbol{W}_{\mathbf{y}}|^{\frac{1}{2}}, \\ E[V_{\boldsymbol{\mu}}(\boldsymbol{Y})]_{PIS} &= \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}(a_{n,p,\gamma})^{p/2}\frac{\mathscr{C}_{n,p}^{2}}{(n-1)^{p/2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}. \qquad (29) \end{aligned}$$

Likewise, the observed and expected volumes of the confidence ellipsoid for $\boldsymbol{\mu}$ (see 14), utilizing synthetic data $\boldsymbol{Z}$ under the PPS method, are presented below.

$$\begin{aligned} V_{\boldsymbol{\mu}}(\boldsymbol{Z})_{PPS} &= \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}(b_{n,p,\alpha,\gamma})^{p/2}|\,\boldsymbol{W}_{\mathbf{z}}|^{\frac{1}{2}}, \\ E[V_{\boldsymbol{\mu}}(\boldsymbol{Z})]_{PPS} &= \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}(b_{n,p,\alpha,\gamma})^{p/2}\mathscr{D}_{n,p}^{2}\mathscr{E}_{n,p,\alpha}|\boldsymbol{\Sigma}|^{\frac{1}{2}}, \qquad (30) \end{aligned}$$

where $\mathscr{D}_{n,p} = \prod_{i=1}^{p}\left[\frac{\sqrt{2}\Gamma(\frac{n-i+1}{2})}{\Gamma(\frac{n-i}{2})}\right]$ and $\mathscr{E}_{n,p,\alpha} = \prod_{i=1}^{p}\left[\frac{\Gamma(\frac{n+\alpha-p-i-2}{2})}{\sqrt{2}\Gamma(\frac{n+\alpha-p-i-1}{2})}\right]$.

In the Bayesian framework, we provide below the observed and expected volumes of credible confidence ellipsoids for $\boldsymbol{\mu}$ within the context of synthetic data generated using the PIS method (see 19),

$$\begin{aligned} V_{\boldsymbol{\mu}}^{B}(\boldsymbol{Y})_{PIS} &= \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}\left(c_{n,p,\delta;\gamma}\right)^{p/2}|\boldsymbol{W}_{\mathbf{y}}|^{1/2}, \\ E[V_{\boldsymbol{\mu}}^{B}(\boldsymbol{Y})]_{PIS} &= \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}\left(c_{n,p,\delta;\gamma}\right)^{p/2}\frac{\mathscr{C}_{n,p}^{2}}{(n-1)^{p/2}}|\boldsymbol{\Sigma}|^{1/2}, \qquad (31) \end{aligned}$$

Table 2: Coefficients of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression under various perturbation schemes ($\gamma = 0.05$).

DIFFERENT SCHEMES	n	p		
		2	3	4
NA DATA ($r = 100$)	25	0.8619	0.9690	1.1219
	50	0.4061	0.2999	0.2231
	100	0.2390	0.1279	0.0683
PIS	25	1.7928	2.8911	5.0991
	50	0.8181	0.8554	0.9147
	100	0.3949	0.2770	0.2017
PPS ($\alpha = 4$)	25	2.2241	5.9769	14.3850
	50	1.2394	1.6682	2.3476
	100	0.5877	0.5294	0.4796
PIS BAYES ($\delta = 10$)	25	1.1445	1.6016	2.3651
	50	0.6672	0.6517	0.6477
	100	0.3572	0.2468	0.1717
PPS BAYES ($\alpha = 1$, $\delta = 10$)	25	1.457	2.4228	5.1696
	50	0.7489	0.7758	0.9171
	100	0.3773	0.2768	0.2032

and under PPS method (see 24),

$$\begin{aligned} V_{\boldsymbol{\mu}}^{B}(\boldsymbol{Z}) &= \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(d_{n,p,\alpha,\delta;\gamma}\right)^{p/2} |\boldsymbol{W}_{\mathbf{z}}|^{1/2}, \\ E[V_{\boldsymbol{\mu}}^{B}(\boldsymbol{Z})] &= \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(d_{n,p,\alpha,\delta,\gamma}\right)^{p/2} \mathscr{D}_{n,p}^{2} \mathscr{E}_{n,p,\alpha} \times |\boldsymbol{\Sigma}|^{1/2}. \end{aligned} \tag{32}$$

6.2. Comparison of expected volumes - all are proportional to $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$

Note that the expected volume expressions presented in equations 29, 30, 31 and 32 for various methods are directly proportional to $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$. The coefficient of $|\boldsymbol{\Sigma} + \boldsymbol{R}|^{\frac{1}{2}}$ in the equation 27 is the same as that of the expected volume under the original data, hence it is immaterial to consider it for the comparison. Rather we compare the expected volume under noise added data when unit level data are not available. We assume $|\boldsymbol{R}|$ to be small enough and calculate the coefficient of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in (28). Consequently, a straightforward comparison of these methods can be made by examining the coefficients of the expected volume expressions, without considering the population parameter $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$. In Table (2), we present the coefficients obtained from various perturbation schemes in different combinations of n and p values. Specifically, we used n values of 25, 50, and 100, and p values of 2, 3, and 4. The parameters $\alpha = 4$ and $\delta = 10$ remain fixed in the frequentist approach, while in the Bayesian framework we used $\alpha = 1$ and $\delta = 10$. Additionally, for data with added noise, we set $r = 100$. Throughout the analysis, we maintain a consistent value of $\gamma = 0.05$.

From Table (2), it is clear that the expected volume decreases as the sample size (n) increases under any schemes, which is quite natural. Also, in all the choices of the pair (n, p), we can

see that the expected volumes under the noise added data (taking $r = 100$) are quite smaller than the other schemes. As anticipated, in both the frequentist and Bayesian frameworks, the expected volumes under PPS exceed those under PIS.

Remark 2: Referring to Remark 1, it is obvious that in case unit level data are available, the expected volume then will be the least among all reported above. Therefore, if one were to make practical recommendations based on the expected volumes only, gathering unit level data and subsequent noise addition will certainly pay off, followed by the same noise addition mechanism based on summary data.

7. Measure of privacy protection

Disclosure risk evaluation

When the original (unit level) microdata is considered to be sensitive and thus hidden through the use of a masked version, it is natural to examine the extent to which sensitivity of a data point has been protected. A slight variation of a popular privacy measure to study the disclosure risk of a single scalar value x_i, given in Klein and Sinha (2016), can be taken as

$$P[|\hat{x}_i - x_i| < \epsilon | X] = \theta_i \tag{33}$$

where X is the entire original data, and $\hat{x}_i$ is an intruder's prediction of x_i based upon seeing the released (artificial/synthetic) data, ϵ be any small positive quantity. Naturally, a high value of the above probability indicates a low level of protection and vice versa. This privacy measure (PM) is computed based on the random mechanism producing the masked data, given the original data X.

In the multivariate case, a generalization of (33) can be taken as

$$\theta_i = P[(\hat{\mathbf{x}}_\mathbf{i} - \mathbf{x}_i)^t A (\hat{\mathbf{x}}_\mathbf{i} - \mathbf{x}_i) \leq \epsilon | X] \tag{34}$$

where A is a positive definite symmetric matrix.

Returning to our specific problem, based on the synthetic multivariate data released by the data producer, a naive intruder's best guess about $\mathbf{x}_i$, the original value for the ith unit, can be discussed under two circumstances: (a) the identities of the perturbed data are released by the data producer and $\mathbf{u}_i, \mathbf{y}_i$ or $\mathbf{z}_i$, the perturbed value of $\mathbf{x}_i$ based on NA/PIS/PPS, corresponding to the identifiable ith unit, is taken as intruder's choice, and (b) the identities of the perturbed data are lost/retained by the data producer in which case $\bar{\mathbf{u}} = [\sum_{i=1}^n \mathbf{u}_i]/n$, $\bar{\mathbf{y}} = [\sum_{i=1}^n \mathbf{y}_i]/n$ or $\bar{\mathbf{z}} = [\sum_{i=1}^n \mathbf{z}_i]/n$ is taken as intruder's choice.

There is also a 3rd case in the multivariate data context in which an intruder may be interested in a particular component, say component 1, of the p vector multivariate data. If original data $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are available, intruder's obvious choice is $\bar{x}_1 = (x_{11} + ... + x_{n1})/n$ where we write $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$, $i = 1, \cdots, n$. In the absence of the original data, we can take $\bar{u}_1$, $\bar{y}_1$ and $\bar{z}_1$ as intruder's choice under NA/PIS/PPS, respectively.

In subsection 7.1 we discuss PP under noise added data, in subsection 7.2 we discuss PP under PIS and PPS is taken up in subsection 7.3.

All the above methods discussed in Sections (7.1), (7.2) and (7.3), are from a naive intruder's

perspective. However, a smart intruder with an excellent training in statistics can think in a different way. We have added a remark to this effect at the end of this section.

7.1. Three cases under noise added (NA) data

7.1.1. Case (a)

Here we assume that the identities of the released perturbed data are known and hence the intruder's best choice of $\mathbf{x}_i$ will be $\mathbf{u}_i$. Recall that $\mathbf{u}_i = \mathbf{x}_i + \mathbf{e}_i$ where $\mathbf{e}_i \sim N_p(\mathbf{0}, \boldsymbol{R})$ is independent of the original data $\boldsymbol{X}$. Note that $\mathbf{e}_i^* = \boldsymbol{R}^{-\frac{1}{2}}\mathbf{e}_i \sim N_p(\mathbf{0}, \boldsymbol{I}_p)$. Define $\boldsymbol{B} = \boldsymbol{R}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{R}^{\frac{1}{2}}$, which is a symmetric positive definite matrix, there exists an orthogonal matrix $\boldsymbol{\Gamma}$ such that $\boldsymbol{B} = \boldsymbol{\Gamma}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}$, where $\boldsymbol{\Lambda} = Diag(\lambda_1, \cdots, \lambda_p)$ be a diagonal matrix with diagonal elements λ_i's $(i = 1, \cdots, p)$, which are the solutions to the equation $|\boldsymbol{B} - \lambda \boldsymbol{I}_p|$. Considering $\mathbf{m}_i = \boldsymbol{\Gamma}\mathbf{e}_i^* \sim N_p(\mathbf{0}, \boldsymbol{I}_p)$ we can deduce the privacy measure (θ_i) corresponding to the i^{th} unit as given by

$$\begin{aligned}
\theta_i &= P\left[(\mathbf{u}_i - \mathbf{x}_i)'\boldsymbol{A}(\mathbf{u}_i - \mathbf{x}_i) \le \epsilon | \boldsymbol{X}\right] \\
&= P\left[\mathbf{e}_i'\boldsymbol{A}\mathbf{e}_i \le \epsilon\right] \\
&= P\left[(\mathbf{e}_i^*)'\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{R}^{\frac{1}{2}}(\mathbf{e}_i^*) \le \epsilon\right] \\
&= P\left[(\mathbf{e}_i^*)'\boldsymbol{B}(\mathbf{e}_i^*) \le \epsilon\right] \\
&= P\left[(\mathbf{e}_i^*)'\boldsymbol{\Gamma}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}(\mathbf{e}_i^*) \le \epsilon\right] \\
&= P\left[\mathbf{m}_i'\boldsymbol{\Lambda}\mathbf{m}_i \le \epsilon\right] \\
&= P\left[\sum_{j=1}^{p} \lambda_j \chi_{1j}^2 \le \epsilon\right] \qquad (35)
\end{aligned}$$

In the above expression χ_{1j}^2, $j = 1, \cdots, p$ are independent central chi square variables each with 1 d.f. Note that the quantity $\theta_i = P\left[\sum_{j=1}^{p} \lambda_j \chi_{1j}^2 \le \epsilon\right] = \theta^*$ is independent of any specific unit i and hence it can be taken as a measure of overall privacy protection. The following are two special cases based on the choice of matrix $\boldsymbol{A}$.

Case 1: $\boldsymbol{A} = \boldsymbol{I}_p \Rightarrow \lambda_1, \cdots, \lambda_p$ are the solutions of $|\boldsymbol{R} - \lambda \boldsymbol{I}_p| = 0$.

Case 2: $\boldsymbol{A} = Diag(a_{11}, \cdots, a_{pp}) \Rightarrow \lambda_1, \cdots, \lambda_p$ are the solutions of $|\boldsymbol{R} - \lambda Diag(\frac{1}{a_{11}}, \cdots, \frac{1}{a_{pp}})| = 0$.

7.1.2. Case (b)

When the identities of the released perturbed data are not known, the intruder's best choice of $\mathbf{x}_i$ $(i = 1, \cdots, n)$ will be $\bar{\mathbf{u}}$. Note that $\bar{\mathbf{u}} - \mathbf{x}_i = \bar{\mathbf{e}} - (\mathbf{x}_i - \bar{\mathbf{x}})$, and for conditionally given $\boldsymbol{X}$, it follows $N_p\left(\mathbf{x}_i - \bar{\mathbf{x}}, \frac{\boldsymbol{R}}{n}\right)$. Define $\mathbf{e}_i^* = \sqrt{n}\boldsymbol{R}^{-\frac{1}{2}}(\bar{\mathbf{u}} - \mathbf{x}_i)$, which implies $\mathbf{e}_i^* | \boldsymbol{X} \sim N_p(\boldsymbol{\delta}_i, \boldsymbol{I}_p)$, where $\boldsymbol{\delta}_i = \sqrt{n}\boldsymbol{R}^{-\frac{1}{2}}(\mathbf{x}_i - \bar{\mathbf{x}})$. Here we take $\boldsymbol{B} = \frac{\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{R}^{\frac{1}{2}}}{n}$, which is a symmetric and positive definite matrix, there exists an orthogonal matrix $\boldsymbol{\Gamma}$ such that $\boldsymbol{B} = \boldsymbol{\Gamma}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}$, where $\boldsymbol{\Lambda} = Diag(\lambda_1, \cdots, \lambda_p)$ be a diagonal matrix with diagonal elements λ_j's $(j = 1, \cdots, p)$, which are the solutions to the equation $|\boldsymbol{B} - \lambda \boldsymbol{I}_p|$. Likewise **Case (a)**, we define $\mathbf{m}_i = \boldsymbol{\Gamma}\mathbf{e}_i^*$, which conditionally for given $\boldsymbol{X}$, follows $N_p(\boldsymbol{\eta}_i, \boldsymbol{I}_p)$, where $\boldsymbol{\eta}_i = \boldsymbol{\Gamma}\boldsymbol{\delta}_i$. We

proceed in a similar fashion as mentioned in **Case (a)** and deduce the privacy measure (θ_i) corresponding to the i^{th} unit as

$$\begin{aligned}
\theta_i &= P\left[(\bar{\mathbf{u}} - \mathbf{x}_i)'\boldsymbol{A}(\bar{\mathbf{u}} - \mathbf{x}_i) \leq \epsilon | \boldsymbol{X}\right] \\
&= P\left[(\mathbf{e}_i^*)'\frac{\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{R}^{\frac{1}{2}}}{n}(\mathbf{e}_i^*) \leq \epsilon | \boldsymbol{X}\right] \\
&= P\left[(\mathbf{e}_i^*)'\boldsymbol{B}(\mathbf{e}_i^*) \leq \epsilon | \boldsymbol{X}\right] \\
&= P\left[(\mathbf{e}_i^*)'\boldsymbol{\Gamma}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}(\mathbf{e}_i^*) \leq \epsilon | \boldsymbol{X}\right] \\
&= P\left[\mathbf{m}_i'\boldsymbol{\Lambda}\mathbf{m}_i \leq \epsilon | \boldsymbol{X}\right] \\
&= P\left[\sum_{j=1}^{p} \lambda_j \chi_{1j}^2(\eta_{ij}^2) \leq \epsilon\right], \qquad (36)
\end{aligned}$$

where $\chi_{1j}^2(\eta_{ij}^2)$, $j = 1, \cdots, p$ are independent noncentral chi-squared variables each with 1 d.f. and noncentrality parameters η_{ij}^2, which is the squared j^{th} component ($j = 1, \cdots, p$) of $\boldsymbol{\eta}_i$.

Unlike **Case (a)**, here θ_i depends on the specific unit i through the noncentrality parameters η_{ij}'s. We can write,

$$\theta_i \leq P[\sum_{j=1}^{p} \lambda_j \chi_{1j}^2 \leq \epsilon] = \theta^* \text{ (say)}.$$

The quantity θ^* is independent of i and can be taken as a measure of overall privacy measure. Two special choices of $\boldsymbol{A}$ as similar to **Case (a)** are given below.

Case 1: $\boldsymbol{A} = \boldsymbol{I}_p \Rightarrow \lambda_1, \cdots, \lambda_p$ are the solutions of $|\frac{\boldsymbol{R}}{n} - \lambda \boldsymbol{I}_p| = 0$.

Case 2: $\boldsymbol{A} = Diag(a_{11}, \cdots, a_{pp}) \Rightarrow \lambda_1, \cdots, \lambda_p$ are the solutions of $|\frac{\boldsymbol{R}}{n} - \lambda Diag(\frac{1}{a_{11}}, \cdots, \frac{1}{a_{pp}})| = 0$.

7.1.3. Case (c)

When an intruder is interested in a particular component, say component 1, of the p-component vector multivariate data, based on the original data $\mathbf{x}_1, \cdots, \mathbf{x}_n$, intruder's obvious choice is $\bar{x}_1 = (x_{11} + ... + x_{n1})/n$ where we write $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$, $i = 1, \cdots, n$. In the absence of the original data, the best choice would be $\bar{u}_1 = \frac{1}{n}\sum_{i=1}^{n} u_{i1}$. Clearly, $\bar{u}_1 - \bar{x}_1 = \bar{e}_1$, independently of the original data $\boldsymbol{X}$, follows $N(0, \frac{r_{11}}{n})$, where r_{11} be the $(1,1)^{\text{th}}$ element of $\boldsymbol{R}$. Therefore the privacy measure (θ) is given by

$$\begin{aligned}
\theta &= P\left[(\bar{u}_1 - \bar{x}_1)^2 \leq \epsilon | \boldsymbol{X}\right] \\
&= P\left[\bar{e}_1^2 \leq \epsilon\right] \\
&= P\left[\chi_1^2 \leq \frac{n\epsilon}{r_{11}}\right] \qquad (37)
\end{aligned}$$

From the above, it readily follows that, more the variability in a particular noise component, more the privacy protection for the same component.

7.2. Three cases under PIS

7.2.1. Case (a)

Since the identities of the released masked data are known, the intruder's choice of $\mathbf{x}_i$ can be taken as $\mathbf{y_i}$, which conditionally given $\boldsymbol{X}$, is $N_p\left(\bar{\mathbf{x}}, \frac{\boldsymbol{W_x}}{n-1}\right)$ according to the PIS scheme. It is interesting to observe that $\mathbf{y}_i$ has no bearing with the index i as far as the PIS scheme is concerned.

Before we compute the PM θ in Case (a), let us look at Case (b).

7.2.2. Case (b)

Since in the absence of the identity of the i^{th} unit $\bar{\mathbf{y}}$ seems to be the intruder's obvious choice of $\mathbf{x}_i$, to compute the PM θ, we proceed as follows. Recall that

$$\theta = P\left[(\bar{\mathbf{y}} - \bar{\mathbf{x}}_\mathbf{i})^t \boldsymbol{A} (\bar{\mathbf{y}} - \bar{\mathbf{x}}_\mathbf{i}) \leq \epsilon | \boldsymbol{X}\right]. \tag{38}$$

Note that under PIS, $\bar{\mathbf{y}}_\mathbf{1}, \cdots, \bar{\mathbf{y}}_\mathbf{n}$ are iid following $N_p\left(\bar{\mathbf{x}}, \frac{\boldsymbol{W_x}}{n-1}\right)$, implying $\bar{\mathbf{y}}|\boldsymbol{X} \sim N_p\left(\bar{\mathbf{x}}, \frac{\boldsymbol{W_x}}{n(n-1)}\right)$. Define $\boldsymbol{D} = \frac{\boldsymbol{W_x}}{n(n-1)}$, we have $(\bar{\mathbf{y}} - \bar{\mathbf{x}}_\mathbf{i})|\boldsymbol{X} \sim N_p\left((\bar{\mathbf{x}} - \bar{\mathbf{x}}_\mathbf{i}), \boldsymbol{D}\right)$, which implies $\boldsymbol{D}^{-1/2}(\bar{\mathbf{y}} - \bar{\mathbf{x}}_\mathbf{i})|\boldsymbol{X} \sim N_p\left(\boldsymbol{D}^{-1/2}(\bar{\mathbf{x}} - \mathbf{x_i}), \boldsymbol{I}_p\right)$. Write $\boldsymbol{Z} = \boldsymbol{D}^{-1/2}(\bar{\mathbf{y}} - \bar{\mathbf{x}}_\mathbf{i})$, then $\theta_i = P[\boldsymbol{Z}^t \boldsymbol{D}^{1/2} \boldsymbol{A} \boldsymbol{D}^{1/2} \boldsymbol{Z} \leq \epsilon | \boldsymbol{X}] = P[\boldsymbol{Z}^t \boldsymbol{B} \boldsymbol{Z} \leq \epsilon | \boldsymbol{X}]$, where $\boldsymbol{D}^{1/2} \boldsymbol{A} \boldsymbol{D}^{1/2} = \boldsymbol{B}$: $p \times p$ symmetric pd and $\boldsymbol{Z}|\boldsymbol{X} \sim N_p(\boldsymbol{\delta}_i, \boldsymbol{I}_p)$ with $\boldsymbol{\delta}_i = \boldsymbol{D}^{-1/2}(\bar{\mathbf{x}} - \mathbf{x_i})$.

Since $\boldsymbol{B}$ is symmetric pd, there exists an orthogonal matrix $\boldsymbol{\Gamma}$ such that $\boldsymbol{\Gamma}^t \boldsymbol{\Lambda} \boldsymbol{\Gamma} = \boldsymbol{B}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix with elements $\lambda_1, \cdots, \lambda_p$ as the characteristic roots of $\boldsymbol{B}$. Let $\boldsymbol{U} = \boldsymbol{\Gamma} \boldsymbol{Z} \sim N_p[\boldsymbol{\eta}_i, \boldsymbol{I}_p]$, where $\boldsymbol{\eta}_i = \Gamma \boldsymbol{\delta}_i$. Then

$$\begin{aligned} \theta_i &= P[\boldsymbol{Z}^t \boldsymbol{\Gamma}^t \boldsymbol{\Lambda} \boldsymbol{Z} \boldsymbol{\Gamma} \leq \epsilon] \\ &= P[\boldsymbol{U}^t \boldsymbol{\Lambda} \boldsymbol{U} \leq \epsilon] \\ &= P[\sum_{j=1}^{p} \lambda_j \chi^2_{1j}(\eta^2_{ij}) \leq \epsilon]. \end{aligned} \tag{39}$$

Note that the roots of $\boldsymbol{B}$ are the solutions of $|\boldsymbol{B} - \lambda \boldsymbol{I}_p| = 0 \iff |\boldsymbol{D}^{1/2} \boldsymbol{A} \boldsymbol{D}^{1/2} - \lambda \boldsymbol{I}_p| = 0 \iff |\boldsymbol{A} - \lambda \boldsymbol{D}^{-1} \boldsymbol{I}_p| = 0 \iff |\boldsymbol{A} - \lambda n(n-1) \mathbf{S}_\mathbf{x}^{-\mathbf{1}}| = \mathbf{0}$. Moreover, $\chi^2_{1j}(\eta^2_{ij})$, $j = 1, \cdots, p$ are independent noncentral chisquare variables each with 1 d.f. and noncentrality parameters as appear above.

For any specific unit i, θ_i above can be taken as a privacy measure. Obviously, for any i

$$\theta_i \leq P[\sum_{j=1}^{p} \lambda_j \chi^2_{1j} \leq \epsilon] = \theta^* \text{ (say)}. \tag{40}$$

We can take the absolute quantity θ^*, which is independent of any specific unit i, as a measure of overall privacy protection. In the above, $\chi^2_{11}, \cdots, \chi^2_{1p}$ are iid central chi-square each with 1 d.f. Here are two special cases:

Case 1: $\boldsymbol{A} = \boldsymbol{I}_p \Rightarrow \lambda_1, \cdots, \lambda_p$ are the solutions of $|\frac{\boldsymbol{W_x}}{n(n-1)} - \lambda \boldsymbol{I}_p| = 0$.

Case 2: $\boldsymbol{A} = Diag(a_{11}, \cdots, a_{pp}) \Rightarrow \lambda_1, \cdots, \lambda_p$ are the solutions of $|\frac{\boldsymbol{W}_{\mathbf{x}}}{n(n-1)} - \lambda Diag(\frac{1}{a_{11}}, \cdots, \frac{1}{a_{pp}})| = 0$. Note that $a_{11}, \cdots, a_{pp}$ can be interpreted as quantities representing relative importance of the p components of the vector $\mathbf{x}$.

Returning now to Case (a), we proceed as in Case (b) and it is easy to check that the PM θ_i simplifies to

$$\theta_i = P[\sum_{j=1}^{p} \lambda_j \chi_{1j}^2(\eta_{ij}^2) \leq \epsilon]. \tag{41}$$

$$\leq P[\sum_{j=1}^{p} \lambda_j \chi_{1j}^2 \leq \epsilon]. \tag{42}$$

where $\lambda_1, \cdots, \lambda_p$ are now the roots of the equation $|\boldsymbol{A} - \lambda(n-1)\boldsymbol{W}_{\mathbf{x}}^{-1}| = 0$ and $\chi_{11}^2, \cdots, \chi_{1p}^2$ are independent central chi-square variables each with 1 d.f. The two special cases of choice of $\boldsymbol{A}$ can be similarly dealt here.

7.2.3. Case (c)

From the conditional multivariate normal distribution of $\bar{\mathbf{y}}|\boldsymbol{X} \sim N_p\left(\bar{\mathbf{x}}, \frac{\boldsymbol{W}_{\mathbf{x}}}{n(n-1)}\right)$, it readily follows that the conditional univariate distribution of $\bar{y}_1$, given X, is normal with mean $\bar{x}_1$ and variance $\frac{\boldsymbol{W}_{\mathbf{x}11}}{n(n-1)} = d$ (say). Therefore the privacy measure (PM) θ, which is $P[(\bar{y}_1 - \bar{x}_1)^2 \leq \epsilon|\boldsymbol{X}]$, can be simplified as

$$\theta = P\left[(\bar{y}_1 - \bar{x}_1)^2 \leq \epsilon|\boldsymbol{X}\right] = P\left[\chi_1^2 \leq \frac{\epsilon}{d}\right]. \tag{43}$$

The implication of the PM in this case is obvious - the component having the maximum sampling variation will offer maximum privacy protection.

7.3. Three cases under PPS

7.3.1. Case (a)

Since the identities of the released masked data are known in this case, the intruder's obvious choice of $\mathbf{x_i}$ is $\mathbf{z_i}$, which (under the PPS scheme) conditionally given $\boldsymbol{X}$ and Σ^*, is $N_p\left(\bar{\mathbf{x}}, (1 + \frac{1}{n})\Sigma^*\right)$ with Σ^* having an Inverted Wishart distribution (see (44) below). Again, as under PIS, here also the unit i has no direct relevance.

Before we compute the PM θ in Case (a), let us look at Case (b).

7.3.2. Case (b)

Recall that $\bar{\mathbf{z}}$ is the intruder's choice of $\mathbf{x_i}$ in this case. To compute the PM θ_i, we proceed as follows.

Recall that under PPS:

$$\bar{\mathbf{z}} - \mathbf{x}_i|\Sigma^*, X \sim N_p\left(\bar{\mathbf{x}} - \mathbf{x}_i, \frac{2}{n}\Sigma^*\right) \text{ and } \Sigma^*|X \sim \text{Wishart}_p^{-1}\left(\boldsymbol{W}_{\mathbf{x}}^{-1}, n + \alpha - p - 2\right) \tag{44}$$

with (Anderson (2003))

$$h(\Sigma^*) \sim e^{-\frac{1}{2}tr\Sigma^{*-1}S_x}|\Sigma^*|^{-(\frac{n+\alpha-1}{2})}|\boldsymbol{W}_{\mathbf{x}}|^{(\frac{n+\alpha-p-2}{2})} \tag{45}$$

Combining (44) and (45), the marginal density of $\bar{\mathbf{z}}$, given X, is readily obtained as:

$$\begin{aligned} f(\bar{\mathbf{z}}|X) &\sim \int_{\Sigma^*} \frac{e^{-\frac{n}{4}(\bar{\mathbf{z}}-\bar{\mathbf{x}})^t\Sigma^{*-1}(\bar{\mathbf{z}}-\bar{\mathbf{x}})}}{|\Sigma^*|^{p/2}} e^{-\frac{1}{2}tr\Sigma^{*-1}\boldsymbol{W}_{\mathbf{x}}}|\Sigma^*|^{-(\frac{n+\alpha-1}{2})}|\boldsymbol{W}_{\mathbf{x}}|^{(\frac{n+\alpha-p-2}{2})}d\Sigma^* \\ &\sim \int_{\Sigma^*} e^{-\frac{1}{2}tr\Sigma^{*-1}[\boldsymbol{W}_{\mathbf{x}}+\frac{n}{2}(\bar{\mathbf{z}}-\bar{\mathbf{x}})(\bar{\mathbf{z}}-\bar{\mathbf{x}})^t]}|\Sigma^*|^{-(\frac{n+\alpha-1+p}{2})}|\boldsymbol{W}_{\mathbf{x}}|^{(\frac{n+\alpha-p-2}{2})}d\Sigma^* \\ &\sim \frac{|\boldsymbol{W}_{\mathbf{x}}|^{\frac{n+\alpha-p-2}{2}}}{|\boldsymbol{W}_{\mathbf{x}}+\frac{n}{2}(\bar{\mathbf{z}}-\bar{\mathbf{x}})^t(\bar{\mathbf{z}}-\bar{\mathbf{x}})|^{\frac{n+\alpha-2}{2}}} \\ &\sim \frac{|\boldsymbol{W}_{\mathbf{x}}|^{-\frac{p}{2}}}{|1+\frac{n}{2}(\bar{\mathbf{z}}-\bar{\mathbf{x}})^t\boldsymbol{W}_{\mathbf{x}}^{-1}(\bar{\mathbf{z}}-\bar{\mathbf{x}})|^{\frac{n+\alpha-2}{2}}} \end{aligned} \tag{46}$$

which is a multivariate t-distribution. The privacy measure (PM) θ_i can then be written as

$$\begin{aligned} \theta_i &= P\Big[(\bar{\mathbf{z}}-\mathbf{x}_i)^t\boldsymbol{A}(\bar{\mathbf{z}}-\mathbf{x}_i) \le \epsilon|\boldsymbol{X}\Big] \\ &= P\Big[\Big\{(\bar{\mathbf{z}}-\bar{\mathbf{x}})+(\mathbf{x}_i-\bar{\mathbf{x}})\Big\}^t\boldsymbol{A}\Big\{(\bar{\mathbf{z}}-\bar{\mathbf{x}})+(\mathbf{x}_i-\bar{\mathbf{x}})\Big\} \le \epsilon|\boldsymbol{X}\Big] \\ &= P\Big[(\mathbf{y}-\boldsymbol{\zeta}_i)^t\boldsymbol{A}(\mathbf{y}-\boldsymbol{\zeta}_i) \le \epsilon|\boldsymbol{X}\Big] \end{aligned} \tag{47}$$

where $\mathbf{y} = \bar{\mathbf{z}} - \bar{\mathbf{x}}$ and $\boldsymbol{\zeta}_i = \mathbf{x}_i - \bar{\mathbf{x}}$. Note from (46) that the pdf of $\mathbf{y}$ can be written as

$$h(\mathbf{y}) \sim |\mathbf{B}|^{p/2}[1+\mathbf{y}^t\mathbf{B}\mathbf{y}]^{-\frac{n+\alpha-2}{2}} \tag{48}$$

where $\mathbf{B} = \frac{n}{2}\boldsymbol{W}_{\mathbf{x}}^{-1}$. It is well known that a multivariate t-distribution is a scale-mixture of normal and gamma. This follows because (48) can be written as

$$\sim \int_0^\infty \Big[e^{-\frac{\mathbf{y}^t\mathbf{B}\mathbf{y}}{2}u}|\mathbf{B}|^{p/2}u^{p/2}\Big]\Big[e^{-\frac{u}{2}}u^{\frac{\nu-p}{2}}\Big]du \tag{49}$$

$$\sim |\mathbf{B}|^{p/2}(1+\mathbf{y}^t\mathbf{B}\mathbf{y})^{-(\frac{\nu}{2}+1)} \text{ where } \nu = n+\alpha-4 \tag{50}$$

$$\mathbf{y}|u \sim N_p\left(\mathbf{0}, \frac{\mathbf{B}^{-1}}{u}\right), \; u \sim e^{-\frac{u}{2}}u^{\frac{\nu-p}{2}}, 0 < u < \infty. \tag{51}$$

Let $\boldsymbol{\Gamma} : p \times p$ be a nonsingular matrix such that $\boldsymbol{\Gamma}\mathbf{B}^{-1}\boldsymbol{\Gamma}^t = \boldsymbol{I}_p \Leftrightarrow \mathbf{B}^{-1} = \boldsymbol{\Gamma}^{-1}(\boldsymbol{\Gamma}^t)^{-1} = (\boldsymbol{\Gamma}^t\boldsymbol{\Gamma})^{-1}$. Then $\mathbf{V}_i \stackrel{\text{def}}{=} \boldsymbol{\Gamma}(\mathbf{y}-\boldsymbol{\zeta})_i|u \sim N_p\left(-\boldsymbol{\Gamma}\boldsymbol{\zeta}_i = \boldsymbol{\delta}_i, \frac{I_p}{u}\right)$.

The privacy measure θ_i from 47) can be expressed as

$$\begin{aligned} \theta_i &= P\Big[(\mathbf{y}-\boldsymbol{\zeta}_i)^t\boldsymbol{A}(\mathbf{y}-\boldsymbol{\zeta}_i) \le \epsilon|\boldsymbol{X}\Big] \\ &= P\Big[\mathbf{V}_i^t\left((\boldsymbol{\Gamma}^{-1})^t\boldsymbol{A}\boldsymbol{\Gamma}^{-1}\right)\mathbf{V}_i \le \epsilon|\boldsymbol{X}\Big]. \end{aligned}$$

Finally, let us write $\mathbf{C} = (\mathbf{\Gamma}^{-1})^t \boldsymbol{A} \mathbf{\Gamma}^{-1}$ and choose an orthogonal matrix $\mathbf{\Lambda}$ satisfying $\mathbf{C} = \mathbf{\Lambda}^t D(\boldsymbol{\lambda}) \mathbf{\Lambda}$, where $D(\boldsymbol{\lambda})$ is a diagonal matrix with the diagonal elements as the roots of $\mathbf{C}$. Then

$$\begin{aligned}
\theta_i &= P\left[\mathbf{V}_i^t \mathbf{\Lambda}^t D(\boldsymbol{\lambda}) \mathbf{\Lambda} \mathbf{V}_i \le \epsilon | \boldsymbol{X}\right] \\
&= P\Big[\mathbf{V}_i^{*t} D(\boldsymbol{\lambda}) \mathbf{V}_i^* \le \epsilon | \boldsymbol{X}\Big], \ \mathbf{V}_i^* = \mathbf{\Lambda} \mathbf{V}_i \sim N_p\left(\boldsymbol{\eta}_i, \frac{1}{u} \boldsymbol{I}_p\right), \text{ where } \boldsymbol{\eta}_i = -\mathbf{\Lambda} \mathbf{\Gamma} \boldsymbol{\zeta}_i \\
&= E_u \left\{ P\left[\sum_{j=1}^{p} \lambda_j \chi_{1j}^2 (u \eta_{ij}^2) \le u\epsilon | u \right] \right\}, \text{ where } \eta_{ij} \text{ be the } j^{th} \text{ component of } \boldsymbol{\eta}_i. \quad (52) \\
&\le P\Big[\sum_{j=1}^{p} \lambda_j \chi_{1j}^2 (\text{central}) \le \epsilon \chi_{\nu-p+2}^2 \Big]. \quad (53)
\end{aligned}$$

Recall that $\lambda_1, \cdots, \lambda_p$ are the roots of $\mathbf{C}$, which are the same as the roots of $\boldsymbol{A}(\mathbf{\Gamma}^t \mathbf{\Gamma})^{-1} = \boldsymbol{A}\mathbf{B}^{-1} = \frac{2}{n}(\boldsymbol{A}\boldsymbol{W}_{\mathbf{x}})$, and $\chi_{11}^2, \cdots, \chi_{1p}^2$ are independent central χ^2 with 1 degree of freedom. The universal upper bound in (53) can be used as a privacy measure for any unit.

Three special cases follow.

Case 1: $\boldsymbol{A} = \mathbf{I}_p \implies \theta \le P\Big[\sum_{i=1}^{p} \lambda_i \chi_{1i}^2 (\text{central}) \le \epsilon \chi_{\nu-p+2}^2 \Big]$, where $\lambda_1, \cdots, \lambda_p$ are the roots of $\frac{2}{n} \boldsymbol{W}_{\mathbf{x}}$.

Case 2: $\boldsymbol{A} = \boldsymbol{W}_{\mathbf{x}}^{-1} \implies \lambda_1 = \cdots = \lambda_p = \frac{2}{n}$, which implies $\theta \le P\Big[\chi_p^2 \le \frac{n}{2} \epsilon \chi_{\nu-p+2}^2\Big]$.

Case 3: $\boldsymbol{A} = Diag(a_1, \cdots, a_p) \implies \lambda_1, \cdots, \lambda_p$ are the roots of $\frac{2}{n} \begin{bmatrix} a_1 & 0 & \dots & 0 \\ 0 & a_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_p \end{bmatrix} \boldsymbol{W}_{\mathbf{x}}$.

Returning now to Case (a), it is easy to verify from the distributional property of $\mathbf{z_i}$ and the derivation under Case (b) that here

$$\theta \le P\Big[\sum_{j=1}^{p} \lambda_j \chi_{1j}^2 (\text{central}) \le \epsilon \chi_{\nu-p+2}^2 \Big] \quad (54)$$

where $\lambda_1, \cdots, \lambda_p$ are now the roots of $(1 + \frac{1}{n}) \boldsymbol{A} \boldsymbol{W}_{\mathbf{x}}$. Three special cases as in Case b can be easily dealt here.

7.3.3. Case (c)

From the derivation under case (a), referring to equation (51) which displays the conditional multivariate normal distribution of $\bar{\mathbf{z}}$, given $\boldsymbol{X}$ and u, it readily follows that the conditional univariate distribution of $\bar{z}_1$, given $\boldsymbol{X}$ and u, is $N\left(\bar{x}_1, (\frac{2}{n}) \boldsymbol{W}_{\mathbf{x}11}\right)$ with the marginal pdf of u as $\sim e^{-u/2} u^{\frac{\nu-p}{2}}$, $0 < u < \infty$. Hence the privacy measure (PM) $P[(\bar{z}_1 - \bar{x}_1)^2 \le \epsilon]$ can be computed as

$$P[(\bar{z}_1 - \bar{x}_1)^2 \le \epsilon] = P\Big[\chi_1^2 \le \frac{n\epsilon}{2\boldsymbol{W}_{\mathbf{x}11}} \chi_{\nu-p+2}^2\Big] = P\left[F_{1,n+\alpha-p-2} \le \frac{n\epsilon(n+\alpha-p-2)}{2\boldsymbol{W}_{\mathbf{x}11}}\right] \quad (55)$$

since $\nu = n + \alpha - 4$.

Remark 3: Smart intruder's case

A smart intruder with sufficient training in statistics is likely to think in a completely different manner than a naive intruder. A general result to predict unobserved $\mathbf{X}$ from an observed $\mathbf{Y}$ is to use the conditional mean formula: $E(\mathbf{X}|\mathbf{Y})$. In our case upon observing the released data ($\mathbf{u}$, $\mathbf{y}$, $\mathbf{z}$) under the three data generation or perturbation schemes, it is possible to compute the conditional means $E(\mathbf{X}|\mathbf{u}$ or $\mathbf{y}$ or $\mathbf{z}]$ although the expressions will be quite complicated in some cases. We do not pursue this aspect here.

8. Applications

In this section, we consider one publicly accessible multivariate dataset obtained from the US Census Bureau website and another multivariate dataset on renal variables from the book by Harris and Boyd (1995). Subsequently, we employ the various data masking procedures described in the prevision sections. The goal is to construct a credible ellipsoid for the unknown mean vector based on the original data and its perturbed versions, and display and compare them. We also study which component of the multivariate data vector is expected to provide least to most privacy protection based on the criterion used in Section 7.

Subsection 8.1 provides a description and summary of the Census Bureau data for $p = 2$, while subsection 8.2 focuses on the renal dataset for $p = 3$, presenting its description and analysis. Privacy protection measures for both datasets are presented in subsection 8.3.

8.1. Description and summary of census bureau data

This subsection provides an overview of the 2023 Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC) data, conducted by the Bureau of the Census for the Bureau of Labor Statistics. The ASEC Supplement includes crucial monthly demographic and labor force data, supplemented by additional details on work experience, income, noncash benefits, health insurance coverage, and migration. Our data analysis focused on the District of Columbia (D.C.) for $p = 2$, we have examined two variables, Total Household Earnings (THHE), which includes Wages and Salary income, and Other Household Earnings (OHHE), encompassing retirement, interest, dividend, and social security income, chosen from a diverse range of available data. The "2023 Annual Social and Economic Supplements" can be accessed at https://www.census.gov/data/datasets/2023/demo/cps/cps-asec-2023.html.

In our analysis for the District of Columbia data from the Census Bureau, utilizing two variables ($p = 2$: THHE and OHHE), we have examined a sample of 171 households with Total Household Earnings (THHE) more than 200,000 USD. The resulting mean vector and dispersion matrix (in thousands) are: $\bar{\mathbf{X}} = \begin{bmatrix} \text{THHE} & \text{OHHE} \\ 347.51113 & 26.44435 \end{bmatrix}$, and $\boldsymbol{S} = \boldsymbol{W}/(n-1) = \begin{bmatrix} 19649.7273 & 548.1169 \\ 548.1169 & 1241.4463 \end{bmatrix}$. Based on the original data, the observed volume (2) of the $(1-\gamma)$

level confidence ellipsoid (1) is 553.25 and the coefficient of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation (3) is 0.11205.

8.1.1. CE Under NA Data: Census bureau data

Case 1: **Unit level data available**

We have taken the noise dispersion matrix as $\boldsymbol{R} = \begin{bmatrix} 1000 & 10 \\ 10 & 100 \end{bmatrix}$ and $r = 100$. If unit level data are available, then for $p = 2$, $n = 171$, a significance level of $\gamma = 5\%$ for type-I error the observed volume (5) under NA data is 577.6583. The coefficient of $|\boldsymbol{\Sigma} + \boldsymbol{R}|^{\frac{1}{2}}$ in the expected volume expression of Equation 6 is the same as for the original data, that is 0.11205. Figure 2 displays the confidence ellipsoid for the unknown mean vector $\boldsymbol{\mu}$ derived from noise added data when unit level data are available.

Case 2: **Unit level data not available**

Likewise the previous case, here we also have taken the noise dispersion matrix as $\boldsymbol{R} = \begin{bmatrix} 1000 & 10 \\ 10 & 100 \end{bmatrix}$ and $r = 100$. If unit level data are not available, then for $p = 2$, $n = 171$, a significance level of $\gamma = 5\%$ for type-I error the observed volume (8) under NA data is 1026.853. The coefficient of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 9 is 0.2012185. Figure 3 displays the confidence ellipsoid for the unknown mean vector $\boldsymbol{\mu}$ derived from noise added data when unit level data are available.

8.1.2. CE Under PIS: Census bureau data

For $p = 2$, $n = 171$, a significance level of $\gamma = 5\%$ for type-I error, the observed volume (11) under PIS is 1140.265. The coefficient of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 12 is 0.2258409. Figure 6 displays the confidence ellipsoid for the unknown mean vector $\boldsymbol{\mu}$ derived from synthetic data using PIS.

8.1.3. CE Under PPS: Census bureau sata

For $p = 2$, $n = 171$, a significance level of $\gamma = 5\%$ for type-I error, $\alpha = 4$, the observed volume (15) under PPS is 1639.902. The coefficient of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 16 is 0.3382968. Figure 7 displays the confidence ellipsoid for the unknown mean vector $\boldsymbol{\mu}$ derived from synthetic data using PPS.

8.1.4. BCCE Under PIS: Census bureau data

For $p = 2$, $n = 171$, a significance level of $\gamma = 5\%$ for type-I error, and a hyperparameter $\delta = 10$ in the prior distribution, the observed volume (20) under PIS is 1062.07. The coefficient of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 21 is 0.2104. Figure 9 displays the credible ellipsoid for the unknown mean vector $\boldsymbol{\mu}$ derived from synthetic data using PIS.

8.1.5. BCCE Under PPS: Census bureau data

For $p = 2$, $n = 171$, a significance level of $\gamma = 5\%$ for type-I error, $\alpha = 1$, and a hyperparameter $\delta = 10$ in the prior distribution, the observed volume (25) under PPS is 1019.4310. The coefficient of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 26 is 0.22239. Figure 10 displays the credible ellipsoid for the unknown mean vector $\boldsymbol{\mu}$ derived from synthetic data using PPS.

8.2. Description and summary of renal data

In this section, we used a renal data set from the book by Harris and Boyd (1995), Appendix 4.2 on page 137. Serum creatinine (SCR), urea nitrogen (BUN), and uric acid (UA) levels were assessed from a single blood specimen collected from a group of male medical students at the University of Virginia between 1987 and 1991 (Harris and Boyd, 1995). To demonstrate the methodologies introduced in this paper, we applied them to a subset of renal data with $p = 3$ (SCR, BUN, and UA) and a sample size of $n = 150$. The resulting mean vector and dispersion matrix are: $\bar{\mathbf{X}} = \begin{bmatrix} \text{BUN} & \text{SCR} & \text{UA} \\ 15.3600 & 1.0967 & 6.4680 \end{bmatrix}$, and $\boldsymbol{S} = \boldsymbol{W}/(n-1) = \begin{bmatrix} 12.9970 & 0.0495 & 0.3478 \\ 0.0495 & 0.0183 & 0.0574 \\ 0.3478 & 0.0574 & 1.5086 \end{bmatrix}$. Based on the original data, the observed volume (2) of the $(1-\gamma)$ level confidence ellipsoid (1) is 0.0294 and the coefficient of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation (3) is 0.0518.

8.2.1. CE Under NA Data: Renal data

Case 1: **Unit level data available**

We have taken the noise dispersion matrix as $\boldsymbol{R} = \begin{bmatrix} 0.7 & -0.3 & -0.3 \\ -0.3 & 0.7 & -0.3 \\ -0.3 & -0.3 & 0.7 \end{bmatrix}$ and $r = 100$. If unit level data are available, then for $p = 3$, $n = 150$, a significance level of $\gamma = 5\%$ for type-I error the observed volume (5) under NA data is 0.24303. The coefficient of $|\boldsymbol{\Sigma} + \boldsymbol{R}|^{\frac{1}{2}}$ in the expected volume expression of Equation 6 is the same as for the original data, that is 0.0518.

Case 2: **Unit level data not available**

Likewise the previous case, here we also have taken the noise dispersion matrix as $\boldsymbol{R} = \begin{bmatrix} 0.7 & -0.3 & -0.3 \\ -0.3 & 0.7 & -0.3 \\ -0.3 & -0.3 & 0.7 \end{bmatrix}$ and $r = 100$. If unit level data are not available, then for $p = 3$, $n = 150$, a significance level of $\gamma = 5\%$ for type-I error the observed volume (8) under NA data is 0.35897. The coefficient of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 9 is 0.10454.

8.2.2. CE Under PIS: Renal data

For $p = 3$, $n = 150$, a significance level of $\gamma = 5\%$ for type-I error, the observed volume (11) under PIS is 0.09941. The coefficient of $|\mathbf{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 12 is 0.14711.

8.2.3. CE Under PPS: Renal data

For $p = 3$, $n = 150$, a significance level of $\gamma = 5\%$ for type-I error, $\alpha = 4$, the observed volume (15) under PPS is 0.15295. The coefficient of $|\mathbf{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 16 is 0.27928.

8.2.4. BCCE Under PIS: Renal data

For $p = 3$, $n = 150$, a significance level of $\gamma = 5\%$ for type-I error, and a hyperparameter $\delta = 10$ in the prior distribution, the observed volume (20) under PIS is 0.0904. The coefficient of $|\mathbf{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 21 is 0.1338.

8.2.5. BCCE Under PPS: Renal data

For $p = 3$, $n = 150$, a significance level of $\gamma = 5\%$ for type-I error, $\alpha = 1$, and a hyperparameter $\delta = 10$ in the prior distribution, the observed volume (25) under PPS is 0.0605. The coefficient of $|\mathbf{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 26 is 0.1482.

The outcomes of observed and expected volumes for both datasets under various perturbation schemes have been summarized into a single, as shown in Table (3).

Table 3: Observed volumes and the coefficients of $|\mathbf{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression (denoted as *Expected) for various perturbation schemes and two data sets ($\gamma = 0.05$).**

DIFFERENT SCHEMES	Volumes	CB Data Set ($n = 171, p = 2$)	Renal Data Set ($n = 150, p = 3$)
NA DATA (Microdata Available)	Observed	577.6583	0.24303
	Expected	0.11205	0.05180
NA DATA (Microdata NOT Available)	Observed	1026.853	0.35897
	Expected	0.20122	0.10454
PIS	Observed	1140.265	0.09941
	Expected	0.22584	0.14711
PPS ($\alpha = 4$)	Observed	1639.902	0.15295
	Expected	0.33830	0.27928
PIS BAYES ($\delta = 10$)	Observed	1062.070	0.09040
	Expected	0.21040	0.13380
PPS BAYES ($\alpha = 1$, $\delta = 10$)	Observed	1019.4310	0.0605
	Expected	0.2239	0.1482

8.3. Privacy protection measures

Here we have obtained privacy protection measures for selective units from both Census Bureau data set ($p = 2$) and Renal data set ($p = 3$). Under noise added data, as it is immaterial to consider the second scenario, that is when the original microdata are not available, we have only considered the scenario when all the units of the original data are available. However we have considered the situation when only the summary statistics corresponding to the perturbed data are available. We have used privacy measures as given in the equations (36), (39) and (52) under NA data, PIS and PPS respectively. Privacy measures for two different data sets have been obtained in the following subsections.

8.3.1. Census bureau dataset

For CB data set, as mentioned in section (8.1), with two variables ($p = 2$: THHE and OHHE), we have examined a sample of 171 households with Total Household Earnings (THHE) more than 200,000 USD. We choose three responses with the values (in thousands) in the two categories as $(214.735, 113.943)$, $(305, 134.217)$ and $(500, 155)$. Under any perturbation scheme, privacy measures for each unit are obtained taking $\epsilon = 0.6(0.05)1$ and for the i^{th} selected unit $\mathbf{x}_i = (x_{i1}, x_{i2})$, the matrix $\boldsymbol{A}$ is chosen as $\boldsymbol{A} = \begin{bmatrix} \frac{1}{x_{i1}^2} & 0 \\ 0 & \frac{1}{x_{i2}^2} \end{bmatrix}$. For noise added data, the noise dispersion matrix is taken as $\boldsymbol{R} = \begin{bmatrix} 10000 & 100 \\ 100 & 1000 \end{bmatrix}$ and $\alpha = 4$ under PPS.

Table 4: Privacy Measures under different schemes of perturbation and for three different units from CB data set.

Units	Schemes	ϵ								
		0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1
Unit 1	NA Data	0	0	0	0	0.0004	0.0085	0.0828	0.3353	0.6920
	PIS	0	0	0	0.0001	0.0030	0.0274	0.1342	0.3662	0.6521
	PPS	0	0.0001	0.0006	0.0051	0.0252	0.0858	0.2117	0.3988	0.6045
Unit 2	NA Data	0.0122	0.3084	0.8799	0.9971	1	1	1	1	1
	PIS	0.0206	0.3183	0.8516	0.9939	1	1	1	1	1
	PPS	0.0691	0.3584	0.7628	0.9597	0.9968	0.9999	1	1	1
Unit 3	NA Data	0	0	0.0012	0.1300	0.7500	0.9919	1	1	1
	PIS	0	0	0.0058	0.1680	0.7126	0.9793	1	1	1
	PPS	0	0.0016	0.0356	0.2449	0.6527	0.9245	0.9926	0.9996	1

8.3.2. Renal dataset

For Renal data set, as mentioned in section (8.2), with three variables ($p = 3$: SCR, BUN and UA), we have examined a sample of 150 male medical students at the University of Virginia between 1987 and 1991 (Harris and Boyd, 1995). We choose three responses with the values in the three categories as $(12, 0.9, 6.1)$, $(15, 1.1, 6.9)$ and $(25, 1.1, 6.6)$. Under any perturbation scheme, privacy measures for each unit are obtained for the choices of $\epsilon \in \{0.005, 0.01, 0.05, 0.1, 0.12, 0.14, 0.145, 0.15, 0.16\}$ and for the i^{th} selected unit $\mathbf{x}_i =$

(x_{i1}, x_{i2}, x_{i3}), the matrix $\boldsymbol{A}$ is chosen as $\boldsymbol{A} = \begin{bmatrix} \frac{1}{x_{i1}^2} & 0 & 0 \\ 0 & \frac{1}{x_{i2}^2} & 0 \\ 0 & 0 & \frac{1}{x_{i3}^2} \end{bmatrix}$. For noise added data, the noise dispersion matrix is taken as $\boldsymbol{R} = \begin{bmatrix} 0.7 & -0.3 & -0.3 \\ -0.3 & 0.7 & -0.3 \\ -0.3 & -0.3 & 0.7 \end{bmatrix}$ and $\alpha = 4$ under PPS.

Table 5: Privacy Measures under different schemes of perturbation and for three different units from Renal data set

Units	Schemes	ϵ								
		0.005	0.01	0.05	0.1	0.12	0.14	0.145	0.15	0.16
Unit 1	NA Data	0	0	0	0.1079	0.3657	0.6169	0.6697	0.7165	0.7961
	PIS	0	0	0	0.0195	0.2549	0.7275	0.8177	0.8850	0.9607
	PPS	0	0	0	0.0684	0.3109	0.6588	0.7331	0.7971	0.8916
Unit 2	NA Data	0.2478	0.7521	0.9995	1	1	1	1	1	1
	PIS	0.5071	0.9755	1	1	1	1	1	1	1
	PPS	0.4415	0.8959	1	1	1	1	1	1	1
Unit 3	NA Data	0	0	0	0	0	0	0.0009	0.3152	0.8991
	PIS	0	0	0	0	0.0003	0.1461	0.3094	0.5236	0.8710
	PPS	0	0	0	0	0.0080	0.2221	0.3560	0.5078	0.7815

9. Conclusion

Referring to Table 2 in Section 6.2, it is evident that the expected volume decreases with increasing sample size (n). Conversely, regardless of the scheme used, the expected volume increases with an increase in the number of components (p). In particular, among all perturbation schemes, the smallest expected volumes are consistently observed with the noise-added data. Moreover, in both frequentist and Bayesian frameworks, PIS resulted in a smaller expected volumes compared to PPS.

We have performed some data analyses in section (8) for Census Bureau data set ($p = 2$) and for the Renal data set ($p = 3$). The observed and expected volumes for both data sets under any scheme are summarized in Table (3). The volumes under noise added data are the smallest among all the schemes and for both the data sets, whereas under two schemes of noise added data (units available and units not available), we can see smaller volumes when units are available. For both data sets, under frequentist setup, PPS is showing larger volumes than PIS. On the other hand, under the Bayes framework, the observed volumes under PPS scheme are marginally smaller than those under the PIS scheme. The diagrams (2, 3, 6, 7, 9, 10) of the ellipsoids obtained for the CB data set under different schemes are given in the appendix. Also some diagrams are obtained overlapping the ellipsoids obtained under two different schemes as, 1. NA 1 and NA 2 (fig : 4), 2. PIS and PPS under the frequentist framework (fig : 8) and 3. PIS and PPS under Bayesian framework (fig : 11). From the diagrams it is clear that one should expect a smaller volume under NA 1 scheme than that under NA 2 scheme as the ellipsoid under NA 2 scheme is containing the ellipsoid under NA 2 scheme. Under frequentist setup, ellipsoid obtained under PPS contains the ellipsoid obtained under PIS. However, in the Bayesian framework, the scenario is not the same, where none of the ellipsoids, under PIS or PPS, contain another.

For privacy protection analysis, as carried out in section (8.3), we have selected three units from both the data sets. Units are so chosen that, one is very close to the sample mean which is happen to be the second unit in both the cases, the third units are a bit distant from the mean and the first units are taken to be extreme. Privacy measures for CB data set are shown in Table (4) and those for Renal data set are shown in Table (5). As expected, the privacy measures for the second units for each data set and under any scheme are very high, which means lower privacy protection. For both data sets, we can see a higher privacy protection for Unit 3 compared to Unit 1. Comparing the perturbation schemes in terms of the privacy measure, we can say that, no scheme can be chosen over others through out for any choices of ϵ. It depends on the choice of specific units and also upon the choices of ϵ.

Acknowledgements

Our sincere thanks are due to Professor Thomas Mathew (UMBC) for his help with Theorem 1 in Section 2.2 as well as for some other critical comments. Bimal Sinha is thankful to Dr. Tommy Wright (CSRM/Census Bureau) for his encouragement and support.

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics, 3rd edition.

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, volume 201. Springer Science & Business Media.

Drechsler, J. and Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, **105**, 1347–1357.

Guin, A., Roy, A., and Sinha, B. (2023). Bayesian analysis of singly imputed synthetic data under the multivariate normal model. *International Journal of Statistical Sciences*, **23**, 1–18.

Gupta, A. and Nagar, D. (1999). *Matrix Variate Distributions*. Chapman and Hall/CRC.

Harris, E. K. and Boyd, J. C. (1995). *Statistical Bases of Reference Values in Laboratory Medicine*. CRC Press.

Kinney, S. K., Reiter, J. P., and Miranda, J. (2014). Synlbd 2.0: improving the synthetic longitudinal business database. *Statistical Journal of International Association for Official Statistics*, **30**, 129–135.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic longitudinal business database. *International Statistical Review*, **79**, 362–384.

Klein, M., Mathew, T., and Sinha, B. (2014). Noise multiplication for statistical disclosure control of extreme values in log-normal regression samples. *Journal of Privacy and Confidentiality*, **6**, 77–125.

Klein, M. and Sinha, B. (2013a). Statistical analysis of noise-multiplied data using multiple imputation. *Journal of Official Statistics*, **29**, 425–465.

Klein, M. and Sinha, B. (2013b). Statistical analysis of noise multiplied data using multiple imputation. *Journal of Official Statistics*, **29**, 425–465.

Klein, M. and Sinha, B. (2015). Inference for singly imputed synthetic data based on posterior predictive sampling under multivariate normal and multiple linear regression models. *Sankhya B*, **77**, 293–311.

Klein, M. and Sinha, B. (2016). Likelihood based finite sample inference for singly imputed synthetic data under the multivariate normal and multiple linear regression models. *Journal of Privacy and Confidentiality*, **7**, 43–98.

Lin, Y.-X. and Wise, P. (2012). Estimation of regression parameters from noise multiplied data. *Journal of Privacy and Confidentiality*, **4**, 61 – 94.

Little, R. J. et al. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, **9**, 407–407.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical science*, , 538–558.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. John Wiley & Sons.

Nayak, T., Sinha, B., and Zayatz, L. (2011). Statistical properties of multiplicative noise masking for confidentiality protection. *Journal of Official Statistics*, **27**, 527–544.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, **19**, 1–16.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, **29**, 181–188.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, **30**, 235–242.

Reiter, J. P. (2005a). Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **168**, 185–205.

Reiter, J. P. (2005b). Significance tests for multi-component estimands multiply imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, **131**, 365–377.

Reiter, J. P. (2005c). Using cart to generate partially synthetic public use microdata. *Journal of Official Statistics*, **21**, 441–462.

Reiter, J. P. and Kinney, S. K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of Official Statistics*, **28**, 583–590.

Reiter, J. P. and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*, **1**, 99–110.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, **102**, 1462–1471.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics*, **9**, 461–468.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*, **91**, 473–489.

Sinha, B., Nayak, T., and Zayatz, L. (2011). Privacy protection and quantile estimation from noise multiplied data. *Sankhya B*, **73**, 297–315.

ANNEXURE

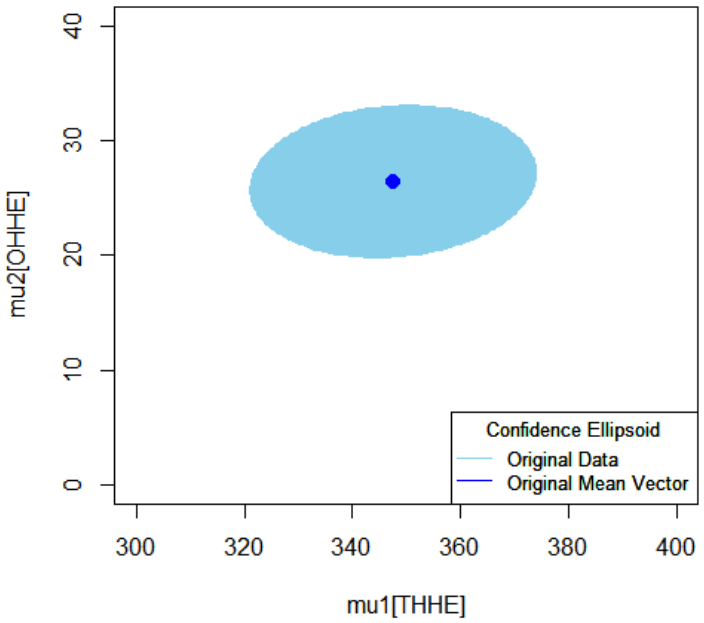


Figure 1: Confidence Ellipsoid for the unknown mean vector using original Data.

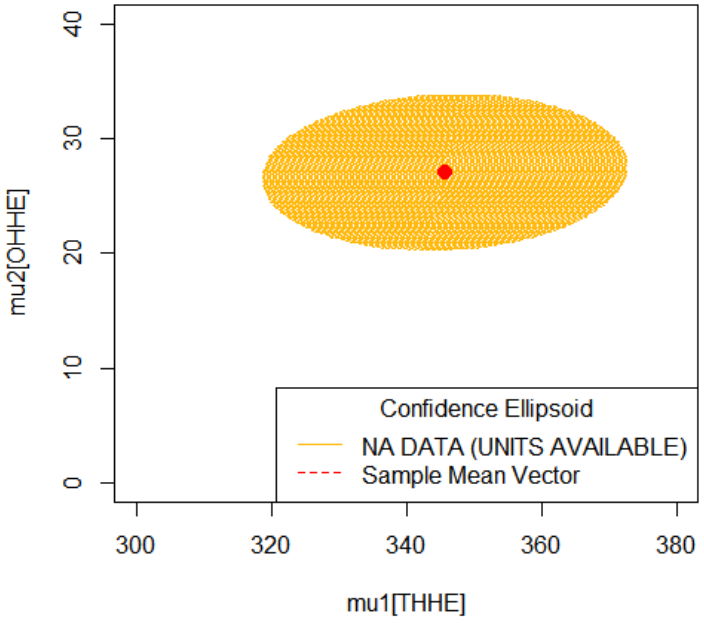


Figure 2: Confidence Ellipsoid for the unknown mean vector using Noise Added Data (Microdata Available).

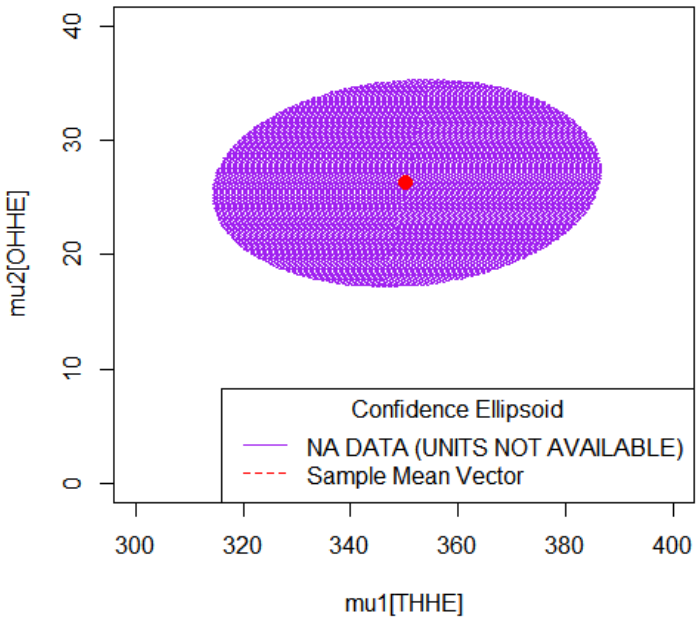


Figure 3: Confidence Ellipsoid for the unknown mean vector using Noise Added Data (Microdata Not Available).

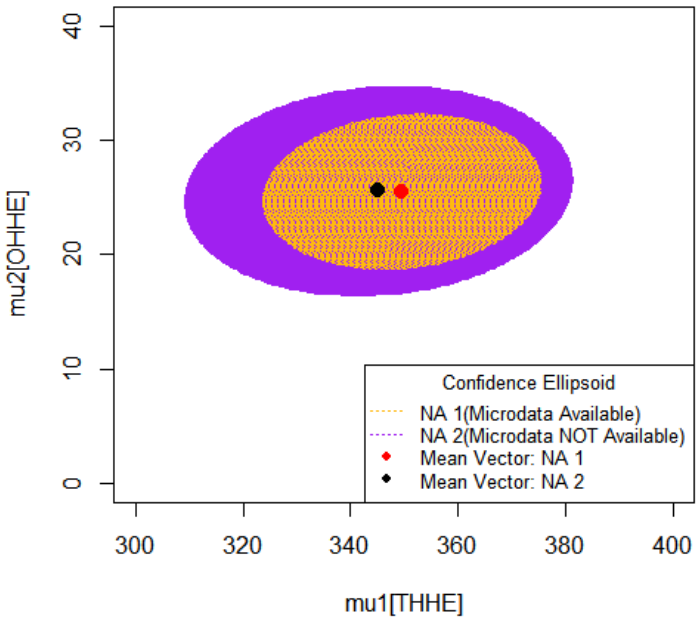


Figure 4: Confidence ellipsoids for the unknown mean vector under NA 1 (Microdata Available) and NA 2 (Microdata Not Available).

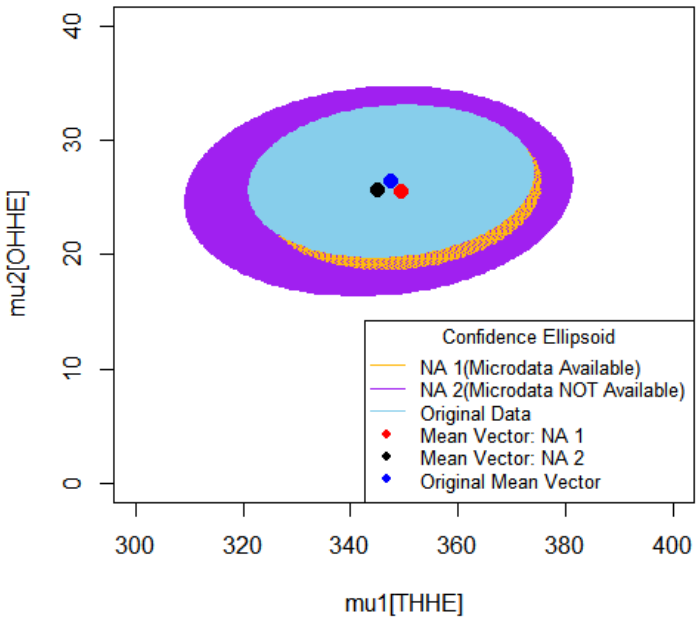


Figure 5: Confidence ellipsoids for the unknown mean vector under Original, NA 1 (Microdata Available) and NA 2 (Microdata Not Available).

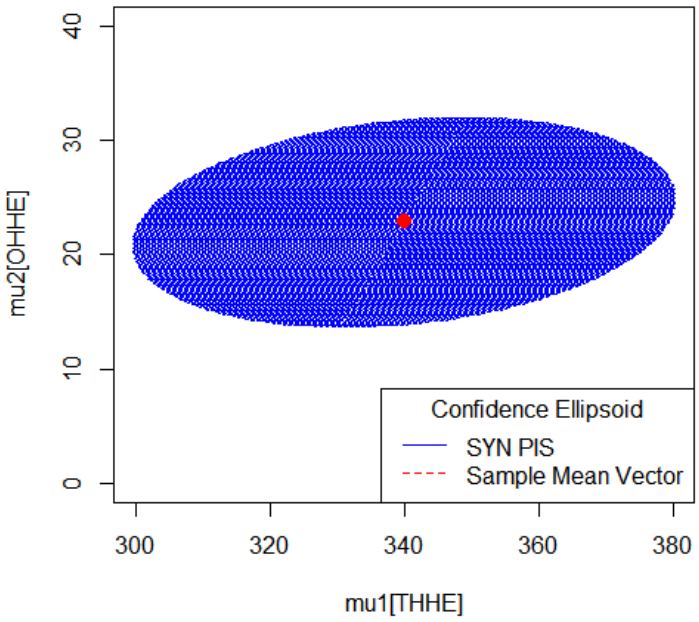


Figure 6: Confidence Ellipsoid for the unknown mean vector using synthetic data under PIS.

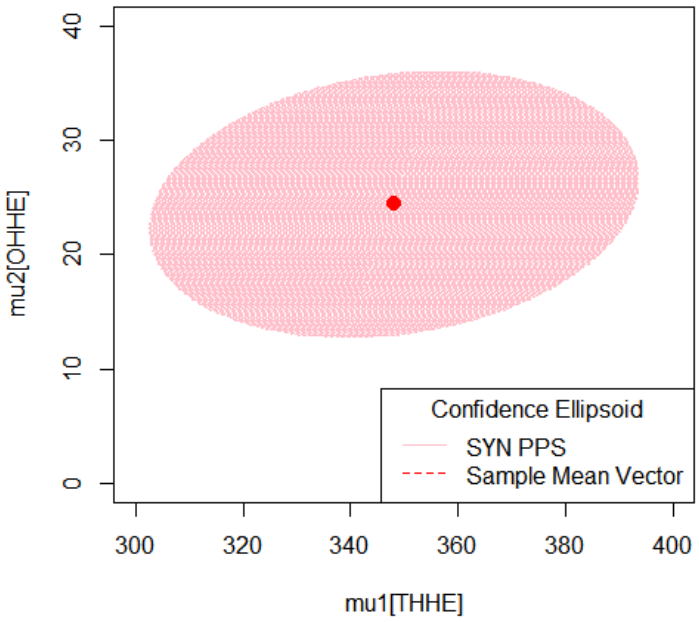


Figure 7: Confidence Ellipsoid for the unknown mean vector using synthetic data under PPS.

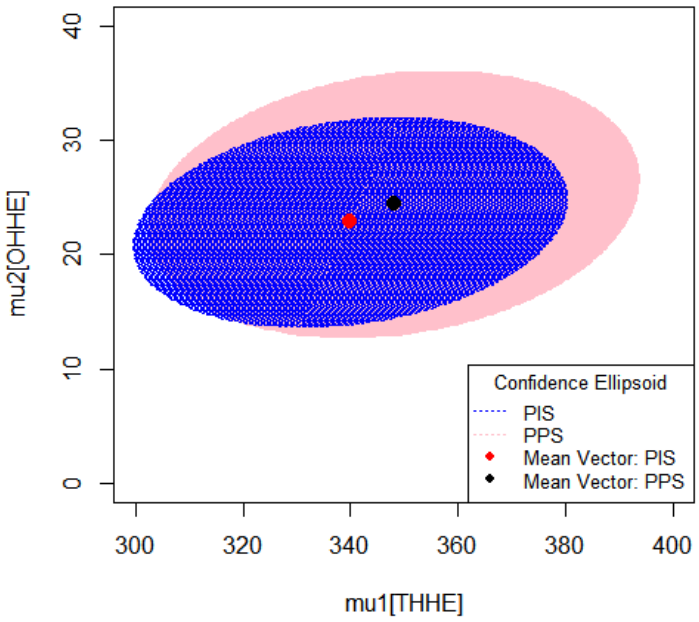


Figure 8: Confidence ellipsoids for the unknown mean vector using synthetic data under PIS and PPS.

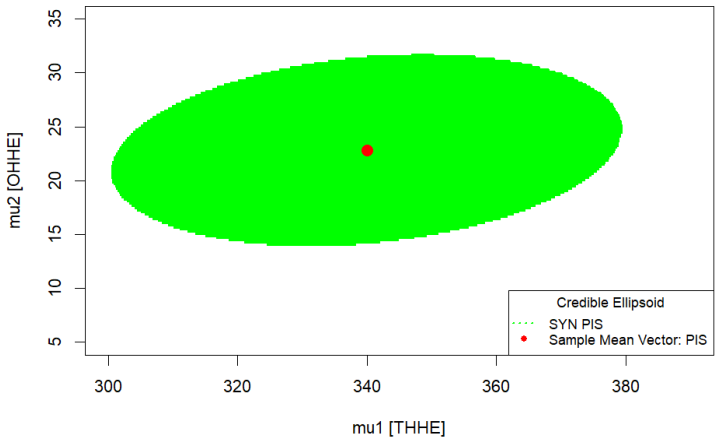


Figure 9: Credible ellipsoid for the unknown mean vector using synthetic data under PIS.

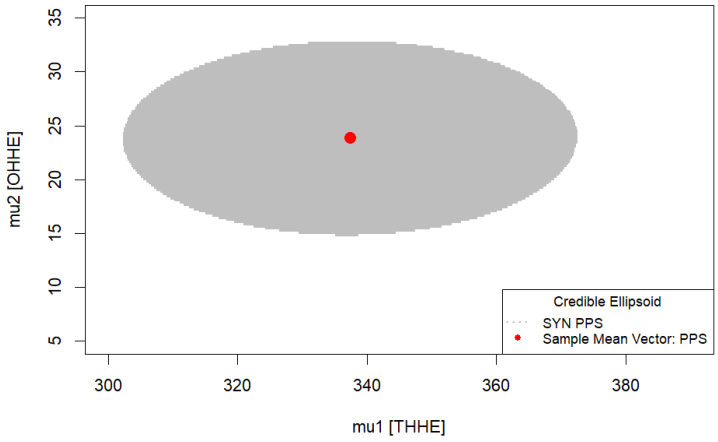


Figure 10: Credible ellipsoid for the unknown mean vector using synthetic data under PPS.

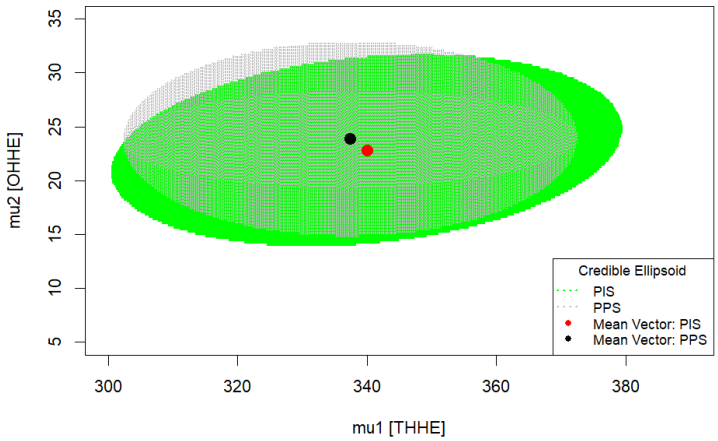


Figure 11: Credible ellipsoids for the unknown mean vector using synthetic data under PIS and PPS.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 153–169
http://www.ssca.org.in/journal

Survey of C.R. Rao's Orthogonal Arrays, Balanced Arrays, and Their Applications

Gour Mohan Saha[1], **Bikas Kumar Sinha**[1] **and Ganesh Dutta**[2]
[1]*Retired Professor of Statistics, Indian Statistical Institute, Kolkata-700108, India*
[2]*Basanti Devi College, 147B Rash Behari Avenue, Kolkata-700029, India*

Received: 13 May 2024; Revised: 15 June 2024; Accepted: 17 June 2024

Abstract

This comprehensive review article on orthogonal arrays (OAs), balanced arrays (BAs) and their practical applications serves as a tribute to the life and ground breaking contributions of the legendary statistician, C.R. Rao (1920-2023). It highlights his profound influence on the field of statistical sciences and explores the significant contributions he made to the realms of OAs and BAs. His work in these areas has left an indelible impact on the domains of experimental design, combinatorial mathematics, and statistical analysis. In this article, we delve into some noteworthy applications of OAs and BAs.

Key words: Orthogonal array; Balanced array; Mixed orthogonal array; Balanced incomplete block design; Partially balanced incomplete block design; Association scheme.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

The foundation for the concepts of Latin squares and mutually orthogonal Latin squares was laid in the early 20th century. Later, these foundational ideas were expanded and generalized to include Latin cubes and hypercubes, as well as orthogonal Latin cubes and hypercubes (*cf.* Kishen (1942, 1949)). These developments marked significant progress in the field of experimental design, as they allowed for the exploration of more complex experimental scenarios with multiple factors and levels.

Rao (1946) further extended these concepts by introducing the notion of arrays with a specific strength. These arrays became a versatile tool for designing experiments with various factors, enabling researchers to investigate complex relationships and interactions efficiently. The pivotal moment in the evolution of these ideas came when Rao (1947) introduced the concept of OAs as a unifying framework that generalized and brought together the previously mentioned structures. This marked a significant leap in the field of experimental design and made it more accessible to practitioners in diverse fields.

Furthermore, Rao (1973) continued to expand his contributions by generalizing OAs to mixed orthogonal arrays (MOAs) of strength d. This development allowed researchers

Corresponding Author: Ganesh Dutta
Email: duttaganesh78@gmail.com

to work with experiments that involved factors at different levels, thus accommodating a broader range of real world scenarios.

The evolution of experimental design, from the early concepts of Latin squares to the sophisticated OAs, represents a remarkable journey of continuous innovation and generalization. C. R. Rao's pioneering work (1946, 1947, 1949, 1961, 1973) has been instrumental in this evolution, establishing these arrays as fundamental tools in experimental design. Rao's contributions have proven invaluable in both industrial and scientific research by enabling highly efficient experiments that require fewer runs. His brilliance is further exemplified by expanding OAs into higher dimensions, thus broadening their applicability across diverse experimental settings. This expansion has notably enhanced the efficiency of experimentation and optimization in various industries, including manufacturing and quality control. Taguchi's work in the 1980s popularized the use of OAs in industry, known as Taguchi methods, which determine optimum combinations of factors to achieve high output and robustness to environmental changes. An article in Forbes Magazine (March 11, 1996, pp. 114-118) highlighted the significance of OAs, dubbing them a "New Mantra" in various U.S. industrial establishments. This recognition underscores the practical utility of OAs in enhancing efficiency and reducing the number of experimental runs required in industrial research. Beyond industry, OAs have also made profound impacts in agricultural and medical sciences, as discussed by Parsad, Gupta, and Gopinath (2020). Additionally, OAs find applications in coding theory, cryptography, and computer experiments. Comprehensive textbooks on this subject, authored by Dey and Mukerjee (1999), Hedayat, Sloane, and Stufken (1999), and Rosa (2017), provide an extensive exploration of these powerful tools, cementing their place as indispensable resources in the realm of experimental design.

BAs, stemming from the foundational work on partially balanced arrays by Chakravarti (1956, 1961) and further advanced by Srivastava and Chopra (1973), epitomize a sophisticated concept within experimental design. Initially termed as partially balanced arrays, the pioneering research by Chakravarti (1956) laid the groundwork for their exploration. Building upon this foundation, Srivastava and Chopra (1973) made significant strides, advocating for the simplification of the term to "balanced arrays", a change we have embraced. This evolution represents a pivotal moment in the realm of experimental design and statistical methodologies. BAs provide a structured and efficient means to investigate the intricate relationships among multiple factors and their respective levels. By systematically varying factors while minimizing confounding effects, these arrays offer a robust framework for achieving statistical efficiency. In essence, they stand as a testament to the ongoing advancement of experimental design, empowering researchers to uncover insights with clarity and precision.

Within the broader framework of OAs, BAs emerge as a noteworthy and important subset. OAs are a specialized type of BAs. They hold a unique and powerful position in the field of experimental design, as they are specifically designed to ensure that the effects of different factors do not interfere with each other. In other words, OAs allow researchers to explore and quantify the impact of various factors on the outcome of interest without undue influence from unrelated factors. In essence, OAs and BAs are intertwined components of experimental design, with BAs serving as a foundational concept and OAs as a refined and focused tool within this framework. Together, they provide researchers with a comprehensive toolkit to design and execute experiments effectively, ensuring that the results obtained are

both reliable and interpretable.

In this review article, we provide definitions of OAs, MOAs, and BAs in Section 2. We focus on the construction of incomplete block designs from BAs and OAs in Section 3. Section 4 discusses the construction of optimum chemical balance weighing designs from BAs. Section 5 covers the construction of second order rotatable designs (SORDs) using BAs. In Section 6, we discuss some methods for constructing BAs. Section 7 considers the construction of orthogonal resolution plans and fractional factorial plans using OAs. Section 8 addresses the application of OAs in Taguchi methods. Section 9 explores other diverse applications of OAs and MOAs in modern research and experimentation. Finally, Section 10 presents the conclusion.

2. Overview of OAs and BAs

In this section, we provide a comprehensive overview of OAs and BAs. While these concepts might be familiar to the audience of this special issue, we briefly revisit them for the sake of completeness and to ensure smooth reading.

2.1. OA

OAs are mathematical structures extensively used in experimental design, coding theory, and quality engineering. They facilitate the systematic testing of different combinations of variables while minimizing the number of experimental runs required.

Definition 1: Consider an array $\mathbf{A}$ of size $k \times N$, where its elements are drawn from a set S comprising s symbols or levels, denoted by $0, 1, \ldots, s-1$. This array $\mathbf{A}$ is termed an OA possessing s levels, with a strength of t, and an index denoted by λ, under the condition that each $t \times N$ subarray within $\mathbf{A}$ contains every t-tuple derived from S precisely λ times as a column.

We denote such an array by $\mathrm{OA}(N, s^k, t)$. Clearly, $N = \lambda s^t$.

Definition 2: A MOA $\mathrm{OA}(N, s_1^{k_1} s_2^{k_2} \ldots s_v^{k_v}, t)$ is an array of size $k \times N$, where $k = k_1 + k_2 + ... + k_v$ is the total number of factors, in which the first k_1 rows have symbols from $\{0, 1, \ldots, s_1 - 1\}$, the next k_2 rows have symbols from $\{0, 1, \ldots, s_2 - 1\}$, and so on. The array has the property that in any $t \times N$ subarray, every possible t-tuple occurs an equal number of times as a column. Of course, if all s_i's are equal, we get the usual $\mathrm{OA}(N, s^k, t)$ as of Definition 1.

For further understanding, readers may refer to authoritative textbooks by Raghavarao (1971), Dey and Mukerjee (1999), Hedayat *et al.* (1999) and Rosa (2017). Additionally, valuable insights can be gained from online resources such as TS-DOC: TS-723 - OAs by WF Kuhfeld, OA testing on Wikipedia, and the design resources server of the Indian Agricultural Statistics Research Institute (IASRI). N. J. A. Sloane's "A Library of OAs" also provides comprehensive information. Furthermore, important references, including works by Bose (1950), Bose and Bush (1952), Bush (1952), Cheng (1980), and Mukhopadhyay (1981), offer deeper insights into the topic.

2.2. BA

BAs are similar to OAs but are specifically designed for applications that require balanced representation across various combinations. This characteristic makes them particularly valuable in the field of combinatorial design.

Definition 3: Let $\mathbf{A}$ be an $k \times N$ array with elements 0,1,2,..., $s-1$. Consider the s^t possible vectors $\mathbf{X}' = (x_1, x_2, \ldots, x_t)$, where each x_i can take any value from $\{0, 1, \ldots, s-1\}$ for $i = 1, 2, \ldots, t$. Associate with each $t \times 1$ vector $\mathbf{X}$ a positive integer $\lambda(x_1, x_2, \ldots, x_t)$, which remains unchanged under permutations of $(x_1, x_2, \ldots, x_t)$. If for every t rowed subarray of $\mathbf{A}$, the s^t distinct $t \times 1$ vectors $\mathbf{X}$ appear as columns exactly $\lambda(x_1, x_2, \ldots, x_t)$ times, then the array $\mathbf{A}$ is called a BA of strength t in N assemblies, with m constraints, s symbols, and the specified $\lambda(x_1, x_2, \ldots, x_t)$ parameters.

It is to be noted that if $\lambda(x_1, x_2, \ldots, x_t) = \lambda$ for all $(x_1, x_2, \ldots, x_t)$, then $\mathbf{A}$ is called an OA of index λ.

The literature on this topic is extensive, making it challenging to cite every relevant work. Therefore, we reference a selection of seminal articles from the early stages, including those by Chakravarti (1956, 1961), Srivastava and Chopra (1973), Rafter and Seiden (1974), and Saha (1981).

3. Constructing incomplete block designs with BAs and OAs

A methodology emerges for constructing BAs, employing the Kronecker product applied to two BAs. This approach leads to the derivation of six distinct balanced incomplete block designs (BIBDs) from a given symmetric balanced incomplete block design (SBIBD). Notably, the method involves operations such as unions, intersections, and difference sets on pairs of blocks of an SBIBD and their complementary designs. Significantly, certain newly generated BIBDs fulfill the minimum replication requirements for specified parameters like v (number of varieties or treatments) and k (block size), showcasing the method's efficiency and efficacy. Expanding beyond its original scope, the study suggests broader applications for this method. It proposes leveraging the new series of SBIBDs to derive additional series of BIBDs, hinting at the potential for an iterative process where new designs build upon established ones, thus enriching the repertoire of available BIBDs. Independent confirmations by Vanstone (1975) and Majindar (1978) regarding the existence of the six BIBDs corresponding to an SBIBD reinforce the method's validity and reliability.

In Saha (1975), the tactical configurations (or t designs) are generalized to G systems of order β, and their equivalence to 2 symbol BAs of strength β is established. This extension confirms their equivalence to 2 symbol BAs of strength β. These findings are then applied to demonstrate that when β is even, $\mathbf{A} \cup \mathbf{A}^c$ yields another 2 symbol BA of strength $\beta + 1$, where $\mathbf{A}$ is a 2 symbol BA of strength β, and $\mathbf{A}^c$ is the complementary array (obtained from $\mathbf{A}$ by interchanging 0s and 1s). This holds true for 2 symbol OAs of strength β when β is even as well. Furthermore, the research identified specific series of 2 symbol OAs with a strength of three, which were obtained from carefully selected 2 symbol BAs with a strength of two. Saha (1975) demonstrated how tactical configurations (t designs) generalize to G systems of order β and established their equivalence to 2 symbol BAs of strength β. It also

sheds light on the behavior of combining such arrays and provides insights into obtaining series of 2 symbol OAs with a higher strength from BAs.

Building upon these foundational works, Saha *et al.* (1985) extend the method to generate s symbol BAs with a strength of t. This advancement is then applied to derive diverse partially balanced incomplete block designs (PBIBDs) characterized by m associate classes. Additionally, their research reveals the coexistence of six distinct series of PBIBDs alongside a linked block PBIBD, showcasing the versatility of the method in addressing various experimental design needs and its significant contributions to advancing the field.

For detailed definitions and further reading on BIBD, SBIBD, PBIBD, and association schemes, several excellent textbooks are available, with Raghavarao (1971) being particularly recommended.

4. Optimal chemical balance weighing designs from BAs and BIBDs

Optimal chemical balance weighing designs are experimental frameworks used in chemical experiments to measure the weights of multiple substances simultaneously with high accuracy. These designs aim to minimize the variance of the estimated weights, ensuring precise and unbiased measurements. The key characteristics of optimal chemical balance weighing designs include efficiency, as they maximize the information obtained from a limited number of weighings; balance, by distributing errors evenly across all measurements to reduce systematic biases; and replication, through repeated measurements to enhance reliability. Additionally, these designs often employ combinatorial structures like BAs and BIBDs to systematically arrange substances on the balance, optimizing the weighing process. In essence, these designs provide a structured approach to achieving high precision and minimal error in the measurement of multiple substances. For further reading, refer to Raghavarao (1971), Silvey (1980), Shah and Sinha (1989) and Pukelsheim (1993).

Dey (1971), Saha (1975), Kageyama and Saha (1983), along with other researchers, initially showcased the derivation of optimal chemical balance weighing designs from the incidence matrices of BIBDs.

Dey (1971) utilized the incidence matrices of BIBDs and balanced ternary designs for constructing optimal chemical balance weighing designs.

Regarding the relationship between BIBDs and optimum chemical balance weighing designs, Saha (1975) proved two significant theorems:

Theorem 1: The existence of a BIBD with parameters v, b, r, k, λ satisfying $b \leq 4(r - \lambda)$ implies the existence of an optimum chemical balance weighing design for v objects in $4(r-\lambda)$ weighings.

Theorem 2: The existence of an affine resolvable BIBD with parameters $v, b = 2r, r, k, \lambda$ implies the existence of an optimum chemical balance weighing design for r objects in v weighings.

Kageyama and Saha (1983) investigated a BIBD with parameters v, b, r, k, λ satisfying $b \leq 4(r-\lambda)$ and tabulated the parameters (in the practical range) of BIBDs which validated the above theorems of Saha (1983).

Expanding upon this foundation, Saha and Kageyama (1984) further developed the methodology by illustrating that optimum weighing designs could also be derived from carefully selected two symbol BAs of strength two. Importantly, these arrays were not limited to being incidence matrices of BIBDs, thus widening the potential design scope. To implement this approach, the first step involves identifying two symbol BAs of strength two with desired properties for optimum chemical balance weighing design construction. These arrays must meet specific criteria to ensure suitability. By leveraging the identified arrays, the optimum chemical balance weighing designs can be generated, utilizing the array's structure and properties. This process involves transforming the array into a design that meets the requirements for the optimum chemical balance weighing. Thus, the findings lead us to construct new optimum chemical balance weighing designs other than the above mentioned methods. This research has far reaching implications for the optimum design of chemical balance experiments and provides a more flexible and versatile framework for developing such designs beyond the limitations of traditional BIBDs.

5. Constructing SORDs using BAs for response surface studies

SORDs are a type of experimental design used primarily in response surface methodology (RSM) to model and optimize processes. These designs are particularly effective when the relationship between the factors and the response variable is quadratic. They accommodate a second order (quadratic) polynomial model, encompassing linear, interaction, and squared terms of the input variables. A design is considered rotatable if the variance of the predicted response at any point depends solely on the distance from the design center, rather than the direction, thereby ensuring uniform precision of prediction at all equidistant points from the center. The most common type of SORD is the central composite design (CCD), which combines a factorial or fractional factorial design with center points and axial (or star) points to estimate curvature. These designs efficiently estimate the coefficients of a second order polynomial, enabling the detection of curvature in the response surface and the identification of optimal conditions. Their flexibility and efficiency in handling multiple factors make SORDs indispensable tools in industrial and scientific research for process optimization. For further details, we refer to Khuri and Cornell (1996).

The integration of BAs has significantly expanded the toolkit available to researchers involved in designing SORD for analyzing response surfaces. A pivotal contribution to this field was made by Das and Saha (1973), who demonstrated the successful construction of SORDs under specific conditions. They outlined requirements for 2 symbol BAs of strength two, which enabled the creation of 4 level SORDs. Leveraging these principles, they introduced several novel series of 4 level SORDs. Notably, they uncovered an intriguing finding: a 4 level SORD can be derived for (i) $v - x$ factors from b magnitude sets, and (ii) v factors from $b + c$ magnitude sets, from a BIBD meeting certain criteria, such as $r > 3\lambda$, or $5r - 2b - 3\lambda > 0$ $(x > 0, c > 0)$. Furthermore, these designs can be augmented by selecting appropriate magnitude sets in addition to those derived from the incidence matrices of BIBDs.

Such designs present researchers with a flexible and adaptable framework, facilitating the conduct of response surface experiments and providing a nuanced exploration into the behavior of intricate systems.

6. Construction of BAs

Association schemes with a large number of associate classes have historically been investigated primarily for their combinatorial significance, without a focus on their application in the development of practical experimental designs. However, Saha (1981) introduced a novel approach by utilizing a new class of cyclic association scheme with m associate classes, referred to as NC_m association scheme. This approach was employed to construct $(m+1)$ symbol BAs of strength two. The resulting BIBDs derived from these arrays were also explored in the same paper.

In a more recent study by Yonglin (2004), association schemes have been employed to investigate their relationship with OAs and frequency squares, which represent a generalization of Latin squares. This research demonstrates the evolving and diverse applications of association schemes in combinatorial design theory, shedding light on their connection to other fundamental structures and concepts.

Researchers made notable contributions for constructing BAs with a strength of two from block designs. For instance, Sinha *et al.* (2002) achieved this by using various types of block designs, including (i) rectangular designs; (ii) group divisible designs; (iii) nested balanced incomplete block designs. These constructions result in BAs, which are useful in experimental design and combinatorial applications.

Balanced nested designs share intricate connections with other combinatorial structures like BAs and balanced n-ary designs. Specifically, the presence of symmetric balanced nested designs mirrors the existence of certain BAs. Delving into this relationship, Fuji-Hara *et al.* (2002) conducted a comprehensive exploration of balanced nested designs. They focused on elucidating the interplay between balanced nested designs and BAs with a strength of two, offering diverse constructions for symmetric balanced nested designs. These constructions proved instrumental in delineating the spectrum of symmetric balanced nested incomplete block designs with block sizes of 3 and 4. Notably, their research unveiled the equivalence between symmetric balanced nested designs and specific categories of BAs. Beyond enriching our understanding of BAs, their work provided invaluable insights into constructing symmetric balanced nested designs, thereby advancing the broader field of combinatorial design theory.

7. Orthogonal resolution and fractional factorial plans with OAs

Orthogonal resolution plans and fractional factorial plans are two types of experimental designs commonly employed in industrial and scientific research to efficiently explore the effects of multiple factors on a response variable while minimizing the number of experimental runs needed. Orthogonal resolution plans are characterized by their ability to provide unbiased estimates of main effects and interactions between factors, even in the presence of confounding. These plans achieve orthogonality by ensuring that each factor is varied independently of the others at different levels, thereby allowing for the unambiguous identification of the effects of individual factors. Additionally, orthogonal resolution plans are designed to have certain desirable properties such as clear aliasing patterns, which aid in the interpretation of results. On the other hand, fractional factorial plans are a subset of orthogonal resolution plans that further reduce the number of experimental runs by systematically selecting a fraction of the total number of possible treatment combinations. Despite this

reduction in the number of experimental runs, fractional factorial plans retain the ability to estimate main effects and selected interactions with minimal loss of information. These plans are particularly useful when the number of factors under investigation is large and conducting a full factorial experiment would be impractical or prohibitively expensive. Overall, orthogonal resolution plans and fractional factorial plans are valuable tools in experimental design, offering efficient and cost effective approaches to exploring complex systems and optimizing processes in various fields.

OAs play a crucial role in the construction of orthogonal resolution plans and subclasses of fractional factorial plans, which are essential for optimizing experimental efficiency and reliability. In orthogonal resolution plans, OAs help organize experiments to ensure clarity and precision in identifying the effects of different factors by minimizing confounding and enhancing the interpretability of results. These plans are categorized by their resolution, with higher resolutions indicating clearer distinctions between main effects and interactions. OAs also aid in constructing fractional factorial plans, which allow researchers to study the most significant factors and interactions using a fraction of the total runs required in a full factorial design. This systematic approach significantly reduces the number of experimental runs needed, saving time and resources while maintaining experimental integrity. By ensuring balanced representation and systematic variation of factor levels, OAs enhance the efficiency and robustness of experimental designs, making them indispensable tools across various scientific and industrial fields. An excellent textbook in this area is authored by Dey and Mukerjee (1999), offering comprehensive insights into the construction and application of OAs and fractional factorial designs in experimental design.

8. Application of OAs in Taguchi methods

Taguchi methods, pioneered by Japanese engineer and statistician Genichi Taguchi, have profoundly influenced quality engineering and process optimization. These methods prioritize robust design, focusing on making products and processes resistant to variations, thus enhancing quality and performance without significant cost increases. Taguchi methods are extensively applied to improve the quality of manufactured goods and refine product and process design. Central to this approach is the optimization of designs to make them robust against various sources of variation, such as manufacturing inconsistencies or environmental changes. This robustness ensures that products and processes perform consistently under diverse conditions. Robust design in Taguchi methods emphasizes reducing the sensitivity of products to variations by identifying and optimizing controllable factors, thus minimizing the impact of uncontrollable noise factors.

A fundamental aspect of Taguchi methods is the design of experiments, which utilizes OAs, a type of fractional factorial design. These arrays enable the efficient study of multiple factors simultaneously, allowing for the identification of main effects and interactions with a minimal number of experimental runs. This efficiency makes Taguchi methods particularly valuable in industries where improving quality and reducing costs are critical, such as automotive, electronics, telecommunications, and manufacturing. Applications of these methods range from enhancing product design robustness to optimizing process parameters for high quality outputs with minimal variability. In quality improvement, Taguchi methods systematically identify and address sources of defects and inconsistencies in production processes.

Balanced repeated replications (BRRs) are crucial for obtaining reliable and generalizable results in experimental design. Taguchi methods inherently support this through the use of OAs, ensuring balanced and systematic experimentation. These arrays are designed to test each factor level an equal number of times across the experiment, thus preventing data skew from imbalance. Incorporating replication and randomization into the experimental design controls for random variations and ensures that observed effects are due to the studied factors rather than external influences. Analysis of variance is often employed to analyze experimental results, identifying significant factors and interactions, thereby ensuring conclusions are based on balanced and reliable data.

In conclusion, Taguchi methods offer a powerful approach to quality engineering and process optimization by emphasizing robust design and systematic experimental designs. The application of BRRs within these methods ensures that experimental results are reliable and generalizable, making them highly valuable across various industries. The use of OAs allows for the efficient examination of multiple parameters in a condensed set of experiments. Determining optimal parameter levels requires an in depth understanding of the process and the consideration of the cost of conducting experiments. By selecting the appropriate OA, based on the number of parameters and levels, researchers can ensure that each variable and setting is tested equally, thereby achieving reliable and comprehensive experimental results. Key references in this field include works by Gupta *et al.* (1982), Taguchi (1987), Taguchi and Konishi (1987), Kacker *et al.* (1991), Sitter (1993) and Rosa (2017).

9. Other diverse applications of OAs and MOAs in modern research

Venturing beyond the discussed domains, let us delve into the diverse realms where OAs and MOAs leave their lasting impression. From coding theory and cryptography to computer experiments and beyond, OAs and MOAs emerge as indispensable tools, enriching modern research and experimentation with precision and efficiency. Join us in this section as we unravel the intricate tapestry of applications where these mathematical constructs play pivotal roles, shaping the landscape of information science, technology, and the design of experiments.

9.1. Coding theory, cryptography and computer experiments

Coding theory, cryptography and computer experiments are three distinct yet interconnected domains at the intersection of mathematics, computer science, and engineering. Coding theory, a fundamental component of information theory, focuses on the design and analysis of error detecting and error correcting codes essential for reliable data transmission and storage in the presence of noise or errors. By systematically encoding data into a form that can withstand errors, coding theory enhances the robustness and integrity of digital communication systems like telecommunications networks and data storage devices. Cryptography, on the other hand, stands as the guardian of communication and information security, employing sophisticated mathematical techniques and algorithms to develop cryptographic protocols and algorithms, orchestrating secure communication and data storage by encoding sensitive information. Cryptography's paramount mission lies in upholding the pillars of confidentiality, integrity, and authenticity within digital communications, serving as a formidable barrier against unauthorized access and malicious intrusions. For deeper insights, readers can explore the works of Kahn (1996) and Stinson and Paterson (2018), along

with the references cited herein. Computer experiments represent a complementary domain leveraging computational methods, simulations, and modeling techniques to study complex systems and phenomena. By designing and conducting simulations or computational models, researchers can explore the behavior, performance, and characteristics of systems under various conditions, providing a cost effective and efficient means of investigating complex systems, enabling validation of theoretical models, optimization of system designs, and exploration of theoretical concepts across diverse fields from engineering and physics to biology and economics. In summary, coding theory, cryptography, and computer experiments each contribute unique insights and methodologies to the broader landscape of information science and technology, forming essential pillars supporting the development of robust and secure communication systems and the exploration and optimization of complex systems across various domains.

OAs serve as invaluable assets in diverse domains, including coding theory, cryptography, and computer experiments. In coding theory, OAs play a pivotal role in the design and analysis of error correcting codes, crucial for reliable data transmission in communication systems. By systematically varying parameters and configurations, OAs aid in constructing codes that can detect and correct errors efficiently, enhancing the robustness and reliability of communication channels. In cryptography, OAs contribute to the development of secure encryption methods by ensuring the resilience of cryptographic algorithms against various attack vectors. Their systematic approach facilitates the design and testing of cryptographic protocols, strengthening the confidentiality and integrity of sensitive information in digital communications. Furthermore, in computer experiments, OAs provide a structured framework for algorithm testing and simulation studies. By enabling systematic exploration of different algorithmic configurations and scenarios, OAs facilitate comprehensive evaluation and optimization of algorithm performance across diverse computational environments. Through their versatility and systematic variation of factors, OAs play a pivotal role in advancing coding theory, cryptography, and computational research, ensuring the development of robust and efficient solutions in today's digital landscape. For further exploration of these topics, notable references include works by Bose and Shrikhande (1959), Niederreiter (1992), Kahn (1996), Hedayat *et al.* (1999), Massey (2002), Adhikari and Bose (2004), Adhikari *et al.* (2007), Bose and Mukerjee (2006, 2010, 2013), Bose *et al.* (2013) and Stinson and Paterson (2018).

9.2. OA based Latin hypercube designs (OALHDs)

OA based Latin hypercube designs (OALHDs) are advanced statistical tools used in computer experiments to ensure space filling properties, which are crucial for comprehensive exploration of the experimental space. OALHDs combine the strengths of OAs and Latin hypercube sampling, facilitating the creation of experimental designs that uniformly cover the entire parameter space. This uniformity ensures that the design points are spread out evenly, preventing clustering and enhancing the reliability of simulation outcomes. The space filling properties of OALHDs are particularly valuable in computer experiments, where they allow researchers to efficiently sample a wide range of input configurations and explore the performance of complex systems under various conditions. By ensuring a thorough and balanced exploration of the input space, OALHDs help in constructing accurate surrogate models, optimizing system performance, and validating theoretical models. Their application spans

numerous fields, including engineering, physics, and environmental science, where robust and efficient design of experiments is critical for gaining insights into complex phenomena and making informed decisions. For detail, readers may consider Sacks *et al.* (1989), Koehler and Owen (1996) and Lin and Tang (2022).

9.3. OAs as BRR structures for variance estimation

A BRR structure is a sampling design commonly used in survey sampling for variance estimation. In BRR, the sample is divided into several balanced replicates, ensuring that each replicate represents the population equally well. Within each replicate, the same survey weights and adjustments are applied as in the original sample. Variance estimation is then performed by computing the variance across these replicates, taking into account both within replicate and between replicate variations. BRR helps to improve the efficiency and accuracy of variance estimation, especially in complex survey designs where traditional methods may be inadequate.

OAs serve as invaluable tools for variance estimation, particularly in the context of large scale complex survey designs where non linear statistics are involved. Acting as BRR structures, OAs provide a systematic and efficient approach to estimating the variance of non linear statistics derived from survey data. By systematically varying factors and configurations within the survey design, OAs ensure balanced representation and systematic variation, thereby capturing the complexities inherent in the survey data. This balanced approach is crucial for accurately estimating the variance of non linear statistics, which may exhibit complex relationships and interactions among survey variables. Additionally, OAs offer the advantage of reducing the computational burden associated with variance estimation in large scale surveys, allowing for efficient and reliable estimation of variance even in complex survey designs. Overall, the utilization of OAs as BRR structures enhances the precision and robustness of variance estimation methods, thereby improving the reliability of survey data analysis in diverse fields. Notable references in this area include works by Gupta *et al.* (1982), Gupta and Nigam (1987), Wu (1991), Sitter (1993), and Parsad and Gupta (2007).

9.4. Optimum covariate designs

In recent years, the quest for experimental units with precisely defined covariate values to achieve optimal precision in regression parameter estimation has garnered significant interest among researchers. The pioneering work by Troya (1982a, 1982b) introduced the concept of optimal covariates designs (OCDs), laying the groundwork for exploring optimal designs to estimate regression parameters associated with controllable covariates. OCDs, renowned for their capacity to offer the most efficient estimation of covariate effects within a presumed linear model, have emerged as indispensable tools in experimental design. Building upon Troya's ground breaking contributions, Das *et al.* (2003) delved into combinatorial solutions, particularly focusing on the estimability of regression coefficients in randomized block designs and certain series of BIBDs.

Rao *et al.* (2003) further elucidated the construction of OCDs derived from MOAs, unraveling the intrinsic relationship between OCDs and experimental designs like completely randomized designs and randomized block designs, both grounded in MOAs. This revelation not only underscores the versatility of MOAs but also expands their application horizons into

experimental design realms. For an in depth exploration of this captivating subject, Das, Dutta, Mandal, and Sinha (2015) offer a comprehensive textbook, serving as an invaluable reference for enthusiasts and practitioners alike in the domain of experimental design.

9.5. Optimizing super absorbent composites: leveraging OAs

At the Indian Agricultural Research Institute (IARI) in New Delhi, a ground breaking experiment was devised to engineer super absorbent composites with optimized water absorption characteristics and improved stability in plant growth media. The objective is to maximize absorbency while minimizing the concentrations of monomer, cross linker, and alkali. This intricate experiment encompassed a multitude of factors, including the nature and concentration of alkali, duration and temperature of exposure, backbone clay ratio, monomer concentration, cross linker concentration, initiator concentration, volume of water, and more. With 3 factors at 3 levels and 6 factors at 5 levels, the experiment constituted a daunting $3^3 \times 5^6$ factorial design, necessitating a staggering 421,875 runs for a single replication − an impractical endeavor given limited resources.

In light of the experimenter's interest in orthogonal estimation of main effects and constrained resources, a MOA of strength two emerged as a pragmatic solution, slashing the number of runs to a manageable 225. Although sacrificing intra effect orthogonality, the MOA ensured sufficient resolution and interaction detection. Furthermore, modifications to the experimental objectives led to the creation of a $3^5 \times 6^8$ factorial design, accommodating additional factors and selected interactions, all within the confines of 72 runs. The strategic utilization of MOAs empowered the experimenter to efficiently explore a diverse array of factors and interactions while upholding the integrity of the experiment.

Additionally, IASRI has harnessed OAs for orthogonal main effect plans in asymmetrical factorials and for variance estimations in large scale complex survey data. These endeavors underscore the versatility and utility of OAs across diverse experimental settings. For further insights, we encourage readers to explore the institute's websites.

10. Conclusion

This review article pays homage to the enduring legacy and profound contributions of the legendary statistician, C. R. Rao (1920-2023), across the realms of experimental design, information science, technology, and industry. Delving into the intricate interplay of OAs and BAs, this article offers readers profound insights into the transformative influence of these arrays, as conceptualized by Professor Rao. Referencing seminal works by Parsad, Gupta, Gopinath (2020), Rao (2020), Kannan and Kundu (2021), and Peddada and Khattree (2023), it invites deeper exploration into Prof. Rao's extraordinary contributions and his profound impact on statistical sciences.

Professor Rao's visionary insights have indelibly shaped experimental design, combinatorial mathematics, and statistical analysis, profoundly influencing these disciplines. This article meticulously navigates through the multifaceted applications of OAs and BAs, eloquently showcasing their versatility and paramount importance across various domains. From the intricate construction of BIBDs to the precision of optimum chemical balance weighing designs, and from SORDs to Taguchi methods, orthogonal resolution plans, frac-

tional factorial plans, coding theory, cryptography, computer experiments, OALHDs, and OCDs, this article unveils the methodological advancements fostered by BAs, OAs, and MOAs.

Through meticulous examination, it elucidates the nuanced relationships between association schemes, OAs, and BAs, revealing their immense potential in both experimental design and combinatorial theory. While acknowledging the remarkable strides made thus far, the article passionately underscores the imperative for ongoing research endeavors to fully unlock the latent capabilities of these abstract mathematical structures and their practical applications in experimental design. Indeed, further exploration and analysis in this domain hold the promise of ushering in more advanced and potent experimental design techniques and strategies, thereby enriching the fabric of scientific inquiry and discovery.

In this article, we choose not to delve into mathematical intricacies, recognizing the extensive literature available on the subject. Condensing such a vast topic into a few pages presents a daunting task, and we are mindful of the challenges it entails. Nevertheless, our objective remains clear to offer a lucid exposition that captivates readers beyond this specialized field, sparking their curiosity and nurturing a deeper interest in the subject matter.

Acknowledgements

We are deeply grateful to a Guest Editor of the journal for his invaluable guidance in enhancing the initial draft of this review manuscript, as well as for his generous suggestions regarding the references. His insightful input has played a significant role in refining the quality and depth of our work, while his comprehensive list of useful references has enriched the scholarly foundation of our article.

References

Adhikari, A. and Bose, M. (2004). Construction of new visual threshold schemes using combinatorial designs. *IEICE Transactions*, **E87-A**, 1198-1202.

Adhikari, A., Bose, M., Kumar, D., and Roy, B. (2007). Applications of partially balanced incomplete block designs in developing (2,n) Visual cryptographic Schemes. *IEICE Transactions*, **E90-A**, 949-951.

Bose, M., Dey, A., and Mukerjee, R. (2013). Key predistribution schemes distributed sensor networks via block designs. *Design, Codes and Cryptography*, **467**, 111–136.

Bose, M. and Mukerjee, R. (2006). Optimal (2, n) visual cryptographic schemes. *Design, Codes and Cryptography*, **40**, 255-267.

Bose, M. and Mukerjee, R. (2010). Optimal (k, n) visual cryptographic schemes for general k. *Designs, Codes and Cryptography*, **55**, 19–35.

Bose, M. and Mukerjee, R. (2013). Union distinct families of sets, with an application to cryptography. *Ars Combinatoria*, **110**, 179–192.

Bose, R. C. (1950). A note on orthogonal arrays. *The Annals of Mathematical Statistics*, **21**, 304–305.

Bose, R. C. and Bush, K.A. (1952). Orthogonal arrays of strength two and three. *The Annals of Mathematical Statistics*, **23**, 508–524.

Bose, R. C. and Shrikhande, S.S. (1959). A note on result in the theory of code construction. *Information and Control*, **2**, 183–194.

Bush, K. A. (1952). Orthogonal arrays of index unity. *The Annals of Mathematical Statistics*, **23**, 426–434.

Chakravarti, I. M. (1956). Fractional replication in asymmetrical factorial designs and partially balanced arrays. *Sankhya*, **17**, 143–164.

Chakravarti, I. M. (1961). On some methods of construction of partially balanced arrays. *The Annals of Mathematical Statistics*, **32**, 1181–1185.

Cheng, C. S. (1980). Orthogonal arrays with variable numbers of symbols. *Annals of Statistics*, **8**, 447–453.

Das, K., Mandal, N. K., and Sinha, B. K. (2003). Optimal experimental designs for models with covariates. *Journal of Statistical Planning and Inference*, **115**, 273-285.

Das, P., Dutta, G., Mandal, N. K., and Sinha, B. K. (2015). *Optimal Covariate Designs: Theory and Applications*. Springer, New York.

Design Resources Server, *Indian Agricultural Statistics Research Institute (ICAR), New Delhi 110 012, India*. Available at: `https://drs.icar.gov.in/Oarray/oa/default.htm`.

Dey, A. (1971). On some chemical balance weighing designs. *Austrian Journal of Statistics*, **13**, 137–141.

Dey, A. and Mukerjee, R. (1999). *Fractional Factorial Plans*. John Wiley and Sons, New York.

Fuji-Hara, R., Kageyama, S., Kuriki, S., Miao, Y., and Shinohara, S. (2002). Balanced nested designs and balanced arrays. *Discrete Mathematics*, **259**, 91–119.

Gupta, V. K., Nigam, A. K., and Dey, A. (1982). Orthogonal main effect plans for asymmetrical factorials. *Technometrics*, **24**, 135–137.

Gupta, V. K. and Nigam, A. K. (1985). A class of asymmetrical orthogonal resolution IV designs. *Journal of Statistical Planning and Inference*, **12**, 381–383.

Gupta, V. K. and Nigam, A. K. (1987). Mixed orthogonal arrays for variance estimation with unequal numbers of primary selections per stratum. *Biometrika*, **74**, 735–742.

Hedayat, A. S., Sloane, N. J. A., and Stufken, J. (1999). *Orthogonal Arrays: Theory and Applications*. Springer, New York.

Kacker, R. N., Lagergren, E. S., and Filliben, J. J. (1991). Taguchi's Orthogonal Arrays are Classical Designs of Experiments. *Journal of Research of the National Institute of Standards and Technology*, **96**, 577–591.

Kageyama, S. and Saha, G. M. (1983). Note on the construction of optimum chemical balance weighing designs. *Annals of the Institute of Statistical Mathematics*, **35**, Part A, 447–452.

Kahn, D. (1996). *The Codebreakers: The Comprehensive History of Secret Communication from Ancient Times to the Internet*. Scribner, New York.

Kannan, N. and Kundu, D. (2021). C. Radhakrishna Rao: A Century in Statistical Science. *International Statistical Review*, **89**, DOI:10.1111/insr.12467.

Khuri, A. I. and Cornell, J. A. (1996). *Response Surface, Design and Analysis* (2nd ed.). Marcel Dekker Inc., New York.

Kishen, K. (1942). On Latin and hyper-graeco-Latin cubes and hyper-cubes. *Current Science*, **11**, 98–99.

Kishen, K. (1949). On the construction of latin and hyper-graeco-latin cubes and hypercubes. *Journal of the Indian Society of Agricultural Statistics*, **2**,20–48.

Koehler, J. R. and Owen, A. B. (1996). 9 Computer experiments, *Handbook of Statistics*, edited by S. Ghosh and C. R. Rao, Elsevier, Amsterdam.

Kuhfeld, W. F. *TS-DOC: TS-723 - Orthogonal Arrays*, Lists of orthogonal arrays, difference schemes, and experimental design tools maintained by Warren F. Kuhfeld. Available at: `https://support.sas.com/techsup/technote/ts723.html`.

Lin, D. and Tang, B. (2022). Latin Hypercubes and Space-filling Designs, arXiv, eprint: 2203.06334 [Online]. Available: `https://arxiv.org/abs/2203.06334`.

Majindar, K. N. (1978). Coexistence of some B.I.B. Designs. *Canadian Mathematical Bulletin*, **21**, 73–77.

Massey, J. L. (2002). Randomness, Arrays, Differences and Duality. *IEEE Transactions on Information Theory*, **48**, 1698–1703.

Mukhopadhyay, A. C. (1981). Construction of some series of orthogonal arrays. *Sankhya B*, **43**, 81–92.

Niederreiter, H. (1992). Orthogonal arrays and other combinatorial aspects in the theory of uniform point distributions in unit cubes. *Discrete Mathematics*, **106-107**, 361–367.

Parsad, R. and Gupta, V. K. (2007). Variance estimation from complex surveys using balanced repeated replication. `https://shorturl.at/aCOid`.

Parsad, R., Gupta, V. K., and Gopinath, P. P. (2020). *Calyampudi Radhakrishna Rao's Life Sketch and its Influence on Designing of Experiments with a Special Reference to Agricultural Sciences*. ICAR-IASRI, New Delhi, 1–18. `http://krishi.icar.gov.in/jspui/handle/123456789/41295`.

Peddada, S. D. and Khattree, R. (2023). C. R. Rao, statistician who transformed data analytics (1920–2023). *Nature*, Oct; 622(7984):691, doi: 10.1038/d41586-023-03200-5.

Pukelsheim, F. (1993). *Optimal Design of Experiments*. John Wiley, New York.

Rafter, J. A. and Seiden, E. (1974). Contributions to the Theory and Construction of Balanced Arrays. *Annals of Statistics*, **2**, 1256– 1273.

Raghavarao, D. (1971). *Constructions and Combinatorial Problems in Design of Experiments*. Wiley, New York.

Rao, B. L. S. P. (2020). C. R. Rao: A Life in Statistics. *Bhāvanā*, 4, `https://bhavana.org.in/c-r-rao-a-life-in-statistics/`.

Rao, C. R. (1946). Hypercubes of strength "d" leading to confounded designs in factorial experiments. *Bulletin of the Calcutta Mathematical Society*, **38**, 67–78.

Rao, C. R. (1947). Factorial experiments derivable from combinatorial arrangements of arrays. *Journal of the Royal Statistical Society* (Supplement), **9**, 128–139.

Rao, C. R. (1949). On a class of arrangements. *Proceedings of the Edinburgh Mathematical Society*, **8**, 119–125.

Rao, C. R. (1961). Combinatorial arrangements analogous to orthogonal arrays. *Sankhya A*, **23**, 283–286.

Rao, C. R. (1973). Some combinatorial problems of arrays and applications to design of experiments. 349-359 of: Srivastava, J. N. (ed.), *A Survey of Combinatorial Theory*, North Holland.

Rao, P. S. S. N. V. P., Rao, S. B., Saha, G. M., and Sinha, B. K. (2003). Optimal designs for covariates' models and mixed orthogonal arrays. *Electronic Notes in Discrete Mathematics*, **15**, 157–160.

Rosa, P. J. (2017). *Taguchi Techniques for Quality Engineering* (2nd ed.). McGraw Hill.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and Analysis of Computer Experiments, *Statistical Science*, **4**, 409–423.

Saha, G. M. and Das, A. R. (1973). Four level second order rotatable designs from partially balanced arrays. *Journal of the Indian Society of Agricultural Statistics*, **25**, 97–102.

Saha, G. M. (1975). Some results in tactical configurations and related topics. *Utilitas Mathematica*, **7**, 223–240.

Saha, G. M. (1975). A note on relations between incomplete block and weighing designs. *Annals of the Institute of Statistical Mathematics*, **27**, 387–390.

Saha, G. M. (1981). Balanced arrays from association schemes and some related results. In: Rao, S.B. (eds) Combinatorics and Graph Theory. *Lecture Notes in Mathematics*, vol 885. Springer, Berlin, Heidelberg.

Saha, G. M. and Kageyama, S. (1984). Balanced arrays and weighing designs. *Austrian Journal of Statistics*, **26**, 119–124.

Saha, G. M. and Samanta, B. K. (1985). A construction of balanced arrays of strength t and some related incomplete block designs. *Annals of the Institute of Statistical Mathematics*, **37**, 337–345.

Shah, K. R. and Sinha, B. K. (1989). *Theory of Optimal Designs*. Lecture Notes in Statistics, vol 54. Springer, New York.

Silvey, S. D. (1980). *Optimal Design*. Chapman and Hall, London.

Sinha, K., Dhar, V., Saha, G. M., and Kageyama, S. (2002). Balanced arrays of strength two from block designs. *Journal of Combinatorial Designs*, **10**, 303–312.

Sitter, R. R. (1993). Balanced repeated replications based on orthogonal multi-arrays. *Biometrika*, **80**, 211–221.

Sloane, N. J. A. (2007). *A Library of Orthogonal Arrays*. Available at: `http://neilsloane.com/oadir/`.

Srivastava, J. N. and Chopra, D. V. (1973). Balanced arrays and orthogonal arrays. *A Survey of Combinatorial Theory*, 411–428, North Holland, Amsterdam.

Stinson, D. R. and Paterson, M. B. (2018). *Cryptography: Theory and Practice*. 4th ed. CRC Press.

Taguchi, G. (1986). *Introduction to Quality Engineering: Design Quality into Products and Processes*. Tokyo: Asian Productivity Organization.

Taguchi, G. (1987). *System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Costs*. MI, American Supplier Institute Inc.

Taguchi, G. and Konishi, S. (1987). *Taguchi Methods Orthogonal Arrays and Linear Graphs: Tools for Quality Engineering*. American Supplier Institute, Dearborn, Michigan.

Troya, L. J. (1982a). Optimal designs for covariates models. *Journal of Statistical Planning and Inference*, **6**, 373-419.

Troya, L. J (1982b). Cyclic designs for covariates models. *Journal of Statistical Planning and Inference*, **7**, 49-75.

Vanstone, S. A. (1975). A note on a construction of b.i.b.d.'s. *Utilitas Mathematica*, **7**, 321–322.

Wikipedia, *Orthogonal array testing - Wikipedia, the free encyclopedia*, Orthogonal array testing is a systematic, statistical way of testing. Orthogonal arrays can be applied in user interface testing, system testing. Available at: `https://en.wikipedia.org/wiki/Orthogonal_array_testing`.

Wu, C. F. J. (1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika*, **78**, 181–188.

Yonglin, Z. (2004). *Combinatorial Designs via Association Scheme*. Ph.D. Thesis, Hong Kong Baptist University, Hong Kong.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 171–197
http://www.ssca.org.in/journal

Model-Free Data Cleaning for Raw Data: An Eigen-Structure Approach

Ravindra Khattree
Department of Mathematics and Statistics
Center for Data Science and Big Data Analytics
Center for Biomedical Research
Oakland University, Rochester, MI, 48309, USA

Received: 13 June 2024; Revised: 10 July 2024; Accepted: 11 July 2024

Abstract

Preprocessing of data at the initial stages before assuming any model for the data is a necessary requirement for observational data. With preliminary data cleaning of raw data in mind, we introduce a model-free approach based on the eigen-structure of the data matrix to assess if a particular observation induces multicollinearity or is excessively outlying within the data. Specifically we look at the eigenvalues and antieigenvalues obtained from the singular value decomposition of the data matrix or a function thereof. We also study detection of the outlier induced multicollinearity or outlier induced masking of multicollinearity present in the data. Usefulness of our approach is illustrated via several examples describing a variety of situations and for several classical data sets. Emphasis is on data matrix of variables rather than model matrix, although these approaches can be later used in model based contexts as well.

Key words: Antieigenvalue; Condition indexes; Eccentricity; Emphasis measure; Multicollinearity; Outlying observations.

AMS Subject Classifications: 62J05, 62P99

1. Introduction

Prof. C. R. Rao was my PhD adviser at the University of Pittsburgh. His teaching and research both have continued to have a lasting impact on my own academic career. Throughout his classes, there was always an implicit but very definite message that any research in statistics should have a definite purpose of understanding and solving some meaningful and practical problem. Reflecting on this philosophy of Prof. Rao, this article on a very fundamental first step of data processing is written as a personal tribute to Prof. Rao and with a purpose of honoring his legacy and place in the world of statistics and science.

Preprocessing of data and data cleaning are essential steps in observational studies and may involve the steps of detecting freak values, identification of outlying observations,

Corresponding Author: Ravindra Khattree
Email: khattree@oakland.edu

choosing the meaningful variables and understanding the underlying dependence among various variables. Inference can be greatly distorted due to some or all of the above issues and a substantial body of work has been done in the past to remedy such problems. See for example, Belsley, Kuh and Welsch (1980), Cook and Weisberg (1982), Belsley (1991), Khattree and Naik (1999), Seber and Lee (2003) and Khattree (2019). All of the above referenced work, except that by Khattree (2019), deal with various model based approaches. While these are perfectly valid approaches for diagnostics, all being model based, they however, tend to not look into the very basic structure of the raw data. We think that in order to gain full understanding of and more insight into data, we must also look at the anomalies in the data at the very fundamental level, the fundamental level being the patterns and differences at the level of $\mathbf{X}'\mathbf{X}$ matrix, where $\mathbf{X}$ is our data matrix of all variables under consideration. This is a fundamental issue in data science, taking a precedence over any subsequent statistical modeling. This article is a step towards addressing this problem by looking at the eigen-structure of the data itself.

In principle and in general, $\mathbf{X}$ may be either a data matrix or a model matrix (in which case, we will use the notation $\mathbf{X}^*$ to distinguish it from the raw data matrix). We here assume all variables to be quantitative. Since the data cleaning at the preprocessing stages must assume no specific model and no specific prespecified choice of a few selected variables, data matrix is a more appropriate context for our work. Therefore it is meaningful that we rely more on the mathematical structure of the data matrix $\mathbf{X}$ than on statistical evaluation of the model to be fitted. This is especially relevant because for large data sets at preprocessing stages, data cleaning is equally important for the explanatory variables as well as response variables and our data matrix may contain both types of variables. This is the approach adopted by Wang and Nyquist (1991) and Khattree (2019). Other approaches not exclusively based on matrix structure or model but based on various other tentative techniques such as aggregate queries are given by Chu *et. al* (2016), Chu and Ilyas (2016) and Ilyas and Chu (2019).

As indicated, our approach here will rely on an evaluation of the eigen-structures of the matrices which are closely related to the singular value decomposition of the whole or parts of the data matrix. We will focus on the evaluation of multicollinearity and the detection of outlying observations by evaluating the changes or deviations in these eigen-structures by the use of eigenvalues and antieigenvalues. While theory of eigenvalues is well established, recent discussions of antieigenvalues along with various applications thereof are available in Khattree (2001, 2002, 2003, 2006, 2010, 2014, 2019) and in Tran and Khattree (2024). Applications have also been presented in Khattree and Bahuguna (2019), Cuntoor and Chellappa (2006) and Guo *et al.* (2018).

We must emphasize that data matrix consists of raw data on variables and not on the mathematical functions thereof. To press that point, although we will not rely on it, suppose the framework was the standard linear model, namely,

$$\mathbf{y}_{n\times 1} = \mathbf{X}^*_{n\times p^*}\beta^*_{p^*\times 1} + \epsilon_{n\times 1} = \mathbf{X}_{n\times p}\beta_{p\times 1} + \mathbf{Z}_{n\times q}\gamma_{q\times 1} + \epsilon_{n\times 1},$$

and suppose the data cleaning was confined to only data on the explanatory variables. The matrix $\mathbf{X}$ may then represent the raw data matrix and $\mathbf{Z}$ contains columns corresponding to other $q(= p^* - p)$ terms in the model such as intercept, and specific mathematical transfor-

mations of columns of $\mathbf{X}$ *e.g.* polynomial or cross products terms. When the type of true model (in terms of which variables, which degree of polynomial or which cross product terms) and/or its dimension are unknown, our interest should be exclusively in the response variable and in the matrix of raw data on explanatory variables namely, $\mathbf{X}$ and what specific model will subsequently be used will be a secondary consideration for a later time. In that sense, we are interested in measuring the multicollinearity and outlying nature of observations within the raw data and we do not concern ourselves as to what specific model is being assumed. The problem of *model induced outlyingness* due to outliers will involve the *model* matrix $\mathbf{X}^* = [\mathbf{X} : \mathbf{Z}]$ and the corresponding response variable. A version of this latter problem, albeit with a very different approach has been discussed in Mason and Gunst (1985).

Clearly, at initial stages, from data cleaning point of view only the former context is relevant and should always take precedence over modeling and model selection. Further, in big data context, one encounters a very large number of observations and a large number of variables, both of which are, presumably, to be used in several different modeling problems in the future. It is thus imperative to have a clean data where "cleanliness" of the data may refer to general robustness of the data, with respect to specific observations and/or specific variables. This is the situation, where we believe our approach has the most currency.

In Section 2, we motivate antieigenvalues of a positive definite matrix as a way to look into the eigen-structure and as the measures of eccentricities of an ellipsoid for various cross-sections which in turn provide us a way to measure the interdependences between a set of variables or multicollinearity.

Section 3 is about identification of outlying observations via antieigenvalues. Towards the end of this section, we also discuss the issues pertaining to $collinearity - outlyingness$ where one of these two may cause the other. We provide illustrations of our approaches using several data sets. Admittedly, to make the understanding of the approach more accessible, we must use data sets which are not excessively large and are readily available. However, that does not disqualify our approach for bigger data sets. In fact, those are the situations where once computationally implemented, these methods will have most utility. Thus in Section 4, we also consider a relatively larger data set consisting of 1599 observations on the quality of red wine. We apply our procedure on this data to make a point that procedure is effective even when we have large data sets and that our approach is able to successfully pick out the observations which are not easy to identify otherwise but whose presence excessively corrupts the data and subsequently also affects the modeling steps. Section 5 provides some concluding remarks.

2. Eccentricities and measurement of multicollinearity

Let, as earlier, $\mathbf{X}_{n\times p}$ be a data matrix. Assume $rank(\mathbf{X}_{n\times p}) = p$ and let $\mathbf{A} = \mathbf{X}'\mathbf{X}$. Consider the quadratic surface, $\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} = c$ where c is a known constant in a $p-$ dimensional space. Since $\mathbf{A}$ is positive definite, this represents an ellipsoid and with an appropriate orthogonal rotation $\mathbf{v} = \mathbf{P}'\mathbf{u}$ where $\mathbf{A} = \mathbf{P}\Lambda\mathbf{P}'$ is the spectral decomposition of $\mathbf{A}$, the surface can be represented as,

$$\mathbf{v}'\Lambda^{-1}\mathbf{v} = c \text{ with } \Lambda = diagonal(\lambda_1, \lambda_2, ..., \lambda_p)$$

or

$$\frac{v_1^2}{\lambda_1} + \frac{v_2^2}{\lambda_2} + ... + \frac{v_p^2}{\lambda_p} = c \text{ where } \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0.$$

The eccentricities of certain two dimensional elliptical cross-sections of this ellipsoid can be quantified in decreasing order as $\sqrt{\frac{\lambda_1}{\lambda_p}} \geq \sqrt{\frac{\lambda_2}{\lambda_{p-1}}} \geq \sqrt{\frac{\lambda_3}{\lambda_{p-2}}} \geq \cdots$. The quantity $e_1 = \sqrt{\frac{\lambda_1}{\lambda_p}}$ is the eccentricity measured respectively via the two most elongated and most compressed directions and hence measures the extreme eccentricity. The next quantity $e_2 = \sqrt{\frac{\lambda_2}{\lambda_{p-1}}}$ represents the comparison of the next two most elongated and most compressed directions and similar comparisons continue for $r = [p/2]$ pairs where $[p/2]$ is the integer part of $p/2$. Clearly, with λ_i considerably larger than $\lambda_{p-i+1}, i = 1, 2, \cdots, r$, e_i will also be large, indicating a particular cross section of the ellipsoid highly elongated thereby indicating the high multicollinearity. A one-to-one monotonically decreasing function of $e_i = \sqrt{\frac{\lambda_i}{\lambda_{p-i+1}}}$ is the i^{th} *antieigenvalue* of the matrix $\mathbf{X}$'$\mathbf{X}$ namely,

$$\eta_i = \frac{2\sqrt{\lambda_i \lambda_{p-i+1}}}{\lambda_i + \lambda_{p-i+1}} = \frac{2}{e_i + e_i^{-1}}, i = 1, 2, ..., r = [p/2]. \tag{1}$$

It can be shown that $0 < \eta_1 \leq \eta_2 \leq ... \leq 1$ are ordered by their magnitudes. Being a monotonic function of $e_i, i = 1, 2, \cdots, r$, these also measure the eccentricities and hence the multicollinearity in the data. To connect this unfamiliar quantity to a familiar context, values close to zero for at least one of the antieigenvalues indicate high multicollinearity while higher values (close to 1) of all $\eta_i, i = 1, 2, \cdots, r$ indicate a lack of multicollinearity. Also, the most ideal situation namely, $\eta_1 = 1$, implies that the matrix $\mathbf{A} = \mathbf{X}$'$\mathbf{X}$ is orthogonal and then there is absolutely no multicollinearity among the columns of $\mathbf{X}$. Further, greater the number of η_i that are close to zero, higher is the number of linear near-dependencies that may exist.

Note that $\mathbf{X}$'$\mathbf{X}$ and $(\mathbf{X}'\mathbf{X})^{-1}$ share the same set of antieigenvalues and hence these r measures of multicollinearity for the two matrices are equal. It is, in some way, a reasonable and desirable property in that, multicollinearity, being synonymous to ill-conditioning of a given matrix, indicates a *computational difficulty* in obtaining an accurate inverse matrix. Intuitively, this computational difficulty should be same for the matrix $\mathbf{A}$ as well as for its *true* inverse $\mathbf{A}^{-1}$ because their eigen-structures are directly related (i^{th} ordered eigenvalue of inverse of a matrix is the reciprocal of the $(p - i + 1)^{th}$ ordered eigenvalue of the original matrix).

A single index of multicollinearity combining all antieigenvalues defined in Equation (1) can be defined as the generalized antieigenvalue (See (Khattree, 2002, 2003)),

$$\Delta = \prod_{i=1}^{r} \frac{2\sqrt{\lambda_i \lambda_{p-i+1}}}{\lambda_i + \lambda_{p-i+1}} = \prod_{i=1}^{r} \eta_i, \text{where } r = [p/2], \tag{2}$$

which is a function of all antieigenvalues and can be interpreted as an overall measure of eccentricity. Clearly Δ is also same for $\mathbf{X}$'$\mathbf{X}$ and $(\mathbf{X}'\mathbf{X})^{-1}$. One may alternatively use the r^{th} root of Δ which would then be the geometric mean of all antieigenvalues.

Belsley, Kuh and Welsch (1980) and Belsley (1991) suggest to look at the *condition number* $\psi = \sqrt{\frac{\lambda_1}{\lambda_p}}$. They also look at the *condition indexes* $\psi_2, \psi_3, \cdots, \psi_p$, where $\psi_i = \sqrt{\frac{\lambda_1}{\lambda_i}}, i = 2, 3, \cdots, p$. Larger values are indicative of possible multicollinearity. Note that $\psi_2, \psi_3, .., \psi_p (= \psi$, the condition number) are all greater than or equal to 1 with no upper bound specified. There is apparently no way to decide what constitutes a large condition index/number. That aside, unlike the sets of antieigenvalues, the two sets of condition indexes, – for $\mathbf{X}$'$\mathbf{X}$ and for $(\mathbf{X}$'$\mathbf{X})^{-1}$, – are different from each others. Specifically these are $\{\sqrt{\lambda_1/\lambda_2}] \leq \sqrt{\lambda_1/\lambda_3} \leq ... \leq \sqrt{\lambda_1/\lambda_p}\}$ and $\{\sqrt{\lambda_{p-1}/\lambda_p}] \leq \sqrt{\lambda_{p-2}/\lambda_p} \leq ... \leq \sqrt{\lambda_1/\lambda_p}\}$ respectively.

How useful and practical are the indexes defined in Equations (1) and (2)? This can be best explained and demonstrated by applying them on real data sets. We will thus illustrate the utility of these indexes by first applying them to four data sets of varying sizes, with different number of variables and of varying features. Specifically, we consider,

(*i*) A data set on properties of soil given by Kendall (1975) with $p = 4, n = 20$. The data collected here is a set of 20 samples of soil for each of which salt content (x_1) clay content (x_2), organic matter (x_3) and acidity on pH scale (x_4) are measured.

(*ii*) A data set by Daniel and Wood (1980) on clinkers with $p = 5, n = 14$. This data set on clinker compounds was collected with a purpose to study their effects on amount and rate at which heat evolves during cement hardening. The independent variables are weight percent of $SiO_2(x_1)$, Al_2O_3 (x_2), Fe_2O_3 (x_3), CaO (x_4) and $MgO(x_5)$. The data are compositional in that ideally their sum should add to hundred percent. However possibly due to impurities and also due to round off errors, these variables do not add to hundred percent (in many cases, in fact, they add to much more than hundred percent and thus cannot be fully explained by round off errors). This data set has been extensively analyzed by Chatterjee and Hadi (1988).

(*iii*) A data set due to Chatterjee, Hadi and Price (2006) with $p = 6, n = 40$. This data set with six predictors (x_1 through x_6) is given in Chatterjee, Hadi and Price (2006) as Table 4.8 (p. 128) and as part of Exercises 4.12-4.14. No detailed description is available. However, data set was used to illustrate the strong presence of multicollinearity.

(*iv*) A data set due to Rao (1948) on cork deposits with $p = 4, n = 28$. This classic data, more easily available in Khattree and Naik (1999), pertains to the cork deposits in four directions (North, East, South and West), denoted respectively by x_1 through x_4 on twenty eight trees in the Himalayan range. These latter authors have extensively studied this data set in various contexts including the detection of outlying observations.

Later in Section 4, we consider a very large dataset as well. The above four data sets, being of manageable size to include here, are given in the appropriate columns of Tables 1-4. In each case, the question is, how well behaved, with respect to multicollinearity, the particular data set is. Thus, we calculate the antieigenvalues η_i as well as generalized antieigenvalue Δ in each case. We will work with r^{th} root of the generalized antieigenvalue as it is essential to bring this measure on equal footing when comparing multicollinearities of various data sets with different number of variables and this measure, as the geometric mean

of all antieigenvalues, does so. To make results more readable, all values of antieigenvalues and generalized antieigenvalue in various tables are multiplied by 100. Smaller values indicate more severe presence of multicollinearity.

Table 1: Detecting multicollinearity, raw data and antieigenvalues η_i values, generalized antieigenvalue Δ and $\Delta^{1/r}$ of $\mathbf{X}_{(-j)}$'$\mathbf{X}_{(-j)}$ [Kendall's data, $r = 2$.]

Deleted Obs. (j)	x_1	x_2	x_3	x_4	η_1	η_2	Δ	$\Delta^{1/r}$
none	**.**	**.**	**.**	**.**	**4.58**	**93.42**	**4.28**	**20.70**
1	13.0	9.7	1.5	6.4	4.56	95.16	4.34	20.82
2	10.0	7.5	1.5	6.5	4.47	94.49	4.22	20.55
3	20.6	12.5	2.3	7.0	4.68	94.39	4.42	21.02
4	**33.8**	**19.0**	**2.8**	**5.8**	**4.15**	**87.77**	**3.64**	**19.07**
5	20.5	14.2	1.9	6.9	4.69	94.68	4.44	21.06
6	10.0	6.7	2.2	7.0	4.58	93.18	4.26	20.65
7	12.7	5.7	2.9	6.7	4.61	91.43	4.22	20.53
8	36.5	15.7	2.3	7.2	4.78	92.81	4.44	21.07
9	37.1	14.3	2.1	7.2	4.58	93.91	4.30	20.74
10	25.5	12.9	1.9	7.3	4.58	93.54	4.29	20.70
11	26.5	14.9	2.4	6.7	4.72	93.07	4.39	20.95
12	22.3	8.4	4.0	7.0	4.39	92.18	4.05	20.12
13	30.8	7.4	2.7	6.4	4.59	95.04	4.36	20.88
14	**25.3**	**7.0**	**4.8**	**7.3**	**4.11**	**90.56**	**3.72**	**19.28**
15	31.2	11.6	2.4	6.5	4.72	94.03	4.44	21.08
16	22.7	10.1	3.3	6.2	4.48	93.34	4.19	20.46
17	31.2	9.6	2.4	6.0	4.69	95.13	4.46	21.11
18	13.2	6.6	2.0	5.8	4.61	93.12	4.29	20.72
19	11.1	6.7	2.2	7.2	4.54	92.75	4.22	20.53
20	20.7	9.6	3.1	5.9	4.48	93.38	4.19	20.46

Remark: Most-outlying observations are highlighted in **bold**. *Top row corresponds to entire data with no deletion. All originally calculated statistics are multiplied by 100.*

Table 5 presents the values of all antieigenvalues along with r^{th} root of generalized antieigenvalue. Based on first antieigenvalue as well as on the r^{th} root of generalized antieigenvalue, the Rao's data largely seems to be relatively well behaved. The Chatterjee, Hadi and Price' data set appears to be suffering from very severe multicollinearity issues. Other two data sets fall in between. The data set by Daniel and Wood does exhibit a certain degree of multicollinearity and reasons for its presence are extensively discussed in Chatterjee and Hadi (1988).

What if the data sets were standardized prior to fitting the model? Needless to say that eigenvalues and hence the antieigenvalues will change. Does that in any way distort the picture in terms of multicollinearity? There is no reason to expect an answer one way or the other since standardization eliminates the differences among variables in terms of degree of variability in relative terms. See Naik and Khattree (1996), Timm (2002) and Johnson and

Wichern (2014) for extensive discussions on this aspect of the data. The standardization would certainly affect the eccentricities of the ellipsoid. Thus, we suggest that one should also analyze the standardized version of $\mathbf{X}$'$\mathbf{X}$ matrix. Table 6 presents the antieigenvalues for the four data sets in this case. While the conclusions are more or less same as those for original unstandardized data, we do notice that in each case corresponding antieigenvalues are larger except in case of cork data and for η_2. Understandably, scaling makes the $\mathbf{X}$'$\mathbf{X}$ matrix more "spherical" compared to what it was for the unscaled data. Thus, standardization seems to help in the sense that standardized data seem to exhibit less multicollinearity.

Note that in this case the most "well behaved" data set among the four is that by Kendall. First and second antieigenvalues as well as the generalized antieigenvalue are all highest for this data set. Rao's cork data has next highest first antieigenvalue as well as the generalized antieigenvalue. As earlier, the data set by Chatterjee, Hadi and Price exhibits a very severe case of multicollinearity as seen by small values.

Table 2: Detecting multicollinearity, raw data and antieigenvalues η_i values, generalized antieigenvalue Δ and $\Delta^{1/r}$ of $\mathbf{X}_{(-j)}$'$\mathbf{X}_{(-j)}$ [Daniel and Wood's data, $r = 2$.]

Deleted Obs. (j)	x_1	x_2	x_3	x_4	x_5	η_1	η_2	Δ	$\Delta^{1/r}$
none	**.**	**.**	**.**	**.**	**.**	**1.95**	**60.77**	**1.18**	**10.88**
1	27.68	3.76	1.98	64.97	2.48	2.02	61.66	1.25	11.17
2	25.96	3.48	5.06	63.15	2.32	2.02	64.30	1.30	11.40
3	**21.86**	**5.75**	**2.77**	**65.02**	**5.04**	**0.25**	**61.92**	**0.16**	**3.96**
4	24.60	5.85	2.80	64.18	2.40	2.02	58.48	1.18	10.86
5	25.04	3.86	2.11	66.57	2.36	1.98	55.75	1.10	10.51
6	22.32	6.17	2.85	66.47	2.43	2.01	62.00	1.24	11.16
7	20.93	4.64	5.74	66.26	2.08	1.95	56.69	1.10	10.51
8	23.54	4.83	7.21	62.03	2.24	2.01	60.36	1.22	11.03
9	21.96	4.65	6.06	64.07	2.32	2.02	60.32	1.22	11.04
10	**21.44**	**8.81**	**1.19**	**66.64**	**2.48**	**1.64**	**56.22**	**0.92**	**9.60**
11	22.48	5.00	7.46	62.72	2.24	2.02	60.39	1.22	11.04
12	21.34	6.07	2.93	67.03	2.56	2.01	62.79	1.26	11.25
13	21.94	5.57	2.68	67.71	2.44	1.99	60.79	1.21	11.01
14	25.72	4.12	6.06	61.05	2.08	2.02	63.92	1.29	11.35

Remark: Most-outlying observations are highlighted in **bold**. *Top row corresponds to entire data with no deletion. All originally calculated statistics are multiplied by 100.*

3. Detection of outlying observations

One of the major investigations during the preprocessing and data cleaning is to identify outlying observations. In a model based approach, this problem is usually dealt by calculating the *leverage values* ($= \mathbf{x}_i$'$(\mathbf{X}$'$\mathbf{X})^{-1}\mathbf{x}_i$ where $\mathbf{x}_i$' is the i^{th} observation, $i = 1, 2, \cdots, n$) of each of the n observations. Investigations in Wang and Nyquist (1991) and Khattree (2019) look at the problem in terms of effect of outlyingness of the observation on the eigen-structure of the data matrix in the sense how it affects the eigen-structure of the

Table 3: Detecting multicollinearity, raw data and antieigenvalues η_i values, generalized antieigenvalue Δ and $\Delta^{1/r}$ of $\mathbf{X}_{(-j)}$'$\mathbf{X}_{(-j)}$ [Chatterjee, Hadi and Price's data, $r = 3$.]

Deleted Obs. (j)	x_1	x_2	x_3	x_4	x_5	x_6	η_1	η_2	η_3	Δ	$\Delta^{1/r}$
none	**.**	**.**	**.**	**.**	**.**	**.**	**0.36**	**42.61**	**56.48**	**0.09**	**9.56**
1	49	79	76	8	15	205	0.37	41.81	57.24	0.09	9.57
2	27	70	31	6	6	129	0.36	43.67	56.82	0.09	9.66
3	115	92	130	0	9	339	0.37	40.03	53.74	0.08	9.25
4	92	62	92	5	8	247	0.37	42.65	56.19	0.09	9.60
5	67	42	94	16	3	202	0.34	42.34	56.62	0.08	9.38
6	31	54	34	14	11	119	0.36	42.86	54.86	0.08	9.47
7	105	60	47	5	10	212	0.36	42.02	59.85	0.09	9.67
8	114	85	84	17	20	285	0.37	42.14	55.16	0.09	9.50
9	98	72	71	12	-1	242	0.37	41.34	56.88	0.09	9.53
10	15	59	99	15	11	174	0.36	42.65	62.72	0.10	9.92
11	62	62	81	9	1	207	0.36	42.31	55.97	0.08	9.46
12	25	11	7	9	9	45	0.35	42.61	54.53	0.08	9.32
13	45	65	84	19	13	195	0.37	42.37	56.33	0.09	9.56
14	92	75	63	9	20	232	0.36	40.48	56.10	0.08	9.38
15	27	26	82	4	17	134	0.35	40.93	58.02	0.08	9.37
16	111	52	93	11	13	256	0.36	43.29	57.51	0.09	9.64
17	78	102	84	5	7	266	0.36	42.65	55.44	0.09	9.50
18	106	87	82	18	7	276	0.37	41.85	56.95	0.09	9.59
19	97	98	71	12	8	266	0.36	43.00	56.86	0.09	9.59
20	67	65	62	13	12	196	0.36	42.62	55.82	0.09	9.50
21	38	26	44	10	8	110	0.35	42.72	55.83	0.08	9.44
22	56	32	99	16	8	188	0.37	43.83	56.94	0.09	9.70
23	54	100	50	11	15	205	0.37	43.93	56.16	0.09	9.67
24	53	55	60	8	0	170	0.35	42.20	56.09	0.08	9.43
25	61	53	79	6	5	193	0.36	42.80	56.23	0.09	9.56
26	60	108	104	17	8	273	0.37	42.33	58.79	0.09	9.73
27	83	78	71	11	8	233	0.37	42.65	56.59	0.09	9.61
28	74	125	66	16	4	265	0.36	44.08	56.54	0.09	9.64
29	89	121	71	8	8	283	0.37	44.22	55.92	0.09	9.67
30	64	30	81	10	10	176	0.37	43.82	56.24	0.09	9.66
31	34	44	65	7	9	143	0.36	42.56	57.09	0.09	9.59
32	71	34	56	8	9	162	0.36	42.83	56.76	0.09	9.61
33	88	30	87	13	0	207	0.36	42.86	57.01	0.09	9.56
34	112	105	123	5	12	340	0.36	41.17	55.36	0.08	9.39
35	57	69	72	5	4	200	0.36	42.60	55.92	0.08	9.47
36	61	35	55	13	0	152	0.36	41.40	56.64	0.09	9.49
37	29	45	47	13	13	123	0.35	42.61	55.07	0.08	9.40
38	82	105	81	20	9	268	0.36	42.32	56.29	0.09	9.49
39	80	55	61	11	1	197	0.37	41.95	56.90	0.09	9.56
40	82	88	54	14	7	225	0.37	42.96	56.76	0.09	9.64

Remark: Most-outlying (none for this data) observations are highlighted in **bold**. *Top row corresponds to entire data with no deletion. All originally calculated statistics are multiplied by 100.*

Table 4: Detecting multicollinearity, raw data and antieigenvalues η_i values, generalized antieigenvalue Δ and $\Delta^{1/r}$ of $\mathbf{X}_{(-j)}$'$\mathbf{X}_{(-j)}$ [C. R. Rao's Cork data]

Deleted Obs. (j)	x_1	x_2	x_3	x_4	η_1	η_2	Δ	$\Delta^{1/r}$
none	**.**	**.**	**.**	**.**	**8.83**	**90.24**	**7.97**	**28.23**
1	72	66	76	77	8.69	90.68	7.88	28.08
2	60	53	66	63	8.94	91.05	8.14	28.53
3	56	57	64	58	8.76	90.05	7.89	28.09
4	41	29	36	38	8.89	89.38	7.94	28.18
5	32	32	35	36	8.70	90.30	7.86	28.03
6	30	35	34	26	8.85	89.65	7.94	28.17
7	39	39	31	27	8.89	91.35	8.12	28.49
8	42	43	31	25	8.82	92.63	8.17	28.58
9	37	40	31	25	8.89	91.62	8.15	28.54
10	33	29	27	36	8.60	88.95	7.65	27.66
11	32	30	34	28	8.89	90.11	8.01	28.30
12	63	45	74	63	8.95	93.76	8.39	28.96
13	54	46	60	52	9.01	90.69	8.17	28.58
14	47	51	52	43	8.88	89.60	7.96	28.21
15	91	79	100	75	8.53	88.95	7.59	27.55
16	**56**	**68**	**47**	**50**	**8.09**	**93.28**	**7.54**	**27.47**
17	79	65	70	61	8.76	90.17	7.90	28.10
18	**81**	**80**	**68**	**58**	**9.06**	**93.89**	**8.51**	**29.17**
19	**78**	**55**	**67**	**60**	**8.05**	**89.64**	**7.21**	**26.86**
20	46	38	37	38	8.91	89.41	7.96	28.22
21	39	35	34	37	8.87	89.67	7.95	28.20
22	32	30	30	32	8.82	90.04	7.94	28.18
23	60	50	67	54	8.96	90.37	8.09	28.45
24	35	37	48	39	8.73	89.43	7.81	27.94
25	39	36	39	31	8.88	90.15	8.01	28.30
26	50	34	37	40	8.55	88.24	7.54	27.46
27	43	37	39	50	8.26	89.09	7.36	27.14
28	48	54	57	43	8.85	88.07	7.79	27.92

Remark: Most-outlying observations are highlighted in **bold**. *Top row corresponds to entire data with no deletion. All originally calculated statistics are multiplied by 100.*

Table 5: Antieigenvalues η_i values, generalized antieigenvalue Δ and $\Delta^{1/r}$: Unstandardized (complete) data. All originally calculated statistics are multiplied by 100.

Data Set	η_1	η_2	η_3	Δ	$\Delta^{1/r}$
Kendall	4.58	93.42	.	4.28	20.70
Daniel and Wood	1.95	60.77	.	1.18	10.88
Chatterjee, Hadi, Price	0.36	42.61	56.48	0.09	9.56
Rao	8.83	90.24	.	7.97	28.23

Table 6: Antieigenvalues η_i values, generalized antieigenvalue Δ and $\Delta^{1/r}$: Standardized (complete) data. All originally calculated statistics are multiplied by 100.

Data Set	η_1	η_2	η_3	Δ	$\Delta^{1/r}$
Kendall	63.84	99.61	.	63.59	79.74
Daniel and Wood	11.27	90.58	.	10.21	31.96
Chatterjee, Hadi, Price	0.00	59.36	93.36	0.00	0.13
Rao	27.25	85.42	.	23.28	48.25

data matrix. We here consider the problem in terms of the sensitivities of the antieigenvalues of $\mathbf{X}$'$\mathbf{X}$ matrix to a particular observation. The premise is that when an observation is outlying, it may be due to considerable changes in the values of one or more variables (or combinations thereof) that will affect the usual pattern among the variables. This will in turn show up in the diagonal and non-diagonal elements of the $\mathbf{X}$'$\mathbf{X}$ matrix. Such changes will then have an effect on the eccentricities of the corresponding ellipsoid. Accordingly, by reverse logic, if the i^{th} observation is not outlying then if $\mathbf{X}_{(-i)}$ is the corresponding $(n-1) \times p$ data matrix obtained by discarding the i^{th} observation from $\mathbf{X}$ matrix, the $p \times p$ matrices $\mathbf{X}$'$\mathbf{X}$ and $\mathbf{X}_{(-i)}$'$\mathbf{X}_{(-i)}$ must not be too different from each other in terms of their eccentricities. Therefore we compare the antieigenvalues of these two matrices for every i and identify the observations to which these eccentricities are very sensitive. It must be emphasized that depending on the situation, the effect of an outlying observation may manifest on different antieigenvalues and hence one must ideally consider all antieigenvalues as well as the generalized antieigenvalue.

To illustrate the procedure, we return to our four data sets discussed earlier. Raw unscaled data are used in each case. For the sake of easy comparison, Tables 1-4 each present the original data along with the antieigenvalues when the particular observation has been deleted. As earlier, all antieigenvalues and the appropriate root of the generalized antieigenvalue have been multiplied by 100.

For the Kendall's data (Table 1), we observe that deleting the observation number 4 results in both antieigenvalues becoming unusually small. A closer look at the particular observation shows that corresponding (x_1, x_2) values are both unusually large for this data point. Another observation which stands out is the observation number 14 for which the first antieigenvalue is the smallest and the second antieigenvalue is second smallest. For this

data point x_3 is unusually large while x_4 is among the several larger values. Clearly, both of these observations have substantial effects on the eigen-structure and eccentricities of the corresponding $\mathbf{X}$'$\mathbf{X}$ matrix.

Daniel and Wood's data set is relatively smaller in size. Yet, it still has two observations which are deemed outlying. These are observations 3 and 10 respectively and they stand out essentially due to relatively large x_5 value (for 3^{rd} observation) and very large x_2 value (for 10^{th} observation) respectively. See Table 2.

In case of Rao's cork data, the first antieigenvalues are somewhat different when the 19^{th} or 16^{th} observations are discarded. When the 18^{th} or 16^{th} observations are removed, the second antieigenvalue shows a considerable change. See Table 4. These observations were identified by Khattree and Naik (1999) as outlying using other techniques. The reasons for them being outlying are also explained there. However the effect is rather mild.

The data set of Chatterjee, Hadi and Price does not seem to contain any outlying observation because for all the antieigenvalues in Table 3, corresponding values are not very different when an observation has been deleted.

To explore further and perhaps in a more definite way than the previous approach where the relative closeness of antieigenvalues was visually assessed in a table, we may yet adopt another criterion (Also, see Khattree, 2019) and directly look at the eigen or antieigenstructures of the matrix

$$\mathbf{G}_i = \mathbf{U}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{U}_i'$$

where $\mathbf{U}_i$ is the upper triangular square root matrix of $\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}$, defined by $\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)} = \mathbf{U}_i'\mathbf{U}_i$.

Ideally, if an observation was not outlying then we must expect $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}$, to be nearly the same. Thus, $\mathbf{G}_i$ in above equation must be "close" to an identity matrix, for which all eigenvalues and hence all antieigenvalues are 1. Therefore, we may argue that an observation is outlying if there is considerable departure of these quantities from unity. It turns out (See Theorem 1 in Appendix) that except possibly for the smallest eigenvalue, all other eigenvalues of $\mathbf{G}_i$ must always be equal to 1. Thus it suffices to look at the departure of the smallest eigenvalue $\delta_{i,p}$ or equivalently the smallest antieigenvalue $\eta_{i,1}$ of $\mathbf{G}_i$ from unity.

All four data sets are subjected to this criterion as well. This criterion results in the identification of all outlying observations found previously by our earlier method. As it turns out, this is a more sensitive criterion in the sense it may also identify many more mildly outlying observations whose presence was perhaps obscured in our earlier approach when two antieigenvalues were compared for similar magnitudes. The results are given in Tables 7-10. Observations have been rearranged by the decreasing magnitudes of the $\eta_{i,1}$ values.

To be specific, Kendall's data shows observations 4 and 14 with significantly smaller values of $\delta_{i,p}$ and $\eta_{i,1}$ compared to other values. This is consistent with our previous evaluation. For Daniel-Wood data, observation numbers 3, 10 and 1 are found to be outlying (due to relatively larger x_5, x_2 and x_1 values). In case of Cork data of Rao, although none of the

observations was determined to be severely outlying in our earlier analysis, a few more mildly outlying observations are now found by this approach (Observation numbers 12, 15, 16, 18, 19). This identification is consistent with what was observed by Khattree and Naik (1999) using various graphical methods such as biplots. Also, in case of Chatterjee, Hadi and Price's data, the observation number 3 (due to x_3 being relatively larger) and observation number 15 (due to x_1 and x_2 being relatively smaller) are also identified although their departures still appear to be subdued. These facts are graphically and more effectively illustrated using the *scree plots* given in Figure 1, where a sudden vertical drop, or lack of it, indicates the presence or absence of outlyingness.

One can also arrive at $\delta_{i,p}$ or $\eta_{i,1}$ as the yardsticks for outlyingness through certain other criteria. Specifically, to measure the distance between $\mathbf{X}$'$\mathbf{X}$ and $\mathbf{X}_{(-i)}$'$\mathbf{X}_{(-i)}$, one may consider as measures, the eigen or antieigenvalues of $\mathbf{X}_{(-i)}$'$\mathbf{X}_{(-i)}$, with respect to $\mathbf{X}$'$\mathbf{X}$ (see Rao, 2005) or alternatively, use the determinant of matrix $\mathbf{X}_{(-i)}$'$\mathbf{X}_{(-i)}(\mathbf{X}$'$\mathbf{X})^{-1}$. In either case, in view of Theorem 1, the final criterion rests on $\delta_{i,p}$ and $\eta_{i,1}$.

Computationally, evaluation of eigenvalues $\delta_{i,p}$ and hence of antieigenvalues $\eta_{i,1}$ is rather straight forward and in fact, does not even require the explicit evaluation of eigenvalues or of the square root matrices – issues which can be a severe computational burden if the data set was large. To be specific, in view of Theorem 1, at most one of the eigenvalues of $\mathbf{G}_i(= \mathbf{U}_i(\mathbf{X}$'$\mathbf{X})^{-1}\mathbf{U}_i$', where $\mathbf{U}_i$ is an upper triangular matrix so that $\mathbf{X}_{(-i)}$'$\mathbf{X}_{(-i)} = \mathbf{U}_i$'$\mathbf{U}_i)$ is not equal to 1 (in fact, less than 1). In view of Theorem 2, $\delta_{i,p} = tr(\mathbf{G}_i) - p + 1 = 1 - \mathbf{x}_i$'$(\mathbf{X}$'$\mathbf{X})^{-1}\mathbf{x}_i$. which is simply a quadratic form in the i^{th} data-row.

When using the $\mathbf{X}$ matrix, unlike the leverage value calculations, our approach to outlying observation detection as outlined here is *not* model based. Polynomial or cross-product terms which may be important in the model have not been considered and columns corresponding to these in the $\mathbf{X}^*$ matrix do play an important role in the computation of leverage values. Thus, $\delta_{i,p} = 1 - \mathbf{x}_i$'$(\mathbf{X}$'$\mathbf{X})^{-1}\mathbf{x}_i$ is in general *not* the model based leverage corresponding to the i^{th} observation (and this may be true in addition to the fact that in any model based approach, the $\mathbf{X}$ matrix contains a constant column corresponding to the intercept of the model). In view of this subtle difference, Khattree (2019) chooses to call $\mathbf{x}_i$'$(\mathbf{X}$'$\mathbf{X})^{-1}\mathbf{x}_i$ as *Emphasis* of the observation $\mathbf{x}_i$ rather than leverage. It will be same as leverage only if the assumed model had no intercept and no polynomial or cross product terms. In fact, that is exactly the point being made here. There may be observations which are simply different from the rest of the data without any consideration of assumed model whatsoever and they should be detected and examined at the very early stages of data cleaning prior to defining any specific model. The clean data may then be intended to be used as a reference dataset for a number of future studies. Consideration of antieigenvalues and *emphasis* measures help us do just that.

Multicollinearity and outlyingness can occasionally go hand in hand. Outlying observations can sometimes introduce or mask the multicollinearity. Fortunately, antieigenvalues can be utilized to assess both and hence provide a useful approach to identify "*collinearity - outlying*" points. We may measure the *collinearity - outlyingness* as the relative change in the antieigenvalues of $\mathbf{X}_{(-i)}$'$\mathbf{X}_{(-i)}$ compared to those of $\mathbf{X}$'$\mathbf{X}$. For $j = 1, 2, \cdots, r$,

$$\gamma_{i,j} = \frac{\eta_{i,j} - \eta_j}{\eta_j} \tag{3}$$

measures this change for $i = 1, 2, \cdots, n$. A similar measure can be defined in terms of the generalized antieigenvalue,

$$\Gamma_i = \frac{\Delta_i - \Delta}{\Delta}. \tag{4}$$

The measures in Equations (3) and (4) are computed for all four data sets. Instead of giving their values in several tables, it may be easier to plot and interpret $\gamma_{i,j}$ and Γ_i graphically. Specifically, we graphically identify the collinearity - outlyingness of individual data points by plotting $\gamma_{i,j}$ (or Γ_i) against theoretical (normal) quantiles in what is equivalent to a Q-Q plot. It is so, since the values of the nonoutlying points are likely to be more or less random (possibly approximately normally distributed). One set of such plots are given in Figure 2 for the four data sets. The statistics used is $\Delta^{1/r}$.

Table 7: Detecting outlyingness, raw data and smallest eigenvalue δ_p values, smallest antieigenvalue $\eta_{j,1}(\times 100)$ for $\mathbf{G}_j$ Matrix [Kendall's Data]

Serial No. No.	Deleted Obs. (j)	x_1	x_2	x_3	x_4	$\delta_{j,p}$	$\eta_{j,1}$
1	18	13.2	6.6	2.0	5.8	0.93	99.93
2	3	20.6	12.5	2.3	7.0	0.90	99.85
3	20	20.7	9.6	3.1	5.9	0.88	99.80
4	10	25.5	12.9	1.9	7.3	0.87	99.77
5	15	31.2	11.6	2.4	6.5	0.87	99.76
6	16	22.7	10.1	3.3	6.2	0.87	99.76
7	11	26.5	14.9	2.4	6.7	0.86	99.72
8	1	13.0	9.7	1.5	6.4	0.85	99.67
9	7	12.7	5.7	2.9	6.7	0.85	99.66
10	5	20.5	14.2	1.9	6.9	0.82	99.49
11	6	10.0	6.7	2.2	7.0	0.82	99.49
12	2	10.0	7.5	1.5	6.5	0.81	99.47
13	8	36.5	15.7	2.3	7.2	0.81	99.44
14	19	11.1	6.7	2.2	7.2	0.81	99.44
15	17	31.2	9.6	2.4	6.0	0.80	99.39
16	12	22.3	8.4	4.0	7.0	0.78	99.26
17	9	37.1	14.3	2.1	7.2	0.74	98.83
18	13	30.8	7.4	2.7	6.4	0.67	98.10
19	**14**	**25.3**	**7.0**	**4.8**	**7.3**	**0.58**	**96.34**
20	**4**	**33.8**	**19.0**	**2.8**	**5.8**	**0.48**	**93.68**

Remark: Most-outlying observations are highlighted in **bold**.

It is important to interpret these graphs in Figure 2 carefully. Larger positive values not falling on the straight line pattern indicate an improvement in terms of multicollinearity when the particular observation is deleted. In other words, these observations when present, tend to introduce multicollinearity. Similarly, observations with values which are more negative and away from the overall straight line pattern will tend to mask the multicollinearity. Thus from the graphs in Figures 2, it is easy to conclude that for Kendall's data, inclusion

of observations numbered 4 and 14 tend to mask the problem, if any, of multicollinearity in the data. A more drastic instance of masking of multicollinearity is vividly seen in Daniel-Wood data. Observation number 3 clearly stands out at the lower left end of that figure. The 10^{th} observation also does so and masks some multicollinearity although it is not as excessive. For Hadi, Chatterjee and Price data, the inclusion of 10^{th} observation somewhat but not excessively increases the multicollinearity. As observed previously, this data set was already found to be highly ill-conditioned. Lastly, in case of Cork data, 18^{th} and 12^{th} observations are towards the higher end. Their inclusion possibly increases the multicollinearity slightly. However, as we have noted earlier, this data set is relatively well behaved in terms of multicollinearity.

It must be emphasized that this analysis of multicollinearity-outlyingness is irrelevant when with or without the particular observation, various measures of multicollinearity indicate a lack of it and is of interest only when the data exhibit a situation when the inclusion/exclusion of an observation drastically alters that situation and makes the data look much better or much worse than otherwise. This scenario is well illustrated by observation number 3 in case of Daniel-Wood data.

4. An evaluation of a large data set: red wine data

So far, we have purposely chosen data sets which could be presented in their entirety and allowed us to see the differences made by certain observations in the eigen-structure of the data. Would the large size of data obscure these changes since a single observation in a large dataset will supposedly have very small fraction of contribution to the overall structure? While this query is difficult to answer theoretically, we apply our approach on a relatively large data set available from the UCI Machine Learning Repository (`https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/`) consisting of 1599 observations on the quality of red wine. The four variables considered here are measurements on, x_1 = fixed acidity, x_2 = chlorides, x_3 = free sulfur dioxide and x_4 = pH for various wine samples.

As a process of data cleaning for such a large data, we will attempt to identify the observations which cause drastic changes in the eigen-structure using the techniques described earlier. We also evaluate the multicollinearity-outlyingness. As a representative subsample, the first ten, middle ten and last ten observations of the dataset are given in Table 11, along with a few cases that we have identified as outlying. Various statistics considered here have been calculated. Specifically, upon deleting one observation at a time for all 1599 observations, we compute the first and second antieigenvalues, and the generalized antieigenvalue. As a reference set to compare, we also compute these values for the entire data without deleting any observation. In view of small magnitudes of certain quantities, whenever needed, we have reported these quantities upto five decimal places.

Analysis results in four observations which stand out. For the observations numbered 82, 107, 152 and 259, the changes in the first antieigenvalue and the generalized antieigenvalue are substantial relative to entire data and compared to other cases of deletion. As can be seen in Table 11 (Columns 6 and 7) for the subsample listed above, compared to entire data, these quantities hardly change in the cases of one at a time deletions of the other 1595 observations.

Columns 8, 9 and 10 of the same table also present the values of $\gamma_{i,1}$, $\gamma_{i,2}$ and Γ_i which measure the collinearity- outlyingness. Clearly for the observations 82, 107, 152 and 259 these values by far stand out. The corresponding negative values indicate that these observations tend to mask the multicollinearity present in the data. For all other observations, the corresponding values are minuscule and practically insignificant.

Why do observations 152 and 259 stand out? It is clear that the value of x_2 is substantially larger while the value x_4 is substantially smaller compared to other observations. For observations 82 and 107, it is their large (but not as large as that for observations 152 and 259) x_2 values which make them outlying.

One may ask, "Why not just use the leverage function as specified in standard textbooks instead of the approach that we suggest?" If the usual leverage measure is used to identify outlying observations and if we use the recommended rule to identify outlying observations as those which have their leverage values higher than twice the mean leverage ($= \frac{2p}{n} = \frac{2\times 4}{1599} = 0.005$), then that procedure ends up identifying a total of 189 observations as outlying! As an alternative recommendation, if we just choose those observations whose leverage values clearly stand out (with an existence of gap between them and rest of the leverage values) then it will identify observations 152 and 259 (leverages = 0.0806 and 0.0801 respectively) but not the observations 82 and 107 whose leverage values (= 0.0420 and 0.0416 respectively) are not as prominent compared to others. To keep these tables manageable, we have not printed all leverage values in Tables 11 and 12. Our approach, although more computer intensive, appears to be much more effective. Automating the above calculations can make such identifications quick and efficient, especially when the data may also contain other nonrandom noises such as freak values due to transcription errors.

Table 12 presents the analysis of the same data minus the observations numbered 82, 107 152 and 259. A comparison of corresponding entries in row 0 (*i.e.*, when "Deleted Obs.=none") in Tables 11 and 12 shows that the removal of above four observations changes the statistics in Columns 6 and 7 of Table 11 substantially. Substantially smaller values in Table 12 suggest that this data set has much more multicolinearity than it originally showed, which was earlier masked by these four observations. Further, deletion of any of the remaining 1595 observations causes little change in the values of $\eta_{i,1}, \Delta_i, \gamma_{i,1}, \gamma_{i,2}$ and Γ_i, leading to the conclusion that there are no more outlying observations in the data set (However, an approach based on leverage values still declares 189 outlying observations!). Also the data set now has no outlying observation induced multicollinearity.

5. Concluding remarks

The work presented here introduces the use of eigen-structure and antieigenvalues for data cleaning early on after data collection but prior to modeling. This helps us identify and evaluate the quality of data and identify the possible anomalies within the data. Our work also supplements existing useful diagnostics techniques and are beneficial in providing the valuable insights into the data. One possibility to calibrate our proposed metrics may be via the probability distribution of the antieigenvalues. Some related work by Martin Singull can be found at `https://users.mai.liu.se/maroh70/pres/iclaa2017.pdf` .

One important recurring question is whether or not to center and/or standardize the data before subjecting them to these tools of analysis of antieigen-structure. It is obvious that

Table 8: Detecting outlyingness, raw data and smallest eigenvalue δ_p values, smallest antieigenvalue $\eta_{j,1}(\times 100)$ for $\mathbf{G}_j$ Matrix [Daniel and Wood's Data]

Serial No.	Deleted Obs. (j)	x_1	x_2	x_3	x_4	x_5	$\delta_{j,p}$	$\eta_{j,1}$
1	6	22.32	6.17	2.85	66.47	2.43	0.86	99.72
2	9	21.96	4.65	6.06	64.07	2.32	0.81	99.47
3	12	21.34	6.07	2.93	67.03	2.56	0.81	99.45
4	4	24.60	5.85	2.80	64.18	2.40	0.79	99.30
5	2	25.96	3.48	5.06	63.15	2.32	0.78	99.23
6	13	21.94	5.57	2.68	67.71	2.44	0.77	99.12
7	8	23.54	4.83	7.21	62.03	2.24	0.71	98.50
8	11	22.48	5.00	7.46	62.72	2.24	0.70	98.40
9	14	25.72	4.12	6.06	61.05	2.08	0.68	98.23
10	5	25.04	3.86	2.11	66.57	2.36	0.64	97.55
11	7	20.93	4.64	5.74	66.26	2.08	0.60	96.80
12	**1**	**27.68**	**3.76**	**1.98**	**64.97**	**2.48**	**0.54**	**95.55**
13	**10**	**21.44**	**8.81**	**1.19**	**66.64**	**2.48**	**0.30**	**83.89**
14	**3**	**21.86**	**5.75**	**2.77**	**65.02**	**5.04**	**0.01**	**23.22**

Remark: Most-outlying observations are highlighted in **bold**.

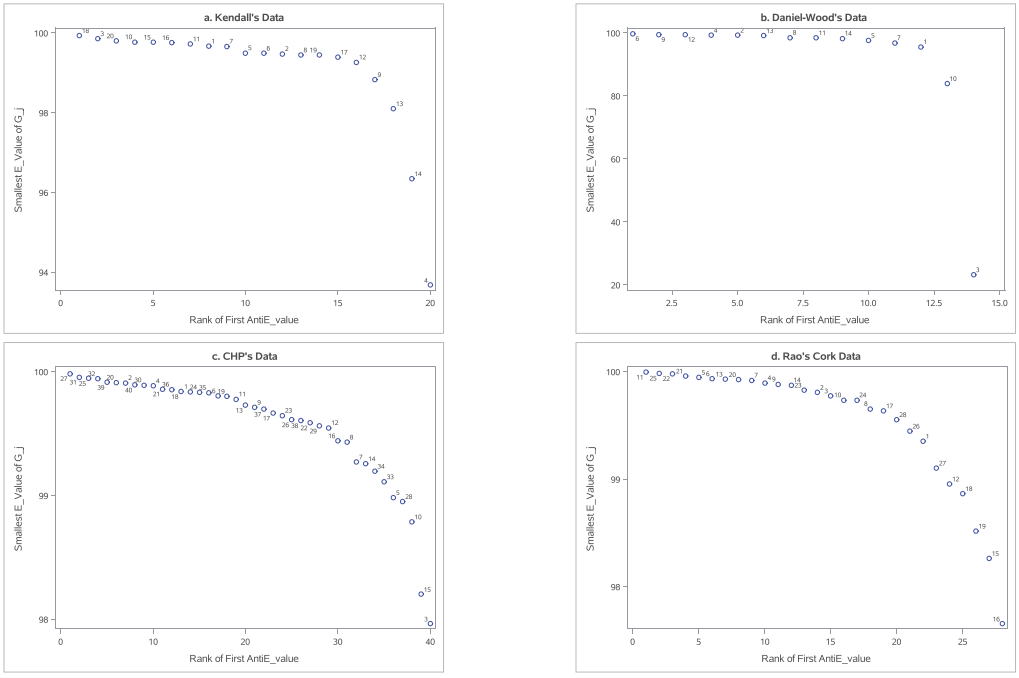


Figure 1: Scree plots for smallest antieigenvalues of $\mathbf{G}_i$ upon deleting an observation

Table 9: Detecting outlyingness, raw data and smallest eigenvalue δ_p values, smallest antieigenvalue $\eta_{j,1}(\times 100)$ for $\mathbf{G}_j$ Matrix [Chatterjee, Hadi and Price's data]

Serial No.	Deleted Obs. (j)	x_1	x_2	x_3	x_4	x_5	x_6	$\delta_{j,p}$	$\eta_{j,1}$
1	27	83	78	71	11	8	233	0.96	99.98
2	31	34	44	65	7	9	143	0.94	99.95
3	25	61	53	79	6	5	193	0.94	99.95
4	32	71	34	56	8	9	162	0.93	99.94
5	39	80	55	61	11	1	197	0.92	99.91
6	20	67	65	62	13	12	196	0.92	99.91
7	2	27	70	31	6	6	129	0.92	99.91
8	40	82	88	54	14	7	225	0.91	99.89
9	30	64	30	81	10	10	176	0.91	99.89
10	4	92	62	92	5	8	247	0.91	99.89
11	21	38	26	44	10	8	110	0.90	99.86
12	36	61	35	55	13	0	152	0.90	99.85
13	18	106	87	82	18	7	276	0.89	99.84
14	1	49	79	76	8	15	205	0.89	99.84
15	24	53	55	60	8	0	170	0.89	99.83
16	35	57	69	72	5	4	200	0.89	99.83
17	6	31	54	34	14	11	119	0.88	99.80
18	19	97	98	71	12	8	266	0.88	99.80
19	11	62	62	81	9	1	207	0.87	99.78
20	13	45	65	84	19	13	195	0.86	99.73
21	9	98	72	71	12	-1	242	0.86	99.71
22	37	29	45	47	13	13	123	0.86	99.70
23	17	78	102	84	5	7	266	0.85	99.66
24	23	54	100	50	11	15	205	0.84	99.64
25	26	60	108	104	17	8	273	0.84	99.61
26	38	82	105	81	20	9	268	0.84	99.60
27	22	56	32	99	16	8	188	0.83	99.59
28	29	89	121	71	8	8	283	0.83	99.56
29	12	25	11	7	9	9	45	0.83	99.55
30	16	111	52	93	11	13	256	0.81	99.44
31	8	114	85	84	17	20	285	0.81	99.43
32	7	105	60	47	5	10	212	0.78	99.27
33	14	92	75	63	9	20	232	0.78	99.26
34	34	112	105	123	5	12	340	0.78	99.19
35	33	88	30	87	13	0	207	0.77	99.11
36	5	67	42	94	16	3	202	0.75	98.98
37	28	74	125	66	16	4	265	0.75	98.95
38	10	15	59	99	15	11	174	0.73	98.79
39	**15**	**27**	**26**	**82**	**4**	**17**	**134**	**0.68**	**98.20**
40	**3**	**115**	**92**	**130**	**0**	**9**	**339**	**0.67**	**97.96**

Remark: Most-outlying observations are highlighted in **bold**.

Table 10: Detecting outlyingness, raw data and smallest eigenvalue δ_p values, smallest antieigenvalue $\eta_{j,1}(\times 100)$ for $\mathbf{G}_j$ Matrix [C. R. Rao's Cork data]

Serial No.	Deleted Obs. (j)	x_1	x_2	x_3	x_4	$\delta_{j,p}$	$\eta_{j,1}$
1	11	32	30	34	28	0.98	99.99
2	25	39	36	39	31	0.97	99.98
3	22	32	30	30	32	0.96	99.98
4	21	39	35	34	37	0.94	99.96
5	5	32	32	35	36	0.94	99.94
6	6	30	35	34	26	0.93	99.93
7	13	54	46	60	52	0.93	99.93
8	20	46	38	37	38	0.93	99.93
9	7	39	39	31	27	0.92	99.92
10	4	41	29	36	38	0.91	99.90
11	9	37	40	31	25	0.91	99.88
12	14	47	51	52	43	0.90	99.87
13	23	60	50	67	54	0.89	99.83
14	2	60	53	66	63	0.88	99.81
15	3	56	57	64	58	0.87	99.77
16	10	33	29	27	36	0.86	99.74
17	24	35	37	48	39	0.86	99.74
18	8	42	43	31	25	0.85	99.65
19	17	79	65	70	61	0.84	99.63
20	28	48	54	57	43	0.83	99.55
21	26	50	34	37	40	0.81	99.45
22	1	72	66	76	77	0.80	99.35
23	27	43	37	39	50	0.76	99.10
24	**12**	**63**	**45**	**74**	**63**	**0.75**	**98.96**
25	**18**	**81**	**80**	**68**	**58**	**0.74**	**98.86**
26	**19**	**78**	**55**	**67**	**60**	**0.71**	**98.51**
27	**15**	**91**	**79**	**100**	**75**	**0.69**	**98.26**
28	**16**	**56**	**68**	**47**	**50**	**0.65**	**97.65**

Remark: Most-outlying observations are highlighted in **bold**.

Table 11: Analysis of red wine data: complete data (n = 1599)

Deleted Obs. (j)	x_1	x_2	x_3	x_4	$\eta_{j,1}$ ($\times 100$)	Δ_j ($\times 100$)	$\gamma_{j,1}$	$\gamma_{j,2}$	Γ_j
none	**.**	**.**	**.**	**.**	**0.462**	**0.124**	**0**	**0**	**0**
1	7.4	0.076	11	3.51	0.462	0.124	0.01259	-0.00722	-0.00360
2	7.8	0.098	25	3.20	0.463	0.124	0.04972	0.05319	0.02659
3	7.8	0.092	15	3.26	0.462	0.124	0.02127	0.02221	0.01111
4	11.2	0.075	17	3.16	0.462	0.124	0.02282	-0.01682	-0.00841
5	7.4	0.076	11	3.51	0.462	0.124	0.01259	-0.00722	-0.00360
6	7.4	0.075	13	3.51	0.462	0.124	0.01615	-0.00458	-0.00229
7	7.9	0.069	15	3.30	0.462	0.124	0.01886	0.02010	0.01005
8	7.3	0.065	15	3.39	0.462	0.124	0.01720	0.00219	0.00110
9	7.8	0.073	9	3.36	0.462	0.124	0.00872	0.01594	0.00797
10	7.5	0.071	17	3.35	0.462	0.124	0.02458	0.01650	0.00825
82	**7.8**	**0.464**	**22**	**3.13**	**0.453**	**0.122**	**-2.00423**	**-2.00403**	**-1.00708**
107	**7.8**	**0.467**	**18**	**3.08**	**0.453**	**0.122**	**-2.06405**	**-2.06336**	**-1.03705**
152	**9.2**	**0.610**	**32**	**2.74**	**0.445**	**0.119**	**-3.83596**	**-3.89969**	**-1.96923**
259	**7.7**	**0.611**	**8**	**3.06**	**0.444**	**0.119**	**-4.02277**	**-4.00309**	**-2.02198**
797	8.7	0.126	24	3.10	0.462	0.124	0.02903	0.01822	0.00911
798	9.3	0.038	21	3.24	0.462	0.124	0.00023	-0.00877	-0.00438
799	9.4	0.082	5	3.29	0.462	0.124	0.00489	0.06907	0.03453
800	9.4	0.082	5	3.29	0.462	0.124	0.00489	0.06907	0.03453
801	7.2	0.082	26	3.25	0.463	0.124	0.05499	0.06207	0.03103
802	8.6	0.068	8	3.23	0.462	0.124	0.00490	0.03901	0.01951
803	5.1	0.044	14	3.56	0.462	0.124	0.00432	-0.12472	-0.06238
804	7.7	0.114	14	3.24	0.462	0.124	0.00626	0.00741	0.00371
805	8.4	0.084	4	3.26	0.462	0.124	0.00423	0.05255	0.02628
806	8.2	0.052	4	3.33	0.462	0.124	-0.01058	0.02851	0.01426
1590	6.6	0.073	29	3.29	0.463	0.124	0.06490	0.08170	0.04084
1591	6.3	0.077	26	3.32	0.463	0.124	0.05366	0.05161	0.02581
1592	5.4	0.089	16	3.67	0.462	0.124	0.02095	-0.10075	-0.05038
1593	6.3	0.076	29	3.42	0.463	0.124	0.06502	0.07259	0.03630
1594	6.8	0.068	28	3.42	0.463	0.124	0.05875	0.06587	0.03293
1595	6.2	0.090	32	3.45	0.463	0.124	0.07723	0.10128	0.05063
1596	5.9	0.062	39	3.52	0.463	0.124	0.10568	0.18234	0.09113
1597	6.3	0.076	29	3.42	0.463	0.124	0.06502	0.07259	0.03630
1598	5.9	0.075	32	3.57	0.463	0.124	0.07692	0.08773	0.04386
1599	6.0	0.067	18	3.39	0.462	0.124	0.02646	-0.02163	-0.01081

Remark: Most-outlying observations are highlighted in **bold**. *Top row corresponds to entire data with no deletion.*

Table 12: Analysis of red wine data: after deleting obs. No. 82, 107, 152 and 259 ($n = 1595$)

Deleted Obs. (j)	x_1	x_2	x_3	x_4	$\eta_{j,1}$ ($\times 100$)	Δ_j ($\times 100$)	$\gamma_{j,1}$	$\gamma_{j,2}$	Γ_j
none	**.**	**.**	**.**	**.**	**0.132**	**0.0793**	**0**	**0**	**0**
1	7.4	0.076	11	3.51	0.132	0.0792	-0.10861	-0.12180	-0.060853
2	7.8	0.098	25	3.20	0.132	0.0793	-0.02016	0.07747	0.038792
3	7.8	0.092	15	3.26	0.132	0.0792	-0.04344	-0.03128	-0.015577
4	11.2	0.075	17	3.16	0.132	0.0792	-0.01573	-0.03490	-0.017385
5	7.4	0.076	11	3.51	0.132	0.0792	-0.10814	-0.12132	-0.060614
6	7.4	0.075	13	3.51	0.132	0.0792	-0.10118	-0.10024	-0.050070
7	7.9	0.069	15	3.30	0.132	0.0792	-0.04437	-0.03356	-0.016718
8	7.3	0.065	15	3.39	0.132	0.0792	-0.07958	-0.06261	-0.031244
9	7.8	0.073	9	3.36	0.132	0.0792	-0.07065	-0.10248	-0.051188
10	7.5	0.071	17	3.35	0.132	0.0792	-0.05981	-0.02949	-0.014680
797	8.7	0.126	24	3.10	0.132	0.0793	0.00877	0.03356	0.016842
798	9.3	0.038	21	3.24	0.132	0.0793	0.01014	0.02349	0.011809
799	9.4	0.082	5	3.29	0.132	0.0793	0.02025	-0.00452	-0.002193
800	9.4	0.082	5	3.29	0.132	0.0793	0.02030	-0.00437	-0.002119
801	7.2	0.082	26	3.25	0.132	0.0793	0.01850	0.04528	0.022701
802	8.6	0.068	8	3.23	0.132	0.0793	0.02001	0.00788	0.004003
803	5.1	0.044	14	3.56	0.132	0.0792	-0.11208	-0.11019	-0.055043
804	7.7	0.114	14	3.24	0.132	0.0793	0.01638	0.01733	0.008728
805	8.4	0.084	4	3.26	0.132	0.0793	0.01609	0.00342	0.001776
806	8.2	0.052	4	3.33	0.132	0.0793	0.01011	-0.00048	-0.000178
1590	6.6	0.073	29	3.29	0.132	0.0793	0.08935	0.04278	0.021454
1591	6.3	0.077	26	3.32	0.132	0.0793	0.09028	0.02157	0.010850
1592	5.4	0.089	16	3.67	0.132	0.0793	0.04197	0.02997	0.015049
1593	6.3	0.076	29	3.42	0.132	0.0793	0.09004	0.03221	0.016167
1594	6.8	0.068	28	3.42	0.132	0.0793	0.08935	0.04278	0.021454
1595	6.2	0.090	32	3.45	0.132	0.0793	0.09028	0.02157	0.010850
1596	5.9	0.062	39	3.52	0.132	0.0793	0.09030	-0.00125	-0.000560
1597	6.3	0.076	29	3.42	0.132	0.0793	0.09051	0.03229	0.016210
1598	5.9	0.075	32	3.57	0.132	0.0793	0.08664	0.01229	0.006207
1599	6.0	0.067	18	3.39	0.132	0.0793	0.08434	0.06733	0.033726

Remark: Upon deletion of observation numbers 82, 107, 152 and 259, there are no more outlying observations. Top row corresponds to entire data (1599 - 4 = 1595 obs.) with no further deletions.

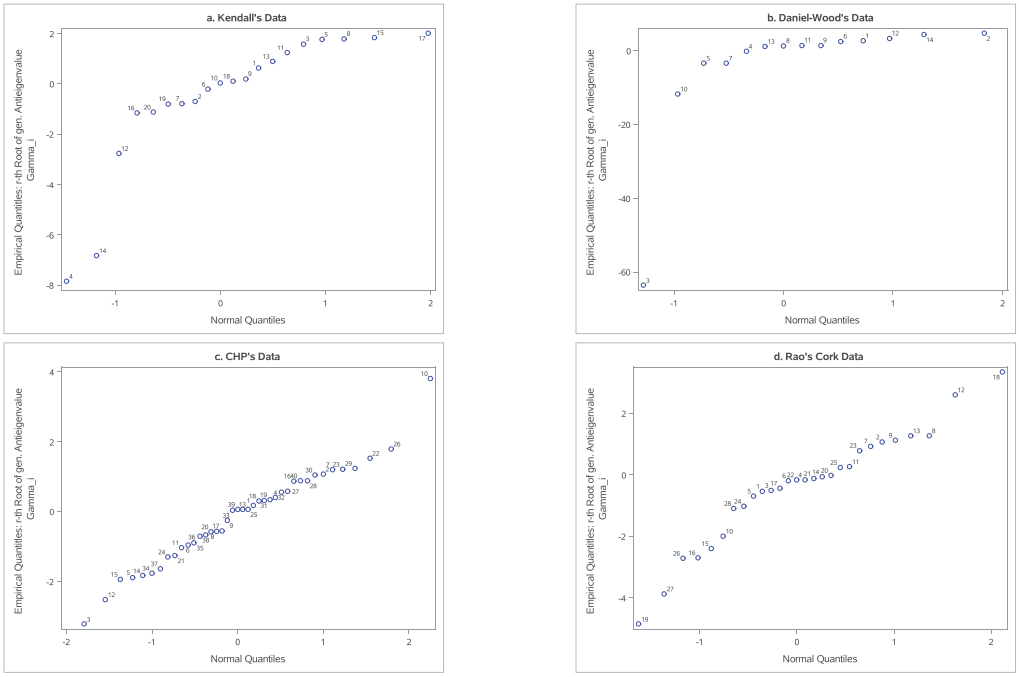


Figure 2: Scree plots for smallest antieigenvalues of $\mathbf{G}_i$ upon deleting an observation

the two sets of antieigenvalues will differ for standardized and unstandardized data. Shifting and scaling have substantial effects on eccentricities of the corresponding $p-$ dimensional ellipsoids. This is especially so since both the means as well as standard deviations can be very sensitive to outliers. Naik and Khattree (1996) provide a detailed discussion of pitfalls of blindly standardizing the data. Our (obvious) suggestion, thus, is to look into both possibilities since the use of unstandardized and standardized data for the subsequent modeling purposes are both acceptable practices.

Another natural question one may pose is, how about identification of a subset of s most outlying observations? How can we assess whether or not a particular subset of observations as a whole, is outlying? Khattree (2019) has extensively dealt with this issue with *Emphasis* as a measure. As it turns out, under that criterion, this problem is equivalent to sequential identification and deletion of observations, one by one, based on the maximum outlyingness using the matrices $\mathbf{G}_i$ and the method described in the present work. Thus, our approach here, at least partially eliminates the problem of first determining an appropriate choice of s and then looking at the computationally intensive task of evaluating the outlyingness of all possible ${}^nC_s$ subsets of s observations. When antieigenvalues are used as the criteria, whether or not such a sequential deletion is possible, is an issue that is still needed to be explored.

The numbers of observations as well as number of variables play important roles and an observation may have substantially less outlying effect on eigen-structure if it was only one of the several thousand observations. We see this in case of red wine data when out of a

total of 1599 observations only four stand out. Thus it is difficult to give a firm and universal threshold value for the determination of an observation's outlyingness. For large data sets, graphical methods and plots such as those given in this work provide a useful approach to identify such patterns.

The approach is admittedly computer intensive and thus there is a genuine need for developing an efficient and effective algorithm based on the methodology presented here. This is an important direction for future work as contemporary data sets are often very large in terms of number of variables as well as number of data points. We have not addressed these algorithmic-efficiency issues in this paper. Further research is needed in this direction.

We must also realize that the context here is that of data cleaning for a large dataset without any assumed model. Such data will often have missing values. With a model-free approach, it will be difficult to incorporate various types of missingness in our approach. However, assuming that the missingness is *completely at random*, our suggestion is to first impute the missing values appropriately and then perform the data cleaning. Since we intend to not assume any model on data, an approach to imputation based on empirical copula has been suggested by Lun and Khattree (2019, 2020, 2024). To what extent the performance may be affected by imputation is yet another aspect which can be explored via simulation studies.

What happens if the number of variables is greater than the number of observations? The genomics data, which very often are inevitably quite noisy, typically have this situation. How to clean data in such a case is admittedly a difficult problem. However, this situation allows us to address a very different problem still relevant in the context. Since our approach is based entirely on eigen-structure and since the nonzero eigenvalues of $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}\mathbf{X}'$ are same, for "$p > n$" situation, our approach can perhaps be adopted to evaluate the quality of variables. Since the interpretations entirely change, this problem needs a further careful consideration.

Acknowledgment and competing interests

The author wishes to thank two referees for their thoughtful comments. This work was supported, in part, by funding from the NSF Award FAIN: 2244091. The author declares no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. Finally, this article is dedicated to the memory of Professor C. R. Rao.

Supplementary Material

The original Wine data are available from the website

`https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality` .

ORCID Information

Ravindra Khattree: 0000-0002-9305-2365

References

Belsley, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. John Wiley and Sons: New York.

Belsley, D. A., Kuh, E., and Welsch, R. E. (2005). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley and Sons: New York.

Chatterjee, S. and Hadi, A. (1988). *Sensitivity Analysis in Linear Regression*. John Wiley and Sons: New York.

Chatterjee, S., Hadi, A., and Price, B. (2006). *Regression Analysis by Example*. John Wiley and Sons: New York.

Chu, X. and Ilyas, I. F. (2016). Qualitative data cleaning. *Proceedings of the VLDB Endowment*, **9**, 1605–1608.

Chu, X., Ilyas, I. F., Krishnan, S., and Wang, J. (2016). Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data*, pages 2201–2206.

Cuntoor, N. P. and Chellappa, R. (2006). Key frame-based activity representation using antieigenvalues. In *Asian Conference on Computer Vision*, pages 499–508. Springer.

Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data: Computer Analysis of Multifactor Data*. John Wiley and Sons: New York.

Guo, C., Jin, M., Guo, Q., and Li, Y. (2018). Antieigenvalue-based spectrum sensing for cognitive radio. *IEEE Wireless Communications Letters*, **8**, 544–547.

Ilyas, I. F. and Chu, X. (2019). *Data Cleaning*. Association for Computing Machinery.

Johnson, R. A. and Wichern, D. W. (2014). *Applied Multivariate Statistical Analysis*. Pearson: London, UK.

Kendall, M. G. (1975). *Multivariate Analysis*. Griffin: London.

Khattree, R. (2001). On the calculation of antieigenvalues and antieigenvectors. *Journal of Interdisciplinary Mathematics*, **4**, 195–199.

Khattree, R. (2002). On generalized antieigenvalue and antieigenmatrix of order r. *American Journal of Mathematical and Management Sciences*, **22**, 89–98.

Khattree, R. (2003). Antieigenvalues and antieigenvectors in statistics. *Journal of Statistical Planning and Inference*, **114**, 131–144.

Khattree, R. (2010). Antieigenvalues provide a bound on realized signal to noise ratio. *Journal of Statistical Planning and Inference*, **140**, 2846–2848.

Khattree, R. (2014). Antieigenvalues and antieigenvectors. *Wiley Stats Ref: Statistics Reference Online*, **1**, 131–134.

Khattree, R. (2019). A note on effects on the eigenstructure of a data matrix when deleting a subset of observations. *Journal of the Indian Society of Agricultural Statistics*, **73**, 11–17.

Khattree, R. and Bahuguna, M. (2019). An alternative data analytic approach to measure the univariate and multivariate skewness. *International Journal of Data Science and Analytics*, **7**, 1–16.

Khattree, R. and Naik, D. N. (1999). *Applied Multivariate Statistics with SAS Software, Second Edition*. SAS Publishing: Cary NC/John Wiley and Sons: New York.

Lun, Z. and Khattree, R. (2019). Multiple imputation for skewed multivariate data: A marriage of the MI and COPULA procedures. In *Proceedings of the SAS Global Forum, Paper 3605-2019*.

Lun, Z. and Khattree, R. (2020). Imputation for non-normal multivariate continuous data using copula transformation. *Proceedings of Joint Statistical Meeting, 2020 - Survey Research Methods Section*, 1922–1930.

Lun, Z. and Khattree, R. (2024). A general approach for imputation of non-normal continuous data based on copula transformation. *Communications in Statistics-Simulation and Computation*, **53**, 567–594.

Mason, R. L. and Gunst, R. F. (1985). Outlier-induced collinearities. *Technometrics*, **27**, 401–407.

Naik, D. N. and Khattree, R. (1996). Revisiting Olympic track records: Some practical considerations in the principal component analysis. *The American Statistician*, **50**, 140–144.

Rao, C. R. (1948). Tests of significance in multivariate analysis. *Biometrika*, **35**, 58–79.

Rao, C. R. (2005). Antieigenvalues and antisingularvalues of a matrix and applications to problems in statistics. *Research Letters in the Information and Mathematical Sciences*, **8**, 53–76.

Timm, N. H. (2002). *Applied Multivariate Analysis*. Springer: Switzerland.

Tran, N. Q. H. and Khattree, R. (2024). Supervised learning via eigen-structures. *Preprint, Under preparation*, .

Wang, S.-G. and Nyquist, H. (1991). Effects on the eigenstructure of a data matrix when deleting an observation. *Computational Statistics and Data Analysis*, **11**, 179–188.

APPENDIX

Appendix 1

Two theorems referred in main text are stated here. These are the special cases of results given in Khattree (2019) Proofs have been omitted.

Theorem 1. *Let* $\mathbf{X}$ *be an* $n \times p$ *matrix with* $n > p$ *and rank* p. *Define* $\mathbf{A} = \mathbf{X}'\mathbf{X}$ *and* $\mathbf{B}_i = \mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}$ *and suppose* $\mathbf{B}_i = \mathbf{U}_i'\mathbf{U}_i$ *where* $\mathbf{U}_i$ *is upper triangular. Then, for the ordered eigenvalues* $\delta_1 \geq \delta_2 \geq ... \geq \delta_p$ *of* $\mathbf{G}_i = \mathbf{U}_i\mathbf{A}^{-1}\mathbf{U}_i'$, $\delta_j = 1$ *for* $j = 1, 2, ..., (p-1)$.

Theorem 2. *The smallest eigenvalue of* $\mathbf{G}_i = \mathbf{U}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{U}_i'$ *where* $\mathbf{U}_i$ *is the upper triangular square root matrix of* $\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}$, *defined by* $\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)} = \mathbf{U}_i'\mathbf{U}_i$ *is* $\delta_{i,p} = 1 - \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$.

Appendix 2

SAS code which generated Table 11

```
/*
data on red wine from :
https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/

Attribute information:

   For more information, read [Cortez et al., 2009].

   Input variables (based on physicochemical tests):
   1 - fixed acidity
   2 - volatile acidity
   3 - citric acid
   4 - residual sugar
   5 - chlorides
   6 - free sulfur dioxide
   7 - total sulfur dioxide
   8 - density
   9 - pH
   10 - sulphates
   11 - alcohol
   Output variable (based on sensory data):
   12 - quality (score between 0 and 10)

*/

data wine; *Red wine data;
infile "C:\Users\Desktop\winequality.txt" ;
input y1-y12;
run;
```

```
data wine; set wine;
x1 = y1; *x1 = fixed acidity;
x2 = y5; *x2 = chlorides;
x3 = y6; *x3 = free sulfur dioxide;
x4 = y9; *x4 = pH;
keep x1-x4;
run;
proc standard data=wine mean=0 std=1
              out=stndtest;
     var x1-x4;
run;
%let mydataset = wine; ***use this to analyze original raw data;
*%let mydataset = stndtest; ***use this to analyze standardized data;


*options nolog;         ***suppresses log file**;


%macro multicol(count = );

%do del_row = 0 %to &count;
data uci; set &mydataset;run;
proc iml; use uci; read all into x;
xpx = x'*x;
if &del_row = 0 then do; smallxpx = xpx; end;
if &del_row > 0 then do;
t&del_row =x[ {&del_row}, ];

smallxpx = xpx - t&del_row'*t&del_row;end;
call eigen(lambda, p, smallxpx);
create evalues from lambda;
append from lambda;close evalues;
quit;
proc transpose data = evalues out = evaluesvar; run;
data evaluesvar&del_row; set evaluesvar;
anti1 =100*2*sqrt(col1*col4)/(col1+col4);
anti2 = 100*2*sqrt(col2*col3)/(col2+col3);
gen_anti = anti1*anti2/100;
rt_gen_anti = (anti1*anti2)**(1/2) ;
obser = &del_row;
run;
proc datasets library=work nolist;
   append base=work.antieig data=work.evaluesvar&del_row force;
run;
proc delete library = work data = evaluesvar&del_row;run;
%end;
```

```
%mend multicol ;
%multicol(count = 1599); ***no. of complete obs = count = 1599;
title;
footnote "Actual values are multiplied by 100";
data wine; set wine; obser = _n_;run;
proc sort data = wine; by obser;run;
proc sort data = antieig; by obser;run;
data combine; merge wine antieig; by obser;run;
******************Calculation of gamma values of  Section 3*********;
data gamma; set antieig;

***The numbers below are obtained from the output
when no observations were deleted.;

anti1all = .46228;;
gamma1 = 100*(anti1-anti1all)/anti1all;
gen_antiall = .12419;
gamma2 = 100*(gen_anti -gen_antiall)/gen_antiall;
rt_gen_antiall = 3.52406;
gamma3 = 100*(rt_gen_anti -rt_gen_antiall)/rt_gen_antiall;
run;
proc sort data = wine; by obser;run;
proc sort data = gamma; by obser;run;
data combine2; merge wine gamma;
by obser;run;
data antieigsmall2; set combine2;
if (obser in (0 82 107 152 259) or obser < 11 or obser gt 1589
or (obser > 796 and obser < 807));
run;
data Table11; set antieigsmall2;
keep obser x1 x2 x3 x4 anti1   gen_anti gamma1 gamma2 gamma3 ;
run;
proc export  data = table11
outfile =
'C:\Users\khattree\Desktop\DataCleaningTable11OfPaper.txt'
replace; ***Output is stored in the .txt file;
run;
```

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 199–209
http://www.ssca.org.in/journal

Horseshoe Prior for Bayesian Linear Regression with Hyperbolic Errors

Shamriddha De and Joyee Ghosh
Department of Statistics and Actuarial Science, The University of Iowa

Received: 31 May 2024; Revised: 18 July 2024; Accepted: 18 July 2024

Abstract

It is well known that squared error loss is not robust to outliers. As an alternative, Huber loss may be used for robust regression; however, Huber loss is not readily amenable to Bayesian computation. It has been shown that hyperbolic loss can be regarded as an approximation to Huber loss, and the hyperbolic distribution can be expressed as a scale mixture of normal distributions, which makes it appealing for Bayesian computation. The idea of Bayesian Huberized lasso was first proposed by Park and Casella (2008), and was formally developed and implemented by Kawakami and Hashimoto (2023). Since the Bayesian Huberized lasso cannot shrink regression coefficients to exactly zero, and has lighter tailed errors than a Cauchy distribution, De and Ghosh (2024) proposed a model that encompasses both hyperbolic and t-errors, with a mixture prior on regression coefficients consisting of two parts, a point mass at zero and a continuous distribution, that can shrink coefficients to exactly zero. The approach of De and Ghosh (2024) could be considered as a gold standard for Bayesian model averaging, but posterior computation with such a point mass mixture prior, popularly known as the spike and slab prior, can be challenging with many covariates. The horseshoe prior is known to mimic some of the desirable properties of spike and slab priors, while being computationally less intensive. Motivated by this attractive property of the horseshoe prior, in this article we develop an algorithm for Bayesian linear regression with hyperbolic errors, and horseshoe priors on the regression coefficients. We illustrate using simulation studies and an analysis of the famous Boston housing dataset, that posterior distributions under horseshoe priors can capture sparsity better than Bayesian lasso priors. For moderate dimensional regression problems, the spike and slab prior performs better than the horseshoe in capturing the sparsity of regression coefficients. However, we find that Markov chain Monte Carlo (MCMC) algorithms with horseshoe priors have improved mixing, which suggests that Bayesian shrinkage with the horseshoe prior and its generalizations, such as the regularized horseshoe prior, could be a promising direction to explore for high dimensional robust regression.

Key words: Bayesian lasso; Markov chain Monte Carlo; Model averaging; Robust regression; Spike and slab prior; Variable selection.

Corresponding Author: Joyee Ghosh
Email: joyee-ghosh@uiowa.edu

1. Introduction

The majority of Bayesian variable selection methods for linear regression have focused on normal errors, which is a challenging problem in its own right, especially for high dimensional problems. Since estimates derived under the normality assumption for errors can be sensitive to outliers, our goal is to robustify the error distribution. The Bayesian variable selection method for linear regression with normal errors can adapt to an unknown degree of sparsity by placing a prior on the unknown inclusion probability of variables. De and Ghosh (2024) developed a model with additional flexibility by allowing the likelihood to simultaneously adapt to an unknown degree of tail heaviness. They focused on the class of scale mixtures of normal densities for robust error distributions. The availability of the scale mixture of normal representation of the heavy tailed error models makes it convenient to implement MCMC sampling algorithms. Let the error distribution have the following form:

$$p(\epsilon) = \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\epsilon^2}{2\sigma^2}} dF(\sigma^2), \tag{1}$$

such that $F(.)$ is a cumulative distribution function (CDF). Then the random error ϵ is said to follow a scale (or variance) mixture of normals, and $F(.)$ is called a mixing distribution. Some popular distributions that can be represented by the scale mixtures of normal representation are the hyperbolic, Student-t, Laplace (double exponential), exponential power *etc.* (Andrews and Mallows (1974); West (1987); Gneiting (1997)).

In particular, the hyperbolic distribution forms a vital point of attention in this article. The said distribution has the following probability density function:

$$p_h(\epsilon; \eta, \rho^2) = \frac{1}{2\sqrt{\eta\rho^2} K_1(\eta)} e^{-\left(\eta\left(\eta + \frac{\epsilon^2}{\rho^2}\right)\right)^{1/2}}, \quad -\infty < \epsilon < \infty, \tag{2}$$

where $K_1(.)$ is a modified Bessel function, $\eta > 0$ is the shape parameter regulating the tail heaviness and $\rho^2 > 0$ is the scale parameter. Gneiting (1997) showed that the above distribution can be represented as a generalized inverse Gaussian (GIG) scale mixture of normally distributed random variables, and thus the hyperbolic distribution belongs to the family of scale mixture of normal distributions, defined in (1). We provide more detail in Section 2 about this representation. In a regression problem, using a hyperbolic error model is equivalent to using a hyperbolic loss function. Additionally, the hyperbolic loss has similarities with the Huber loss (Park and Casella (2008)). The Huber loss is popular for robust regression in the frequentist literature but it is computationally difficult to handle in a Bayesian set up. Accordingly, in a Bayesian setting, we focus on the hyperbolic loss as an alternative to the Huber loss, like previous authors (Park and Casella (2008); Kawakami and Hashimoto (2023); De and Ghosh (2024)).

Bayesian variable selection with two component mixture priors used by De and Ghosh (2024) leads to a vast model space, when the number of covariates is large. An alternative strategy that has been shown to perform favorably is using a continuous shrinkage prior to replace the mixture priors. For example, the Bayesian lasso (Park and Casella (2008)) is a continuous shrinkage prior, which has been implemented by Kawakami and Hashimoto (2023), for regression models with hyperbolic errors. Another well known technique is to use the Bayesian horseshoe prior (Carvalho *et al.* (2010), Makalic and Schmidt (2015), Bhadra

et al. (2017)), which also belongs to the family of continuous shrinkage priors and has been demonstrated to perform very well for shrinking noise variables to practically zero, while keeping signals almost intact for the normal means problem. For a $p \times 1$ vector $\boldsymbol{\beta}$ of regression coefficients, the horseshoe prior is defined as

$$(\beta_j \mid \lambda_j, \tau^2, \rho^2) \overset{\text{ind}}{\sim} N(0, \lambda_j^2 \tau^2 \rho^2), \quad \lambda_j \overset{\text{ind}}{\sim} C^+(0,1), \quad \tau \sim C^+(0,1), \tag{3}$$

for $j = 1, 2, \ldots, p$, where ρ^2 is the scale of the error distribution, and $C^+(0,1)$ represents the standard half-Cauchy distribution with the density function

$$p(x) = \frac{2}{\pi(1+x^2)}, \quad x > 0.$$

In the context of regression models with heavy tailed errors, the horseshoe prior has been utilized by Hamura *et al.* (2022). However, the focus of their paper is on super-heavy tailed error distributions, in comparison to which even the Student-t distribution is regarded as a thin tailed distribution. Hamura *et al.* (2022) considered the horseshoe prior for illustration for some applications, but most of the paper focuses on multivariate normal priors for regression coefficients. In contrast, this article focuses on hyperbolic errors as a proxy to the Huberized loss function. The main question of interest that we try to investigate is how methods based on horseshoe priors compare with those based on lasso, and spike and slab priors, under varying levels of sparsity.

The article is organized as follows. In Section 2, we introduce the hyperbolic distribution and horseshoe priors, and develop an algorithm for posterior computation. In Section 3 we conduct simulation studies with the true model having hyperbolic errors, and compare the results of the posterior estimates from the horseshoe prior versus the spike and slab prior (De and Ghosh (2024)) and the Bayesian lasso prior (Kawakami and Hashimoto (2023)). In Section 4, we analyze the famous Boston housing data with the three aforementioned priors, after adding noise variables to the original dataset. Finally, in Section 5, we provide a summary of the results, and discuss some future directions.

2. Hyperbolic error model with horseshoe prior

Let $\mathbf{Y}$, $\mathbf{X}$ and $\boldsymbol{\beta}$ denote the $n \times 1$ vector of response variables, the $n \times p$ design matrix, and the $p \times 1$ vector of regression coefficients, respectively. We consider a regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ is the $n \times 1$ vector of errors, such that $\epsilon_i \overset{\text{iid}}{\sim} p_h(\epsilon_i; \eta, \rho^2)$, $i = 1, 2, \ldots, n$, where $p_h(.; \eta, \rho^2)$ is the hyperbolic density with parameters η and ρ^2 defined in (2). Park and Casella (2008) showed that the normal scale-mixture representation by Gneiting (1997) leads to the representation of (4) in a computationally convenient form as

$$\mathbf{Y} \mid \boldsymbol{\beta}, \sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2 \;\sim\; \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{D}), \tag{5}$$

$$p(\sigma_1^2, \ldots, \sigma_n^2 | \eta, \rho^2) \;=\; \prod_{i=1}^{n} \frac{1}{2K_1(\eta)\rho^2} e^{-\frac{\eta}{2}\left(\frac{\sigma_i^2}{\rho^2} + \frac{\rho^2}{\sigma_i^2}\right)}, \tag{6}$$

where $\boldsymbol{D} = \mathsf{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$, and $\mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{D})$ denotes the multivariate normal distribution with mean and variance covariance matrix as $\mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{D}$, respectively. The diagonal elements of $\boldsymbol{D}$ have independent GIG distributions. Marginalizing out these scale parameters will yield a likelihood with independent hyperbolic errors with the form given in (2). The aforementioned normal scale-mixture representation of the hyperbolic error model is an important computational trick for developing a Gibbs sampling algorithm for posterior computation in our subsequent Bayesian analysis.

A Bayesian approach requires putting suitable priors on all unknown parameters. For the above model, this requires putting priors on the vector of regression coefficients, $\boldsymbol{\beta}$, as well as on the two other model parameters, namely, η and ρ^2, which correspond to the error distribution. In this article, our goal is to use the horseshoe prior on regression coefficients in conjunction with an hyperbolic error model. To that end, we use the following hierarchical representation of the horseshoe prior in (3), proposed by Makalic and Schmidt (2015), which facilitates posterior computation via Gibbs sampling. In particular, this hierarchical representation leads to closed form full conditional distributions for all unknown parameters, which is a crucial step for the subsequent Bayesian analysis.

$$\begin{aligned}
p(\boldsymbol{\beta}|\lambda_1^2, \lambda_2^2, \ldots, \lambda_p^2, \tau^2, \rho^2) &= \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\rho^2\tau^2\lambda_j^2}} e^{-\frac{1}{2}\left(\frac{\beta_j^2}{\rho^2\tau^2\lambda_j^2}\right)}, \\
p(\lambda_1^2 \ldots, \lambda_p^2|\nu_1, \ldots \nu_p) &= \prod_{j=1}^{p} \frac{(1/\nu_j)^{1/2}}{\Gamma(1/2)} e^{-\frac{1}{\lambda_j^2\nu_j}} \left(\frac{1}{\lambda_j^2}\right)^{\frac{1}{2}+1}, \\
p(\nu_1, \ldots, \nu_p) &= \prod_{j=1}^{p} \frac{1}{\Gamma(1/2)} e^{-\frac{1}{\nu_j}} \left(\frac{1}{\nu_j}\right)^{\frac{1}{2}+1}, \\
p(\tau^2|\xi) &= \frac{(1/\xi)^{1/2}}{\Gamma(1/2)} e^{-\frac{1}{\tau^2\xi}} \left(\frac{1}{\tau^2}\right)^{\frac{1}{2}+1}, \\
p(\xi) &= \frac{1}{\Gamma(1/2)} e^{-\frac{1}{\xi}} \left(\frac{1}{\xi}\right)^{\frac{1}{2}+1}. \qquad (7)
\end{aligned}$$

The hierarchical prior structure in (7) is equivalent to the horseshoe prior in (3) on the regression coefficients, upon marginalization over ν_j's ($j = 1, 2, \ldots, p$) and ξ. As far as the scale parameter ρ^2 and the shape parameter η of the error density are concerned, we put the following priors:

$$\begin{aligned}
p(\rho^2) &= \frac{b^a}{\Gamma(a)} (\rho^2)^{-(a+1)} e^{-b/\rho^2}, \\
p(\eta) &= \frac{1}{K}, \text{ for } \eta \in \{\eta_1, \ldots, \eta_K\}. \qquad (8)
\end{aligned}$$

To reduce ambiguity about different forms of parametrizations, we have directly specified the probability density function (pdf) or probability mass function (pmf) in the above prior specification. In particular, we have specified a conditional normal prior on the regression coefficients, inverse gamma priors on the scale parameters, and a discrete uniform prior on the tail heaviness parameter η. The full conditional distributions corresponding to the above priors lead to standard distributions from which sampling is straightforward. We use Gibbs sampling to approximately sample from the joint posterior distribution.

3. Simulation study

In this section, we generate data from models with hyperbolic errors, and compare the performances of posterior distributions under the horseshoe, lasso, and spike and slab priors. We consider two cases as follows.

3.1. Sparse true model

We first consider an example with $n = 100$ observations and $p = 15$ (excluding the intercept). We generate the errors from a hyperbolic distribution with $\eta = 0.5$ and $\rho^2 = 2$. We set the intercept equal to 2 and specify a relatively sparse model with 5 nonzero regression coefficients, all equal to 3. We generate 100 datasets from this model. We set the priors and hyperparameters for lasso and spike and slab priors following De and Ghosh (2024), denoted by them as HBL (Bayesian Huberized lasso) and HEM (hyperbolic error model), respectively. For the priors proposed by us in this article, given in (7) and (8), we use the same hyperparameters for the tail heaviness and scale parameters, η and ρ^2, respectively, as the other two priors. In particular, we standardize the response variables and each column of the design matrix, to have mean and standard deviation equal to 0 and 1, respectively. We set $a = 2.1$ and $b = 0.1$ for the hyperparameters of the inverse gamma prior on ρ^2, to have most of the prior mass between 0 and 1. This choice is not unreasonable as the response variables have been standardized to have variance equal to 1. For the tail heaviness parameter η, we specify the support points as $\{0.05, 0.1, 0.2, 0.3, \ldots, 1, 2, 5, 10, 20, 50\}$, following De and Ghosh (2024), to have a wide range of tail heaviness parameters. We run the MCMC algorithms for 100,000 iterations, after a burnin of 10,000 iterations. We estimate the regression coefficients using posterior medians of the MCMC samples.

The results are summarized in Figures 1 and 2. The top left panel in Figure 1 shows the root mean squared error (RMSE) for signals (nonzero regression coefficients, excluding the intercept term), that is

$$\sqrt{\sum_{\substack{j=1 \\ \beta_j \neq 0}}^{p} \left(\beta_j - \hat{\beta}_j\right)^2 / 5},$$

where $\hat{\beta}_j$ is the estimate of β_j, $j = 1, \ldots p$, and there are 5 nonzero regression coefficients. This RMSE is similar for the horseshoe and lasso, and somewhat better for the spike and slab prior. The top right panel shows the RMSE for 10 noise variables (zero regression coefficients), that is

$$\sqrt{\sum_{\substack{j=1 \\ \beta_j = 0}}^{p} \left(\beta_j - \hat{\beta}_j\right)^2 / 10}.$$

This is where the spike and slab prior shines, and the horseshoe is significantly better than the lasso, though not as good as the spike and slab prior. The bottom left panel shows the overall RMSE in estimating all regression coefficients (including the intercept β_0), given by

$$\sqrt{\sum_{j=0}^{p} \left(\beta_j - \hat{\beta}_j\right)^2 / (p+1)}.$$

The bottom right panel shows the overall RMSE for each method, relative to the RMSE of the method that has the smallest RMSE for that dataset. The relative RMSE for the spike and slab prior is concentrated around 1, which shows it is the best method overall, followed by the horseshoe, which also seems significantly better than the lasso prior.

Figure 2 shows the effective sample size (ESS), for the MCMC samples of the regression coefficients. ESS can be used to quantify the mixing in the Markov chain, and larger values are preferable. For example, for independent Monte Carlo sampling, the values of ESS would be equal to 100,000, the actual Monte Carlo sample size. For both signals and noise variables, the lasso has the largest ESS, followed by the horseshoe, and the spike and slab priors. Spike and slab priors are known to have slow mixing, so the results are in agreement with this well known fact.

3.2. Non-sparse true model

We next turn to a non-sparse data generating model, with many nonzero regression coefficients. The spike and slab and horseshoe priors are not expected to have as much of an advantage over lasso, in this set up, as they had enjoyed in the previous sparse set up. Here we set $n = 200$ observations and $p = 30$. We set the intercept equal to 2 as earlier, and specify a non-sparse model with 20 nonzero regression coefficients, all equal to 0.8. Everything else is specified as in the earlier simulation study.

The results are presented in Figures 3 and 4. The advantage of the horseshoe prior over the lasso prior disappears in this example, and while the spike and slab prior seems to be the best overall from the bottom right panel in Figure 3, its gains over the other methods is much reduced in this example. This is expected, due to the relatively non-sparse nature of this example. Figure 4 shows that lasso still has the largest values of ESS for both signals and noise variables. Thus this example illustrates scenarios where the lasso prior could be preferable, compared to spike and slab and horseshoe priors.

4. Application to boston housing dataset

We use the Boston housing dataset, available from the `MASS` package in `R`. This dataset is known to be heavy tailed compared to a normal distribution, and has been extensively used as a benchmark dataset in the literature, to illustrate the performance of methods in robust regression. The dataset has $n = 506$ observations and $p = 13$ covariates. The response variable is the median value of occupied homes in Boston, and the covariates are crime rate, property tax, distance to Boston employment centers, access to highways *etc.* We use a log transformation on the response variable, so that the distribution of residuals is roughly symmetric.

A preliminary frequentist linear regression analysis with the usual assumption of normal errors shows that most of the variables have significant p-values; or in other words, the vector of regression coefficients is not sparse. We have illustrated in the second example of the simulation study, that spike and slab and horseshoe priors are are not expected to have much of an advantage in such non-sparse scenarios. To make the application more interesting, we add 30 noise variables, generated from a normal distribution with mean 0 and standard deviation 1, to the original dataset, to have a total of $p = 43$ covariates.

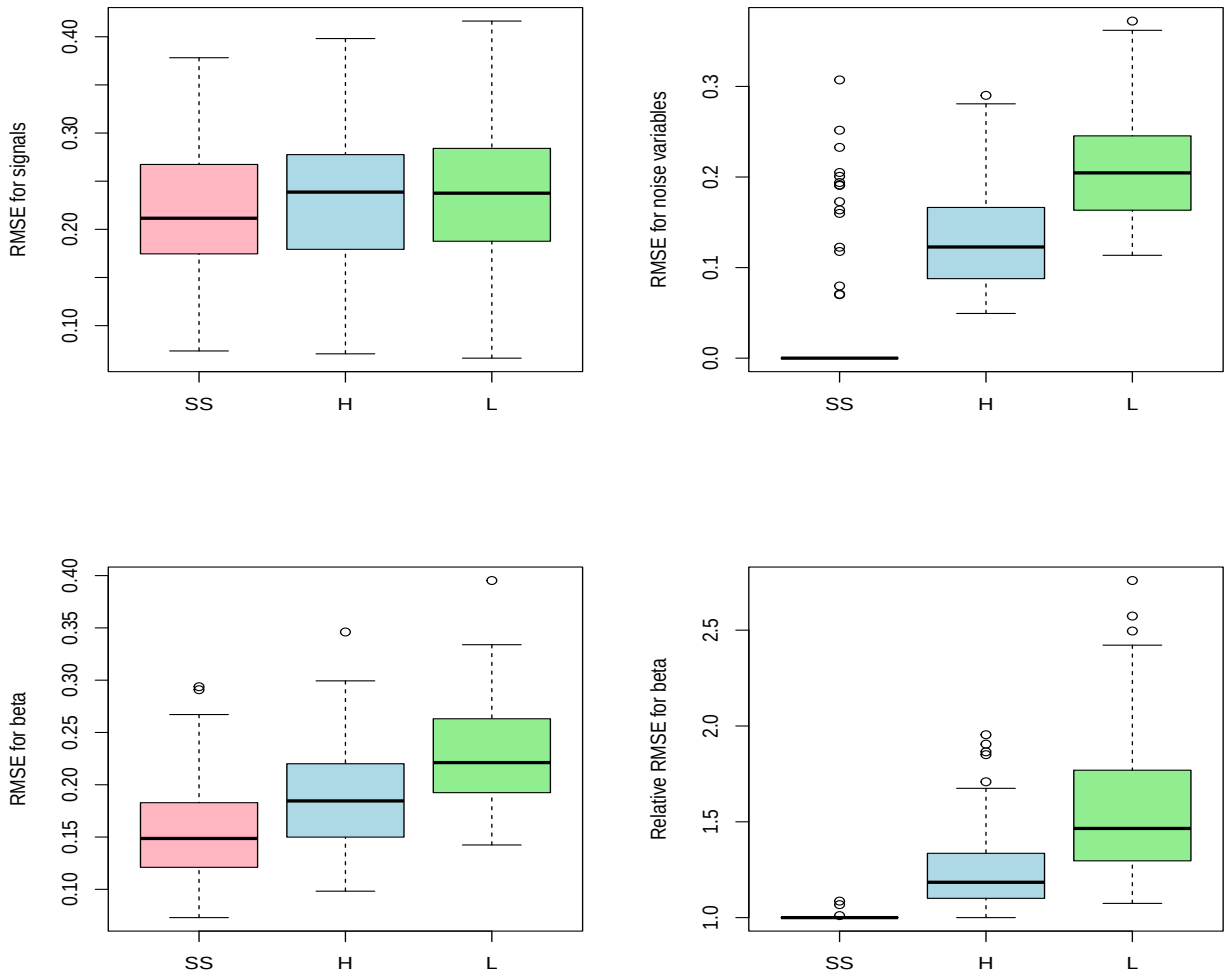


Figure 1: Results for simulation study under sparse true model with $p = 15$ and $n = 100$. Box plots in the top panel show the root mean squared error (RMSE) corresponding to spike and slab (SS), horseshoe (H), and lasso (L) priors for 100 datasets, for signals and noise variables respectively. Box plots in the bottom left panel show the overall RMSE in estimating all the regression coefficients, including the intercept. Box plots in the bottom right panel show the overall RMSE relative to the RMSE of the best method; values of relative RMSE close to 1 indicate that the method is frequently the best.

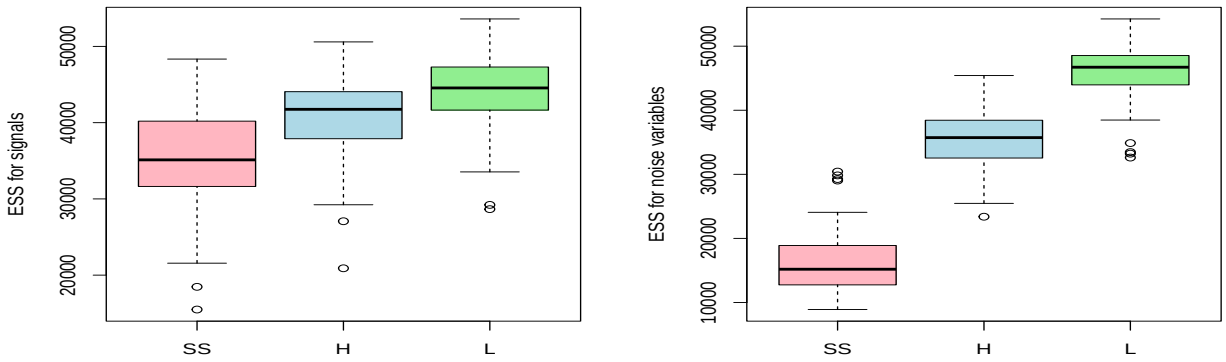


Figure 2: Results for simulation study under sparse true model with $p = 15$ and $n = 100$. Box plots show the effective sample size (ESS) of the MCMC samples for the regression coefficients, corresponding to spike and slab (SS), horseshoe (H), and lasso (L) priors for 100 datasets, for signals and noise variables respectively.

We randomly choose 50% of the observations to include in a training dataset, and use the remaining 50% as a test dataset, to evaluate out of sample predictive performance of the methods with different priors. To reduce sensitivity of the results to a specific choice of training and test data split, we repeat the process 100 times, to create 100 different training and test datasets.

We evaluate the predictive performance using both point and interval estimates. For point estimate for prediction, we use the median of the posterior predictive distribution. For each of the 253 observations in a test dataset, we compute the absolute difference between the observed value of the response variable and its predicted value (both on log scale), and then compute the median of these differences, which we refer to as Median absolute deviation (MAD) for prediction. For 100 test datasets, we get 100 values of MAD. For each test dataset, the method with the smallest MAD is deemed to have the best MAD, and the MAD for the other methods are compared relative to the best MAD. This is repeated 100 times, and presented in the left panel of Figure 5. If a method has values close to 1, that indicates the method has the smallest MAD frequently. The difference between the methods based on different priors is not large, but overall, the spike and slab prior seems to be the best with smallest values of MAD, followed by the horseshoe, and then by the lasso prior. We next consider interval estimates for prediction by estimating 90% equal-tailed prediction intervals for each observation in the test datasets. The resulting frequentist coverage of the prediction intervals is shown in the right panel of Figure 5. All methods seem to have coverage close to 90%, shown by the dashed line, though there is some variability around 90%. The diamond in each box plot shows the overall coverage across 100 test datasets, which seems fairly close to 90%.

5. Discussion

In this article, we have introduced an algorithm based on the horseshoe prior, for robust regression with hyperbolic errors, as an alternative to existing methods that rely on

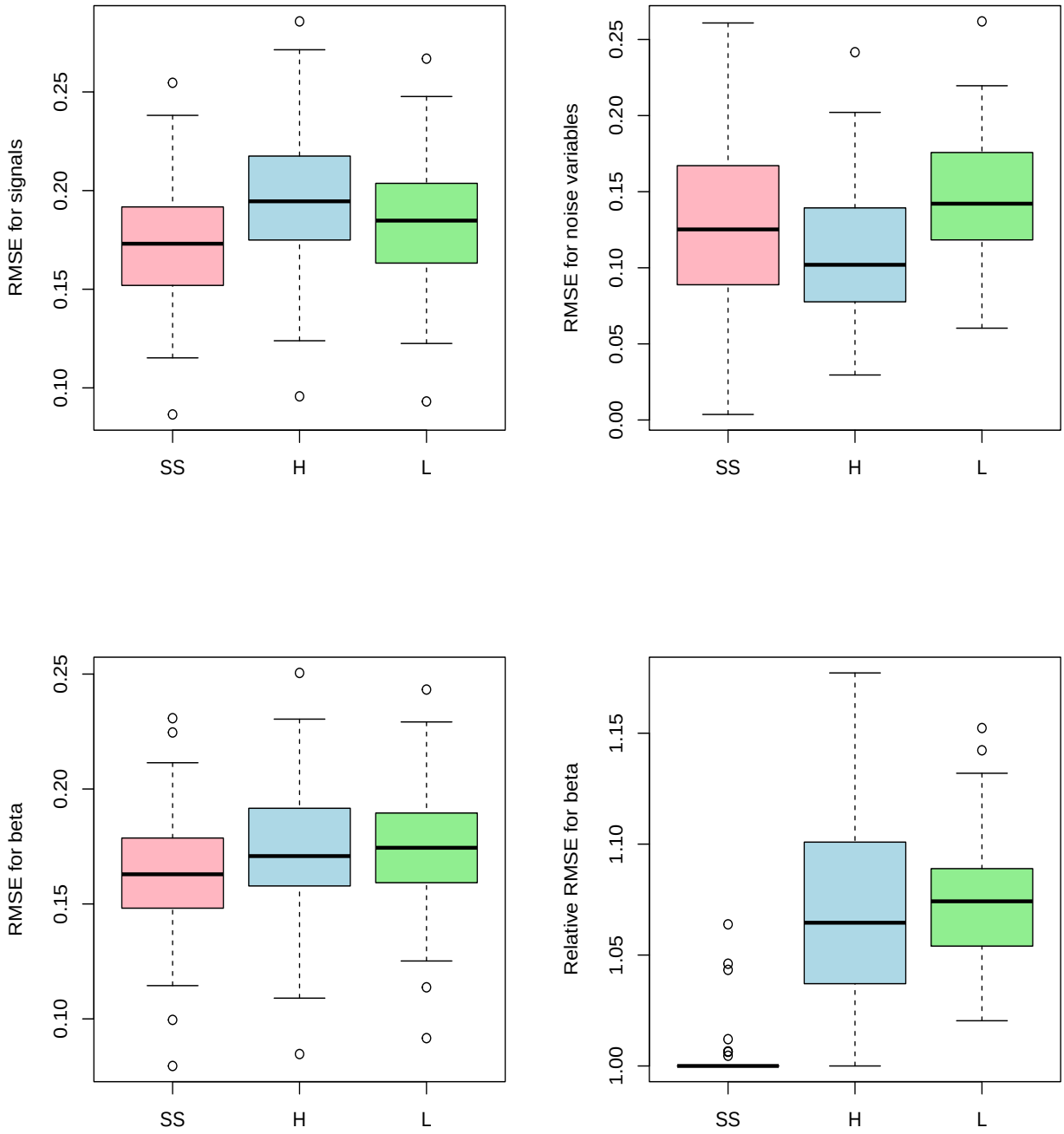


Figure 3: Results for simulation study under non-sparse true model with $p = 30$ and $n = 200$. Box plots in the top panel show the root mean squared error (RMSE) corresponding to spike and slab (SS), horseshoe (H), and lasso (L) priors for 100 datasets, for signals and noise variables respectively. Box plots in the bottom left panel show the overall RMSE in estimating all the regression coefficients, including the intercept. Box plots in the bottom right panel show the overall RMSE relative to the RMSE of the best method; values of relative RMSE close to 1 indicate that the method is frequently the best.

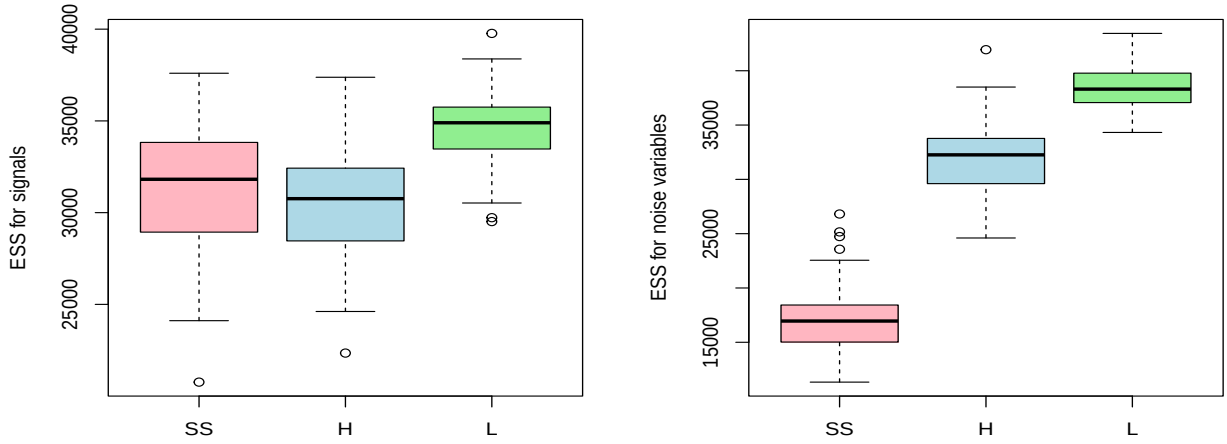


Figure 4: Results for simulation study under non-sparse true model with $p = 30$ and $n = 200$. Box plots show the effective sample size (ESS) of the MCMC samples for the regression coefficients, corresponding to spike and slab (SS), horseshoe (H), and lasso (L) priors for 100 datasets, for signals and noise variables respectively.

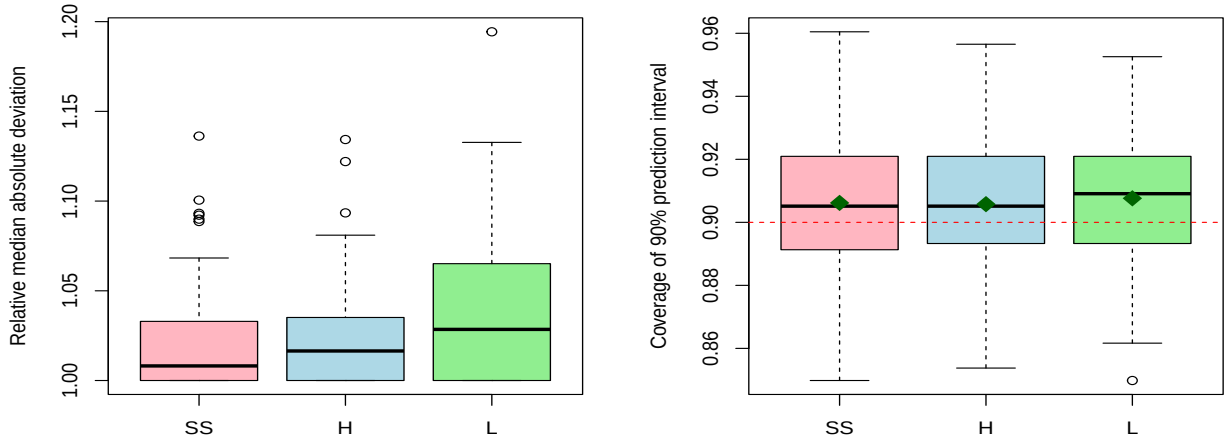


Figure 5: Results for the Boston housing dataset, after adding 30 noise variables to the original dataset with $p = 13$ covariates. Box plots in the left panel show the median absolute deviation (MAD) for out of sample prediction, corresponding to spike and slab (SS), horseshoe (H), and lasso (L) priors relative to the method with the least MAD for that dataset, for 100 test datasets. The right panel shows the corresponding coverage for 90% prediction intervals; the dashed line is at 0.9 and the diamonds represent the overall mean coverage over 100 test datasets.

the spike and slab or lasso priors. Our results based on simulation studies suggest that the horseshoe prior can improve upon the lasso prior in estimating sparsity. The horseshoe prior seems to be outperformed by the spike and slab prior, in terms of accuracy in recovering true parameters, for moderate dimensional problems that we investigated in this article.

We found that the mixing in the Markov chain for the horseshoe prior is consistently better than that of spike and slab priors. It is well known that posterior computation for Bayesian variable selection with spike and slab priors, does not scale well with high dimensions. So for large p, the horseshoe prior could offer an alternative approach, given its improved mixing. For hyperbolic regression, we found computation under the horseshoe prior to be somewhat unstable for large p, due to having to invert large $p \times p$ matrices. Further investigation is needed regarding how to make the computation more stable. One possible direction is using the regularized horseshoe prior of Piironen and Vehtari (2017).

Acknowledgements

Joyee Ghosh's research was supported by NSF Grant DMS-1612763. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation. The authors thank the Editors for helpful suggestions that improved the quality of the paper.

References

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B*, **36**, 99–102.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2017). The horseshoe estimator of ultra-sparse signals. *Bayesian Analysis*, **12**, 1105–1131.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97**, 465–480.

De, S. and Ghosh, J. (2024). Robust Bayesian model averaging for linear regression models with heavy-tailed errors. Technical report.

Gneiting, T. (1997). Normal scale mixtures and dual probability densities. *Journal of Statistical Computation and Simulation*, **59**, 375–384.

Hamura, Y., Irie, K., and Sugasawa, S. (2022). Log-regularly varying scale mixture of normals for robust regression. *Computational Statistics and Data Analysis*, **173**, 107517.

Kawakami, J. and Hashimoto, S. (2023). Approximate Gibbs sampler for Bayesian huberized lasso. *Journal of Statistical Computation and Simulation*, **93**, 128–162.

Makalic, E. and Schmidt, D. F. (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, **23**, 179–182.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, **103**, 681–686.

Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, **11**, 5018 – 5051.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, **74**, 646–648.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 211–225
http://www.ssca.org.in/journal

Testing with Cubic Smoothing Splines

Tapio Nummi[1], Jyrki Möttönen[2] and Jianxin Pan[3]
[1]*Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland*
[2]*Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland*
[3]*Faculty of Science and Technology, BNU-HKBU United International College, Zhuhai, Guangdong 519087, PR China*

Received: 28 May 2024; Revised: 3 August 2024; Accepted: 10 August 2024

Abstract

In this paper, we present some possible ways to perform estimation and testing for cubic smoothing splines. Special emphasis is placed on the analysis of correlated data, when using semi-parametric regression models (Schimek, 2000), and the so-called spline growth model (Nummi and Koskela, 2008; Nummi *et al.*, 2017), an extension of the basic growth curve model (Potthoff and Roy, 1964; Rao, 1965). Furthermore, practical applications in fields such as medicine and animal breeding are introduced, highlighting the versatility and efficacy of cubic smoothing splines in real-world applications.

Key words: Covariance structures; Eigenvalue decomposition; Growth curves; Semi-parametric regression.

AMS Subject Classifications: 62J05, 62J10

1. Introduction

In our paper, we specifically delve into the intricacies of cubic smoothing splines. One of the standout advantages inherent in smoothing splines is their adaptability, granting precise control over the delicate balance between interpolating data points and maintaining the overall smoothness of the curve. This control is facilitated by a smoothing parameter, empowering researchers to fine-tune the model for optimal performance. For statistical inference with smoothing splines and semi-parametric regression we can refer to the books by Eubank and Spiegelman (1990), Green and Silverman (1993), Ruppert *et al.* (2003), Wu and Zhang (2006), Harezlak *et al.* (2018) and Stasinopoulos *et al.* (2017), for example.

The notable flexibility of smoothing splines extends beyond their ability to capture intricate data patterns. They also boast a range of theoretical properties that significantly enhance their utility. In various scenarios, smoothing splines emerge as a compelling alternative to parametric models. This preference arises from the inherent challenge of justifying

Corresponding Author: Tapio Nummi
Email: tapio.nummi@tuni.fi

the choice of a parametric function, which often lacks a clear rationale or relies on a rough approximation of the true underlying function form.

Characterized by their high flexibility, splines offer an advantageous choice by providing a flexible and accurate approximation of the true function form. This is particularly valuable in situations where a clear parametric alternative may prove elusive or is based on a rough approximation. The limitations of parametric models become especially evident when testing different competing models against each other, as they typically also provide a limited set of possible alternative hypotheses. In contrast, cubic smoothing splines offer a very broad family of alternative model choices. When pitted against corresponding parametric models, they not only showcase their adaptability but also present a more comprehensive and versatile set of alternatives for a more robust model comparison. In this context papers by Speckman (1988), Eubank and Hart (1992), Azzalini and Bowman (1993), Cantoni and Hastie (2002), Härdle *et al.* (1998), Lin and Zhang (1999), Verbyla *et al.* (1999), Schimek (2000),Zhang and Lin (2003), Liu and Wang (2004), Nummi *et al.* (2011), and Nummi *et al.* (2013) serve as valuable references. This paper concentrates on the inference of cubic smoothing splines and semi-parametric regression. Our methods exhibits flexibility also in the sense that they apply also under correlated data, further extending its utility for testing growth curves (Koskela *et al.*, 2006; Nummi and Mesue, 2013; Nummi *et al.*, 2017), for example.

In Section 2.1, we present some methods used to estimate cubic smoothing splines and corresponding semi-parametric regression models. Subsequently, in Section 3, we elucidate techniques for accurately approximating the spline fit, and introduce a comprehensive set of hypotheses and tests relevant to semi-parametric regression models. Furthermore, we illustrate these methods with an example of medical testing, demonstrating their practical application potential. In Sections 4 and 5, we focus on estimation and testing in a spline growth model and its multivariate extension. These methods are illustrated with a practical application on animal breeding. In Section 6, some concluding remarks are provided.

2. Cubic smoothing splines and semi-parametric regression

2.1. Cubic smoothing splines

Consider the vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$, observed at measuring points $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top$ on the interval $[a, b]$, where $a < x_1 < x_2 < \cdots < x_n < b$. A cubic smoothing spline can be expressed as

$$\mathbf{y} = \mathbf{g} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{g} = (g(x_1), g(x_2), \ldots, g(x_n))^\top$ represents a vector of the smooth, twice-differentiable curve $g(\cdot)$. The term $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top \sim N_n(\mathbf{0}, \sigma^2\mathbf{R})$ accounts for normally distributed errors, where $\mathbf{R}$ is a covariance matrix characterized by parameters within the vector $\boldsymbol{\theta}$.

The estimation of cubic smoothing splines $\mathbf{g}$ can be achieved through a penalized least squares criterion (PLS). This process commences by defining the roughness matrix $\mathbf{K} = \nabla\boldsymbol{\Delta}^{-1}\nabla^\top$, wherein the non-zero elements of the banded $n \times (n-2)$ matrix ∇ and the $(n-2) \times (n-2)$ matrix $\boldsymbol{\Delta}$ are given by

$$\nabla_{k,k} = \frac{1}{h_k}, \quad \nabla_{k+1,k} = -\left(\frac{1}{h_k} + \frac{1}{h_{k+1}}\right), \quad \nabla_{k+2,k} = \frac{1}{h_{k+1}}$$

and

$$\Delta_{k,k+1} = \Delta_{k+1,k} = \frac{h_{k+1}}{6}, \quad \Delta_{k,k} = \frac{h_k + h_{k+1}}{3},$$

where $k = 1, 2, \ldots, (n-2)$ and $h_j = x_{j+1} - x_j$, with $j = 1, 2, \ldots, (n-1)$. The penalized least squares criterion at points $x_1, x_2, \ldots, x_n$ is then expressed as

$$Q_1 = (\mathbf{y} - \mathbf{g})^\top \mathbf{R}^{-1}(\mathbf{y} - \mathbf{g}) + \alpha \mathbf{g}^\top \mathbf{K}\mathbf{g} \tag{2}$$

The minimum with a fixed positive smoothing parameter α is a cubic smoothing spline (*e.g.* Green and Silverman (1993))

$$\tilde{\mathbf{g}} = (\mathbf{H} + \alpha\mathbf{K})^{-1}\mathbf{H}\mathbf{y} = \mathbf{S}_\alpha \mathbf{y}, \tag{3}$$

where we denote $\mathbf{H} = \mathbf{R}^{-1}$ and $\mathbf{S}_\alpha = (\mathbf{H} + \alpha\mathbf{K})^{-1}\mathbf{H}$ is the so-called *smoother matrix*. It is easily seen that if the covariance matrix $\mathbf{R}$ satisfies the equation

$$\mathbf{R}\mathbf{K} = \mathbf{K}, \text{ or equivalently, } \mathbf{K} = \mathbf{H}\mathbf{K}, \tag{4}$$

the smoother matrix reduces to the form $\mathbf{S}_\alpha = (\mathbf{I} + \alpha\mathbf{K})^{-1}$. The resulting spline estimator in this case becomes as Nummi and Koskela (2008),

$$\hat{\mathbf{g}} = (\mathbf{I} + \alpha\mathbf{K})^{-1}\mathbf{y}. \tag{5}$$

It can be seen that this estimator does not depend on the covariance matrix $\mathbf{R}$. It is demonstrated in Nummi *et al.* (2011) that certain important covariance structures used in the analysis of repeated measures or longitudinal data satisfy condition (4). These structures include the uniform covariance structure $\mathbf{R} = \mathbf{I} + d^2\mathbf{1}\mathbf{1}^\top$ and the linear structure $\mathbf{R} = \mathbf{I} + \mathbf{X}\mathbf{D}\mathbf{X}^\top$, where $d^2 > 0$, $\mathbf{D}$ is positive definite, and $\mathbf{X} = (\mathbf{1}, \mathbf{x})$, for example. It is worth noting that in this scenario, when the smoothing parameter α is fixed, the estimated splines become simple linear functions of the observations $y_1, y_2, \ldots, y_n$, and further this offers also the possibility to use the methodology in the case of correlated data, which will be tackled in particular in Section 4.

2.2. Semi-parametric regression

The spline model in (1) seamlessly extends into a semi-parametric regression model

$$\mathbf{y} = \mathbf{U}\mathbf{b} + \mathbf{g} + \epsilon, \tag{6}$$

where $\mathbf{U}\mathbf{b}$ represents the linear component, with $\mathbf{U}$ being a full-rank $n \times k$ matrix of values of k explanatory variables (excluding the constant term), and $\mathbf{b}$ a k-vector of unknown parameters. Semi-parametric regression models have been considered in Nummi *et al.* (2013), Green and Silverman (1993), Schimek (2000), and Wu and Zhang (2006), for example. The PLS criterion for this case is expressed as

$$Q_2 = [\mathbf{y} - (\mathbf{U}\mathbf{b} + \mathbf{g})]^\top \mathbf{H}[\mathbf{y} - (\mathbf{U}\mathbf{b} + \mathbf{g})] + \alpha \mathbf{g}^\top \mathbf{K}\mathbf{g}. \tag{7}$$

Minimizing with respect to $\mathbf{b}$ and $\mathbf{g}$ leads to the following estimates (Green and Silverman, 1993)

$$\tilde{\mathbf{b}} = [\mathbf{U}^\top \mathbf{H}(\mathbf{I} - \mathbf{S}_\alpha)\mathbf{U}]^{-1}\mathbf{U}^\top \mathbf{H}(\mathbf{I} - \mathbf{S}_\alpha)\mathbf{y} \tag{8}$$

and

$$\tilde{\mathbf{g}} = \mathbf{S}_\alpha(\mathbf{y} - \mathbf{U}\tilde{\mathbf{b}}), \tag{9}$$

where $\mathbf{S}_\alpha = (\mathbf{H} + \alpha\mathbf{K})^{-1}\mathbf{H}$. It can be shown that if the condition (4) holds, the estimates simplify to (Nummi *et al.*, 2013)

$$\hat{\mathbf{b}} = [\mathbf{U}^\top(\mathbf{I} - \mathbf{S}_\alpha)\mathbf{U}]^{-1}\mathbf{U}^\top(\mathbf{I} - \mathbf{S}_\alpha)\mathbf{y}, \tag{10}$$

where $\hat{\mathbf{g}} = \mathbf{S}_\alpha(\mathbf{y} - \mathbf{U}\hat{\mathbf{b}})$, $\mathbf{S}_\alpha = (\mathbf{I} + \alpha\mathbf{K})^{-1}$ and the fitted semi-parametric curve can be obtained as

$$\hat{\mu} = \mathbf{M}\mathbf{y}, \tag{11}$$

where $\mathbf{M} = \mathbf{S}_\alpha + \tilde{\mathbf{U}}[\tilde{\mathbf{U}}^\top\mathbf{U}]^{-1}\tilde{\mathbf{U}}^\top$ and $\tilde{\mathbf{U}} = (\mathbf{I} - \mathbf{S}_\alpha)\mathbf{U}$, respectively. It appears that, once the smoothing parameter α is fixed, the estimation process for both the cubic smoothing spline and the semi-parametric model becomes quite straightforward. In the upcoming chapter, we will delve into the methodologies employed for hypothesis testing.

3. Testing

3.1. Approximate fit

Testing in the context of cubic splines poses challenges, primarily because the smoother matrix inherently lacks the properties of a projector matrix. Consequently, established methods, such as those developed for linear models, do not seamlessly apply to cubic splines. Here we outline a few potential avenues and methodologies for conducting tests related to various hypotheses.

Our approach is centered around approximating the smoother matrix $\mathbf{S}_\alpha$ with a matrix possessing the properties of a projector matrix. This approximation not only yields a highly accurate representation of a cubic smoothing spline fit but also generates a cubic spline itself, as it is rooted in a linear combination of cubic splines (Nummi *et al.*, 2011). It can be demonstrated that $\mathbf{S}_\alpha$ can be decomposed as (see also Hastie (1996))

$$\mathbf{S}_\alpha = \mathbf{T}(\mathbf{I} + \alpha\mathbf{\Lambda})^{-1}\mathbf{T}^\top, \tag{12}$$

where the matrix of eigenvectors $\mathbf{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_n)$ can be directly calculated from the roughness matrix $\mathbf{K}$, and the eigenvalues $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_n)$ of $\mathbf{K}$ are interrelated with $\mathbf{S}_\alpha$ such that the eigenvalues of $\mathbf{S}_\alpha$ are given by $\gamma_j = 1/(1 + \alpha\lambda_i)$, indicating a reverse order of eigenvectors of $\mathbf{K}$ and $\mathbf{S}_\alpha$. Intriguingly, the sequence of eigenvectors of $\mathbf{S}_\alpha$ appears to increase in complexity like a sequence of orthogonal polynomials (see *e.g.*, Ruppert *et al.* (2003)), and the eigenvalues $\gamma_j \in (0, 1)$ show how much dumping is made for each $\mathbf{t}_j$ when the smoother is applied. We can effectively approximate $\mathbf{S}_\alpha$ by

$$\mathbf{M}_c = \mathbf{T}_c\mathbf{T}_c^\top, \tag{13}$$

where $\mathbf{T}_c = (\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_c)$ denotes the first c eigenvectors of $\mathbf{T}$, which can be chosen using modified generalized cross-validation criteria (Nummi and Mesue, 2013)

$$GCV_1(c) = \frac{\frac{1}{n}\sum_{i=i}^{n}[y_i - \bar{y}_i]^2}{(1 - \frac{c}{n})^2}, \tag{14}$$

where $\bar{y}_i$ is now computed using the formula (5) with $\mathbf{S}_\alpha$ replaced by $\mathbf{M}_c$, for instance. It was demonstrated in Nummi *et al.* (2011) that this yields a pretty good approximation especially if the number of effective degrees of freedom is not unreasonably large. Further decomposition of $\mathbf{T}_c$ $(c > 2)$ takes the form $\mathbf{T}_c = (\mathbf{T}_2, \mathbf{T}_{c-2})$, where $\mathbf{T}_2$ encompasses the first two eigenvectors, and $\mathbf{T}_{c-2}$ comprises the remaining eigenvectors. Note that we can take $\mathbf{T}_2 = (\mathbf{t}_1, \mathbf{t}_2)$, where $\mathbf{t}_1 = \mathbf{1}/\sqrt{n}$ and $\mathbf{t}_2$ is given by $t_{2i} = (x_i - \bar{x})/S_x^2$, where $\bar{x}$ is the mean of $x_i s$ and $S_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2$ (Nummi *et al.*, 2011). It is easy to see that $\mathbf{t}_1$ and $\mathbf{t}_2$ span a straight-line model.

We can now approximate $\hat{\mathbf{g}}$ for model (1) by

$$\tilde{\mu} = \mathbf{M}_c\mathbf{y} = (\mathbf{M}_1 + \mathbf{M}_2)\mathbf{y}, \tag{15}$$

where $\mathbf{M}_1 = \mathbf{T}_2\mathbf{T}_2^\top$ and $\mathbf{M}_2 = \mathbf{T}_{c-2}\mathbf{T}_{c-2}^\top$ and further for the model (11) we have

$$\tilde{\mu} = \tilde{\mathbf{M}}\mathbf{y} = (\mathbf{M}_c + \mathbf{M}_3)\mathbf{y}, \tag{16}$$

where $\mathbf{M}_3 = \bar{\mathbf{U}}[\bar{\mathbf{U}}^\top\mathbf{U}]^{-1}\bar{\mathbf{U}}^\top$ and $\bar{\mathbf{U}} = (\mathbf{I} - \mathbf{M}_c)\mathbf{U}$, respectively.

3.2. Hypotheses and test statistics

Testing is based on sums of squares as defined in this paragraph. It is first noted that if we have the correlation model $\mathbf{R} = \mathbf{I} + \mathbf{X}\mathbf{D}\mathbf{X}^\top$, for example, we have $\tilde{\mathbf{M}}\mathbf{X}\mathbf{D}\mathbf{X}^\top = \mathbf{X}\mathbf{D}\mathbf{X}^\top$ and therefore

$$(\mathbf{I} - \tilde{\mathbf{M}})(\mathbf{I} + \mathbf{X}\mathbf{D}\mathbf{X}^\top)(\mathbf{I} - \tilde{\mathbf{M}}) = (\mathbf{I} - \tilde{\mathbf{M}}). \tag{17}$$

We can further note that, under normality and the assumed correlation model, the following relationships hold (Nummi *et al.*, 2013)

$$\sigma^{-2}\mathbf{y}^\top(\mathbf{I} - \tilde{\mathbf{M}})\mathbf{y} = \sigma^{-2}S_{min} \sim \chi^2_{n-c-k}. \tag{18}$$

Similarly, we can define

$$\sigma^{-2}\mathbf{y}^\top(\mathbf{I} - \mathbf{M}_c)\mathbf{y} = \sigma^{-2}S_{spl} \sim \chi^2_{n-c}, \tag{19}$$

$$\sigma^{-2}\mathbf{y}^\top(\mathbf{I} - \mathbf{M}_1)\mathbf{y} = \sigma^{-2}S_{lin} \sim \chi^2_{n-2} \tag{20}$$

and

$$\sigma^{-2}\mathbf{y}^\top(\mathbf{I} - \mathbf{P}_i)\mathbf{y} = \sigma^{-2}S_{reg,i} \sim \chi^2_{n-k-i}, i = 1, 2, \tag{21}$$

where $\mathbf{P}_i = \mathbf{U}_i(\mathbf{U}_i^\top\mathbf{U}_i)^{-1}\mathbf{U}_i$, where for $i = 1, \mathbf{U}_1 = (\mathbf{1}, \mathbf{U})$ and $i = 2, \mathbf{U}_2 = (\mathbf{X}, \mathbf{U})$, and where $\mathbf{X} = (\mathbf{1}, \mathbf{X})$, respectively. These sum-of-squares expressions can now be utilized for the hypothesis testing of different special cases of the basic semi-parametric model. We can now formulate a set of compelling hypotheses each designed to assess various aspects of the models introduced. Note that the tests introduced in this section are applicable also to correlated data, provided an appropriate form of covariance matrix is employed.

3.2.1. Test 1: Cubic smoothing spline

The first test introduced here aims to scrutinize whether the basic linear model is applicable when compared to the assumed cubic smoothing spline alternative (model (1)). The hypotheses are formulated as follows

$$\mathbf{H}_0\colon\ \mu = \mathbf{X}\mathbf{b}_2,$$

where $\mathbf{X} = [\mathbf{1}, \mathbf{x}]$ and $\mathbf{b}_2$ is a vector of two regression coefficients. The alternative hypothesis is

$$\mathbf{H}_a\colon\ \mu = \mathbf{g},$$

where $\mathbf{g}$ represents the assumed spline model. Since $\mathbf{M}_c\mathbf{M}_1 = \mathbf{M}_1$ (columns $\mathbf{M}_1$ are in the span of $\mathbf{M}_c$) it is observed that $(\mathbf{I} - \mathbf{M}_c)(\mathbf{M}_c - \mathbf{M}_1) = \mathbf{0}$ and therefore S_{spl} and $S_{lin} - S_{spl}$ are independent and

$$F_1 = \frac{(S_{lin} - S_{spl})/(c-2)}{S_{spl}/(n-c)} \sim F(c-2, n-c). \tag{22}$$

Then observing a larger F_1 than quantile $F_{1-\alpha}(c-2, n-c)$ yields the rejection of the null hypothesis. It was shown in a power study of Nummi *et al.* (2011) that this test performed very well when compared to other alternatives.

3.2.2. Tests 2: Semi-parametric model

A) Testing the significance of linear covariates in the full model

Suppose the full semi-parametric model may include a set of linear covariates, denoted as $\mathbf{U}$. We first test the significance of this set in the full model. The null hypothesis is

$$\mathbf{H}_0\colon\ \mu = \mathbf{g},$$

and the alternative hypothesis, a full semi-parametric model, is

$$\mathbf{H}_a\colon\ \mu = \mathbf{U}\mathbf{b} + \mathbf{g},$$

where $\mathbf{U}\mathbf{b}$ is a linear term and $\mathbf{g}$ is a smoothing spline term. Using similar arguments as before, we get

$$F_{2A} = \frac{(S_{spl} - S_{min})/k}{S_{min}/(n-k-c)} \sim F(k, n-k-c). \tag{23}$$

If the observed F_{2A} is larger than the critical value $F_{1-\alpha}(k, n-k-c)$, we reject the null hypothesis.

B) Assessing the fit of the model with linear model

This test evaluates whether the assumed linear model provides a better fit compared to a semi-parametric alternative. The hypotheses are defined as follows

$$\mathbf{H}_0\colon\ \mu = \mathbf{U}_{k+2}\mathbf{b}_{k+2},$$

where $\mathbf{U}_{k+2} = [\mathbf{X}, \mathbf{U}]$, $\mathbf{X} = [\mathbf{1}, \mathbf{x}]$, and $\mathbf{b}_{k+2}$ is a vector of $k+2$ regression coefficients. The alternative hypothesis remains the same as in part A. The test statistic for this hypothesis becomes

$$F_{2B} = \frac{(S_{reg,2} - S_{min})/(c-2)}{S_{min}/(n-k-c)} \sim F(c-2, n-k-c). \tag{24}$$

If the observed F_{2B} exceeds the critical value $F_{1-\alpha}(c-2, n-k-c)$, we reject the null hypothesis. According to Nummi *et al.* (2013), the power of this test was investigated through a simulation study. The study found that estimating c from the observed data results in only a minimal loss of power compared to the scenario where c is known.

3.2.3. Test 3: Linear model

Ultimately, we can explore the need to include the variable $\mathbf{x}$, which was initially presumed to be a smooth term ($c > 2$), as a linear term alongside other linear terms within a full linear model. The hypotheses are formulated as

$$\mathbf{H}_0\colon\ \mu = \mathbf{U}_{k+1}\mathbf{b}_{k+1},$$

where $\mathbf{U}_{k+1} = [\mathbf{1}, \mathbf{U}]$, and $\mathbf{b}_{k+1}$ is a $k+1$ vector of regression coefficients. The alternative hypothesis is

$$\mathbf{H}_a\colon\ \mu = \mathbf{U}_{k+2}\mathbf{b}_{k+2},$$

where $\mathbf{U}_{k+2} = [\mathbf{X}, \mathbf{U}]$ and this can be tested as

$$F_3 = \frac{(S_{reg,1} - S_{reg,2})}{S_{reg,2}/(n-k-2)} \sim F(1, n-k-2). \tag{25}$$

Then observing a larger F_3 than quantile $F_{1-\alpha}(1, n-k-2)$ yields the rejection of the null hypothesis.

Example 1: PSA testing

As an illustration, we utilized part of the dataset gathered for the Finnprostate Study VII conducted by Professor Teuvo L. J. Tammela in Finland in 1990-2000 at Tampere University. The primary objective of this study was to examine individuals susceptible to prostate cancer. It is important to note that in this article, we will refrain from delving into the medical intricacies of the subject matter. Instead, our focus is solely on employing this dataset to exemplify the methodologies presented.

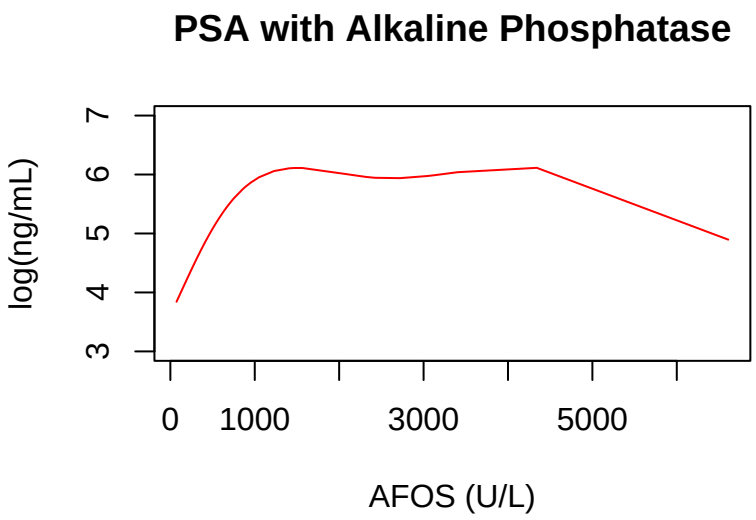


Figure 1: Plot of approximated spline fit for the values of prostate-specific antigen (log(ng/mL) as a function of values of alkaline phosphatase test (U/L).

In this instance, our examination of 537 individuals is centered on the variables Prostate-Specific Antigen (PSA, ng/mL), Body Mass Index (BMI, kg/m^2), Prostate Length (Length, cm), and Alkaline Phosphatase test (AFOS, U/L). The primary aim of our study is to construct a model for the log(PSA) value utilizing the variables AFOS, BMI, and Length. To commence, we explore the relationship between PSA and AFOS, assuming that a suitable spline model would best describe this connection. Employing the criteria $GCV^*(c)$, where c ranges from 1 to 6, we obtain the values 6.3152, 0.9808, 0.9042, 0.8791, 0.8755, and 0.8769. Consequently, our preferred choice is $c = 5$. It should be noted that for some measuring points $x_1, \ldots, x_n$, we have multiple values and therefore we need to replace the smoother matrix $\mathbf{S}_\alpha$ by

$$\mathbf{S}_\alpha = \mathbf{N}(\mathbf{N}^\top\mathbf{N} + \alpha\mathbf{K})^{-1}\mathbf{N}^\top, \tag{26}$$

where $\mathbf{N}$ is an incidence matrix of corresponding measuring times. The approximated spline fit is depicted in Figure 1. Upon subjecting this to a linear model test (Test 1), we obtain $F_1 = 22.45$, with the corresponding quantile $F_{0.95}(3, 532) = 2.622$. This unequivocally rejects the null hypothesis concerning a linear association.

Subsequently, we delve into semi-parametric model 6. Our preliminary analysis suggests that BMI, Length, and the interaction term Alkaline×OI can be utilized as explanatory variables in the $\mathbf{U}$-matrix, where OI is the obesity indicator ($OI = 1$ if $BMI > 30$, and 0 otherwise). Alkaline with $c = 5$ is used in (16) for model fitting and testing. Using the test statistic F_{2A}, we evaluated the significance of this set of covariates within the full semi-parametric model. The resulting value,$F_{2A} = 17.06$, exceeds the corresponding critical value $F_{0.95}(3, 529) = 2.62$, indicating clear significance. Additionally, the value of the test statistic F_{2B} is 28.63, which also surpasses the critical value $F_{0.95}(3, 529) = 2.62$. Consequently, the null hypothesis is firmly rejected, confirming that the model is semi-parametric rather than fully linear. Test 3 is not executed in this scenario, as it only becomes relevant if the null hypothesis from Test 2B is accepted.

4. Testing for growth data

In certain cases, growth modeling can be grounded in a theoretical framework, enabling the derivation of a parametric model for developmental processes. However, more frequently, such a theoretical foundation may be lacking, necessitating the adoption of alternative approximations. We found that cubic smoothing splines for many cases provide a well justified alternative since they quite accurately follow the true growth function. In the following, we outline the methodology for testing some relevant hypotheses when employing cubic smoothing splines to model the growth function.

The growth curve model of Potthoff and Roy (Potthoff and Roy, 1964) can be written as

$$\mathbf{Y} = \mathbf{T}\mathbf{B}\mathbf{A}^\top + \mathbf{E}, \tag{27}$$

where $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$ is the $q \times n$ matrix of independent $q \times 1$ response vectors, $\mathbf{T}$ is a $q \times p$ within-individual design matrix, $\mathbf{A}$ is an $n \times m$ between-individual design matrix, $\mathbf{B}$ is an unknown $p \times m$ parameter matrix to be estimated and $\mathbf{E}$ is a $q \times n$ matrix of random errors. It is assumed that the columns $\mathbf{e}_1, \ldots, \mathbf{e}_n$ of $\mathbf{E}$ are independently normally distributed as $\mathbf{e}_i \sim N_q(\mathbf{0}, \mathbf{\Sigma}),\ i = 1, \ldots, n$.

We can write model (27) in a more general way by using cubic smoothing splines. Let

$$\mathbf{Y} = \mathbf{G}\mathbf{A}^\top + \mathbf{E}, \tag{28}$$

where $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_m)$ is the matrix of smooth mean growth curves at time points $t_1, t_2, \ldots, t_q$. We further assume that $\mathbf{\Sigma}$ is a parsimonious covariance structure $\mathbf{\Sigma} = \sigma^2\mathbf{R}(\boldsymbol{\theta})$ with covariance parameters $\boldsymbol{\theta}$. Model (28) is referred to as the spline growth model (SGM). Note that we get the Potthoff and Roy model as a special case by setting $\mathbf{G} = \mathbf{T}\mathbf{B}$. The smooth solution for the matrix of mean growth curves $\mathbf{G}$ can be obtained by minimizing the following penalized least squares (PLS) objective function

$$Q = \mathrm{tr}[(\mathbf{Y} - \mathbf{G}\mathbf{A}^\top)^\top\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{G}\mathbf{A}^\top) + \alpha\mathbf{A}\mathbf{G}^\top\mathbf{K}\mathbf{G}\mathbf{A}^\top], \tag{29}$$

where α is a fixed smoothing parameter and $\mathbf{K}$ is the roughness matrix defined in Section 2.1. It can be easily seen that Q can be rewritten n the form

$$Q = \mathrm{tr}[(\mathbf{G}\mathbf{A}^\top - (\mathbf{H} + \alpha\mathbf{K})^{-1}\mathbf{H}\mathbf{Y})^\top(\mathbf{H} + \alpha\mathbf{K})(\mathbf{G}\mathbf{A}^\top - (\mathbf{H} + \alpha\mathbf{K})^{-1}\mathbf{H}\mathbf{Y})] + w, \tag{30}$$

where $\mathbf{H} = \mathbf{R}^{-1}$, $(\mathbf{H}+\alpha\mathbf{K})$ is a positive definite matrix and $w = \mathrm{tr}[\mathbf{Y}^\top\mathbf{H}^{-1}(\mathbf{H}+\alpha\mathbf{K})^{-1}\mathbf{H}^{-1}\mathbf{Y} - \mathbf{Y}^\top\mathbf{H}\mathbf{Y}]$ does not depend on $\mathbf{G}$. The function Q is minimized for fixed values of α and $\mathbf{H}$ when $\mathbf{G}\mathbf{A}^\top = (\mathbf{H} + \alpha\mathbf{K})^{-1}\mathbf{H}\mathbf{Y}$. Multiplying both sides of the equation on the right by $\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}$ gives the spline estimator

$$\tilde{\mathbf{G}} = (\mathbf{H} + \alpha\mathbf{K})^{-1}\mathbf{H}\mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}. \tag{31}$$

The estimator $\tilde{\mathbf{G}}$ has one drawback when thinking about practical applications. The matrix $\mathbf{H}$ is unknown, so it should be estimated from the data. However, in some special cases the estimator is simplified to a form that does not depend on the covariance matrix. Suppose that the matrix $\mathbf{H}$ fulfills the condition $\mathbf{K} = \mathbf{H}\mathbf{K}$ (or equivalently $\mathbf{R}$ fulfills the condition $\mathbf{K} = \mathbf{R}\mathbf{K}$). Then the spline estimator (31) simplifies to

$$\hat{\mathbf{G}} \;=\; (\mathbf{I}_q + \alpha\mathbf{K})^{-1}\mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1} = \mathbf{S}\mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}, \tag{32}$$

where $\mathbf{S} = (\mathbf{I}_q + \alpha\mathbf{K})^{-1}$ is the smoother matrix. The smoothing parameter α can be chosen by using the generalized cross-validation criteria

$$GCV_2(\alpha) = \frac{\frac{1}{nq}\mathrm{tr}[(\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^\top]}{(1-\frac{m\cdot edf}{nq})^2}, \tag{33}$$

where $\hat{\mathbf{Y}} = \hat{\mathbf{G}}\mathbf{A}^\top$ and $edf = \mathrm{tr}(\mathbf{S})$ is the effective degrees of freedom of the smoother matrix $\mathbf{S}$.

As in Section 3, for testing we need to approximate the smoother matrix with a matrix that has the properties of a projection matrix. We can approximate the spline estimate (32) with

$$\breve{\mathbf{G}} \;=\; \mathbf{M}_c\mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}, \tag{34}$$

where $\mathbf{M}_c = \mathbf{T}_c\mathbf{T}_c^\top$ and $\mathbf{T}_c$ contains the c first eigenvectors of the smoother matrix $\mathbf{S}$. The number of eigenvectors c can be easily estimated using a modified generalized cross-validation criterion obtained by replacing $\hat{\mathbf{Y}}$ and edf in formula (14) with $\breve{\mathbf{Y}} = \breve{\mathbf{G}}\mathbf{A}^\top$ and c, respectively. We can now approximate the fitted spline curves with

$$\breve{\mathbf{Y}} \;=\; \breve{\mathbf{G}}\mathbf{A}^\top = \mathbf{T}_c\mathbf{T}_c^\top\mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top = \mathbf{T}_c\hat{\mathbf{\Omega}}\mathbf{A}^\top, \tag{35}$$

where $\hat{\mathbf{\Omega}} = \mathbf{T}_c^\top\mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}$. The matrix $\hat{\mathbf{\Omega}}$ contains all the relevant information for testing mean curves and it is also an unbiased estimate of the parameter matrix $\mathbf{\Omega}$ of the statistical model $\mathbf{Y} = \mathbf{T}_c\mathbf{\Omega}\mathbf{A}^\top + \mathbf{E}$. Therefore, we will henceforth focus on testing linear hypotheses of the form

$$H_0 : \mathbf{C}\mathbf{\Omega}\mathbf{D} = \mathbf{0},$$

where $\mathbf{C}$ and $\mathbf{D}$ are known $\nu \times c$ and $m \times g$ matrices with ranks ν and g respectively. It is shown in Nummi and Mesue (2013) that testing can be based on

$$F = \frac{Q_*/\nu g}{\hat{\sigma}^2} \sim F[\nu g, n(q-c)], \tag{36}$$

where

$$Q_* = \mathrm{tr}([\mathbf{D}^\top(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{D}]^{-1}[\mathbf{C}\hat{\mathbf{\Omega}}\mathbf{D}]^\top[\mathbf{C}\mathbf{T}_c^\top\mathbf{R}\mathbf{T}_c\mathbf{C}^\top]^{-1}[\mathbf{C}\hat{\mathbf{\Omega}}\mathbf{D}]) \tag{37}$$

and

$$\hat{\sigma}^2 = \frac{1}{n(q-c)}\mathrm{tr}[\mathbf{Y}^\top(\mathbf{I}_q - \mathbf{M}_c)\mathbf{Y}]. \tag{38}$$

In real-life applications, the matrix $\mathbf{R}$ contains parameters to be estimated and therefore the distribution of the F-statistic is only approximate. However, if we are only interested in progression in time we can drop the constant term by using $\mathbf{C} = [\mathbf{0}, \mathbf{I}_{c-1}]$, and if the uniform covariance model $\mathbf{R} = d^2\mathbf{1}_q\mathbf{1}_q^\top + \mathbf{I}_q$ is assumed, the test statistic Q_* simplifies to

$$Q_{**} = \mathrm{tr}([\mathbf{C}\hat{\mathbf{\Omega}}\mathbf{D}][\mathbf{D}^\top(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{D}]^{-1}[\mathbf{C}\hat{\mathbf{\Omega}}\mathbf{D}]^\top). \tag{39}$$

It can be shown that the distribution of the test statistics Q_{**} is exact. This is an important result since the uniform covariance model is quite common and a good approximation in many situations. In Nummi and Mesue (2013) other kinds of situations are discussed, that give an exact version of the F-test introduced here.

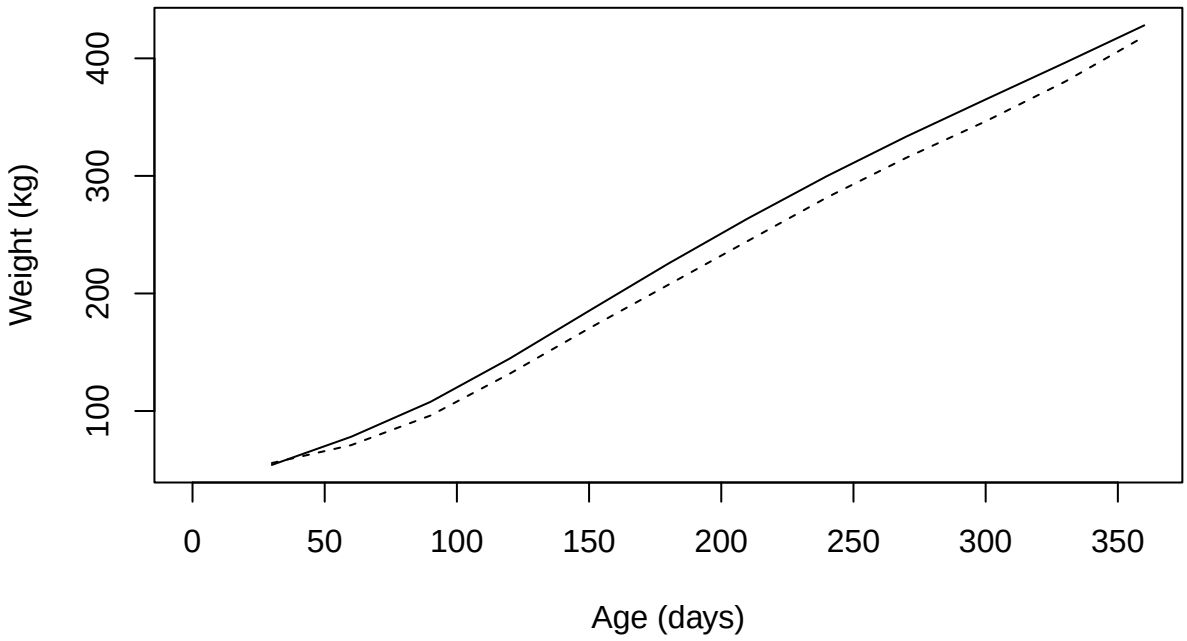


Figure 2: Plot of approximated spline fits for Finncattle bulls (solid curve) and Ayrshire bulls (dashed curve).

Example 2: Testing bulls at a research station in Finland

In this example, we present our methodology using a subset of data pertaining to 2712 bulls tested at an experimental station in Finland during the years 1965 to 1977. The original dataset comprised three breeds: Ayrshire, Finncattle, and Frisian. However, for the purposes of this illustration, we focused on a specific subset consisting of 208 bulls born in 1966, with 168 Ayrshire and 40 Finncattle bulls. The bulls underwent regular weighing, conducted every 30 days starting from the age of 30 days. For more comprehensive details, see the references Lindström and Maijala (1970) and Liski (1987).

To set up the spline growth model the between-individual design matrix $\mathbf{A}$ was defined as follows. For the Finncattle bulls, the rows of $\mathbf{A}$ are $(1, 0)$ and for the Ayrshire bulls the rows of $\mathbf{A}$ are $(0, 1)$. Using the generalized cross-validation criteria (33), we got the smoothing parameter $\alpha = 4142$. The number of eigenvectors c was then estimated using the modified generalized cross-validation criteria (33). The function $GCV_2(c)$ was minimized at $c = 7$. Figure 2 gives the approximated spline fits for the Finncattle bulls (solid curve) and the Ayrshire bulls (dashed curve).

To test if the progression is the same in both groups, we used the 6×7 matrix $\mathbf{C} = (\mathbf{0}, \mathbf{I}_6)$ and 2×1 vector $\mathbf{D} = (1, -1)^\top$. The value of the F-test statistic is

$$F = 102.1803,$$

which gives the P-value $\mathbb{P}(F_{6,1040} \geq 102.1803) \approx 0$. Therefore, the null hypothesis of equal progression of the response variable in the two test groups (Finncattle and Ayrshire) is clearly rejected. We also calculated the P-value of the permutation test. We randomly permuted the rows of matrix $\mathbf{A}$ and re-calculated the value of the F-statistic using the permuted matrix $\mathbf{A}$. After permuting $\mathbf{A}$ and re-calculating F-statistic $N = 100,000$ times, we got the estimated permutation test P-value

$$\frac{\#\{F_i \geq 102.1803\}}{N} = 0.00086.$$

Therefore, it can be affirmed that testing of the growth curves against each other can be readily implemented also using computational methods.

5. Testing in the multivariate spline growth model

The testing of the spline growth model can be generalized straightforwardly to a multivariate response case. The multivariate spline growth curve model can be written as

$$\mathbf{Y} = \mathbf{G}\mathbf{A}^\top + \mathbf{E}, \tag{40}$$

where

$$\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n) = \begin{pmatrix} \mathbf{y}_{11} & \mathbf{y}_{21} & \cdots & \mathbf{y}_{n1} \\ \mathbf{y}_{12} & \mathbf{y}_{22} & \cdots & \mathbf{y}_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}_{1s} & \mathbf{y}_{2s} & \cdots & \mathbf{y}_{ns} \end{pmatrix}$$

is a $qs \times n$ matrix of the vectors of measurements of s responses and

$$\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_m) = \begin{pmatrix} \mathbf{g}_{11} & \mathbf{g}_{21} & \cdots & \mathbf{g}_{m1} \\ \mathbf{g}_{12} & \mathbf{g}_{22} & \cdots & \mathbf{g}_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{g}_{1s} & \mathbf{g}_{2s} & \cdots & \mathbf{g}_{ms} \end{pmatrix}$$

is the corresponding $qs \times m$ matrix of smooth mean curves. See Nummi *et al.* (2017) for more details. For the covariance matrix $\mathbf{R}$ we can take, for example, a multivariate version of the uniform structure

$$\begin{aligned} \mathbf{R} &= (\mathbf{I}_s \otimes \mathbf{1}_q)\mathbf{D}(\mathbf{I}_s \otimes \mathbf{1}_q)^\top + \mathbf{I}_{qs} \\ &= \begin{pmatrix} d_1^2\mathbf{1}_q\mathbf{1}_q^\top + \mathbf{I}_q & d_{12}\mathbf{1}_q\mathbf{1}_q^\top & \cdots & d_{1s}\mathbf{1}_q\mathbf{1}_q^\top \\ d_{21}\mathbf{1}_q\mathbf{1}_q^\top & d_2^2\mathbf{1}_q\mathbf{1}_q^\top + \mathbf{I}_q & \cdots & d_{2s}\mathbf{1}_q\mathbf{1}_q^\top \\ \vdots & \vdots & \ddots & \vdots \\ d_{s1}\mathbf{1}_q\mathbf{1}_q^\top & d_{s2}\mathbf{1}_q\mathbf{1}_q^\top & \cdots & d_s^2\mathbf{1}_q\mathbf{1}_q^\top + \mathbf{I}_q \end{pmatrix}. \end{aligned} \tag{41}$$

If we now define the roughness part of the fitting criteria as

$$\mathbf{K}_s = \mathbf{W} \otimes \mathbf{K},$$

where $\mathbf{W} = \mathrm{diag}(\alpha_1, \ldots, \alpha_s)$ is a diagonal matrix of smoothing parameters $\alpha_1, \ldots, \alpha_s$ and $\mathbf{K}$ is the roughness matrix computed using the time points $t_1, \ldots, t_q$, then the roughness matrix $\mathbf{K}_s$ meets the condition

$$\mathbf{R}\mathbf{K}_s = \mathbf{K}_s \tag{42}$$

and the unweighted spline estimator becomes

$$\begin{aligned}\hat{\mathbf{G}} &= (\mathbf{I}_{qs} + \mathbf{W} \otimes \mathbf{K})^{-1}\mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1} \\ &= \begin{pmatrix} \mathbf{S}(\alpha_1) & \mathbf{O} & \mathbf{O} & \dots & \mathbf{O} \\ \mathbf{O} & \mathbf{S}(\alpha_2) & \mathbf{O} & \dots & \mathbf{O} \\ \vdots & & \ddots & & \vdots \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \dots & \mathbf{S}(\alpha_s) \end{pmatrix} \mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}, \end{aligned} \tag{43}$$

where $\mathbf{S}(\alpha_j) = (\mathbf{I}_q + \alpha_j\mathbf{K})^{-1}$, for $j = 1, \dots, s$. If we use the approximation technique introduced earlier we get

$$\hat{\mathbf{G}} = \begin{pmatrix} \mathbf{M}_{\bullet 1}\mathbf{M}_{\bullet 1}^\top & \mathbf{O} & \mathbf{O} & \dots & \mathbf{O} \\ \mathbf{O} & \mathbf{M}_{\bullet 2}\mathbf{M}_{\bullet 2}^\top & \mathbf{O} & \dots & \mathbf{O} \\ \vdots & \vdots & \ddots & & \vdots \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \dots & \mathbf{M}_{\bullet s}\mathbf{M}_{\bullet s}^\top \end{pmatrix} \mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}, \tag{44}$$

where $\mathbf{M}_{\bullet j}\mathbf{M}_{\bullet j}^\top = \mathbf{P}_j$ is an approximation matrix for the jth variable. Note that the dimensions needed can be estimated using the generalized cross-validation criteria introduced in 33. A straightforward generalization of the earlier considerations gives us an estimator

$$\hat{\mathbf{\Omega}} = \mathbf{M}_\bullet^\top\mathbf{Y}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}, \tag{45}$$

where $\mathbf{M}_\bullet = \text{diag}(\mathbf{M}_{\bullet 1}, \mathbf{M}_{\bullet 2}, \dots, \mathbf{M}_{\bullet s})$, of the multivariate growth curve model

$$\mathbf{Y} = \mathbf{M}_\bullet\Omega\mathbf{A}^\top. \tag{46}$$

Testing can be based on the linear hypothesis

$$H_0 : \mathbf{C}\mathbf{\Omega}\mathbf{D} = \mathbf{0},$$

where $\mathbf{C}$ and $\mathbf{D}$ are known $\nu \times c$ and $m \times g$ matrices with ranks ν and g, respectively, with

$$F = \frac{Q_*/\nu g}{\hat{\sigma}^2} \sim F[\nu g, n(sq - c_{tot})], \tag{47}$$

where $c_{tot} = c_1 + \cdots + c_s$ and

$$Q_* = \text{tr}\{[\mathbf{D}^\top(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{D}]^{-1}[\mathbf{C}\hat{\Omega}\mathbf{D}]^\top[\mathbf{C}\mathbf{M}_\bullet^\top\mathbf{R}\mathbf{M}_\bullet\mathbf{C}^\top]^{-1}[\mathbf{C}\hat{\Omega}\mathbf{D}]\} \tag{48}$$

and

$$\hat{\sigma}^2 = \sum_{l=1}^{s} \frac{1}{n(q - c_l)}\text{tr}[\mathbf{Y}_l^\top(\mathbf{I}_q - \mathbf{P}_l)\mathbf{Y}_l]. \tag{49}$$

If we are interested in testing the equality of the progression of spline curves, then we can choose

$$\mathbf{C} = \text{diag}([\mathbf{0}, \mathbf{I}_{c_1 - 1}], \dots, [\mathbf{0}, \mathbf{I}_{c_s - 1}]) \quad \text{and} \quad \mathbf{D} = [\mathbf{1}_{m-1}, -\mathbf{I}_{m-1}]^\top$$

and, furthermore, if we assume that the covariance matrix has a uniform structure (41), the test statistic simplifies to the form

$$Q_{**} = \text{tr}\{[\mathbf{C}\hat{\Omega}\mathbf{D}][\mathbf{D}^\top(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{D}]^{-1}[\mathbf{C}\hat{\Omega}\mathbf{D}]^\top\}, \tag{50}$$

which does not depend on the covariance matrix $\mathbf{R}$. The F statistic is then distributed as $F[df_1, df_2]$ with degrees of freedoms $df_1 = (c_{tot} - s)(m - 1)$ and $df_2 = n(sq - c_{tot})$.

6. Concluding remarks

In this paper, we explored various methodologies for estimating and testing cubic smoothing splines. We place particular emphasis on analyzing correlated data within semiparametric regression models, as well as the spline growth model, an extension of the basic growth curve model. Additionally, we introduced practical applications including medicine and animal breeding. These examples underscore the versatility and effectiveness of cubic smoothing splines in real-world scenarios.

Acknowledgements

We are deeply honored to have had the opportunity to contribute our publication to the special issue dedicated to C. R. Rao, one of the most esteemed statisticians in history. This is an immense privilege for us. We would also like to that the anonymous referee for the comments that led to improvements of the paper.

References

Azzalini, A. and Bowman, A. (1993). On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society. Series B (Methodological)*, **55**, 549–557.

Cantoni, E. and Hastie, T. (2002). Degrees-of-freedom tests for smoothing splines. *Biometrika*, **89**, 251–263.

Eubank, R. L. and Hart, J. D. (1992). Testing goodness-of-fit in regression via order selection criteria. *The Annals of Statistics*, **1**, 1412–1425.

Eubank, R. L. and Spiegelman, C. H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *Journal of the American Statistical Association*, **85**, 387–392.

Green, P. J. and Silverman, B. W. (1993). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall/CRC, London, 1st ed. edition.

Härdle, W., Mammen, E., and Müller, M. (1998). Testing parametric versus semiparametric modeling in generalized linear models. *Journal of the American Statistical Association*, **93**, 1461–1474.

Harezlak, J., Ruppert, D., and Wand, M. P. (2018). *Semiparametric Regression With R*, volume 109. Springer.

Hastie, T. (1996). Pseudosplines. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 379–396.

Koskela, L., Nummi, T., Wenzel, S., and Kivinen, V. P. (2006). On the analysis of cubic smoothing spline-based stem curve prediction for forest harvesters. *Canadian Journal of Forest Research*, **36**, 2909–2919.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **61**, 381–400.

Lindström, U. and Maijala, K. (1970). Evaluation of performance test results for a.i. bulls. *Acta Agriculturae Scandinavica*, **20**, 207–218.

Liski, E. P. (1987). A growth curve analysis for bulls tested at station. *Biometrical Journal*, **29**, 331–343.

Liu, A. and Wang, Y. (2004). Hypothesis testing in smoothing spline models. *Journal of Statistical Computation and Simulation*, **74**, 581–597.

Nummi, T. and Koskela, L. (2008). Analysis of growth curve data by using cubic smoothing splines. *Journal of Applied Statistics*, **35**, 681–691.

Nummi, T. and Mesue, N. (2013). Testing of growth curves with cubic smoothing splines. In Dasgupta, R., editor, *Advances in Growth Curve Models*, pages 49–59, New York, NY. Springer.

Nummi, T., Möttönen, J., and Tuomisto, M. T. (2017). Testing of multivariate spline growth model. In Chen, D. G., Jin, Z., Li, G., Li, Y., Liu, A., and Zhao, Y., editors, *New Advances in Statistics and Data Science*, pages 75–85. Springer International Publishing, Cham.

Nummi, T., Pan, J., and Mesue, N. (2013). Testing linearity in semiparametric regression models. *Statistics and Its Interface*, **6**, 3–8.

Nummi, T., Pan, J., Siren, T., and Liu, K. (2011). Testing for cubic smoothing splines under dependent data. *Biometrics*, **67**, 871–875.

Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313–326.

Rao, C. R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, **52**, 447–458.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York.

Schimek, M. G. (2000). Estimation and inference in partially linear models with smoothing splines. *Journal of Statistical Planning and Inference*, **91**, 525–540.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **50**, 413–436.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. (2017). *Flexible Regression and Smoothing: Using GAMLSS in R*. CRC Press.

Verbyla, A. P., Cullis, B. R., Kenward, M. G., and Welham, S. J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **48**, 269–311.

Wu, L. and Zhang, J.-T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches*. Wiley, Hoboken, New Jersey.

Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, **4**, 57–74.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 227–242
http://www.ssca.org.in/journal

Some Combinatorial Structures and Their Applications in Cryptography

Mausumi Bose
Indian Statistical Institute and St. Xavier's College, Kolkata

Received: 20 June 2024; Revised: 15 August 2024; Accepted: 17 August 2024

Abstract

The science of cryptography makes use of knowledge from several areas of mathematics including number theory, algebraic and combinatorial structures, probability, linear algebra, information theory and others. In this article we give a brief and selected review of some combinatorial structures and highlight their applications in some cryptograhic schemes. Among these structures are the orthogonal arrays, which were introduced by Prof C. R. Rao more than seventy years ago for applications in statistics. Their use in this new field of cryptography is yet another example of the versatility and power of these arrays.

Key words: Block designs; Hadamard matrix; Orthogonal arrays; Error correcting codes; Key predistribution schemes; Visual cryptography.

AMS Subject Classifications: 05B05, 94A60

1. Introduction and preliminaries

Cryptography is an ancient subject, with its early forms traceable to the Pharaonic period of ancient Egypt. The early forms of cryptography were basically some forms of a 'substitution cipher' in which the sender substituted each letter in the plain text of the message by another letter according to some substitution rule g. This substitution rule was known by the receiver and so he could use $d = g^{-1}$ to get the original plain text back from the cipher text. It is known that different versions of this method of cryptography were used by the Romans in Caesar's time, the Indians in ancient times, and in the world wars as the rotor-cipher machines, *e.g.*, the Enigma machine (*cf.* Kahn, 1996).

Over time, these encryption and decryption methods have grown in sophistication and complexity. Currently, cryptography is a field of fundamental importance for protecting the confidentiality and integrity in communication. Various ideas from mathematics and statistics, for instance, number theory, specially prime numbers and finite fields, combinatorics and designs of experiments, sampling methods, results from probability theory, are used to design a variety of cryptographic schemes.

Corresponding Author: Mausumi Bose
Email: mausumi.bose@gmail.com

In this article, we first describe some selected combinatorial structures which are used by statisticians in the context of designs of experiments and then, we give a flavour of the usefulness of these structures in the context of cryptography. This selection is purely subjective and for the sake of brevity, many other interesting uses of these and other combinatorial structures in cryptographic schemes could not be covered.

Section 2 describes the combinatorial structures with examples. Section 3 introduces some areas in cryptography and uses the examples of Section 2 to illustrate how these combinatorial structures are useful in building the schemes for these areas. Throughout, we avoid mathematical details and give references where the details may be found.

2. Some combinatorial structures

In this section we briefly review some of the combinatorial structures which are used in statistics and give examples which are later used in Section 3. For details on these structures we refer to the books by Raghavarao (1971) and Street and Street (1987). More details on the applications of these structures to cryptographic schemes can be found in the books by Stinson (2004), and Stinson and Paterson (2018), and in the references cited.

We shall write 1_n and I_n to denote the unit vector of order n and the identity matrix of order n, respectively. Let $J_{a\times b} = 1_a 1_b'$ and $O_{a\times b}$ be an $a \times b$ null matrix. A finite field of order s will be denoted by GF(s), where s is a prime or prime power. We write Ω to denote a set of s symbols, labeled by $0, 1, \ldots, s-1$.

2.1. Hadamard matrix

Definition 1: An $n \times n$ matrix H_n, with elements -1 and 1 is called a *Hadamard matrix* if $H_n H_n' = nI_n$.

As seen below, H_1 and H_2 exist. For $n > 2$, it is known that H_n exists only if n is an integral multiple if 4. According to the Hadamard conjecture, the converse is also true. Clearly, any two distinct rows of H_n differ in exactly $n/2$ positions. Also, if we multiply all elements in a row (or column) of H_n by -1, the matrix still remains a Hadamard matrix. So, without loss of generality, we can write all Hadamard matrices in the *standard form*, *i.e.*, with all entries in first row and first column being equal to 1.

There are many methods for constructing these matrices, the simplest one is due to Sylvester (1867) who showed that if H_n is a Hadamard matrix, then $H_2 \otimes H_n$ is also a Hadamard matrix of order $2n$ where $\otimes$ denotes Kronecker product. Hence, with $H_1 = (1)$, and $H_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, a Hadamard matrix of order 2^k can be constructed for every non-negative integer k. These matrices are in standard form and all rows (columns) except the first row (column) have $+1$ in exactly $n/2$ positions and -1 in the remaining $n/2$ positions.

Example 1: H_8 constructed by Sylvester's method is as follows:

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}$$

In statistics, Hadamard matrices are used in constructing designs for experiments, *e.g.*, optimal weighing designs. These matrices have also found applications in signal processing and telecommunication. (*cf.* Yarlagadda and Hershey (1997) and Serberry, Wysocki and Wysocki(2005)). In section 3.2 we highlight the use of these matrices in constructing error-correcting codes.

2.2. Orthogonal arrays

In a series of landmark papers (1946, 1947, 1949) C R Rao proposed some combinatorial structures with applications to statistics, and gave their constructions. Since then, these structures have been widely studied and the entire class of these structures has been called Orthogonal arrays (OAs).

Definition 2: An $M \times k$ array with entries from a set Ω of s symbols is an *orthogonal array* (OA) with M runs, s symbols, strength $t(0 \leq t \leq k)$ and index λ if every $M \times t$ sub-array of this array contains each t-tuple of elements from Ω exactly λ times as a row. Clearly, $M = \lambda s^t$. Such an array will be denoted by $OA(\lambda s^t, k, s, t)$.

From Definition 2.2 it follows that an OA of strength t and index λ is also an OA of strength t' $(0 \leq t' < t)$ and index $\lambda s^{t-t'}$. Also, if a Hadamard matrix H_{4n} exists, then writing it in the standard form and then deleting the first column, one can easily obtain an $OA(4n, 4n-1, 2, 2)$.

Rao (1946, 1947) gave a method for obtaining the OA$(s^n, \frac{s^n-1}{s-1}, s, 2)$ whenever $n \geq 2$, over GF(s). For this, we write all n-tuples from GF(s) as rows to get an $s^n \times n$ array with columns $C_1, C_2, \ldots, C_n$. Then the columns of the OA are of the form $\sum_{i=1}^{n} z_i C_i$ where $z_i \, \epsilon \, GF(s)$, z_i not all zero, and the first non-zero z_i is unity. This method of construction gives what are known as linear OAs and is illustrated in Example 2.

Example 2: The following is an OA(8,7,2,2), where the first 3 columns are written first and then the next 4 columns follow as described above.

0	0	0	0	0	0	0
0	0	1	1	1	0	1
0	1	0	0	1	1	1
0	1	1	1	0	1	0
1	0	0	1	1	1	0
1	0	1	0	0	1	1
1	1	0	1	0	0	1
1	1	1	0	1	0	0

Many other methods of construction of orthogonal arrays are available in the literature. Rao (1947) gave bounds for N for OA$(N = \lambda s^t, k, s, t)$ and Bush (1952) gave improved bounds for arrays of index unity. Arrays of index unity are of special interest and as shown in Bush (1952), if $s(\geq 2)$ is a prime power then an OA$(s^t, s+1, s, t)$ of index unity exists whenever $s > t$. Furthermore, an OA$(s^3, s+2, s, 3)$ exists if s is a power of 2 and Example 3 gives such an OA.

Example 3: OA(8,4,2,3) of index unity.

$$\begin{array}{cccc} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{array}$$

For a comprehensive exposition on OAs we refer to Hedayat, Stufken and Sloane (1999). In statistics, OAs are used in constructing designs, specially in the context of fractional factorial designs, (*cf.* Mukerjee and Wu (2006)). In Section 3 we describe the use of OAs in constructing codes and threshold schemes.

2.3. Binary block designs

Definition 3: A *block design* is an arrangement of v symbols in b blocks or sets of sizes $k_1, \ldots, k_b$, the i^{th} symbol occurring r_i times in the design, $1 \leq i \leq v$. The *incidence matrix* N of the design is a $v \times b$ matrix, such that its $(i,j)^{th}$ element equals the number of times the i^{th} symbol occurs in the j^{th} block, $1 \leq i \leq v, 1 \leq j \leq b$. For $1 \leq i_1 < i_2 \leq v$, let $\lambda_{i_1 i_2}$ denote the number of blocks containing symbols i_1 and i_2.

The *design is binary* if the symbols in each block are distinct, *i.e.*, N is binary. A binary design with $r_1 = \ldots = r_v = r$ and $k_1 = \ldots = k_b = k$ shall be written as a binary (v, b, r, k) block design.

2.4. Balanced incomplete block designs (BIBDs)

Definition 4: A binary (v, b, r, k) block design with $k < v$ and $\lambda_{i_1 i_2}$ all equal (=λ, say) is called a *Balanced Incomplete Block Design* (BIBD).

We write such a design as BIBD(v, b, r, k, λ). It follows from Definition 2.3 that $b \geq v, vr = bk$ and $r(k-1) = \lambda(v-1)$. A BIBD with $v = b, r = k$ is called a *symmetric BIBD*. BIBDs with $k = 3$ (or *Steiner's triple systems*) have been specially studied and one is shown in Example 4.

Example 4: A BIBD$(v = b = 7, r = k = 3, \lambda = 1)$, with blocks as columns.

$$\begin{array}{ccccccc} 2 & 1 & 1 & 2 & 1 & 4 & 3 \\ 3 & 3 & 4 & 6 & 2 & 5 & 5 \\ 4 & 6 & 7 & 7 & 5 & 6 & 7 \end{array}$$

Definition 5: A BIBD (v, b, r, k, λ) is said to be *resolvable* if its blocks can be partitioned into r sets, each set containing b/r blocks, such that every set contains each treatment exactly once.

Example 5: A resolvable BIBD($v = 9, b = 12, r = 4, k = 3, \lambda = 1$) with blocks grouped into 4 sets of 3 blocks is shown below, blocks within each set written as rows:

1	2	3	1	4	7	1	6	8	1	5	9
4	5	6	2	5	8	2	4	9	2	6	7
7	8	9	3	6	9	3	5	7	3	4	8

Several methods of construction of BIBDs are known, e.g., if a Hadamard matrix H_n exists and is in standard form, then deleting its first row and column and replacing -1 by 0, we get the incidence matrix of a symmetric BIBD with parameters $(v = b = 4t - 1, r = k = 2t - 1, \lambda = t - 1)$. Again, if we delete a block from this BIBD and delete all symbols that occur in this deleted block, we get the residual design as a BIBD$(2t, 4t - 2, 2t - 1, t, t - 1)$.

Example 6: Starting from H_{12}, we can construct a BIBD$(11, 11, 5, 5, 2)$. Then the residual design obtained from this is a BIBD$(6, 10, 5, 3, 2)$.

In statistics, BIBDs are well studied and known to be A-,D-,E-optimal for a full set of orthonormal treatment contrasts in the class of all block designs with v treatments in b blocks of size k each. In Section 3 we illustrate their use in obtaining anonymous threshold schemes and visual cryptographic schemes.

2.5. Pairwise balanced designs (PBDs)

Definition 6: A block design with r_i all equal (= r, say) and $\lambda_{i_1 i_2}$ all equal (= λ, say), is called a *Pairwise Balanced Design*. It will be written as PBD$(v, \{k_1, \ldots, k_p\}, \lambda)$ where $k_1, \ldots, k_p$ are the possible block sizes.

Example 7: A PBD(5, {3, 2}, 2) with blocks shown as columns

1	2	1	2	1	2	3	4	1	1
2	3	3	4	3	3	5	5	2	4
5	4	4	5	5					

In Section 3 we illustrate their use in obtaining visual cryptographic schemes.

2.6. Partially balanced incomplete block designs (PBIBDs)

Definition 7: A binary (v, b, r, k) block design with $k < v$ and $\lambda_{i_1 i_2}$ taking only 2 values, (= λ_1, or λ_2, say) for all $1 \leq i_1 < i_2 \leq v$. will be called a *Partially Balanced Incomplete Block Design* (PBIBD) with two associate classes.

Such a design will be written as PBIBD $(v, b, r, k, \lambda_1, \lambda_2)$. There are various association schemes underlying PBIBDs, these schemes determining which pair of symbols occur together λ_1 times and which occur λ_2 times. For simplicity, we do not elaborate on association schemes and refer to Raghavarao (1971), pp 121-127, for details.

Example 8: A PBIBD $(6, 4, 2, 3, 0, 1)$ with blocks as columns

$$\begin{array}{cccc} 1 & 1 & 2 & 3 \\ 2 & 4 & 4 & 5 \\ 3 & 5 & 6 & 6 \end{array}$$

In Section 3 we illustrate their use in obtaining visual cryptographic schemes and also mention their use in key predistibution networks.

3. Applications in cryptography

3.1. Codes and error-correcting codes

Definition 8: Let Ω be a set of elements or symbols. A set of k-tuples of the symbols in Ω, where $k \geq 1$ is an integer, is called a code C over the alphabet Ω. Each k-tuple in C is called a codeword. If $\Omega = \mathrm{GF}(2)$, then C is a binary code.

The *Hamming weight* of a codeword is the number of ones in it. The *Hamming distance* between any two codewords is the number of positions in which they differ. The *distance* of a code C, denoted by d, is the minimal Hamming distance between any two distinct codewords in C. A code with N codewords, each of length k over an alphabet Ω consisting of s elements, and having distance d may be written as a (N, k, d, s) code.

An error-correcting code can correct errors incurred during the transmission of data over noisy channels. A linear block code takes a sequence of m symbols from Ω and encodes it as a sequence of $k(> m)$ symbols. The redundant elements are added to the original message to facilitate recovery of the message. The ability of a code to detect and correct errors is measured by its distance d; a code with distance d can correct up to a maximum of e errors where $e = \lfloor \frac{(d-1)}{2} \rfloor$ and can detect up to $d - 1$ errors.

3.1.1. Hadamard codes

Error-correcting codes obtained from Hadamard matrices have the maximal error correcting ability for a given length of codeword and so these are useful when a message is transmitted over a noisy or unreliable channel. For instance, as described in Serberry *et al.* (2005), these codes were used in the 1960's in the Mariner and Voyager space probes to encode information transmitted back to the earth and due to the powerful error-correction capabilities of these codes, it was possible to decode properly the pictures of Jupiter, Saturn, Uranus, Neptune and their moons.

Hadamard matrices obtained from Sylvetser's method of construction are usually used for obtaining Hadamard codes as they lead to linear codes, but Hadamard matrices constructed by other methods lead to codes too, though not necessarily linear. These latter codes were first studied by Bose and Skrikhande (1959) in connection with symmetrical block code designs.

Let $n = 2^k$. A Hadamard code C is obtained from a Hadamard matrix H_{2^k} by replacing -1 by 1 and 1 by 0. The rows of C are the 2^k codewords, each of length 2^k. As discussed in Section 2.1, all rows of H_n, other than the first row, have $+1$ in exactly $n/2$

positions and any two distinct rows of H_n differ in exactly $n/2$ positions. So, with $n = 2^k$, each non-zero codeword in C has Hamming weight 2^{k-1} and any two codewords in C have Hamming distance equal to 2^{k-1}.

Example 9: Let a message x be represented as a binary vector of length 3. So $m = 3$ and we use a code based on H_8 shown in Example 1, by replacing -1 by 1 and 1 by 0. Table 1 shows the original messages and the corresponding encoded messages given by codewords of length 8 obtained from rows of H_8.

Incidentally, it may be mentioned here that the rows on the right side of Table 1 can also be obtained as a linear transform of the array on the left, over GF(2). Then, on deleting the first column of zeros, we get an orthogonal array OA (8,7,2,2) which is isomorphic to the one shown in Example 2. So an OA (8,7,2,2), with a column of zeros added to it, can also give the same code as in Table 1, but the construction of the code from H_8 is simpler.

Table 1: Original and Encoded messages

Original message	Encoded message
0 0 0	0 0 0 0 0 0 0 0
1 0 0	0 1 0 1 0 1 0 1
0 1 0	0 0 1 1 0 0 1 1
1 1 0	0 1 1 0 0 1 1 0
0 0 1	0 0 0 0 1 1 1 1
1 0 1	0 1 0 1 1 0 1 0
0 1 1	0 0 1 1 1 1 0 0
1 1 1	0 1 1 0 1 0 0 1

For a message x, the encoded message is transmitted and the received message is a vector of length 8, say y, with a possible error, *i.e.*, flipping of 0 and 1. To decode y, we find the Hamming distance between y and the 8 codewords, the message corresponding to the codeword with the least Hamming distance will be the original message. For example, if y = 0 1 0 0 0 1 1 0, then the Hamming distance between y and the 8 codewords in their order of Table 1 are 3,3,5,1,3,3,5,5. The least value 1 corresponds to the codeword 0 1 1 0 0 1 1 0. So the original message is decoded correctly as 1 1 0. Thus, one error can be corrected.

A Hadamard code has a large block length (= 2^k) compared to the message length k. However, it can correct $2^{k-2} - 1$ errors in a 2^k-bit encoded message, which is extremely good.

Moreover, we can improve upon code C by writing the code as $C = \begin{pmatrix} H_{2^k} \\ -H_{2^k} \end{pmatrix}$ and then replacing -1 by 1 and 1 by 0 as before. So by this method, the code in Example 9 can be improved upon. It is easy to see from Definition 2.1 that such a code C can accommodate $k + 1$ messages while still having block length 2^k and distance 2^{k-1}. This code C is also sometimes called a *Hadamard code* and it is the same as the *first order Reed-Muller code over the binary alphabet.*

3.1.2. Codes from orthogonal arrays

For a (N, k, d, s) code, the Singleton Bound is $N \leq s^{k-d+1}$. Codes for which N attains this bound are called the *maximum distance separable (MDS) codes* as they have the maximum possible distance between codewords. It can be shown that orthogonal arrays with index unity are equivalent to MDS codes, *i.e.*, an (s^{k-d+1}, k, d, s) code is equivalent to an OA(s^t, k, s, t), where $t = k-d+1$. So MDS codes can be obtained from orthogonal arrays of strength unity which were discussed in Section 2.2.

Example 10: When $t \geq 2$ is an integer and s is a prime power, an OA(s^t, s, s, t) exists and this gives an MDS $(s^t, s, s-t+1, s)$ code which is the well known *Reed Solomon code*, being optimal with respect to the Singleton bound.

There is another bound on N called the Sphere-packing Bound and a code for which this bound is satisfied with equality is called a *perfect* code. It can be shown that for s a prime power and an integer $n \geq 2$, if we start from the linear OAs$(s^n, \frac{s^{n-1}}{s-1}, s, 2)$ constructed as in Example 2 and then take the code which is the orthogonal complement of this OA, we will get a perfect $(s^m, \frac{s^{n-1}}{s-1}, 3, s)$ code, where $m = \frac{s^n-1}{s-1} - n$. These codes are known as *Hamming codes*. When $s = 2$, the code is $(2^{2^n-n-1}, 2^n - 1, 3, 2)$.

Example 11: Starting from an OA(8,7,2,2), shown in Example 2, the orthocomplement of this array gives a *binary Hamming code* or a (16,7,3,2) perfect code.

Interestingly, from a binary Hamming code C if we form a matrix A with its columns being the codewords of C with weight 3, then it can be seen that A gives the *incidence matrix of a symmetric BIBD* with 2^{n-1} treatments and blocks of size 3, *i.e.* a *Steiner's triple system*. So the (16,7,3,2) code constructed as above will give the incidence matrix of the BIBD (7, 7, 3, 3,1) as shown in Example 4.

3.2. Threshold schemes

Let a and b be two integers, $2 \leq a \leq b$. Suppose there is a secret K and a set of b participants $\mathcal{P}$. A dealer who does not belong to $\mathcal{P}$, assigns each participant a 'share', *i.e.*, some partial information about K.

The method of assigning these shares is called a (a, b) *threshold scheme* if any a participants can compute the secret K by pooling their shares, and no set of $a-1$ participants can recover K from their shares. The secret K can be chosen from a set of secrets $\mathcal{K}$ and each share is chosen from a specified share set $\mathcal{S}$.

These schemes have many uses, e.g., there may be a set of 5 individuals, each of whom hold a key to a safe but the safe can be opened only if 3 or more persons use their keys together, it cannot be opened by any single person or 2 persons. This is a (3,5) threshold scheme.

An (a, b) threshold scheme is called an *anonymous* (a, b) *threshold scheme* if the participants receive distinct shares and the recovery of the secret can be done by a participants without knowing which participant holds which share.

3.2.1. Threshold (a, b) schemes from orthogonal arrays

A (t, k) threshold scheme can be obtained from an OA$(s^t, k+1, s, t)$. For assigning the shares, the rows of the OA are first rearranged so that they can be grouped in sets of s^{t-1} rows, the last column of every row in the ith group having the symbol i, $0 \leq i \leq s-1$. The scheme will have s secrets each secret corresponding to one group and s shares, each share corresponding to one symbol of the OA. The participants know the OA which is used in the scheme.

If the dealer wants to assign secret i, he chooses one row at random from the ith group and assigns the element in the jth column to the jth participant, $1 \leq j \leq k$. This will be a (t, k) threshold scheme. Since the OA has strength t and index unity, if t participants combine their shares, there will be a unique row of the OA which will match with their t shares in the corresponding t columns, and so the secret will be revealed. This is because, with the knowledge of the OA, participants will be knowing the group that this unique row comes from. There will not be any such unique row if $t-1$ or fewer participants combine their shares, thus making the scheme secure. This scheme is not anonymous since in order to reveal the secret, it must be known which participant held which share.

Example 12: The OA(8,4,2,3) in Example 3 gives a (3,3) threshold scheme where all participants have to get together in order to reveal the secret.

3.2.2. Anonymous (a, b) threshold schemes from resolvable BIBDs

An anonymous $(2, k)$ threshold scheme can be obtained from a resolvable BIBD with $\lambda = 1$. To see this, consider a resolvable BIBD$(v, b, r, k, 1)$. From Definition 2.4, $r = \frac{v-1}{k-1}$ and the b blocks can be divided into r disjoint sets say $L_1, \ldots, L_r$, each set having $\frac{b}{r}$ blocks. The dealer can share r secrets, each secret associated with a set and there will be v shares, each share associated with one symbol. There can be k participants, the resolvable design being known to all participants. Suppose the dealer picks a set L_i as the secret and chooses any random block from this set. He allocates the k symbols in this block to the k participants, each getting one symbol. This will be a $(2, k)$ anonymous threshold scheme as illustrated below.

Example 13: Consider the resolvable BIBD with $\lambda = 1$ in Example 5. It is divided into 4 sets $L_1, \ldots, L_4$, and so, there are 4 secrets. Since $k = 3$ there can be 3 participants. Suppose the secret is L_2 and the dealer chooses the block {2,5,8} and allocates $2, 5$, and 8 to the 3 participants, respectively. Now if participants 2 and 3 get together, their combined share is {5,8} and since the BIBD has $\lambda = 1$, there is a unique block which can have the symbols 5 and 8 together. So they can identify the secret L_2 uniquely, while this identification cannot be done by any single participant. Thus this is a (2,3) anonymous threshold scheme. This scheme is anonymous as it it not necessary to know which participant held which share.

3.3. Visual cryptographic schemes

Visual cryptographic schemes (VCS) are threshold schemes which encode a secret image or text in such a way that the decoding can be done simply by the human eye, without any computations. Naor and Shamir (1994) introduced VCS for black and white

images. In a (k, n) VCS with n participants, a secret image, or text is encrypted into n shares, each share being printed on a transparency sheet. Each participant is given one share, and if k of them stack their shares one on top of another, the secret is discernible visually. If less than k participants stack their shares, the secret is not visible.

During encryption, each pixel of the original image is encrypted into a number of subpixels, say m. This number m is called the *pixel expansion* of the VCS. The clarity with which a reconstructed image is visible is measured by the *relative contrast* of the VCS. The aim is to keep m small and relative contrast high.

For simplicity, we will only elaborate on VCS for black and white images.

Suppose the n participants are labeled as $1, \ldots, n$. Given a Boolean matrix A, let A_i denote its ith row and A_{ij} denote the Boolean 'or' of rows A_i and A_j. Let $w(T)$ be the weight of a Boolean vector T. We assume that the secret image is a collection of black and white pixels, and a black(white) pixel will be represented by 1(0).

Definition 9: A $(2, n)$ VCS, with n participants and pixel expansion m, is defined by two $n \times m$ Boolean basis matrices S^1 and S^0, respectively, for black amd white pixels such that (i) S^1 and S^0 are equal up to a column permutation, *i.e.*, $w(S^1_i) = w(S^0_i)$, $1 \leq i \leq n$, and (ii) $w(S^1_{i_1,i_2}) > w(S^0_{i_1,i_2}), 1 \leq i_1 < i_2 \leq n$.

Let π be a random permutation of $\{1, \ldots, m\}$. While encryption, if a pixel in the secret image is black(white), then π is applied to the columns of $S^1(S^0)$ and row i of the permuted matrix forms the share of the ith participant. Thus each pixel of the image is encrypted and distributed into n shares, each of which consists of m subpixels. The random permutation used in allocating shares together with condition (i) of Definition 3.2 ensures that no single participant can recover the image. Moreover, for any $i_i < i_2$, if shares i_1 and i_2 are stacked together by aligning the subpixels, and the combined share is obtained by taking the Booelan 'or' of these 2 shares, then condition (ii) of Definition 3.3 guarantees that the grey level of a black pixel is darker than that of a white pixel and this makes the recovered image discernible. For any $i_1 < i_2$, the quantity $\xi_{i_1,i_2} = m^{-1}\{w(S^1_{i_1,i_2}) - w(S^0_{i_1,i_2})\}$ is positive in view of (ii), and it is called the relative contrast of the recovery of the image by participants i_1 and i_2.

A VCS is said to be *balanced* if the $\xi_{i_1,i_2}(1 \leq i_1 < i_2 \leq n)$ values are all equal (=ξ, say) and *unbalanced* otherwise. In any *balanced* $(2, n)$ VCS, for given n, the relative contrast ξ is bounded above by $(= \frac{\lfloor n/2 \rfloor \lceil n/2 \rceil}{n(n-1)}) = \xi_0$, say.

3.3.1. Obtaining $(2, n)$ VCS from BIBDs

Blundo, De Santis and Stinson (1999) gave the following three constructions of *balanced* $(2, n)$ VCS, and for all of these the relative contrasts attains the upper bound ξ_0 for given n.

If a BIBD(n, b, r, k, λ) exits, then there exists a balanced $(2, n)$ VCS with $m = b$ and $\xi = \xi_0 = (r - \lambda)/b$ with S^1 as the incidence matrix of the BIBD and $S^0 = [J_{n\times r},\ O_{n\times(b-r)}]$.

Example 14: From the BIBD in Example 4 we have:

$$S^1 = \begin{matrix} 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{matrix} \qquad S^0 = \begin{matrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{matrix}$$

After permuting the columns of S^1 and S^0, when a row of $S^1(S^0)$ is assigned as a subpixel corresponding to a black(white) pixel to a participant, he cannot know what the original pixel was since in both cases he gets a Boolean vector with weight 3. But when any 2 participants combine their shares, 'or' of any 2 rows of $S^1(S^0)$ give a Boolean vector with weight 5(3). So, on recovery, if the original pixel was black, it appears darker to the human eye than if the original pixel was white. So, $\xi = (5-3)/7 = 2/7$.

Suppose n is even. Then a $(2, n)$ VCS exists with optimal ξ and smallest possible pixel expansion ($m = 2n-2$) if there exists a BIBD$(n, 2(n-1), n-1, n/2, n/2-1)$.

Example 15: The BIBD $(6, 10, 5, 3, 2)$ of Example 6 will give a $(2, 6)$ VCS with optimal $\xi = 3/10$ and smallest possible pixel expansion for this value of n as $m = 10$.

3.3.2. Obtaining $(2, n)$ VCS from PBDs

Suppose n is odd. Then there exists a $(2, n)$ VCS with pixel expansion m and optimal ξ if there exists a PBD $(n, \{(n+1)/2, (n-1)/2\}, r - m(n+1)/4n)$ with exactly m blocks, where r is the common replication number of each symbol. As before, S^1 will be the $n \times b$ incidence matrix of the PBD and $S^0 = [J_{n\times m},\ O_{n\times(m-r)}]$.

Example 16: The PBD(5, {3,2},2) of Example 7 is a PBD with parameters as above with $n = 5, m = 10$ and $r = 5$. So this will give a $(2, 5)$ VCS with $m = 10$ and $\xi = 3/10$.

3.3.3. Obtaining $(2, n)$ VCS from PBIBDs

Adhikary and Bose (2004) and Adhikary, Bose, Kumar and Roy (2005) used *Latin squares* and PBIBDs to show that one can get *unbalanced* $(2, n)$ VCS where the relative contrast for some pairs of participants are more than the optimal bound of ξ for the balanced case. Moreover, given v, since PBIBDs require only partial balance, they have fewer blocks than BIBDs, and hence lead to VCS with smaller pixel expansion (m) than those from BIBDs. We illustrate their method with PBIBDs below:

Example 17: The PBIBD $(6, 4, 2, 3, 0, 1)$ in Example 8 gives $(2, 6)$ VCS with S^1 as the incidence matrix of this design and S^0 as shown below. This is an unbalanced $(2, 6)$ VCS and it can be checked that for some pairs of symbols the relative contrasts are $\xi_{1,6} = \xi_{2,5} =$

$\xi_{3,4} = 1/2$ while for other pairs it is $1/4$.

$$S^1 = \begin{matrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{matrix} \qquad S^0 = \begin{matrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{matrix}$$

Given n, we may choose a suitable PBIBD from the tables of Clatworthy (1973) with small b and large λ_1 to get a $(2, n)$ VCS with $n = v$ and pixel expansion $m = b$. If a PBIBD with $v = n$ is not available, we can choose a PBIBD with $v > n$, construct S^1 from its incidence matrix and then delete $v - n$ suitable rows from it to get S^1 for the $(2, n)$ VCS.

3.3.4. Optimal $(2, n)$ VCS through general binary block designs

We now consider the scenario where both the given n and the allowable m are held fixed and then we optimize with respect to the relative contrasts. As seen above, unbalanced VCS play a crucial role in optimizing over the ξ_{i_1,i_2}, but given n and m, there often does not exist any VCS that maximizes each ξ_{i_1,i_2} separately.

Let $\mathcal{V}(n, m)$ be the class of all $(2, n)$ VCS, balanced or unbalanced, with n participants and pixel expansion m. Then, in the spirit of *A-and E-optimality* in statistical design theory, (*cf.* Shah and Sinha (1989)), Bose and Mukerjee (2006) introduced the notion of *optimal* VCS that maximize the average, say $\bar{\xi}$, or the minimum, say ξ_{min}, of the ξ_{i_1,i_2}, $1 \le i_1 < i_2 \le n$, over $\mathcal{V}(n, m)$.

Such optimal VCS were called *Type I optimal* and *Type II optimal*, respectively. A VCS which is both Type I and Type II optimal will be called Type III optimal. Indeed, given n, m, a VCS, say V^* which is Type I optimal, is also admissible in the sense that there cannot exist another VCS, say V, with same n, m such that each ξ_{i_1,i_2} under V is greater than or equal to the corresponding ξ_{i_1,i_2} under V^*, the inequality being strict for some $i_1 < i_2$.

Bose and Mukerjee showed that the following binary block designs lead to Type III optimal $(2, n)$ VCS:

(i) Suppose m is odd. Let $n = m$. If there exists a binary (v, b, r, k) block design with $v = n = m$ such that $r = (m-1)/2$, $k = (m-1)/2$, and no two of the $\lambda_{ij} (1 \le i < j \le n)$ differ by more than unity, then with basis matrices S^1 as the incidence matrix of this design and $S^0 = [J_{n\times(m-1)/2},\ O_{n\times(m+1)/2)}]$ we get an optimal Type III $(2, n)$ VCS.

(ii) Suppose m is even. If there exists a binary block design with $v = n$, $b = m$, $r_1 = \ldots = r_v = m/2$, $k_j = (n - \delta)/2$ $(1 \le j \le m/2)$, $k_j = (n + \delta)/2$ $(m/2 \le j \le m)$, where $\delta = 1$ if n is odd and $= 0$ if n is even, and no two of the $\lambda_{ij} (1 \le i < j \le n)$ differ by more than unity, then with basis matrices S^1 as the incidence matrix of this design and $S^0 = [J_{n\times m/2},\ O_{n\times m/2)}]$ we get an optimal Type III $(2, n)$ VCS.

There are several broad classes of block designs which satisfy the conditions in (i) and (ii) above. These include *BIBDs, PBIBDs, symmetrical unequal block designs* and *regular*

graph designs with appropriately chosen parameters. We give examples based on BIBD and PBIBD below. For more examples and results we refer to Bose and Mukerjee (2006).

Example 18: For $n = m = 4t - 1$, the BIBD $(v = b = 4t - 1, r = k = 2t - 1, \lambda = t - 1)$ discussed before Example 6, gives a Type III optimal VCS in $\mathcal{V}(4t - 1, 4t - 1)$. Again, the residual BIBD $(2t, 4t - 2, 2t - 1, t, t - 1)$ obtained from the earlier BIBD gives a Type III optimal VCS in $\mathcal{V}(2t, 4t - 2)$. Again, if we delete the last rows of the matrices S^1 and S^0 from those used for the optimal VCS from the residual design, we get a Type III optimal VCS in $\mathcal{V}(2t - 1, 4t - 2)$. So the designs in Example 6 gives optimal Type III VCS.

Example 19: For $n = m = 2t$, where $2 \leq t \leq 12$, $t \neq 7$, as shown in Table 1 of Bose and Mukerjee (2006), we can find an initial block T of cardinality t, such that among the ordered differences (mod n) arising out of the elements of T, each of $1, 2, \ldots, n - 1$ occurs either ρ or $\rho+1$ times, where $\rho = \lfloor t(t - 1)/n - 1) \rfloor$. Upon development of T we get a regular graph design which leads to a Type III optimal VCS in $\mathcal{V}(2t, 2t), 2 \leq t \leq 12, t \neq 7$. When $t = 2$ or 3, this design is also a PBIBD. Moreover, if we delete the last rows of S^1 and S^0 of the optimal VCS in $\mathcal{V}(2t, 2t)$ as obtained above, then we get a Type III optimal VCS in $\mathcal{V}(2t - 1, 2t)$.

3.3.5. Optimal (k, n) VCS from block designs

Bose and Mukerjee (2010) studied (k, n) VCS and gave conditions for their existence and also methods for getting optimal (k, n)VCS. For simplicity, we only give two examples with $k = 3$, one obtained from BIBD and another from PBIBD.

For a binary $(v = n, b, r, k)$ block design let λ_{i_1,i_2,i_3} denote the number of blocks containing symbols i_1, i_2, i_3, and $\xi(i_1, i_2, i_3)$ be the relative contrast for the recovery of the image by the 3 participants $i_1, i_2, i_3, 1 \leq i_2 < i_2 < i_3 \leq n$. We call a $(3, n)$ VCS optimal if it maximizes the average of $\xi(i_1, i_2, i_3)$ over $1 \leq i_2 < i_2 < i_3 \leq n$. Given a design with incidence matrix N, we take $S^1 = [J_{n\times(b-2r)}\ N]$ and $S^0 = [O_{n\times(b-2r)}\ \bar{N}]$ where $\bar{N}$ is obtained from N by interchanging its elements 1 and 0.

Example 20: Let $n = 13$. Then the BIBD $(13, 26, 6, 3, 1)$ (given in Takeuchi (1962)) will lead to an optimal (unbalanced) $(3, 13)$ VCS. This will have pixel expansion $m = 2(b - r) = 40$ and $\xi(i_1, i_2, i_3) = 5/40$ if i_1, i_2, i_3 occur together in a block and $3/40$ otherwise. It also maximizes the minimum possible value of $\xi(i_1, i_2, i_3)$ among all $(3, 13)$VCS with $m = 40$.

Example 21: Let $n = 20$. The PBIBD $(20, 16, 4, 5, 0, 1)$ (=design SR58 in Clatworthy (1973) tables) will lead to an optimal (unbalanced) $(3, 20)$ VCS. This will have pixel expansion $m = 2(b - r) = 24$ and the smallest value of $\xi(i_1, i_2, i_3) = 1/24$. It also maximizes the minimum possible value of $\xi(i_1, i_2, i_3)$ among all $(3, 20)$ VCS with $m = 24$.

It may also be noted that the two VCS in Examples 20 and 21 substantially reduce the pixel expansion compared to the corresponding *balanced* $(3, n)$ VCS as in Blundo et. al (2003), which have $m = 440$ and $m = 23256$, respectively.

3.4. Key predistribution schemes for distributed sensor networks using block designs

Key predistribution schemes (KPS) is another area of cryptography where block designs have been effectively used to get good schemes. These schemes are used in various applications, for instance, in a military operation, where sensor nodes with secret keys installed in them, may be distributed in a random manner over a sensitive area and, once deployed, these nodes are required to communicate with each other through secure keys in order to gather and relay information.

The two main metrics for a KPS are the *network size*, *i.e.*, number of nodes (n) and the *key storage* k, *i.e.*, the number of keys stored per node. Any two nodes within a neighbourhood can communicate with each other if they have $q(\geq 1)$ common keys, where q is the *intersection threshold* of the network. If two nodes do not have q keys in common then they can still communicate through multiple secure links if there is a sequence of one or more intermediate nodes connecting them such that every pair of adjacent nodes in this sequence share q common keys.

If some nodes are captured in an attack, all keys in them are lost but the remaining nodes can still communicate (*i.e.*, be resilient) using the remaining keys. For more details on the applications, the security framework and models for these distributed sensor networks (DSNs) we refer *e.g.*, to Roman et al. (2005) and Du *et al.* (2005) and Martin (2009).

Key assignment schemes based on *combinatorial designs* is specially useful since using the combinatorial structures of the underlying designs, one can study the connectivity and resiliency properties of the scheme, and also carry out shared-key discovery and path-key establishment in a structured manner. Camtepe and Yener (2004) used *finite projective planes* and *generalized quadrangles* and Dong et al. (2008) used *3-designs* to construct KPS with $q = 1$. Lee and Stinson (2008) used *transversal designs* to construct KPS and give schemes separately for $q = 1$ and for $q = 2$.

Bose, Dey and Mukerjee (2013) suggested one general construction method for KPS for any given q and by varying the choices of the designs, this resulted in KPS for networks with varying numbers of nodes, key-pool sizes and numbers of keys per node, thus providing more flexibility in choosing a scheme suitable for the requirements of a situation. The designs used were *BIBDs, PBIBDs based on the group-divisible, Latin square and triangular association schemes,* and suitable duals of these designs. This method works for general q and can cover a wide variety of values of n.

The complexity of the KPS scheme and its various metrics leads to involved algebra and so we refrain from elaborating further in this area.

4. Conclusion

In this article, an endeavour has been made to highlight the fact that combinatorial designs have a wide applicability in various areas of cryptography.

We have mainly focused on Hadamard matrices, orthogonal arrays, pairwise balanced designs, balanced incomplete block designs and partially balanced incomplete block designs and their applications in various areas of cryptography. There are many other topics that

could not be covered, *e.g.* a combinatorial structure used often in the context of threshold access structures is the perfect hash family (PHF). Long *et al.* (2006) and Martin and Ng (2007) used generalized cumulative arrays focusing on the situation where all participants have the same probability of being selected for activation. Bose and Mukerjee (2014) gave a method where an unequal probability scheme given by PHFs leads to better levels of group and participant anonymity, and also showed that BIBDs can be used to get schemes in this context too.

The Anti-collusion Digital Fingerprinting Codes is another area of cryptography where combinatorial designs have been used effectively, for instance, Trappe *et al.*(2003) used BIBDs, Kang *et al.*(2006) used PBIBDs, Yagi *et al.*(2009) used finite geometries, Li *et al.*(2009) used OAs, Bose and Mukerjee (2010) used partially cover-free families.

To conclude, our aim in this article is to highlight that there are various areas of cryptography where combinatorial designs and structures give effective and efficient schemes. We hope that this article will generate sufficient interest among statisticians who are already familiar with these structures, to take up research in this area.

Acknowledgements

This work has been supported by the National Board for Higher Mathematics, Dept of Atomic Energy, Govt of India. The author sincerely thanks the two referees for their valuable comments which have led to an improvement in the presentation of the paper.

References

Adhikari, A. and Bose, M. (2004). Construction of new visual threshold schemes using combinatorial designs. *IEICE Transactions*, **E87-A**, 1198-1202.

Adhikari, A., Bose, M., Kumar, D., and Roy, B. (2007). Applications of partially balanced incomplete block designs in developing (2,n) Visual cryptographic Schemes. *IEICE Transactions*, **E90-A**, 949-951.

Blundo, C., De Santis, A., and Stinson, D. R. (1999). On the contrast in visual cryptography schemes. *Journal of Cryptology*, **12**, 261-289.

Bose, M., Dey, A., and Mukerjee, R. (2013). Key predistribution schemes distributed sensor networks via block designs. *Designs, Codes and Cryptography*, **467**, 111-136.

Bose, M. and Mukerjee, R. (2006). Optimal $(2, n)$ visual cryptographic schemes. *Designs, Codes and Cryptography*, **40**, 255-267.

Bose, M. and Mukerjee, R. (2010). Optimal (k, n) visual cryptographic schemes for general k. *Designs, Codes and Cryptography*, **55**, 19-35.

Bose, M. and Mukerjee, R. (2013). Union distinct families of sets, with an application to cryptography. *Ars Combinatoria,* **110** 179-192.

Bose, M. and Mukerjee, R. (2014). An unequal probability scheme for improving anonymity in shared key operations. *Journal of Statistical Theory and Practice*, **8**, 100-112.

Bose, R. C. and Shrikhande, S. S. (1959). A note on result in the theory of code construction. *Information and Control*, **2**, 183–194.

Bush, K. A. (1952). Orthogonal arrays of index unity. *Annals of Mathematical Statistics*, **23**, 416-434.

Clatworthy, W. H. (1973). *Tables of Two-associate Partially Balanced Designs.* National Bureau of Standards, Applied Maths, series no **63**, Washingto D.C.

Climato, S., Prisco, R. D., and Santis, A. D. (2005). Optimal coloured threshold visual cryptography schemes. *Designs Codes and Cryptography*, **35**, 311-335.

Hedayat, A. S., Stufken, J., and Sloane, N. J. A. (1999). *Orthogonal Arrays: Theory and Applications.* Springer-Verlag, New York.

Ibrahim, D. R., Teh, J. S., and Abdullah, R. (2021). An overview of visual cryptography techniques. *Multimedia Tools and Applications*, **80**, 31927-31952.

Kahn, D. (1996). *The Codebreakers: The Comprehensive History of Secret Communication from Ancient Times to the Internet.* Scribner, New York.

Kang, I., Arce, G., and Lee, H. K. (2011). Color Extended Visual Cryptography Using Error Diffusion. *IEEE Transactions on Image Processing*, **20**, 132-145.

Kang, I., Sinha, K., and Lee, H. K. (2006). New digital fingerprint code scheme using group-divisible design. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Science*, **E89-A**, 3732-3735.

Mukerjee, R. and Wu, C. F. J. (2006). *A Modern Theory of Factorial Design.* Springer, New York.

Naor, M. and Shamir, A. (1994). Visual cryptography. Advances in Cryptology, Eurocrypt'94. Lecture Notes in Computer Science, **950**, 1-12, Springer-Verlag.

Plotkin M. (1960). Binary codes with specified minimum distance, *IRE Transactions*, **IT-6** 445–450.

Raghavarao, D. (1971). *Construction and Combinatorial Problems in Design of Experiments.* New York, Wiley.

Rao, C. R. (1946). Hypercubes of strength 'd' leading to confounded designs in factorial experiments. *Bulletin of Calcutta Mathematical Society*, **38**, 67-78.

Rao, C. R. (1947). Factorial experiments derivable from combinatorial arrangements of arrays. *Journal of Royal Statistical Society*, **9**, 128-139.

Rao, C. R. (1949). On a class of arrangements. *Proceedings of Edinburgh Mathematical Society*, **8**, 119-125.

Serberry J, Wysocky B. J., and Wysocki T. A. (2005). On some applications of Hadamard matrices. *Metrika*, **62**, 221-239.

Shah, K. and Sinha, B.K. (1989). *Theory of Optimal Designs.* Springer-Verlag, New York.

Stinson, D. R. (2004). *Combinatorial Designs: Constructions and Analysis,* Springer, ISBN 978-0-387-95487-5.

Stinson, D. R. and Paterson, M. B. (2018). *Cryptography: Theory and Practice*, 4th ed. CRC Press.

Street, A. P. and Street, D. J. (1987). *Combinatorics of Experimental Design.* Oxford. Clarendon Press.

Takeuchi, K. (1962). A table of difference sets generating balanced incomplete block designs. *Review of the International Statistical Institute*, **30**, 361-366.

Trappe, W., Wu, M, Wang, Z. J., and Liu, K. J. R. (2003). Anti-collusion finger-printing for multimedia. *IEEE Transactions on Signal Processing*, **51**, 1069-1087.

Yagi, H., Matsushima, T., and Hirasawa, S. (2007). Improved collusion-secure codes for digital fringerprinting based on finite geometries. *IEEE International Conference on Systems, Man and Cybernetics*, 948-953.

Yarlagadda, R. K. and Hershey, J. E. (1997). *Hadamard Matrix Analysis and Synthesis with Applications to Communications and Signal: Image Processing.* Kluwer.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 243–257
http://www.ssca.org.in/journal

Split-plot Designs with Main Plot Treatments in Incomplete Blocks

B. N. Mandal[1], Rajender Parsad[2] and Sukanta Dash[2]
[1]*ICAR-Indian Agricultural Research Institute, Jharkhand 825405, India*
[2]*ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110012, India*

Received: 14 January 2024; Revised: 16 August 2024; Accepted: 20 August 2024

Abstract

Split-plot designs are widely used in agricultural experiments because of its ability to allocate different factors to plots of different sizes. In standard split-plot designs, main plot treatments are allocated either in a completely randomized design or in a randomized complete block design and subplot treatments are allocated within each main plot. In this paper, we consider split-plot designs where main plot treatments are allocated in a connected incomplete block design. We propose a method of construction and present a catalogue of such designs. We also propose a method of analysis of such split-plot designs. We have implemented proposed construction and analysis methods using R language.

Key words: Split-plot design; Main plot; Whole plot; Subplot; Construction; Analysis.

AMS Subject Classifications: 62K10, 62K15

Prologue

Today, we are all united in our desire to pay our respect to Late Prof. Calyampudi Radhakrishna Rao. Prof. Rao, an Oracle in the field of Statistics, left an indelible mark on the fields of statistics, mathematics, and scientific research worldwide. His groundbreaking contributions have influenced diverse areas, including economics, genetics, anthropology, and medicine. Rao received numerous accolades, including the US National Medal of Science in 2002, and was awarded the International Prize in Statistics in 2023 - a distinction often likened to the 'statistics' equivalent of the Nobel Prize. His legacy continues to inspire generations, and he remains one of the most influential statisticians of all time. It gives us immense pleasure to know that the Society of Statistics, Computer and Applications has decided to bring out a Special Issue of the Statistics and Applications in memory of Late Prof. C R. Rao. This paper is a tribute in honour and loving memory of Late Prof. C R Rao who had a strong bondage with ICAR-Indian Agricultural Statistics Research Institute (ICAR-IASRI), New Delhi and the Indian Society of Agricultural Statistics. He visited the Institute during 2001 to receive Sankhyiki Bhusan Title conferred upon him by the Indian Society of Agricultural Statistics. His keynote address on 'Has Statistics a Future ? If So, in

Corresponding Author: Rajender Parsad
Email: rajender.parsad@icar.gov.in

What Form?' during the 60th Annual Conference of Indian Society of Agricultural Statistics and International Conference on Statistics and Informatics organized by ICAR-IASRI, New Delhi was published in the Journal of the Indian Society of Agricultural Statistics. The paper had set the tone for the requirement of transformation in Statistics in the era of Information and Communication Technology and Big Data. He has made monumental contributions to Design of Experiments. We have also prepared a Technical Bulletin entitled 'CR Rao's Life Sketch and its Influence on Designing of Experiments with a special reference to Agricultural Sciences' available at `http://krishi.icar.gov.in/jspui/handle/123456789/41295`.

By giving us an opportunity to contribute to the Special Issue, we have been given a chance to say thank you, Prof Rao, for paving the way and developing the playground of Statistics where all statisticians like us are working. We express our profound thankfulness to the Guest Editors of this Special Issue and the Chair Editor of Statistics and Applications for giving this opportunity to contribute in such an invaluable Special Issue.

1. Introduction

A split-plot design is a special kind of design in which two factors A and B with m and s levels, respectively, are allocated such that m levels of factor A (also called main plot treatments) are allocated in main plots using a suitable design and s levels of factor B (also called subplot treatments) are allocated to s smaller subplots within each main plot. These designs were originally developed by Fisher (1925). Popular choice of suitable design for levels of factor A is either a completely randomized design or a randomized complete block design. In a split-plot design, the main effect of B and interaction AB are estimated with higher precision and main effect of A are estimated with lesser precision. The main advantage of a split plot design is that the design can accommodate two different plot sizes for two different factors and is, thus, used in many agricultural and other experiments where one of the factor requires comparatively bigger plot size than the other factor. For example, consider an experiment involving irrigation methods (factor A) and fertilizer doses (factor B). It is possible to apply fertilizer doses in smaller plots but application of irrigation methods require bigger plots. So one can apply irrigation methods to bigger plots first and then each bigger plot is subdivided into smaller plots for application of different fertilizer doses. Other such experiments include study of tillage systems (factor A) and various management practices such as doses of fertilizer, pesticides *etc.* as factor B. Split plot designs are adopted in all such experiments where it is not practical to apply both the levels of factor A and B to plots of same size.

In certain situations it may not be possible to allocate all the m levels of factor A in a randomized complete block design and the number of main plots in each block may be restricted to k such that $k < m$. When m is moderately large, then it may not be possible to maintain homogeneity within the blocks with m main plots as these plots are bigger in size. Hence, it is advisable to use lesser number of main plots in such cases and as a result, an incomplete split-plot designs with blocks being incomplete with respect to main plot treatments is preferable. Robinson (1970) pioneered the idea of incomplete split-plot designs in which he arranged the levels of factor A and B in balanced incomplete block (BIB) designs. Bhargava and Shah (1975) considered incomplete split-plot design with main plot treatments in an incomplete block design where they considered unequal block sizes for main plot treatments and mainly studied tests for main effects of factor B and interaction AB.

Mathew and Sinha (1992) went a step further and presented various optimum and exact tests under fixed, random and mixed effects models in the case of unbalanced split-plot designs where main plot treatments are replicated unequal number of times. Mejza (1985) considered incomplete split-plot designs with main plot treatments in incomplete blocks and presented an analysis procedure with a different model than we study here.

There are some other works on incomplete split plot designs where particular classes of incomplete block designs were used either to allocate main plot and / or subplot treatments. Ozawa *et al.* (2004) obtained incomplete split-plot designs using Kronecker product of two component designs, one for levels of factor A and another for levels of factor B. Ozawa and Kuriki (2006) constructed incomplete split-plot designs using semi-Kronecker product of two types of α-resolvable designs. Kuriki and Nakajima (2007) constructed incomplete split-plot designs by semi-Kronecker product of two resolvable designs with second design being a square lattice design for factor B. Kristensen (2012) proposed four methods of constructing incomplete split-plot designs using α-designs. Works on incomplete split-plot designs considering subplot treatments in an incomplete block design are also available, see, for example, Robinson (1967); Mejza and Mejza (1984) and Mandal *et al.* (2020).

In this article, we consider incomplete split-plot designs where m levels of factor A are arranged in a connected incomplete block design with blocks of each of size k such that $k < m$ and s levels of factor B are allocated in s subplots within each main plot. We propose a methodology of analysis of data from experiments conducted using such designs following the standard fixed effects additive linear model approach. Since in agricultural experiments, generally factors and their levels are only a carefully chosen entities among which comparisons are desired, and blocks are also not a random sample from bigger population of blocks, random effects and mixed effects models for analysis of split-plot designs are not considered here and thus, we restrict ourselves to fixed effects model only.

2. Construction

In this section, we present construction of incomplete split plot designs where m levels of factor A are arranged in a connected proper binary incomplete block design with blocks of same size and s levels of factor B are arranged randomly within each main plot. To construct a design, take a binary connected proper incomplete block design D with number of treatments m, number of blocks b and block size $k < m$. Arrange the m levels of factor A using design D. Within each level of factor A, apply s subplot treatments at random. Obtained design is an incomplete split plot design where blocks are incomplete with respect to factor A and whole plots are complete with respect to factor B.

We illustrate the construction with an example.

Example 1: Let $m = 5, s = 5, b = 5, k = 3$. So a connected binary proper incomplete block design D for factor A is

$$\begin{array}{ccc} (1 & 4 & 5) \\ (2 & 3 & 5) \\ (1 & 3 & 4) \\ (2 & 3 & 4) \\ (1 & 2 & 5) \end{array}$$

In D, there are 5 blocks and in each block, three main plot treatments are allocated. Now, randomly assign each of the s levels of factor B in each of the main plots. We get the following incomplete split-plot design.

Block 1	1 (5 4 3 1 2)	4 (5 4 3 1 2)	5 (2 3 4 5 1)
Block 2	2 (3 4 2 1 5)	3 (2 3 4 5 1)	5 (4 1 5 3 2)
Block 3	1 (3 4 1 5 2)	3 (3 4 1 5 2)	4 (3 4 2 1 5)
Block 4	2 (3 4 1 5 2)	3 (3 4 2 1 5)	4 (4 1 5 3 2)
Block 5	1 (4 1 5 3 2)	2 (2 3 4 5 1)	5 (5 4 3 1 2)

Remark 1: We recommend that the design D should be so chosen that it has high A- and D-efficiency. One can use the available efficient incomplete block designs in literature for this purpose. We utilized A-efficient incomplete block designs generated by the R package *ibd* (Mandal, 2019). If the design D is equireplicate with r replications for each of the m levels of factor A, then in the incomplete split-plot design, each AB combination appears r times. Had a complete split-plot design with b blocks been chosen, each AB treatment combination would have appeared b times. Since number of main plots in an incomplete split-plot design is k in each block, it is expected that blocks would be more homogeneous than a block containing m main plots. This will increase precision of comparisons among main effects of factor A. Further, whenever $r < b$, incomplete split-plot designs is expected to be more resource efficient because then they will require lesser number of main plots. For example, consider an experiment conducted by Pandey *et al.* (2000) who used $m = 5$ levels of irrigation regimes as factor A and $s = 5$ levels of Nitrogen doses as factor B and they used complete split-plot design with four blocks. This experiment required 20 main plots and 100 subplots in total. Had an incomplete split-plot design as given in Example 1 with 5 blocks with block size 3 been used, only 15 main plots and 75 subplots would have been required.

We have used the method to construct incomplete split-plot designs in the restricted parametric range of $m \leq 6, s \leq 6$ and $b \leq 10$. The list of parameters for which design has been generated is available, see Mandal *et al.* (2019c). However, the proposed method is general and works for any m, s, b, k provided a suitable connected incomplete block design D with parameters (m, b, k) exists and is available in literature.

3. Analysis

In this section, we present a methodology for analysis of data from experiments conducted using incomplete split-plot designs considered in this paper. We consider fixed effect additive linear model for this purpose:

$$y_{jil} = \mu + \rho_j + \alpha_i + \gamma_{ji} + \beta_l + \delta_{il} + \epsilon_{jil} \tag{1}$$

where y_{jil} denote the observation from the experimental unit in jth block receiving ith level of factor A and lth level of factor B, μ is the general mean, ρ_j is the effect of jth block, α_i is the main effect of ith level of factor A, γ_{ji} is the interaction terms between blocks and ith level of factor A, β_l is the main effect of lth level of factor B, δ_{il} is the interaction effect of ith level of factor A and lth level of factor B and ϵ_{jil} is the random subplot error with zero mean and constant variance $\sigma^2, j = 1, 2, ..., b; i = 1, 2, ..., m; l = 1, 2, ..., s$. Here all the effects are fixed effects except subplot error. Note here that data do not exist for

all (j, i, l) combinations since all levels of factor A do not appear within each block. Here it may be mentioned that Mathew and Sinha (1992) also considered a model similar to (1). However, they considered unbalanced cases, *i.e.*, the blocks may be of unequal sizes and may contain different number of main plot treatments and they also considered cases of random and mixed effects scenarios. In our case, each block is of constant size k and contains $k < m$ main plot treatments and we do not consider random and mixed effect models.

In matrix notation, the model (1) may be represented as

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}_1\boldsymbol{\rho} + \mathbf{X}_2\boldsymbol{\alpha} + \mathbf{X}_3\boldsymbol{\gamma} + \mathbf{X}_4\boldsymbol{\beta} + \mathbf{X}_5\boldsymbol{\delta} + \boldsymbol{\epsilon} \tag{2}$$

where $\mathbf{y}$ denotes the vector of n observations, $\mathbf{1}$ denotes the vector of ones, $\mathbf{X}_1$ denotes $n \times b$ observation versus block incidence matrix, $\boldsymbol{\rho}$ denotes $b \times 1$ vector of block effects, $\mathbf{X}_2$ denotes $n \times m$ observation versus factor A incidence matrix, $\boldsymbol{\alpha}$ denotes $m \times 1$ vector of main effects of factor A, $\mathbf{X}_3$ denotes $n \times bk$ observation versus block-A incidence matrix, $\boldsymbol{\gamma}$ denotes $bk \times 1$ vector of block versus factor A interactions, $\mathbf{X}_4$ denotes $n \times s$ observation versus factor B incidence matrix, $\boldsymbol{\beta}$ denotes $s \times 1$ vector of main effects of factor B, $\mathbf{X}_5$ denotes $n \times ms$ observation versus AB interaction incidence matrix, $\boldsymbol{\delta}$ denotes $ms \times 1$ vector of AB interaction effects and $\boldsymbol{\epsilon}$ denotes $n \times 1$ vector of errors. We assume that errors are i.i.d. normal with $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n$. Under the given set-up of design construction, $n = bks$ and $\mathbf{X}_1 = \mathbf{I}_b \otimes \mathbf{1}_{ks}$. The model (2) can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{3}$$

where $\mathbf{X} = (\mathbf{1} : \mathbf{X}_1 : \mathbf{X}_2 : \mathbf{X}_3 : \mathbf{X}_4 : \mathbf{X}_5)$ and $\boldsymbol{\theta} = (\mu, \boldsymbol{\rho}', \boldsymbol{\alpha}', \boldsymbol{\gamma}', \boldsymbol{\beta}', \boldsymbol{\delta}')'$.

Normal equations are given by

$$\mathbf{X}'\mathbf{X}\boldsymbol{\theta} = \mathbf{X}'\mathbf{y}$$

where

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{X}_1 & \mathbf{1}'\mathbf{X}_2 & \mathbf{1}'\mathbf{X}_3 & \mathbf{1}'\mathbf{X}_4 & \mathbf{1}'\mathbf{X}_5 \\ \mathbf{X}_1'\mathbf{1} & \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 & \mathbf{X}_1'\mathbf{X}_3 & \mathbf{X}_1'\mathbf{X}_4 & \mathbf{X}_1'\mathbf{X}_5 \\ \mathbf{X}_2'\mathbf{1} & \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 & \mathbf{X}_2'\mathbf{X}_3 & \mathbf{X}_2'\mathbf{X}_4 & \mathbf{X}_2'\mathbf{X}_5 \\ \mathbf{X}_3'\mathbf{1} & \mathbf{X}_3'\mathbf{X}_1 & \mathbf{X}_3'\mathbf{X}_2 & \mathbf{X}_3'\mathbf{X}_3 & \mathbf{X}_3'\mathbf{X}_4 & \mathbf{X}_3'\mathbf{X}_5 \\ \mathbf{X}_4'\mathbf{1} & \mathbf{X}_4'\mathbf{X}_1 & \mathbf{X}_4'\mathbf{X}_2 & \mathbf{X}_4'\mathbf{X}_3 & \mathbf{X}_4'\mathbf{X}_4 & \mathbf{X}_4'\mathbf{X}_5 \\ \mathbf{X}_5'\mathbf{1} & \mathbf{X}_5'\mathbf{X}_1 & \mathbf{X}_5'\mathbf{X}_2 & \mathbf{X}_5'\mathbf{X}_3 & \mathbf{X}_5'\mathbf{X}_4 & \mathbf{X}_5'\mathbf{X}_5 \end{pmatrix}. \tag{4}$$

Now, following relations can be verified:

$$\mathbf{X}_1'\mathbf{1} = ks\mathbf{1}_b \qquad \mathbf{X}_2'\mathbf{1} = s\mathbf{r}$$
$$\mathbf{X}_3'\mathbf{1} = s\mathbf{1}_{bk} \qquad \mathbf{X}_4'\mathbf{1} = bk\mathbf{1}_s$$
$$\mathbf{X}_5'\mathbf{1} = \mathbf{r} \otimes \mathbf{1}_s \qquad \mathbf{X}_1'\mathbf{X}_1 = ks\mathbf{I}_b$$
$$\mathbf{X}_2'\mathbf{X}_1 = s\mathbf{N}_1, \text{ say} \qquad \mathbf{X}_3'\mathbf{X}_1 = s\mathbf{1}_k \otimes \mathbf{I}_b'$$
$$\mathbf{X}_4'\mathbf{X}_1 = k\mathbf{1}_s\mathbf{1}_b' \qquad \mathbf{X}_5'\mathbf{X}_1 = \mathbf{N}_3, \text{ say}$$
$$\mathbf{X}_2'\mathbf{X}_2 = s\mathbf{R} \qquad \mathbf{X}_3'\mathbf{X}_2 = \mathbf{N}_2, \text{ say}$$
$$\mathbf{X}_4'\mathbf{X}_2 = \mathbf{r}' \otimes \mathbf{1}_s \qquad \mathbf{X}_5'\mathbf{X}_2 = \mathbf{R} \otimes \mathbf{1}_s$$
$$\mathbf{X}_3'\mathbf{X}_3 = s\mathbf{I}_{bk} \qquad \mathbf{X}_4'\mathbf{X}_3 = \mathbf{1}_s\mathbf{1}_{bk}', \text{ say}$$
$$\mathbf{X}_5'\mathbf{X}_3 = \mathbf{N}_4, \text{ say} \qquad \mathbf{X}_4'\mathbf{X}_4 = bk\mathbf{I}_s$$
$$\mathbf{X}_5'\mathbf{X}_4 = \mathbf{r} \otimes \mathbf{I}_s, \text{ say} \qquad \mathbf{X}_5'\mathbf{X}_5 = \mathbf{R} \otimes \mathbf{I}_s$$

with $\mathbf{r}$ being the vector of replications of levels of factor A and $\mathbf{R}$ being diagonal matrix with elements of $\mathbf{r}$. Therefore, equation (4) can be written as

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} bks & ks\mathbf{1}_b' & s\mathbf{r}' & s\mathbf{1}_{bk}' & bk\mathbf{1}_s' & \mathbf{r}' \otimes \mathbf{1}_s' \\ ks\mathbf{1}_b & ks\mathbf{I}_b & s\mathbf{N}_1' & s\mathbf{1}_k' \otimes \mathbf{I}_b & k\mathbf{1}_b\mathbf{1}_s' & \mathbf{N}_3' \\ s\mathbf{r} & s\mathbf{N}_1 & s\mathbf{R} & \mathbf{N}_2' & \mathbf{r} \otimes \mathbf{1}_s' & \mathbf{R} \otimes \mathbf{1}_s' \\ s\mathbf{1}_{bk} & s\mathbf{1}_k \otimes \mathbf{I}_b & \mathbf{N}_2 & s\mathbf{I}_{bk} & \mathbf{1}_{bk}\mathbf{1}_s' & \mathbf{N}_4' \\ bk\mathbf{1}_s & k\mathbf{1}_s\mathbf{1}_b' & \mathbf{r}' \otimes \mathbf{1}_s & \mathbf{1}_s\mathbf{1}_{bk}' & bk\mathbf{I}_s & \mathbf{r}' \otimes \mathbf{I}_s \\ \mathbf{r} \otimes \mathbf{1}_s & \mathbf{N}_3 & \mathbf{R} \otimes \mathbf{1}_s & \mathbf{N}_4 & \mathbf{r} \otimes \mathbf{I}_s & \mathbf{R} \otimes \mathbf{I}_s \end{pmatrix}. \tag{5}$$

It may be seen that

$$\mathbf{X}'\mathbf{y} = (y... : \mathbf{y}_{B..}' : \mathbf{y}_{.M.}' : \mathbf{y}_{BM.}' : \mathbf{y}_{.S.}' : \mathbf{y}_{.MS}')'$$

where $y...$ denote the gross total of all observations, $\mathbf{y}_{B..}$ is the vector of block totals, $\mathbf{y}_{.M.}$ is the vector of totals for m levels of factor A, $\mathbf{y}_{BM.}$ is the vector of totals corresponding to block-factor A combinations, $\mathbf{y}_{..S}$ is the vector of totals for s levels of factor B and $\mathbf{y}_{.MS}$ is the vector of totals corresponding to AB combinations.

One can verify that the number of rows in $\mathbf{X}'\mathbf{X}$ is $1 + b + m + bk + s + ms$, but there are total $1 + 1 + (m + b - 1) + 1 + (m + s - 1) = 1 + b + 2m + s$ linearly dependent rows and they are as follows: sum of 2nd to $(b + 1)$th row is equal to the first row, sum of $(b + 2)$th row to $(b + m + 2)$th row is equal to the first row, summing rows for each level of γ_{ji} over i keeping j fixed gives row corresponding to jth $(j = 1, 2, ..., b)$ block and similarly summing rows for each level of γ_{ji} over j keeping i fixed gives row corresponding to row of ith $(i = 1, 2, ..., m)$ level of factor A, summing of rows corresponding to s levels of factor B gives the first row, summing rows for each level of δ_{il} over i keeping l fixed gives row corresponding to lth $(l = 1, 2, ..., s)$ level of factor B, summing rows for each level of δ_{il} over l keeping i fixed gives row corresponding to ith level of factor A. Therefore, to get a solution to the normal equations (3), one can set $(1 + b + 2m + s)$ parameter estimates to zero. We set $\hat{\mu} = 0, \hat{\rho}_j = 0, \hat{\alpha}_i = 0, \hat{\beta}_l = 0 \forall j, i, l$ and we also set every sth component of $\hat{\boldsymbol{\delta}}$ as zero, *i.e.*, $\hat{\delta}_s = 0, \hat{\delta}_{2s} = 0, ..., \hat{\delta}_{ms} = 0$. As a result, we get,

$$\begin{pmatrix} s\mathbf{I}_{bk} & \tilde{\mathbf{N}}_4' \\ \tilde{\mathbf{N}}_4 & \mathbf{R} \otimes \mathbf{I}_{s-1} \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\gamma}} \\ \hat{\boldsymbol{\delta}}_{(-m)} \end{pmatrix} = \begin{pmatrix} \mathbf{y}_{BM.} \\ \tilde{\mathbf{y}}_{.MS} \end{pmatrix} \tag{6}$$

where $\tilde{\mathbf{N}}_4$ is the matrix obtained after removing every sth row of $\mathbf{N}_4$, $\hat{\boldsymbol{\delta}}_{(-m)}$ is the vector after removing every sth element of $\hat{\boldsymbol{\delta}}$ and $\tilde{\mathbf{y}}_{.MS}$ is the vector obtained after removing every sth element of $\mathbf{y}_{.MS}$. From (6), we get,

$$\hat{\boldsymbol{\gamma}} = \frac{1}{s}\left(\mathbf{y}_{BM.} - \tilde{\mathbf{N}}_4'\hat{\boldsymbol{\delta}}_{(-m)}\right).$$

After a little algebra, it may be seen that

$$\hat{\boldsymbol{\delta}}_{(-m)} = \mathbf{C}_{MS}^{-1}\mathbf{Q}_{MS}$$

where $\mathbf{C}_{MS} = \left(\mathbf{R} \otimes \mathbf{I}_{s-1} - \frac{1}{s}\tilde{\mathbf{N}}_4\tilde{\mathbf{N}}_4'\right)$ and $\mathbf{Q}_{MS} = \left(\tilde{\mathbf{y}}_{.MS} - \frac{1}{s}\tilde{\mathbf{N}}_4\mathbf{y}_{BM.}\right)$.

Denoting the model sum of squares due to fitting parameters $\mu, \boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\delta}$ with $R(\mu, \boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\delta})$, we get

$$R(\mu, \boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\delta}) = \hat{\boldsymbol{\gamma}}'\mathbf{X}_3'\mathbf{y} + \hat{\boldsymbol{\delta}}'\mathbf{X}_5'\mathbf{y} = \frac{1}{s}\mathbf{y}_{BM.}'\mathbf{y}_{BM.} + \mathbf{Q}_{MS}'\mathbf{C}_{MS}^{-1}\mathbf{Q}_{MS}.$$

Similarly, it may be verified that

$$R(\mu, \boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \hat{\boldsymbol{\gamma}}'\mathbf{X}_3'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}_4'\mathbf{y} = \frac{1}{s}\mathbf{y}_{BM.}'\mathbf{y}_{BM.} + \mathbf{Q}_S'\mathbf{C}_S^{-1}\mathbf{Q}_S$$

where $\mathbf{C}_S = bk\mathbf{I}_{s-1} - \frac{bk}{s}\mathbf{1}_{s-1}\mathbf{1}_{s-1}'$ and $\mathbf{Q}_S = \tilde{\mathbf{y}}_{..S} - \frac{y_{...}}{s}\mathbf{1}_{s-1}$

$$R(\mu, \boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \hat{\boldsymbol{\rho}}'\mathbf{X}_1'\mathbf{y} + \hat{\boldsymbol{\alpha}}'\mathbf{X}_2'\mathbf{y} + \hat{\boldsymbol{\gamma}}'\mathbf{X}_3'\mathbf{y} = \frac{1}{s}\mathbf{y}_{BM.}'\mathbf{y}_{BM.}$$

$$R(\mu, \boldsymbol{\rho}, \boldsymbol{\alpha}) = \hat{\boldsymbol{\rho}}'\mathbf{X}_1'\mathbf{y} + \hat{\boldsymbol{\alpha}}'\mathbf{X}_2'\mathbf{y} = \frac{1}{ks}\mathbf{y}_{B..}'\mathbf{y}_{B..} + \mathbf{Q}_M'\mathbf{C}_M^{-1}\mathbf{Q}_M$$

where $\mathbf{C}_M = s\mathbf{R}_{m-1} - \frac{s}{k}\tilde{\mathbf{N}}_1\tilde{\mathbf{N}}_1'$ and $\mathbf{Q}_M = \tilde{\mathbf{y}}_{.M.} - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{y}_{B..}$

$$R(\mu, \boldsymbol{\rho}) = \hat{\boldsymbol{\rho}}'\mathbf{X}_1'\mathbf{y} = \frac{1}{ks}\mathbf{y}_{B..}'\mathbf{y}_{B..}$$

and

$$R(\mu) = \frac{1}{bks}{y...}^2$$

Residual sum of squares after fitting the model (3) is given by

$$SSE = \mathbf{y}'\mathbf{y} - \frac{1}{s}\mathbf{y}_{BM.}'\mathbf{y}_{BM.} - \mathbf{Q}_{MS}'\mathbf{C}_{MS}^{-1}\mathbf{Q}_{MS}. \tag{7}$$

Theorem 1: Under model (3), $SSE/\sigma^2 \sim \chi^2_{bks-bk-ms+m}$.

Proof: It is well known that in a fixed effects linear model (3), $SSE/\sigma^2 \sim \chi^2_{n-rank(\mathbf{X})}$. Here, $rank(\mathbf{X}) = rank(\mathbf{X}'\mathbf{X}) = bk + ms - m$. So the result follows. □

3.1. Testing significance of interactions between A and B

Consider the null hypothesis $H_0 : \delta_{i1} = \delta_{i2} = \cdots = \delta_{is} \forall i = 1, 2, ..., m$ versus H_1 : At least two of them are different. Under the null hypothesis, the reduced model is

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}_1\boldsymbol{\rho} + \mathbf{X}_2\boldsymbol{\alpha} + \mathbf{X}_3\boldsymbol{\gamma} + \mathbf{X}_4\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

The residual sum of squares under reduced model is

$$SSE_1 = \mathbf{y}'\mathbf{y} - R(\mu, \boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - \frac{1}{s}\mathbf{y}_{BM.}'\mathbf{y}_{BM.} - \mathbf{Q}_S'\mathbf{C}_S^{-1}\mathbf{Q}_S.$$

Theorem 2: $SSE_1/\sigma^2 \sim \chi^2_{bks-bk-s+1}$.

Proof: The rank of the model matrix $\mathbf{X}_{r1} = (\mathbf{1} : \mathbf{X}_1 : \mathbf{X}_2 : \mathbf{X}_3 : \mathbf{X}_4)$ is $bk + s - 1$ since out of $(1 + b + m + bk + s)$ rows of $\mathbf{X}'_{r1}\mathbf{X}_{r1}$, there are $b + m + 2$ dependencies. Hence, the result follows. □ Now, $SSE_1 - SSE = \mathbf{Q}'_{MS}\mathbf{C}^{-1}_{MS}\mathbf{Q}_{MS} - \mathbf{Q}'_S\mathbf{C}^{-1}_S\mathbf{Q}_S$. Therefore, the test statistic for testing $H_0 : \delta_{i1} = \delta_{i2} = \cdots = \delta_{is} \forall i = 1, 2, ..., m$ is

$$F_1 = \frac{(SSE_1 - SSE)/(m-1)(s-1)}{SSE/(bks - bk - ms + m)}$$
$$= \frac{(\mathbf{Q}'_{MS}\mathbf{C}^{-1}_{MS}\mathbf{Q}_{MS} - \mathbf{Q}'_S\mathbf{C}^{-1}_S\mathbf{Q}_S)/(m-1)(s-1)}{SSE/(bks - bk - ms + m)} \sim F_{(m-1)(s-1),(bks-bk-ms+m)}$$

under null hypothesis. Null hypothesis is rejected whenever calculated value of $F_1 > F_{\alpha,(m-1)(s-1),(bks-bk-ms+m)}$ where $F_{\alpha,(m-1)(s-1),(bks-bk-ms+m)}$ denotes the upper α percent point of an F-distribution with $(m-1)(s-1)$ and $(bks - bk - ms + m)$ degrees of freedom.

3.2. Testing significance of main effects of factor B

Assuming that interactions between A and B is absent, we consider the null hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_s = \beta$, say versus the alternative H_1 : At least two of them are different. Consider the following test statistic

$$F_2 = \frac{\mathbf{Q}'_S\mathbf{C}^{-1}_S\mathbf{Q}_S/(s-1)}{SSE/(bks - bk - ms + m)}$$

which follows $F_{(s-1),(bks-bk-ms+m)}$, See Appendix for proof. One can reject the null hypothesis when calculated value of $F_2 > F_{\alpha,(s-1),(bks-bk-ms+m)}$.

3.3. Testing significance of main effects of factor A

Since main effects of A can be tested when interactions of A with B and with blocks is absent, we assume that interactions between A and B is absent and then we consider the null hypothesis $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_m = \alpha$, say versus the alternative H_1 : At least two of them are different. One can see that

$$F_3 = \frac{\mathbf{Q}'_M\mathbf{C}^{-1}_M\mathbf{Q}_M/(m-1)}{SSW/(bk - b - m + 1)} \sim F_{(m-1),(bk-b-m+1)}.$$

$SSW = R(\boldsymbol{\gamma}|\mu, \boldsymbol{\rho}, \boldsymbol{\alpha}) = R(\mu, \boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) - R(\mu, \boldsymbol{\rho}, \boldsymbol{\alpha}) = \frac{1}{s}\mathbf{y}'_{BM.}\mathbf{y}_{BM.} - \frac{1}{ks}\mathbf{y}'_{B..}\mathbf{y}_{B..} - \mathbf{Q}'_M\mathbf{C}^{-1}_M\mathbf{Q}_M.$

Above results can be summarized in the form of analysis of variance (ANOVA) table as given in Table 1 where,

$SSR = \frac{1}{ks}\mathbf{y}'_{B..}\mathbf{y}_{B..} - \frac{1}{bks}y...^2$ $\quad SSA = \mathbf{Q}'_M\mathbf{C}^{-1}_M\mathbf{Q}_M$
$SSW = \frac{1}{s}\mathbf{y}'_{BM.}\mathbf{y}_{BM.} - \frac{1}{ks}\mathbf{y}'_{B..}\mathbf{y}_{B..} - \mathbf{Q}'_M\mathbf{C}^{-1}_M\mathbf{Q}_M.$ $\quad SSB = \mathbf{Q}'_S\mathbf{C}^{-1}_S\mathbf{Q}_S$
$SSAB = \mathbf{Q}'_{MS}\mathbf{C}^{-1}_{MS}\mathbf{Q}_{MS} - \mathbf{Q}'_S\mathbf{C}^{-1}_S\mathbf{Q}_S$ $\quad SST = \mathbf{y}'\mathbf{y} - \frac{1}{bks}y...^2$

Remark 2: The model formulation under a split-plot design often involves a whole-plot error and a split-plot error, both of which are assumed to be random, satisfying the usual normality assumptions. This formulation leads to two ANOVA tables: a whole-plot ANOVA and a split-plot ANOVA. The present work considers a model that includes only one random

Table 1: ANOVA table depicting analysis of incomplete split plot designs

Source	Degrees of freedom	Sum of squares	Mean squares	F
Blocks	$b-1$	SSR	-	-
A	$m-1$	SSA	$MSA = SSA/(m-1)$	$F_3 = MSA/MSW$
Block × A	$bk-b-m+1$	SSW	$MSW = SSW/(bk-b-m+1)$	
B	$s-1$	SSB	$MSB = SSB/(s-1)$	$F_2 = MSB/MSE$
AB	$(m-1)(s-1)$	$SSAB$	$MSAB = SSAB/(m-1)(s-1)$	$F_1 = MSAB/MSE$
Error	$bks-bk-ms+m$	SSE	$MSE = SSE/(bks-bk-ms+m)$	-
Total	$bks-1$	SST	-	-

error term. For testing the significance of the main effects due to whole plot factor, we assume that A × B interaction and block × A interactions are absent and then use the mean square due to block × A interaction in the denominator of the F ratio and this F ratio coincides with the corresponding F ratio in the whole-plot ANOVA.

3.4. Estimation of treatment contrasts

First we consider estimation of treatment contrasts of factor B. It may be seen that

$$\hat{\boldsymbol{\beta}}_{s-1} = \mathbf{C}_S^{-1}\mathbf{Q}_S.$$

Thus, we can write

$$\hat{\boldsymbol{\beta}} = \mathbf{C}_S^{*-}\mathbf{Q}_S^* \tag{8}$$

where $\mathbf{C}_S^* = bk\mathbf{I}_s - \frac{bk}{s}\mathbf{1}_s\mathbf{1}_s'$ and $\mathbf{Q}_S^* = \mathbf{y}_{..S} - \frac{y_{...}}{s}\mathbf{1}_s$.

Theorem 3: Let $\mathbf{p}'\boldsymbol{\beta}$ be a linear parametric function such that $\mathbf{p}'\mathbf{1} = 0$. Then $\mathbf{p}'\boldsymbol{\beta}$ is estimable.

Proof: Consider the estimator $\mathbf{p}'\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is given by equation (8). Then,

$$\begin{aligned}
E(\mathbf{p}'\hat{\boldsymbol{\beta}}) &= E(\mathbf{p}'\mathbf{C}_S^{*-}\mathbf{Q}_S^*) \\
&= \mathbf{p}'\mathbf{C}_S^{*-}E(\mathbf{y}_{..S} - \frac{y_{...}}{s}\mathbf{1}_s) \\
&= \mathbf{p}'\mathbf{C}_S^{*-}(\mathbf{X}_4' - \frac{1}{s}\mathbf{X}_4'\mathbf{X}_3\mathbf{X}_3')E(\mathbf{y}) \\
&= \mathbf{p}'\mathbf{C}_S^{*-}(\mathbf{X}_4' - \frac{1}{s}\mathbf{X}_4'\mathbf{X}_3\mathbf{X}_3')(\mu\mathbf{1} + \mathbf{X}_1\boldsymbol{\rho} + \mathbf{X}_2\boldsymbol{\alpha} + \mathbf{X}_3\boldsymbol{\gamma} + \mathbf{X}_4\boldsymbol{\beta}) \\
&= \mathbf{p}'\mathbf{C}_S^{*-}\mathbf{C}_S^*\boldsymbol{\beta} \text{ (after simplification)} \\
&= \mathbf{p}'\boldsymbol{\beta}
\end{aligned}$$

since $\mathbf{p}'\mathbf{1} = 0$. This completes the proof. □

It is easy to see that $v(\mathbf{p}'\hat{\boldsymbol{\beta}}) = \mathbf{p}'\mathbf{C}_S^{*-}\mathbf{p}\sigma^2$ where $v(.)$ denotes variance. So under normality of errors in model (1), $\mathbf{p}'\hat{\boldsymbol{\beta}} \sim N(\mathbf{p}'\boldsymbol{\beta}, \mathbf{p}'\mathbf{C}_S^{*-}\mathbf{p}\sigma^2)$. Thus, testing of hypothesis

$H_0 : \mathbf{p}'\boldsymbol{\beta} = b$ can be performed using the test statistic

$$F_s = \frac{(\mathbf{p}'\hat{\boldsymbol{\beta}} - b)^2/(\mathbf{p}'\mathbf{C}_S^{*-}\mathbf{p})}{SSE/(bks - bk - ms + m)}.$$

Under null hypothesis $F_s \sim F_{1,bks-bk-ms+m}$.

Exactly on similar lines, it can be proved that for testing $H_0 : \mathbf{q}'\boldsymbol{\delta} = d$, one can use the test statistic

$$F_{ms} = \frac{(\mathbf{q}'\hat{\boldsymbol{\delta}} - d)^2/(\mathbf{q}'\mathbf{C}_{MS}^{*-}\mathbf{q})}{SSE/(bks - bk - ms + m)}$$

which follows F distribution with 1 and $(bks - bk - ms + m)$ degrees of freedom under null hypothesis. Here, $\hat{\boldsymbol{\delta}} = \mathbf{C}_{MS}^{*-}\mathbf{Q}_{MS}^{*}$ with $\mathbf{C}_{MS}^{*} = \mathbf{R} \otimes \mathbf{I}_{s-1} - \frac{1}{s}\mathbf{N}_4\mathbf{N}_4'$ and $\mathbf{Q}_{MS}^{*} = \mathbf{y}_{.MS} - \frac{1}{s}\mathbf{N}_4\mathbf{y}_{BM.}$.

Now consider a treatment contrast $\mathbf{w}'\boldsymbol{\alpha}$ of main effects of factor A. An estimator of this treatment contrast is given by $\mathbf{w}'\hat{\boldsymbol{\alpha}} = \mathbf{w}'\mathbf{C}_M^{*-}\mathbf{Q}_M^{*}$ where $\mathbf{C}_M^{*} = s\mathbf{R}_m - \frac{s}{k}\mathbf{N}_1\mathbf{N}_1'$ and $\mathbf{Q}_M^{*} = \mathbf{y}_{.M.} - \frac{1}{k}\mathbf{N}_1\mathbf{y}_{B..}$. To test $H_0 : \mathbf{w}'\hat{\boldsymbol{\alpha}} = a$, the following test statistic can be used:

$$F_m = \frac{(\mathbf{w}'\hat{\boldsymbol{\alpha}} - a)^2/(\mathbf{w}'\mathbf{C}_M^{*-}\mathbf{w})}{SSW/(bk - b - m + 1)}.$$

Under null hypothesis, $F_m \sim F_{1,(bk-b-m+1)}$ and inferences can be made accordingly.

4. Concluding remarks

In this paper, we have proposed a method of construction of incomplete split-plot designs where main plot treatments are allocated using a connected proper incomplete block design. We have also presented an analysis methodology for the proposed designs. We have implemented the proposed methods of construction and analysis using R language and the same is available as part of an R package 'ispd' which can be accessed on https://cran.r-project.org/web/packages/ispd/index.html, see (Mandal *et al.*, 2019a). Further, we have also implemented the construction and analysis methodology as part of an web application which is available on http://drsr.icar.gov.in/ISPD/Home.jsp, see (Mandal *et al.*, 2019b). The will enable the experimenters and statisticians to use these designs with ease.

Acknowledgements

The authors sincerely acknowledges the financial support received in the form of Extra Mural Research Grant by Science and Engineering Research Board (SERB), Department of Science and Technology, India. We are thankful to the Chair Editor and General Editors for their valuable guidance in improving the manuscript. We are also thankful to the anonymous reviewers for their enlightening remarks which led to considerable improvement in the presentation of the article.

References

Bhargava, R. and Shah, K. (1975). Analysis of some mixed-models for block and split-plot designs. *Annals of the Institute of Statistical Mathematics*, **27**, 365–375.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Kristensen, K. (2012). Incomplete split-plot designs based on α-designs: a compromise between traditional split-plot designs and randomised complete block design. *Euphytica*, **183**, 401–413.

Kuriki, S. and Nakajima, K. (2007). Square lattice designs in incomplete split-plot designs. *Journal of Statistical Theory and Practice*, **1**, 417–426.

Mandal, B. N. (2019). *ibd: Incomplete Block Designs*. R package version 1.5.

Mandal, B. N., Dash, S., and Parsad, R. (2019a). *ispd: Incomplete Split-Plot Designs*. R package version 0.1.

Mandal, B. N., Parsad, R., and Dash, S. (2019b). Incomplete split plot designs : Construction and analysis. http://drsr.icar.gov.in/ISPD/how_to.jsp.

Mandal, B. N., Parsad, R., and Dash, S. (2019c). Incomplete split-plot designs: Blocks are incomplete, main plots are complete : Design resources server. http://drs.icar.gov.in.

Mandal, B. N., Parsad, R., and Dash, S. (2020). Incomplete split-plot designs: Construction and analysis. *Statistics and Probability Letters*, **166**, 108869.

Mathew, T. and Sinha, B. K. (1992). Exact and optimum tests in unbalanced split-plot designs under mixed and random models. *Journal of the American Statistical Association*, **87**, 192–200.

Mejza, I. and Mejza, S. (1984). Incomplete split plot designs. *Statistics and Probability Letters*, **2**, 327–332.

Mejza, S. (1985). A split-plot design with wholeplot treatments in an incomplete block design. In *Linear Statistical Inference*, pages 211–222. Springer.

Ozawa, K. and Kuriki, S. (2006). Incomplete split-plot designs generated from α-resolvable designs. *Statistics and Probability Letters*, **76**, 1245–1254.

Ozawa, K., Mejza, S., Jimbo, M., Mejza, I., and Kuriki, S. (2004). Incomplete split-plot designs generated by some resolvable balanced designs. *Statistics and probability letters*, **68**, 9–15.

Pandey, R., Maranville, J., and Admou, A. (2000). Deficit irrigation and nitrogen effects on maize in a sahelian environment: I. grain yield and yield components. *Agricultural Water Management*, **46**, 1–13.

Robinson, J. (1967). Incomplete split plot designs. *Biometrics*, **23**, 793–802.

Robinson, J. (1970). Blocking in incomplete split plot designs. *Biometrika*, **57**, 347–350.

Appendix

Proof of F_2 following $F_{(s-1),(bks-bk-ms+m)}$:
First we prove that $\mathbf{Q}_S'\mathbf{C}_S^{-1}\mathbf{Q}_S/\sigma^2 \sim \chi^{2'}_{s-1}$ with non-centrality parameter $\boldsymbol{\theta}'\mathbf{X}'(\tilde{\mathbf{X}}_4-\frac{1}{s}\mathbf{X}_3\mathbf{1}_{bk}\mathbf{1}'_{s-1})\mathbf{C}_s^{-1}(\tilde{\mathbf{X}}_4'-\frac{1}{s}\mathbf{1}_{s-1}\mathbf{1}'_{bk}\mathbf{X}_3')\mathbf{X}\boldsymbol{\theta}/2\sigma^2$.

Note that $\mathbf{Q}_S = \left(\tilde{\mathbf{y}}_{..S} - \frac{y_{...}}{s}\mathbf{1}_{s-1}\right) = \tilde{\mathbf{X}}_4'\mathbf{y} - \frac{1}{s}\tilde{\mathbf{X}}_4'\mathbf{X}_3\mathbf{X}_3'\mathbf{y}$ and hence $\mathbf{Q}_S'\mathbf{C}_S^{-1}\mathbf{Q}_S = \mathbf{y}'(\tilde{\mathbf{X}}_4-\frac{1}{s}\mathbf{X}_3\mathbf{X}_3'\tilde{\mathbf{X}}_4)\mathbf{C}_s^{-1}(\tilde{\mathbf{X}}_4-\frac{1}{s}\mathbf{X}_3\mathbf{X}_3'\tilde{\mathbf{X}}_4)'\mathbf{y}$. Now, $(\tilde{\mathbf{X}}_4-\frac{1}{s}\mathbf{X}_3\mathbf{X}_3'\tilde{\mathbf{X}}_4)\mathbf{C}_s^{-1}(\tilde{\mathbf{X}}_4-\frac{1}{s}\mathbf{X}_3\mathbf{X}_3'\tilde{\mathbf{X}}_4)'$ is idempotent because

$$
\begin{aligned}
&(\tilde{\mathbf{X}}_4-\frac{1}{s}\mathbf{X}_3\mathbf{X}_3'\tilde{\mathbf{X}}_4)\mathbf{C}_s^{-1}(\tilde{\mathbf{X}}_4-\frac{1}{s}\mathbf{X}_3\mathbf{X}_3'\tilde{\mathbf{X}}_4)'(\tilde{\mathbf{X}}_4-\frac{1}{s}\mathbf{X}_3\mathbf{X}_3'\tilde{\mathbf{X}}_4)\mathbf{C}_s^{-1}(\tilde{\mathbf{X}}_4-\frac{1}{s}\mathbf{X}_3\mathbf{X}_3'\tilde{\mathbf{X}}_4)'\\
&=(\tilde{\mathbf{X}}_4-\frac{1}{s}\mathbf{X}_3\mathbf{1}_{bk}\mathbf{1}'_{s-1})\mathbf{C}_s^{-1}(\tilde{\mathbf{X}}_4'-\frac{1}{s}\mathbf{1}_{s-1}\mathbf{1}'_{bk}\mathbf{X}_3')(\tilde{\mathbf{X}}_4-\frac{1}{s}\mathbf{X}_3\mathbf{1}_{bk}\mathbf{1}'_{s-1})\mathbf{C}_s^{-1}(\tilde{\mathbf{X}}_4'-\frac{1}{s}\mathbf{1}_{s-1}\mathbf{1}'_{bk}\mathbf{X}_3')\\
&=(\tilde{\mathbf{X}}_4-\frac{1}{s}\mathbf{X}_3\mathbf{1}_{bk}\mathbf{1}'_{s-1})\mathbf{C}_s^{-1}\mathbf{C}_s\mathbf{C}_s^{-1}(\tilde{\mathbf{X}}_4'-\frac{1}{s}\mathbf{1}_{s-1}\mathbf{1}'_{bk}\mathbf{X}_3')
\end{aligned}
$$

since

$$
\begin{aligned}
&(\tilde{\mathbf{X}}_4'-\frac{1}{s}\mathbf{1}_{s-1}\mathbf{1}'_{bk}\mathbf{X}_3')(\tilde{\mathbf{X}}_4-\frac{1}{s}\mathbf{X}_3\mathbf{1}_{bk}\mathbf{1}'_{s-1})\\
&=\tilde{\mathbf{X}}_4'\tilde{\mathbf{X}}_4-\frac{1}{s}\tilde{\mathbf{X}}_4'\mathbf{X}_3\mathbf{1}_{bk}\mathbf{1}'_{s-1}-\frac{1}{s}\mathbf{1}_{s-1}\mathbf{1}'_{bk}\mathbf{X}_3'\tilde{\mathbf{X}}_4+\frac{1}{s^2}\mathbf{1}_{s-1}\mathbf{1}'_{bk}\mathbf{X}_3'\mathbf{X}_3\mathbf{1}_{bk}\mathbf{1}'_{s-1}\\
&=bk(\mathbf{I}_{s-1}-\frac{1}{s}\mathbf{1}_{s-1}\mathbf{1}'_{s-1}-\frac{1}{s}\mathbf{1}_{s-1}\mathbf{1}'_{s-1}+\frac{1}{s}\mathbf{1}_{s-1}\mathbf{1}'_{s-1})\\
&=bk((\mathbf{I}_{s-1}-\frac{1}{s}\mathbf{1}_{s-1}\mathbf{1}'_{s-1})\\
&=\mathbf{C}_s.
\end{aligned}
$$

Now, under H_0,

$$
\begin{aligned}
\mathbf{X}\boldsymbol{\theta}&=\mu\mathbf{1}+\mathbf{X}_1\boldsymbol{\rho}+\mathbf{X}_2\boldsymbol{\alpha}+\mathbf{X}_3\boldsymbol{\gamma}+\beta\mathbf{X}_4\mathbf{1}_s\\
&=\mu\mathbf{1}+\mathbf{X}_1\boldsymbol{\rho}+\mathbf{X}_2\boldsymbol{\alpha}+\mathbf{X}_3\boldsymbol{\gamma}+\beta\mathbf{1}\\
&=(\mu+\beta)\mathbf{1}+\mathbf{X}_1\boldsymbol{\rho}+\mathbf{X}_2\boldsymbol{\alpha}+\mathbf{X}_3\boldsymbol{\gamma}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&(\tilde{\mathbf{X}}_4'-\frac{1}{s}\mathbf{1}_{s-1}\mathbf{1}'_{bk}\mathbf{X}_3')\mathbf{X}\boldsymbol{\theta}\\
&=(\tilde{\mathbf{X}}_4'-\frac{1}{s}\mathbf{1}_{s-1}\mathbf{1}'_{bk}\mathbf{X}_3')((\mu+\beta)\mathbf{1}+\mathbf{X}_1\boldsymbol{\rho}+\mathbf{X}_2\boldsymbol{\alpha}+\mathbf{X}_3\boldsymbol{\gamma})\\
&=(\mu+\beta)\tilde{\mathbf{X}}_4'\mathbf{1}+\tilde{\mathbf{X}}_4'\mathbf{X}_1\boldsymbol{\rho}+\tilde{\mathbf{X}}_4'\mathbf{X}_2\boldsymbol{\alpha}+\tilde{\mathbf{X}}_4'\mathbf{X}_3\boldsymbol{\gamma}-\frac{\mu+\beta}{s}\mathbf{1}_{s-1}\mathbf{1}'_{bk}\mathbf{X}_3'\mathbf{1}\\
&\qquad\qquad-\frac{1}{s}\mathbf{1}_{s-1}\mathbf{1}'_{bk}\mathbf{X}_3'\mathbf{X}_1\boldsymbol{\rho}-\frac{1}{s}\mathbf{1}_{s-1}\mathbf{1}'_{bk}\mathbf{X}_3'\mathbf{X}_2\boldsymbol{\alpha}-\frac{1}{s}\mathbf{1}_{s-1}\mathbf{1}'_{bk}\mathbf{X}_3'\mathbf{X}_3\boldsymbol{\gamma}\\
&=(\mu+\beta)bk\mathbf{1}_{s-1}+k\mathbf{1}_{s-1}\mathbf{1}'_b\boldsymbol{\rho}+\mathbf{r}'\otimes\mathbf{1}_{s-1}\boldsymbol{\alpha}+\mathbf{1}_{s-1}\mathbf{1}'_{bk}\boldsymbol{\gamma}-\\
&\qquad\frac{\mu+\beta}{s}\mathbf{1}_{s-1}\mathbf{1}'_{bk}s\mathbf{1}_{bk}-\frac{1}{s}\mathbf{1}_{s-1}\mathbf{1}'_{bk}s(\mathbf{1}_k\otimes\mathbf{I}_b)\boldsymbol{\rho}-\frac{1}{s}\mathbf{1}_{s-1}\mathbf{1}'_{bk}\mathbf{N}_2\boldsymbol{\alpha}-\frac{1}{s}\mathbf{1}_{s-1}\mathbf{1}'_{bk}s\mathbf{I}_{bk}\boldsymbol{\gamma}
\end{aligned}
$$

$$= \mathbf{r}' \otimes \mathbf{1}_{s-1}\boldsymbol{\alpha} - \frac{1}{s}\mathbf{1}_{s-1}s\mathbf{r}'\boldsymbol{\alpha}$$
$$= \mathbf{0}$$

As a result, the non-centrality parameter is zero. Thus, $\mathbf{Q}_S'\mathbf{C}_S^{-1}\mathbf{Q}_S/\sigma^2 \sim \chi^2_{s-1}$ under H_0. Here, the degrees of freedom is equal to the rank of the matrix of the quadratic form $(\tilde{\mathbf{X}}_4 - \frac{1}{s}\mathbf{X}_3\mathbf{X}_3'\tilde{\mathbf{X}}_4)\mathbf{C}_s^{-1}(\tilde{\mathbf{X}}_4 - \frac{1}{s}\mathbf{X}_3\mathbf{X}_3'\tilde{\mathbf{X}}_4)'$ and the rank of this matrix is clearly $s-1$.

To check independence of $\mathbf{Q}_S'\mathbf{C}_S^{-1}\mathbf{Q}_S$ and SSE, we know that $SSE = \mathbf{y}'(\mathbf{I} - \mathbf{X}\mathbf{G}\mathbf{X}')\mathbf{y}$ where $\mathbf{G}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$. Now,

$$\begin{aligned}
&(\mathbf{I} - \mathbf{X}\mathbf{G}\mathbf{X}')(\tilde{\mathbf{X}}_4 - \frac{1}{s}\mathbf{X}_3\mathbf{X}_3'\tilde{\mathbf{X}}_4)\mathbf{C}_s^{-1}(\tilde{\mathbf{X}}_4 - \frac{1}{s}\mathbf{X}_3\mathbf{X}_3'\tilde{\mathbf{X}}_4)' \\
&= (\tilde{\mathbf{X}}_4 - \frac{1}{s}\mathbf{X}_3\mathbf{X}_3'\tilde{\mathbf{X}}_4 - \mathbf{X}\mathbf{G}\mathbf{X}'\tilde{\mathbf{X}}_4 + \frac{1}{s}\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X}_3\mathbf{X}_3'\tilde{\mathbf{X}}_4)\mathbf{C}_s^{-1}(\tilde{\mathbf{X}}_4 - \frac{1}{k}\mathbf{X}_3\mathbf{X}_3'\tilde{\mathbf{X}}_4)' \\
&= \mathbf{0}
\end{aligned}$$

because $\mathbf{X}\mathbf{G}\mathbf{X}'\tilde{\mathbf{X}}_4 = \tilde{\mathbf{X}}_4$ and $\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X}_3\mathbf{X}_3'\tilde{\mathbf{X}}_4 = \mathbf{X}_3\mathbf{X}_3'\tilde{\mathbf{X}}_4$ due to properties of generalized inverse matrix $\mathbf{G}$. Hence, two quadratic forms $\mathbf{Q}_S'\mathbf{C}_S^{-1}\mathbf{Q}_S$ and SSE are independent. Hence, F_2 under null hypothesis follows F-distribution with $(s-1)$ and $(bks - bk - ms + m)$ degrees of freedom. □ Proof of $F_3 \sim F_{(m-1),(bk-b-m+1)}$: where

$$SSW = R(\boldsymbol{\gamma}|\mu,\boldsymbol{\rho},\boldsymbol{\alpha}) = R(\mu,\boldsymbol{\rho},\boldsymbol{\alpha},\boldsymbol{\gamma}) - R(\mu,\boldsymbol{\rho},\boldsymbol{\alpha}) = \tfrac{1}{s}\mathbf{y}_{BM.}'\mathbf{y}_{BM.} - \tfrac{1}{ks}\mathbf{y}_{B..}'\mathbf{y}_{B..} - \mathbf{Q}_M'\mathbf{C}_M^{-1}\mathbf{Q}_M.$$

First we prove that $\mathbf{Q}_M'\mathbf{C}_M^{-1}\mathbf{Q}_M/\sigma^2 \sim \chi^2_{m-1}$ under null hypothesis.

$$\begin{aligned}
\mathbf{Q}_M'\mathbf{C}_M^{-1}\mathbf{Q}_M &= (\tilde{\mathbf{y}}_{.M.} - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{y}_{B..})'\mathbf{C}_M^{-1}(\tilde{\mathbf{y}}_{.M.} - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{y}_{B..}) \\
&= (\tilde{\mathbf{X}}_2'\mathbf{y} - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1'\mathbf{y})'\mathbf{C}_M^{-1}(\tilde{\mathbf{X}}_2'\mathbf{y} - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1'\mathbf{y}) \\
&= \mathbf{y}'(\tilde{\mathbf{X}}_2 - \frac{1}{k}\mathbf{X}_1\tilde{\mathbf{N}}_1')\mathbf{C}_M^{-1}(\tilde{\mathbf{X}}_2' - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1')\mathbf{y} \\
&= \mathbf{y}'\mathbf{A}\mathbf{y}
\end{aligned}$$

where $\mathbf{A} = (\tilde{\mathbf{X}}_2 - \frac{1}{k}\mathbf{X}_1\tilde{\mathbf{N}}_1')\mathbf{C}_M^{-1}(\tilde{\mathbf{X}}_2' - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1')$. Now,

$$\mathbf{A}\mathbf{A} = (\tilde{\mathbf{X}}_2 - \frac{1}{k}\mathbf{X}_1\tilde{\mathbf{N}}_1')\mathbf{C}_M^{-1}(\tilde{\mathbf{X}}_2' - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1')(\tilde{\mathbf{X}}_2 - \frac{1}{k}\mathbf{X}_1\tilde{\mathbf{N}}_1')\mathbf{C}_M^{-1}(\tilde{\mathbf{X}}_2' - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1').$$

It may be seen that

$$\begin{aligned}
(\tilde{\mathbf{X}}_2' - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1')(\tilde{\mathbf{X}}_2 - \frac{1}{k}\mathbf{X}_1\tilde{\mathbf{N}}_1') &= \tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2 - \frac{1}{k}\tilde{\mathbf{X}}_2'\mathbf{X}_1\tilde{\mathbf{N}}_1' - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1'\tilde{\mathbf{X}}_2 + \frac{1}{k^2}\tilde{\mathbf{N}}_1\mathbf{X}_1'\mathbf{X}_1\tilde{\mathbf{N}}_1' \\
&= \tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2 - \frac{s}{k}\tilde{\mathbf{N}}_1\tilde{\mathbf{N}}_1' - \frac{s}{k}\tilde{\mathbf{N}}_1\tilde{\mathbf{N}}_1' + \frac{ks}{k^2}\tilde{\mathbf{N}}_1\tilde{\mathbf{N}}_1' \\
&= s\mathbf{R}_{m-1} - \frac{s}{k}\tilde{\mathbf{N}}_1\tilde{\mathbf{N}}_1' \\
&= \mathbf{C}_M.
\end{aligned}$$

As a result,

$$\begin{aligned}\mathbf{AA} &= (\tilde{\mathbf{X}}_2 - \frac{1}{k}\mathbf{X}_1\tilde{\mathbf{N}}_1')\mathbf{C}_M^{-1}\mathbf{C}_M\mathbf{C}_M^{-1}(\tilde{\mathbf{X}}_2' - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1')\\ &= (\tilde{\mathbf{X}}_2 - \frac{1}{k}\mathbf{X}_1\tilde{\mathbf{N}}_1')\mathbf{C}_M^{-1}(\tilde{\mathbf{X}}_2' - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1')\\ &= \mathbf{A}\end{aligned}$$

Hence, the matrix of the quadratic form of $\mathbf{Q}_M'\mathbf{C}_M^{-1}\mathbf{Q}_M/\sigma^2$ is idempotent. Thus, $\mathbf{Q}_M'\mathbf{C}_M^{-1}\mathbf{Q}_M/\sigma^2 \sim \chi^2_{m-1}$ with non-centrality parameter $\frac{1}{2\sigma^2}\boldsymbol{\theta}\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\theta}$ where $\mathbf{X}\boldsymbol{\theta} = \mu\mathbf{1} + \mathbf{X}\boldsymbol{\rho} + \mathbf{X}_2\boldsymbol{\alpha}$ since the rank of the matrix $\mathbf{A}$ is $m-1$. Under null-hypothesis, $\boldsymbol{\alpha} = \alpha\mathbf{1}_m$. So the non-centrality parameter can be shown to be zero as

$$\begin{aligned}\mathbf{AX}\boldsymbol{\theta} &= (\tilde{\mathbf{X}}_2 - \frac{1}{k}\mathbf{X}_1\tilde{\mathbf{N}}_1')\mathbf{C}_M^{-1}(\tilde{\mathbf{X}}_2' - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1')(\mu\mathbf{1} + \mathbf{X}_1\boldsymbol{\rho} + \alpha\mathbf{X}_2\mathbf{1}_m)\\ &= (\tilde{\mathbf{X}}_2 - \frac{1}{k}\mathbf{X}_1\tilde{\mathbf{N}}_1')\mathbf{C}_M^{-1}(\tilde{\mathbf{X}}_2' - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1')\{(\mu+\alpha)\mathbf{1} + \mathbf{X}_1\boldsymbol{\rho}\}\\ &= (\tilde{\mathbf{X}}_2 - \frac{1}{k}\mathbf{X}_1\tilde{\mathbf{N}}_1')\mathbf{C}_M^{-1}\left\{(\mu+\alpha)\tilde{\mathbf{X}}_2'\mathbf{1} + \tilde{\mathbf{X}}_2\mathbf{X}_1\boldsymbol{\rho} - \frac{\mu+\alpha}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1'\mathbf{1} - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1'\mathbf{X}_1\boldsymbol{\rho}\right\}\\ &= (\tilde{\mathbf{X}}_2 - \frac{1}{k}\mathbf{X}_1\tilde{\mathbf{N}}_1')\mathbf{C}_M^{-1}\left\{(\mu+\alpha)s\mathbf{r}_{m-1} + s\tilde{\mathbf{N}}_1\boldsymbol{\rho} - \frac{\mu+\alpha}{k}\tilde{\mathbf{N}}_1 ks\mathbf{1}_b - \frac{1}{k}\tilde{\mathbf{N}}_1 ks\mathbf{I}_b\boldsymbol{\rho}\right\}\\ &= (\tilde{\mathbf{X}}_2 - \frac{1}{k}\mathbf{X}_1\tilde{\mathbf{N}}_1')\mathbf{C}_M^{-1}\left\{(\mu+\alpha)s\mathbf{r}_{m-1} + s\tilde{\mathbf{N}}_1\boldsymbol{\rho} - (\mu+\alpha)s\mathbf{r}_{m-1} - s\tilde{\mathbf{N}}_1\boldsymbol{\rho}\right\}\\ &= \mathbf{0}.\end{aligned}$$

Thus, $\mathbf{Q}_M'\mathbf{C}_M^{-1}\mathbf{Q}_M/\sigma^2 \sim \chi^2_{m-1}$.

Now note that

$$\begin{aligned}SSW &= \frac{1}{s}\mathbf{y}_{BM.}'\mathbf{y}_{BM.} - \frac{1}{ks}\mathbf{y}_{B..}'\mathbf{y}_{B..} - \mathbf{Q}_M'\mathbf{C}_M^{-1}\mathbf{Q}_M\\ &= \frac{1}{s}\mathbf{y}'\mathbf{X}_3\mathbf{X}_3'\mathbf{y} - \frac{1}{ks}\mathbf{y}'\mathbf{X}_1\mathbf{X}_1'\mathbf{y} - \mathbf{y}'\mathbf{A}\mathbf{y}\\ &= \mathbf{y}'\mathbf{B}\mathbf{y}\end{aligned}$$

where $\mathbf{B} = \frac{1}{s}\mathbf{X}_3\mathbf{X}_3' - \frac{1}{ks}\mathbf{X}_1\mathbf{X}_1' - \mathbf{A}$.

To check independence of $\mathbf{Q}_M'\mathbf{C}_M^{-1}\mathbf{Q}_M$ and SSW, we need to prove that $\mathbf{AVB} = \mathbf{0}$ where $\mathbf{A}, \mathbf{B}$ are as defined above and here $\mathbf{V} = \sigma^2\mathbf{I}$. So it suffices to show that $\mathbf{AB} = \mathbf{0}$. Now,

$$\mathbf{AB} = (\tilde{\mathbf{X}}_2 - \frac{1}{k}\mathbf{X}_1\tilde{\mathbf{N}}_1')\mathbf{C}_M^{-1}(\tilde{\mathbf{X}}_2' - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1')(\frac{1}{s}\mathbf{X}_3\mathbf{X}_3' - \frac{1}{ks}\mathbf{X}_1\mathbf{X}_1' - \mathbf{A}).$$

The last two terms of $\mathbf{AB}$ may be simplified as

$$\begin{aligned}&(\tilde{\mathbf{X}}_2' - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1')(\frac{1}{s}\mathbf{X}_3\mathbf{X}_3' - \frac{1}{ks}\mathbf{X}_1\mathbf{X}_1' - \mathbf{A})\\ &= \frac{1}{s}\tilde{\mathbf{X}}_2'\mathbf{X}_3\mathbf{X}_3' - \frac{1}{ks}\tilde{\mathbf{X}}_2'\mathbf{X}_1\mathbf{X}_1' - \tilde{\mathbf{X}}_2'\mathbf{A} - \frac{1}{ks}\tilde{\mathbf{N}}_1\mathbf{X}_1'\mathbf{X}_3\mathbf{X}_3' + \frac{1}{k^2s}\tilde{\mathbf{N}}_1\mathbf{X}_1'\mathbf{X}_1\mathbf{X}_1' + \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1'\mathbf{A} \quad (9)\\ &= \frac{1}{s}\tilde{\mathbf{X}}_2'\mathbf{X}_3\mathbf{X}_3' - \frac{1}{ks}s\tilde{\mathbf{N}}_1\mathbf{X}_1' - \tilde{\mathbf{X}}_2'\mathbf{A} - \frac{1}{ks}\tilde{\mathbf{N}}_1\mathbf{X}_1'\mathbf{X}_3\mathbf{X}_3' + \frac{1}{k^2s}ks\tilde{\mathbf{N}}_1\mathbf{X}_1' + \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1'\mathbf{A}.\end{aligned}$$

It may be checked that

$$\begin{aligned}\tilde{\mathbf{X}}_2'\mathbf{A} &= (\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2 - \frac{1}{k}\tilde{\mathbf{X}}_2'\mathbf{X}_1\tilde{\mathbf{N}}_1')\mathbf{C}_M^{-1}(\tilde{\mathbf{X}}_2' - \frac{1}{k}\tilde{\mathbf{N}}_1'\mathbf{X}_1') \\ &= (s\mathbf{R}_{m-1} - \frac{1}{k}s\tilde{\mathbf{N}}_1\tilde{\mathbf{N}}_1')\mathbf{C}_M^{-1}(\tilde{\mathbf{X}}_2' - \frac{1}{k}\tilde{\mathbf{N}}_1'\mathbf{X}_1') \\ &= \tilde{\mathbf{X}}_2' - \frac{1}{k}\tilde{\mathbf{N}}_1'\mathbf{X}_1'.\end{aligned}$$

Also,

$$\begin{aligned}\frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1'\mathbf{A} &= \frac{1}{k}\tilde{\mathbf{N}}_1(\mathbf{X}_1'\tilde{\mathbf{X}}_2 - \frac{1}{k}\mathbf{X}_1'\mathbf{X}_1\tilde{\mathbf{N}}_1')\mathbf{C}_M^{-1}(\tilde{\mathbf{X}}_2' - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1') \\ &= \frac{1}{k}\tilde{\mathbf{N}}_1(s\tilde{\mathbf{N}}_1' - \frac{1}{k}ks\tilde{\mathbf{N}}_1')\mathbf{C}_M^{-1}(\tilde{\mathbf{X}}_2' - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1') \\ &= \mathbf{0}.\end{aligned}$$

Hence, equation (9) can be simplified as

$$\begin{aligned}&(\tilde{\mathbf{X}}_2' - \frac{1}{k}\tilde{\mathbf{N}}_1\mathbf{X}_1')(\frac{1}{s}\mathbf{X}_3\mathbf{X}_3' - \frac{1}{ks}\mathbf{X}_1\mathbf{X}_1' - \mathbf{A}) \\ &= \frac{1}{s}\tilde{\mathbf{X}}_2'\mathbf{X}_3\mathbf{X}_3' - \tilde{\mathbf{X}}_2' + \frac{1}{k}\tilde{\mathbf{N}}_1'\mathbf{X}_1' - \frac{1}{ks}\tilde{\mathbf{N}}_1'\mathbf{X}_1'\mathbf{X}_3\mathbf{X}_3' \\ &= \tilde{\mathbf{X}}_2' - \tilde{\mathbf{X}}_2' + \frac{1}{k}\tilde{\mathbf{N}}_1'\mathbf{X}_1' - \frac{1}{k}\tilde{\mathbf{N}}_1'\mathbf{X}_1' \\ &= \mathbf{0}.\end{aligned}$$

Thus, two quadratic forms $\mathbf{Q}_M'\mathbf{C}_M^{-1}\mathbf{Q}_M$ and SSW are independent and hence, under null hypothesis, the test statistic F_3 follows F-distribution with $(m-1)$ and $(bk-b-m+1)$ degrees of freedom. Null hypothesis should be rejected whenever calculated value of F_3 exceeds $F_{\alpha,(m-1),(bk-b-m+1)}$ where $F_{\alpha,(m-1),(bk-b-m+1)}$ denotes the upper α percent point of an F-distribution with $(m-1)$ and $(bk-b-m+1)$ degrees of freedom.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 259–283
http://www.ssca.org.in/journal

Meta Analysis for Rare Events

Dulal K. Bhaumik[1,2], **Anup K. Amatya**[1,2] **and Soumya Sahu**[1]
[1]*Division of Epidemiology and Biostatistics, University of Illinois Chicago*
[2]*Department of Psychiatry, University of Illinois Chicago*

Received: 30 April 2024; Revised: 09 July 2024; Accepted: 28 August 2024

Abstract

Meta-analysis has become a widely used tool for evaluating the efficacy and safety of medical interventions, offering numerous advantages and utilities. However, recent studies have raised questions about the accuracy of commonly used moment-based meta-analytic methods, particularly for rare binary outcomes. This issue is further complicated in studies with heterogeneous effect sizes. Likelihood-based mixed-effects modeling provides an alternative to moment-based methods, such as inverse-variance weighted fixed- and random-effects estimators. In this review paper, we discuss several meta-analysis methods specifically designed for analyzing rare event data. We elaborate on the use of continuity correction for studies with zero total events, taking into account study heterogeneity. The problem is motivated, and results are illustrated using a well-known meta-analysis study. By exploring and comparing these different methodologies, researchers can gain insights into the most appropriate approaches for analyzing rare event data in meta-analytic studies.

Key words: Conditional likelihood; Mantel-Haenszel method; The Peto method; Confidence distribution methods; Odds ratio.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Meta-analysis is a powerful statistical tool used to combine results from multiple studies, particularly useful for making robust inferences about rare events, which require large sample sizes due to their low frequency (less than 0.1%). Traditional clinical trials often lack sufficient power to draw sound conclusions about rare adverse events, such as those associated with pharmaceutical agents. The challenge lies in incorporating studies with few or no observed adverse events into the analysis. While fixed-effect and random-effect meta-analyses are common, Bayesian methodologies and confidence distribution approaches offer alternatives. Each method has unique strengths and weaknesses, and the optimal approach for analyzing rare events remains a topic of ongoing research. We try to clarify the idea that rare event meta analysis may end up with some studies with zero total events. However, those studies with zero events are also informative and should be included in the analysis

Corresponding Author: Dulal K. Bhaumik
Email: dbhaumik@uic.edu

and this demands the concept of continuity correction. Data analysis of meta analysis of rare events is developed addressing these concerns and difficulty of zero total events depending on what type methodology is used.

Meta-analysis is a convenient statistical tool that combines results from multiple trials and makes a robust inference regarding the parameter of our interest. To make a valid statistical inference for a rare event requires a trial with a large sample size. Regular clinical trials are not sufficiently powered to draw a statistically sound conclusion regarding events that often occur at a rate of less than 0.1%. Low-frequency events are commonly encountered in the investigation of adverse events (*e.g.*, suicide) associated with a pharmaceutical agent (*e.g.*, antidepressants). A further complication arises in rare adverse event studies because clinical trials are typically designed to assess the efficacy rather than the safety of the product. As a result, the chance of observing a reasonable number of such adverse events in a single trial or study is relatively low. Quite often, in such situations, not even a single adverse event is observed in efficacy trials. Utilizing such studies meaningfully in the analysis is the greatest challenge in the meta-analysis of rare events. Several strategies have been proposed to make a valid decision regarding the parameters of our interest, incorporating all available studies. However, none of those is universally accepted, and as a result, the issue is still open. Traditional methods of meta-analysis either treat the underlying treatment effect as a fixed parameter across multiple studies or assume individual study treatment effect as a random sample from a hypothetical pool of treatment effects. The first form of meta-analysis is called fixed-effect meta-analysis, and the second form is called random-effect meta-analysis. Bayesian methodologies are also used to allow hierarchical modeling with a greater opportunity for sensitivity analysis. Recently, the third method, based on the concept of confidence distribution, has been put forth as an attractive alternative for meta-analysis of rare events. For each methodology, there are several estimation techniques with respective strengths and weaknesses. This article discusses some characteristics of rare event studies and provides an overview of meta-analytic methods suitable for the analysis of rare events, along with the issues pertaining to those methods.

2. Zero total event studies

A study in which no outcome event is observed is called a zero-total event study. Studies where outcome events are observed in one arm but not in the other arm are called zero-cell studies. Zero total event studies in rare event analysis are contentious due to their lack of events in one or both treatment arms, but recent literature suggests they should not be excluded, despite challenges in variance computation and continuity correction. The ubiquitous characteristic of rare event studies is the absence of events in either one or both treatment arms. The answers have been contentious, and inconclusive. The core of the issue is the argument that the zero total event studies do not contribute any information towards the estimation of the effect and, hence, are irrelevant and should be removed from the analysis Whitehead and Whitehead (1991); Sweeting *et al.* (2004). However, in general, a zero total event study with a large sample size is expected to provide stronger evidence for any hypothesized effect compared to a smaller sample size zero total study Friedrich *et al.* (2007); Liu (2012); Kuss (2014). Furthermore, recent publications are providing theoretical support to the relevance of the zero total event studies Liu *et al.* (2014); Xie *et al.* (2014). Therefore, zero total event studies should not be excluded. The major obstacle to the inclusion of

zero total event studies in a traditional meta-analysis is the numerical ill-conditioning for the computation of variance of the effect size using ratio measures. Analysts are addressing this problem by proposing the concept of Continuity Correction, even though there is no consensus on what exact value we should use for the continuity correction. Additional complexity arises when a significant heterogeneity exists among studies.

Example: Zero total event studies in meta-analysis of rositaglitazone and risk of cardiovascular events

On November 14, the Food and Drug Administration (FDA) put a "black box warning" on Rositaglitazone's product to inform consumers of such risks. On September 23, 2010, the FDA limited access to Rositaglitazone because of concerns about increased cardiovascular risk. The most prominent study that led to the action by the FDA was the meta-analysis conducted by Nissen and Wolski (2007). As part of the analysis, 42 trials were selected from the published literature, the FDA website, and a clinical trial registry maintained by the drug manufacturer (GlaxoSmithKline). Table 1 reports the myocardial infarction (MI) events and deaths from cardiovascular causes (CVD) that were reported in the 42 clinical trials included in the study. Of those 42 studies, four studies (9.5%; study 20, 31, 33, and 38) are zero total event studies for MI endpoint, and 19 studies (45%; study 2–4, 6, 7, 9, 10, 12, 14, 17, 21–24, 29, 31, 36–38) are zero total event studies for CVD endpoint. Overall, there were 86 MIs and 39 CVDs in the rosiglitazone group and 72 MIs and 39 CVDs in the control group.

2.1. Conditional likelihood

This section explains the practical reasons for not favoring the zero total event studies, mainly because of computational difficulty under the set up of conditional likelihood. The most compelling argument for supporting the exclusion of zero total event studies comes from the conditional likelihood inference perspective. The conditional maximum likelihood estimation procedure estimates the parameter of interest by maximizing conditional likelihood given the minimal sufficient statistics for the nuisance parameters. Consider a sequence of observations $\{x_{t1}, x_{t2}, \ldots, x_{tk}\}$, and $\{x_{c1}, x_{c2}, \ldots, x_{ck}\}$ from k studies/trials with $\{n_{t1}, n_{t2}, \ldots, n_{tk}\}$, and $\{n_{c1}, n_{c2}, \ldots, n_{ck}\}$ treatment, and control group sample sizes respectively. Observations from an individual study form the following 2×2 table given in Table 2.

For a fixed observed event total t_i, only random variable in the ith table is X_i ($i = 1, 2, \ldots, k$) (count in the upper left cell), which follows a hyper-geometric distribution. The corresponding conditional likelihood function given $T_i = t_i$ is as follows:

$$L_{x_{ti}|t_i}(\theta) = P_\theta(X_{ti} = x_{ti}|T_i = t_i) = \frac{\binom{n_{ti}}{x_{ti}}\binom{n_{ci}}{t_i - x_{ti}}\psi^{x_{ti}}}{\sum_{\nu=u_i}^{l_i}\binom{n_{ti}}{\nu}\binom{n_{ci}}{t_i-\nu}\psi^{\nu}}, \tag{1}$$

and the joint conditional likelihood function is given by the following expression:

$$\phi(x_{t1}, x_{t2}, \ldots, x_{tk}|t_1, t_2, \ldots, t_k) = \prod L_{x_{ti}|t_i}(\theta), \tag{2}$$

Table 1: Example data: Rositaglitazone and the risk of cardiovascular events Nissen and Wolski (2007)

study	Rosiglitazone			Control		
	Total	MI	CVD	Total	MI	CVD
1	357	2	1	176	0	0
2	391	2	0	207	1	0
3	774	1	0	185	1	0
4	213	0	0	109	1	0
5	232	1	1	116	0	0
6	43	0	0	47	1	0
7	121	1	0	124	0	0
8	110	5	3	114	2	2
9	382	1	0	384	0	0
10	284	1	0	135	0	0
11	294	0	2	302	1	1
12	563	2	0	142	0	0
13	278	2	0	279	1	1
14	418	2	0	212	0	0
15	395	2	2	198	1	0
16	203	1	1	106	1	1
17	104	1	0	99	2	0
18	212	2	1	107	0	0
19	138	3	1	139	1	0
20	196	0	1	96	0	0
21	122	0	0	120	1	0
22	175	0	0	173	1	0
23	56	1	0	58	0	0
24	39	1	0	38	0	0
25	561	0	1	276	2	0
26	116	2	2	111	3	1
27	148	1	2	143	0	0
28	231	1	1	242	0	0
29	89	1	0	88	0	0
30	168	1	1	172	0	0
31	116	0	0	61	0	0
32	1172	1	1	377	0	0
33	706	0	1	325	0	0
34	204	1	0	185	2	1
35	288	1	1	280	0	0
36	254	1	0	272	0	0
37	314	1	0	154	0	0
38	162	0	0	160	0	0
39	442	1	1	112	0	0
40	394	1	1	124	0	0
41	2635	15	12	2634	9	10
42	1456	27	2	2895	41	5

Table 2: 2×2 **contingency table for the** ith **trial/study**

	Event		
	yes	no	total
Treatment	x_{ti}	$n_{ti} - x_{ti}$	n_{ti}
Control	x_{ci}	$n_{ci} - x_{ci}$	n_{ci}
Total	t_i	$(n_{ti} + n_{ci}) - t_i$	$n_{ti} + n_{ci}$

for $l_i \leq x_{ti} \leq u_i$, where $l_i = max(0, t_i - n_{ci})$, $u_i = min(n_{ti}, t_i)$, and $\psi = \exp\theta$ is the odds ratio which is assumed to be the same across k studies involved in the analysis. The value of θ that maximizes this conditional likelihood is called the conditional maximum likelihood estimate (CMLE) for ψ. Note that $L_i\psi(x_{ti}|t_i) = 1$, for zero total event studies, does not directly contribute to the joint conditional likelihood. An asymptotic property of CLME can be proved under a reasonable set of conditions. However, unlike the direct maximum likelihood estimate, the CMLE, in general, is not efficient except for some special (but important) situations, where the asymptotic variance attends the Cramer-Rao lower bound (Andersen, 1970, see). Unfortunately, the CLME obtained from (2) is not derived from one of those special situations and is not an efficient estimator of ψ. Thus, the most reasonable basis for the exclusion of zero total event studies is based on a procedure that maintains asymptotic properties and does not use all information contained in the data for the parameter of interest. Furthermore, Xie *et al.* (2014) has shown conclusively that the zero total event studies do contain information that is a function of ψ, π_{ci}, and sample sizes n_i.

The basic idea behind the derivation outlined by Xie *et al.* (2014) is as follows. Suppose that X_{ti} and X_{ci} are independent binomial random variables following $B(\pi_{ti}, n_{ti})$, and $B(\pi_{ci}, n_{ci})$, respectively. The full (unconditional) likelihood function is given as:

$$L_{\boldsymbol{x_t},\boldsymbol{x_c}}(\theta, \boldsymbol{\pi_c}) = L_{\boldsymbol{x_t},\boldsymbol{x_c}}(\boldsymbol{\pi_t}, \boldsymbol{\pi_c}) = \prod_{i=1}^{k} \binom{n_{ti}}{x_{ti}} \binom{n_{ci}}{x_{ci}} \pi_{ti}^{x_{ti}} (1-\pi_{ti})^{n_{ti}-x_{ti}} \pi_{ci}^{x_{ci}} (1-\pi_{ci})^{n_{ci}-x_{ci}}. \quad (3)$$

Under the assumption that the odds ratio is the same across k studies, π_{ti} and π_{ci} satisfy a constraint $\{\pi_{ti}/(1-\pi_{ti})\}/\{\pi_{ci}/(1-\pi_{ci})\} = e^{\theta}$. From the likelihood principle, it follows that the above likelihood function contains all information relevant for making an inference for the parameter of interest. The full likelihood (3) can be rewritten as

$$L_{\boldsymbol{x_t},\boldsymbol{x_c}}(\theta, \boldsymbol{\pi_c}) = L_{\boldsymbol{x_t}|\boldsymbol{t}}(\theta) D_t(\theta, \boldsymbol{\pi_c}), \quad (4)$$

where $D_t(\theta, \boldsymbol{\pi_c}) = \dfrac{L(\theta, \boldsymbol{\pi_c})}{L_{x|t}(\theta)}$. Xie *et al.* (2014) showed that $D_t(\theta, \boldsymbol{\pi_c})$ is a function of both θ and $\boldsymbol{\pi_c}$. Therefore, they argued that the conditional likelihood inference and full likelihood inference are different, suggesting that "the conditional likelihood approach can incur omission or distortion of information". Clearly, the zero total event studies contribute $\prod_{\{i:t_i=0\}} (1-\pi_{ti})^{n_{ti}} (1-\pi_{ci})^{n_{ci}}$ portion of information to the full likelihood. But that portion of information, which is also a function of both θ and $\boldsymbol{\pi_c}$, is omitted from conditional likelihood. As a result, inferences under conditional likelihood that effectively omit zero total event studies will be weaker and less reliable.

The conditional likelihood (1) is developed under a specific assumption that t_i's are fixed in addition to the same assumption on n_{ti} and n_{ci} for each study. However, in general, studies or trials that are included in meta-analysis do not have control over observed total events. Consequently, the hypothesis testing under the assumption of fixed t_i is conservative and loses power when only the n_{ti} and n_{ci} are fixed. Thus, Xie *et al.* (2014) concluded that zero total event studies do contain information on the intervention effect.

As described above, arguments for and against of excluding zero total event studies are generally put forth by assuming a common odds ratio across studies. However, in contrast to the conservative findings under such assumptions, simulation studies have suggested that methods that exclude zero total event studies can have an inflated type I error rate when odds ratios vary between studies Bhaumik *et al.* (2012). Furthermore, popular methods used in practice have a tendency to overestimate the true odds ratio and underestimate the between study heterogeneity. This also indicates that the zero total event studies do contain relevant information on the parameters of our interest. In what follows, we discuss how to include the zero total event studies in a meaningful way in meta-analysis.

3. Moment matching methods

In this section we discuss some frequently used meta analysis methods based on weighted average estimates with the continuity correction when applied in sparse data. Traditional meta-analysis methods are perhaps the most useful methods that are used in practice. Those are derived based on the moment-matching approach. These methods include various forms of weighted average estimates of the overall intervention effect. The inverse variance weighted method, Mantel-Haenszel method, and Peto method are the three most widely used methods under this category. These methods typically require some form of adjustment when applied to sparse data. Although intended for different purposes in the context of a chi-square test, such adjustment made in individual cells of 2×2 tables in meta-analysis is known as the continuity correction.

3.1. Continuity correction

The controversy over continuity correction in meta-analysis of rare event studies persists, with alternative correction factors proposed to mitigate bias and coverage issues, while recent developments suggest methods avoiding continuity correction altogether could be possible. As mentioned before, continuity correction is a controversial topic. There are competing views on the appropriateness of the use of continuity correction in meta-analysis. In the context of traditional analysis, there is no other choice but to discard zero total event studies or to use a Bayesian approach without any continuity correction. The value that has received the most attention for the continuity correction is 1/2. It was accepted as the value for continuity correction on the basis of the argument put forth in Cox (1970). According to Cox, when using the odds as the effect measure, choosing a correction factor of 1/2 gives the least biased estimator of the true log odds in a single treatment group situation. The factor 1/2 is also used to improve the approximation of a discrete distribution by a continuous distribution (i.e. 1-degrees of freedom chi-square), or to obtain an approximation to the product hypergeometric probability. However, adding a constant continuity correction such as 1/2 can create some undesirable problems, including reversal of the effect direction, particularly if the treatment arms are unbalanced Rücker *et al.* (2009). An investigation by

Sweeting *et al.* (2004) concluded that using the continuity correction of 1/2 may be outperformed in terms of both bias and coverage by other choices of correction factor when studying the odds ratio between two groups. An important point to be noted here is that the study was conducted under the assumption of fixed intervention effect across studies, and excluding zero total event studies. They noted that the application of their alternative continuity correction factor might not be applicable when using random-effect models. Two alternative correction strategies that Sweeting *et al.* (2004) showed to be outperforming, under fixed-effect assumption, are (1) to add a factor of the reciprocal of the size of the opposite treatment arm to the cells, and (2) to use empirical continuity correction. However, they also cautioned that not a single correction factor or method is superior in all situations, and recommended to perform sensitivity analysis using several different correction factors. See Sweeting *et al.* (2004) for details on the aforementioned alternative approaches for continuity correction.

Recent efforts on methodological development and validation studies suggest that the issue of continuity correction can potentially be avoided altogether using those methods (in the frequentist domain) that do not require continuity correction. Furthermore, these methods allow the inclusion of all studies in meta-analysis. Nonetheless, the Mantel-Hanszel and Peto methods are widely used for meta-analysis of rare events. Therefore, these popular classical methods, along with a somewhat underutilized but highly relevant method using arcsine risk difference measure, are briefly described in the following sections.

3.2. Mantel-Haenszel method

The Mantel-Haenszel method for meta-analysis adjusts for potential confounding factors and uses weighted averages to estimate the combined odds ratio, with alternative continuity corrections recommended to mitigate bias and improve coverage. The Mantel-Haenszel method was originally developed for stratified analysis adjusting for the third potential confounding factor. The fixed-effect meta-analysis can be viewed as a stratified design where each individual study is treated as a stratum. Based on the Mantel-Haenazel method, the pooled odds ratio across all K studies is estimated using the following expression:

$$\widehat{OR}_{MH} = \frac{\sum_{i=1}^{K} x_{Ti}(n_{Ci} - x_{Ci})/N_i}{\sum_{i=1}^{K} x_{Ci}(n_{Ti} - x_{Ti})/N_i}. \tag{5}$$

Equation (5) can be rewritten as a weighted average estimate as follows:

$$\widehat{OR}_{MH} = \frac{\sum_{i=1}^{K} w_i \widehat{OR}_i}{\sum_{i=1}^{K} w_i}, \tag{6}$$

$$\text{where } \; w_i = \frac{x_{Ci}(n_{Ti} - x_{Ti})}{N_i}, \text{ and } \widehat{OR}_i = \frac{x_{Ti}(n_{Ci} - x_{Ci})}{x_{Ci}}(n_{Ti} - x_{Ti}) \tag{7}$$

It is clear from equation (5) that zero cell studies contribute to the estimation of a combined odds ratio. However, zero total event studies are implicitly excluded from the computation unless a continuity correction is added. The weights in equation (7) are not reciprocals of the variances of odds ratio estimates from individual studies. Therefore, the variance estimate

of the combined odds ratio is not as straightforward as in the inverse variance method. The Robins-Breslow-Greenland method is generally accepted as an easy-to-use variance estimator for $ln(\widehat{OR}_{MH})$. It has the following expression:

$$\widehat{Var}[ln(\widehat{OR}_{MH})] = \frac{S_3}{2S_1^2} + \frac{S_5}{2S_1S_2} + \frac{S_4}{2S_2^2}, \tag{8}$$

where $S_1 = \sum_{i=1}^{K} \frac{x_{Ti}(n_{Ci} - x_{Ci})}{N_i}$, $S_2 = \sum_{i=1}^{K} \frac{x_{Ci}(n_{Ti} - x_{Ti})}{N_i}$,
$S_3 = \sum_{i=1}^{K} \frac{x_{Ti}(n_{Ci} - x_{Ci})(x_{Ti} + n_{Ci} - x_{Ci})}{N_i^2}$, $S_4 = \sum_{i=1}^{K} \frac{x_{Ci}(n_{Ti} - x_{Ti})(x_{Ci} + n_{Ti} - x_{Ti})}{N_i^2}$,
and $S_5 = \sum_{i=1}^{K} \frac{x_{Ci}(n_{Ti} - x_{Ti})(x_{Ti} + n_{Ci} - x_{Ci}) + x_{Ti}(n_{Ci} - x_{Ci})(x_{Ci} + n_{Ti} - x_{Ti})}{N_i^2}$. A null hypothesis of equal odds in treatment and control subjects, i.e., $OR_{MH} = 1$, may be tested by the following χ^2-test:

$$X^2_{MH} = \left[\sum_{i=1}^{K} \frac{x_{Ti}(n_{Ci} - x_{Ci}) - x_{Ci}(n_{Ti} - x_{Ti})}{N_i}\right]^2. \tag{9}$$

The Mantel-Haenszel method with the continuity correction of 1/2 produces biased estimates and low coverage rates for event rates below 1 percent Bradburn *et al.* (2007). Therefore, under fixed-effect conditions, an alternative continuity correction is recommended instead of 1/2 to reduce bias and improve coverage characteristics of this estimator Sweeting *et al.* (2004).

3.3. The Peto method

The Peto method in meta-analysis of moderately rare events excludes zero total event studies automatically and estimates the pooled log odds ratio based on weighted differences from individual tables, with limitations in unbalanced data and close-to-1 odds ratios. The Peto method is popular for meta-analysis of moderately rare events. Similar to the Mantel-Haenszel method, this method does not require artificial continuity correction when events are not observed in one of the treatment arms. However, the zero total event studies are automatically given zero weight and effectively are excluded from the analysis. When marginal totals in Table 2 are fixed, the following two quantities are the mean and variance of hypergeometric distribution under the null hypothesis that the odds ratio is one.

$$E_i = \frac{(x_{Ti} + x_{Ci})n_{Ti}}{N_i}, \tag{10}$$

and

$$V_i = \frac{(x_{Ti} + x_{Ci})(N_i - x_{Ti} - x_{Ci})n_{Ti}n_{Ci}}{N_i^2(N_i - 1)}. \tag{11}$$

Based on E_i and V_i, the Peto estimate of pooled log odds ratio from K independent tables has the following expression:

$$ln(\widehat{OR})_{Peto} = \frac{\sum_{i=1}^{K}(x_{Ti} - E_i)}{\sum_{i-1}^{K} V_i}, \tag{12}$$

and

$$Var[ln(\widehat{OR})_{Peto}] = \frac{1}{\sum_{i-1}^{K} V_i}. \tag{13}$$

The Peto estimator of the combined odds ratio is not a consistent estimator and can provide severely biased results when applied to unbalanced data Greenland and Salvan (1990). The validity of the Peto estimator in a meta-analysis of rare event studies is limited to the analysis of reasonably balanced studies that have odds ratio close to 1.

3.4. Arcsine transformation

The arcsine transformation method is needed when the objective is to include all studies in the meta-analysis, including those with very rare events or zero total events, while stabilizing variance estimates to provide more accurate intervention effect estimates. Zero event in either or both arms of a given study/trial does not necessarily indicate that the true probability of an event is 0. On the contrary, it indicates that the event probability is very small, and the sample size in the study is not large enough to observe an event. The arcsine transform method estimates the combined effect by combining all studies including the zero total event studies. The arcsine difference (AS) measure of intervention effect for the ith study is defined as:

$$AS_i = arcsin\sqrt{p_{Ti}} - arcsin\sqrt{p_{Ci}}, \tag{14}$$

and its asymptotic variance given in equation (15) is finite and non-zero and depends only on the study sample size.

$$\sigma^2_{AS_i} = \frac{1}{4n_{Ti}} + \frac{1}{4n_{Ci}}. \tag{15}$$

Similar to the MH and Peto methods, the combined AS is a weighted mean of the individual AS_i's, where the $w_i = 1/\sigma^2_{AS_i}$ are the weights. Therefore,

$$\widehat{AS} = \frac{\sum_{i=1}^{K} w_i AS_i}{\sum_{i=1}^{K} w_i}. \tag{16}$$

Rücker *et al.* (2009) has recommended using $0.42/n$ instead of $1/4n$ in equation (15) to estimate the variance conservatively for small event probabilities. Simulation studies of Rücker *et al.* (2009) suggest that the bias of the estimate is slightly higher than the other two methods mentioned above. The key advantage of this method is the variance stabilizing property of the arcsine transformation, which leads to more robust estimation, even for the rare events Rücker *et al.* (2009). Nevertheless, a lack of direct interpretation has limited its wider use as a measure of intervention effect.

3.5. Heterogeneity

Fixed-effect methods like Mantel-Haenszel and Peto can significantly overestimate treatment effects in meta-analysis of rare events, especially in the presence of heterogeneity, leading to inflated type I error rates and biases. Historically, the treatment effect heterogeneity has not received sufficient attention in the context of meta-analysis of rare events. Using a continuity correction, majority of simulation studies are performed under the assumption of fixed treatment effect Sweeting *et al.* (2004); Bradburn *et al.* (2007), or a small heterogeneity Rücker *et al.* (2009). The rationale behind selecting the fixed treatment effect is the negligible heterogeneity. The zero estimate, however, is not always due to the absence of heterogeneous treatment effects but mainly due to the unavailability of adequate methods. On the other hand, studies that evaluate heterogeneity are conducted for moderate event rates. As a result, the performance of fixed-effect methods in the presence of heterogeneity is not well understood, particularly for low event rates. Although those methods are expected to perform poorly, only a few studies have extensively explored specific characteristics of the poor performance. For example, Bhaumik *et al.* (2012) showed that the asymptotic bias of combined odds ratio (with constant continuity correction "a") in the presence of treatment heterogeneity to be

$$\begin{aligned} Bias(\hat{\theta}_{wa}) = &-\frac{(p_{t|\epsilon} - q_{t|\epsilon})}{n(q_{t|\epsilon}p_{t|\epsilon})}\left\{a + \frac{p_c q_c - p_{t|\epsilon}q_{t|\epsilon}}{2(p_{t|\epsilon}q_{t|\epsilon} + p_c q_c)}\right\} \\ &+ \frac{(p_c - q_c)}{n(p_c q_c)}\left\{a + \frac{p_{t|\epsilon}q_{t|\epsilon} - p_c q_c}{2(p_{t|\epsilon}q_{t|\epsilon} + p_c q_c)}\right\}, \end{aligned} \tag{17}$$

where $p_{t|\epsilon}$ and p_c are unobservable underlying true event rates, $\hat{\theta}_{wa} = \sum_{i=1}^{k} \hat{w}_i(\tau^2)\hat{\theta}_{ia} / \sum_{i=1}^{k} \hat{w}_i(\tau^2)$, and $\hat{w}_i(\tau^2) = \dfrac{1}{\hat{\sigma}_i^2(\tau^2)}$. Their simulation study suggests that, for low event rates, the Mantel-Haenszel and Peto methods can grossly overestimate the treatment effect (see Figure 1) and produce an unacceptably high type I error rate. Therefore, in the presence of heterogeneity, the behavior of fixed-effect methods does not follow the patterns demonstrated in the simulation studies without heterogeneity. The bias of the treatment effect is reduced when random-effects methods with 1/2 continuity correction are used along with the improved estimates of heterogeneity parameters. However, even with alternative methods, an estimate of heterogeneity may not produce a non-zero value when event rates are extremely low (*e.g.*, 1/1000). As the true state of heterogeneity is unknown a priori, a large bias and an inflated type I error rate (see Figure 1) are potential threats associated with the validity of estimates of treatment effects obtained from weighted average methods, including Mantel-Henszel, Peto, and Dersimonial-Liard. These undesirable characteristics become more pronounced for low event rates. Shuster (2010) has also raised similar concerns regarding the validity of empirically based weighting in random effects and demonstrated that empirical weighting produces substantial bias for the DerSimonian-Laird approach.

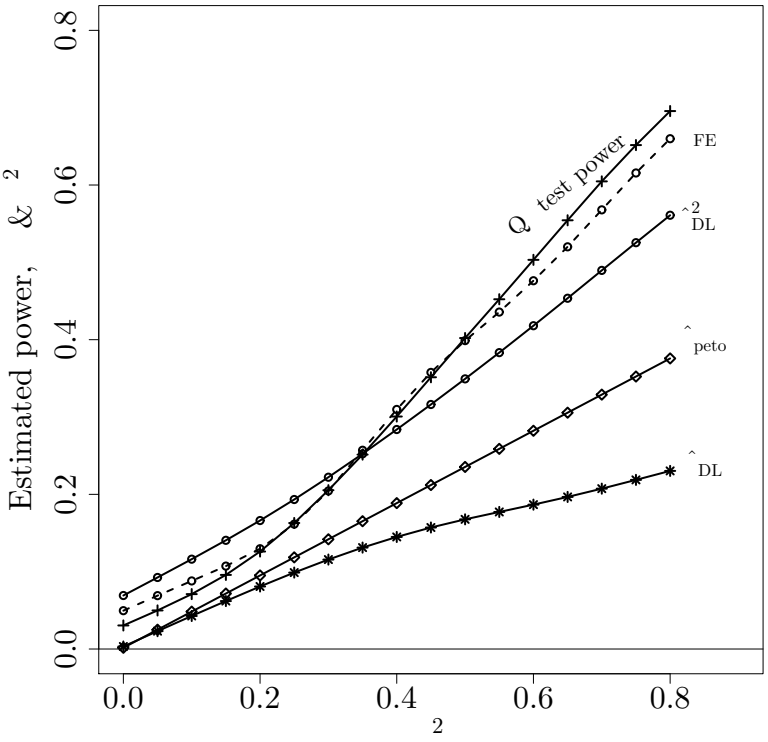


Figure 1: Power curve of Q-test as a function of τ^2 for a low event rate (0.4%). The true value of θ is set at 0. The α_{FE} is a type I error rate for testing the null hypothesis using the fixed-effect (Peto) method.

4. Likelihood based methods

4.1. Maximum marginal likelihood methods

The maximum marginal likelihood (MML) method in meta-analysis allows for simultaneous estimation of treatment effects and heterogeneity parameters, accommodating studies with zero total events without requiring continuity corrections. The MML method is model-based, an alternative to the moments matching methods. The major advantage of the MML approach over traditional methods is that the zero total events studies can be included without any artificial continuity corrections. It does have the flexibility of estimating both the overall treatment effect, and the heterogeneity parameter(s) simultaneously.

Consider an observed 2×2 Table 2 for the ith study for a meta-analysis of k studies. Suppose the probability of observing an event in the ith study is p_{ti} for the treatment group and p_{ci} for the control group. The log-odds of adverse events in group $j \in \{T, C\}$ can be modeled as follows.

$$ln\left(\frac{p_{ji}}{1-p_{ji}}\right) = \mu + \epsilon_{1i} + (\theta + \epsilon_{2i})T_{ji} \tag{18}$$

where T_{ji} is the treatment indicator variable defined as $T_{ji} = 0$ for $j = c$ and $T_{ji} = 1$ for $j = t$; and $\epsilon_1 \sim N(0, \sigma_\mu^2)$ and $\epsilon_2 \sim N(0, \tau^2)$ are the random-effects associated with mean

log-odds of an event in control group μ, and treatment effect θ, such that

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\mu^2 & \rho\sigma_\mu\tau \\ \rho\sigma_\mu\tau & \tau^2 \end{pmatrix} \right\}. \tag{19}$$

Therefore, this model allows for heterogeneity in both the baseline risk and the treatment effect. Conditional on the random effects, the likelihood function for the ith study is

$$l(\boldsymbol{x_i}|\epsilon) = p_{ti}^{x_{ti}} q_{ti}^{n_{ti}-x_{ti}} p_{ci}^{x_{ci}} q_{ci}^{n_{ci}-x_{ci}}, \tag{20}$$

where $\boldsymbol{x_i} = (x_{ti}, x_{ci})$ is the vector pattern of responses from study i. The models (18)-(19) involve three parameters μ, θ, and Σ, where Σ denotes covariance matrix on the right hand side of Equation (19). The marginal likelihood function for these parameters is obtained by integrating the conditional likelihood (20) over the distribution of random effects as follows

$$h(\boldsymbol{\beta}; \boldsymbol{x_i}) = h(\boldsymbol{x_i}) = \int_\epsilon l(\boldsymbol{x_i}|\epsilon) g(\epsilon) d\epsilon, \tag{21}$$

where $g(\epsilon)$ represents the related bivariate normal density. As studies are assumed to be independent, the full log-likelihood for k studies can be expressed as

$$\log L = \sum_{i=1}^{k} \log h(\boldsymbol{x_i}), \tag{22}$$

and for a parameter vector $\boldsymbol{\beta} = (\mu, \theta, \Sigma)$, the first derivatives of the log-likelihood with respect to $\boldsymbol{\beta}$ are

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{k} \frac{1}{h(\boldsymbol{x_i})} \frac{\partial h(\boldsymbol{x_i})}{\partial \boldsymbol{\beta}}, \tag{23}$$

where

$$\frac{\partial h(\boldsymbol{x_i})}{\partial \boldsymbol{\beta}} = \int_\epsilon \frac{\partial \log l(\boldsymbol{x_i}|\epsilon)}{\partial \boldsymbol{\beta}} l(\boldsymbol{x_i}|\epsilon) g(\epsilon) d\epsilon \,. \tag{24}$$

A close-form solution of (24) is generally not available for nonlinear models. Therefore, numerical techniques such as Gauss-Hermite quadrature are required for the integration of the random effect space (i.e., ϵ). The marginal likelihood equation in (21) can be approximated numerically to any practical degree of accuracy by summing on a specified number of quadrature nodes and the corresponding quadrature weights. Commercial software packages such as SAS, STATA, SuperMix . can easily fit MML models, and the GLIMMIX procedure in SAS or the glmer package in R can be used to fit alternative linearized approximation to (24).

The MML models offer a variety of modeling strategies in the context of meta-analysis. Treatment effect may be estimated with a single random effect (background incidence or treatment effect) or a model with two correlated random effects. However, this flexibility to construct a model with a combination of multiple random effects also creates room for model mis-specifications. The detailed analysis of the impact of such model misspecification on the characteristics and testing of the overall effect estimator and the heterogeneity parameter has shown that the models that allow heterogeneity in both baseline rate and treatment effect

across studies have low type I and type II error rates, and are the least biased compared to other model specifications Amatya *et al.* (2015).

4.2. Beta-binomial model

The beta-binomial model offers a Bayesian framework for meta-analysis, allowing for estimation of treatment effects and correlations between event probabilities across studies. The beta-binomial model is another alternative to the moment-based methods. In the Bayesian setup, a meta-analysis of binary events can be performed in two ways using the beta-binomial model. The first way is to adopt a univariate approach, where event probabilities p_{Ti} and p_{Ci} are assumed to be independent. However, the individual binary observations within the jth arm of the ith study (which add up to x_{ji} for $j =\in \{T, C\}$) are allowed to be correlated by imposing $p_{ji} \sim beta(\alpha_j, \beta_j)$ as a prior. As a result, $E(p_j) = \mu_j = \frac{\alpha_j}{\alpha_j+\beta_j}$, $Var(p_j) = \mu(1-\mu)\theta/(1+\theta)$ with $\theta = 1/(\alpha_j + \beta_j)$, and the correlation between observations within jth arm of each study is $\rho_j = 1/(\alpha_j + \beta_j + 1)$, and the marginal distribution of x_{ji} is the beta-binomial distribution with the following log-likelihood function:

$$\begin{aligned} l_{ji}(\alpha_j, \beta_j) =& ln\Gamma(n_{ji}+1) + ln\Gamma(x_{ji}+\alpha_j) + ln\Gamma(n_{ji}-x_{ji}+\beta_j) \\ &+ ln\Gamma(\alpha_j+\beta_j) - ln\Gamma(x_{ji}+1) - ln\Gamma(n_{ji}-y_{ji}+1) \\ &- ln\Gamma(n_{ji}+\alpha_j+\beta_j) - ln\Gamma(\alpha_j) - ln\Gamma(\beta_j), \end{aligned} \tag{25}$$

and the joint log-likelihood function is:

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{K} \sum_{j\in\{T,C\}} l_{ji}(\alpha_j, \beta_j). \tag{26}$$

The number of parameters is reduced further by modeling the mean function $g(\mu_j) = b_0 + b_1 x_j$, where g is a link function as in the generalized linear model, and $x_j = 1$, if $j = T$ and $x_j = 0$, if $j = C$. A specific link function for g determines the type of effect. For example, the logit link function gives the log odds ratio, and the log link function measures log relative risk. Kuss (2014) recommends avoiding the identity link to estimate the risk difference and suggests to use the estimated event probabilities $\hat{p}_C = g^{-1}(\hat{b}_0)$ and $\hat{p}_T = g^{-1}(\hat{b}_0 + \hat{b}_1)$ from the logit model for the control and treatment groups, respectively.

The second approach is to use the bivariate beta-binomial model which addresses the correlation between the event probabilities of two treatment arms of the studies. The correlation between control event rates (proportion) and treatment effects has been identified in studies by various authors (Schmid *et al.*, 1998, and references therein). Unlike the MML, the bivariate beta-binomial model implies a linear relationship between p_T and p_C on the original scale. Chu *et al.* (2012) described a beta-binomial model in two stages. In the first stage, X_{ji} is assumed to be independently binomially distributed, such that

$$P(X_{Ti} = x_{Ti}, X_{Ci} = x_{Ci} | n_{Ti}, n_{Ci}, p_{Ti}, p_{Ci}) = \prod_{j\in\{T,C\}} \binom{n_{ji}}{x_{ji}} (p_{ji})^{x_{ji}} (1-p_{ji})^{n_{ji}-x_{ji}}. \tag{27}$$

In the second stage, the joint distribution of p_{Ti}, and p_{Ci} is specified using a Sarmanov beta prior distribution as follows (see Luo *et al.*, 2014):

$$\begin{aligned} p_{Ti}, p_{Ci}|\alpha_T, \alpha_C, \beta_T, \beta_C &\sim f(p_T, p_C; \alpha_T, \alpha_C, \beta_T, \beta_C) \\ &= beta(p_T; \alpha_T, \beta_T) beta(p_C; \alpha_C, \beta_C) \left(1 + \rho \frac{(p_T - \mu_T)(P_C - \mu_C)}{\delta_T \delta_C}\right), \end{aligned} \tag{28}$$

where ρ is the correlation coefficient between p_{Ti} and p_{Ci}; $beta(p; \alpha, \beta) = [\mathrm{B}(\alpha, \beta)]^{-1} p^{\alpha-1}(1-p)^{\beta-1}$ with $\mathrm{B}(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt$; and $\mu_j = \alpha_j/(\alpha_j+\beta_j)$, $\delta_j^2 = \mu_j(1-\mu_j)/(\alpha_j+\beta_j+1)$, and $j \in \{T, C\}$. As a result, the log marginalized likelihood function for the unknown hyperparameters $(\alpha_T, \alpha_C, \beta_T, \beta_C, \rho)$ is

$$\begin{aligned} &log\ L(\alpha_T, \alpha_C, \beta_T, \beta_C, \rho) \\ &= \sum_{i=1}^{k} log\left[P_{BB}(x_{Ti}; n_{Ti}, \alpha_T, \beta_T) P_{BB}(x_{Ci}; n_{Ci}, \alpha_C, \beta_C)\right] \\ &+ log\left\{1 + \rho \frac{\left(\frac{x_{Ti}+\alpha_T}{n_{Ti}+\alpha_T+\beta_T} - \mu_T\right)\left(\frac{x_{Ci}+\alpha_C}{n_{Ci}+\alpha_C+\beta_C} - \mu_T\right)}{\delta_T \delta_C}\right\}, \end{aligned} \tag{29}$$

where $P_{BB}(x; n; \alpha; \beta)$ is the probability mass function of a beta-binomial distribution, such that

$$P_{BB}(x; n; \alpha; \beta) = \binom{n}{x} \frac{B(x+\alpha, n-x+\beta)}{B(\alpha, \beta)}. \tag{30}$$

The maximum likelihood estimates $(\hat{\alpha}_T, \hat{\alpha}_C, \hat{\beta}_T, \hat{\beta}_C, \hat{\rho})$ is obtained by maximizing likelihood function (29). Based on these estimates, three overall effect measures are estimated as follows:

$$\text{Odds Ratio} = \widehat{OR} = \frac{\hat{\mu}_T/(1-\hat{\mu}_T)}{\hat{\mu}_C/(1-\hat{\mu}_C)} = \frac{\hat{\alpha}_T \hat{\beta}_C}{\hat{\alpha}_C \hat{\beta}_T}, \tag{31}$$

$$\text{Relative Risk} = \widehat{RR} = \frac{\hat{\mu}_T}{\hat{\mu}_C} = \frac{\hat{\alpha}_T/(\hat{\alpha}_T+\hat{\beta}_T)}{\hat{\alpha}_C/(\hat{\alpha}_C+\hat{\beta}_C)}, \tag{32}$$

$$\text{Risk difference} = \widehat{RD} = \hat{\mu}_T - \hat{\mu}_C = \frac{\hat{\alpha}_T}{(\hat{\alpha}_T+\hat{\beta}_T)} - \frac{\hat{\alpha}_C}{(\hat{\alpha}_C+\hat{\beta}_C)}. \tag{33}$$

The variances of these estimates are calculated using the delta method.

5. Confidence distribution methods

Confidence distribution, in meta-analysis refers to a statistical method where the uncertainty about a parameter (such as an effect size) is represented by a distribution rather than a single point estimate. This distribution integrates information from multiple studies, accommodating varying study sizes and results, including studies with zero total events. Xie *et al.* (2011) have developed a unified framework for meta-analysis by combining confidence distributions (CD) from individual studies. The combined CD function is obtained

by appropriately weighting the individual distribution estimators. This is in contrast to the traditional meta-analysis, where a combined estimate is obtained by averaging individual point estimates with appropriate weights. The combined CD does have various optimality conditions. This method also allows straightforward integration of data from all studies including zero total events.

Suppose that the CD function $H_i(\theta) = H_i(\boldsymbol{X}_i, \theta)$, $i = 1, \ldots, k$ for the parameter θ can be obtained from each study with corresponding samples $\boldsymbol{X}_i$ of size n_i. A combined confidence distribution function across k studies (H_c) is constructed as

$$H_c = G_c\{g_c(H_1(\theta), \ldots, H_k(\theta))\}, \tag{34}$$

where $g_c(u_1, \ldots, u_k) = w_1 F_0^{-1}(u_1) + \ldots + w_k F_0^{-1}(u_k)$ is a monotonic function that has the cumulative distribution function $G_c(t) = P(g_c(U_1, \ldots, U_k) \leq t)$ for $U_i \sim U[0, 1]$. The transformation function $F_0(\cdot)$ is weighted by fixed positive weights $w_i \geq 0$. The conventional fixed- and random-effect meta-analysis approaches can be easily derived using the recipe in (34) (see Xie *et al.*, 2011).

5.1. Odds ratio

Meta-analysis of rare event studies using odds ratio under the CD framework was developed by Liu (2012). This method uses exact p-values based on mid-p adaptation of Fisher's exact test for the odds ratio as the CD functions for individual studies and combines them by applying the general CD combination method as described in (34). Using this exact test, the p-value function for the odds ratio Ψ is obtained as follows:

$$p_i(\Psi) \equiv p_i(\Psi; x_{Ti}, x_{Ti}) = Pr_\Psi(X_{Ti} > x_{Ti} | T_i = t_i) + \frac{1}{2} Pr_\Psi(X_{Ti} = x_{Ti} | T_i = t_i), \tag{35}$$

where, the hypothesis of interest is

$$H_0 : \Psi = \Psi^0 vs. H_1 > \Psi^0.$$

The X_{Ti}'s are assumed to follow a hypergeometric distribution conditional on $T_i = X_{Ti} + X_{Ci}$. Then, for $L_i = max(0, t_i - n_{Ci})$, and $U_i = min(n_{Ti}, t_i)$. It follows that

$$Pr_\Psi(X_{Ti} = x_{Ti} | T_i = t_i) = \frac{\binom{n_{Ti}}{x_{Ti}} \binom{n_{Ci}}{t_i - x_{Ti}} \Psi^{x_{Ti}}}{\sum_{s=L_i}^{U_i} \binom{n_{Ti}}{s} \binom{n_{Ci}}{t_i - s} \Psi^s}, \quad L_i \leq x_{Ti} \leq U_i. \tag{36}$$

The statistic $p_i(\Psi^0)$ asymptotically follows $U(0, 1)$. However, for the meta-analysis of rare events, the asymptotic conditions are seldom valid, causing a substantial deviation of $p_i(\Psi^0)$ from $U(0, 1)$. Nonetheless, Liu (2012) has shown that the general idea of a CD combining algorithm can still be used in the finite sample setting after some adjustments. They also showed that zero total event studies can provide meaningful contributions in the presence of uncertainty. The impact of zero total event studies is appropriately accounted for in the

sample size computation of the corresponding studies by using the weights:

$$w_i \propto \left[\{n_{Ti}\pi_{Ti}(1-\pi_{Ti})\}^{-1} + \{n_{Ci}\pi_{Ci}(1-\pi_{Ci})\}^{-1}\right]^{-1/2}, \tag{37}$$

which requires estimates of π_{Ci} and π_{Ti}. To improve an efficiency of the overall estimate, Liu *et al.* (2014) proposed to model π_{Ci} using a beta(β_1, β_2) distribution. The parameters of this beta distribution are estimated as follows:

$$(\hat{\beta}_1, \hat{\beta}_2, \hat{\Psi}) = \underset{\beta_1,\beta_2,\Psi}{\arg\max} \sum_{i=1}^{k} Log \int_0^1 f_\psi(x_{Ci}, x_{Ti}|\pi_{Ci}) f_{\beta_1,\beta_2}(\pi_{Ci}) d\pi_{Ci}, \tag{38}$$

where $f_{\beta_1,\beta_2}(\pi_{Ci}) = \pi_{Ci}^{\beta_1-1}(1-\pi_{Ci})^{\beta_2-1} / \int_1^0 \pi_{Ci}^{\beta_1-1}(1-\pi_{Ci})^{\beta_2-1} d(x_{Ci})$, $f_\psi(x_{Ci}, x_{Ti}|\pi_{Ci}) = c(x_{Ci}, x_{Ti})\pi_{Ci}^{x_{Ci}}(1-\pi_{Ci})^{n_{Ci}-x_{Ci}}\pi_{Ti}^{x_{Ti}}(1-\pi_{Ti})^{n_{Ti}-x_{Ti}}$, and $\pi_{Ti} = (\Psi\pi_{Ci})/(1-\pi_{Ci}+\Psi\pi_{Ci})$. The mean of the empirical conditional density of π_{Ci} is used as an estimate of π_{Ci} and an estimate of π_{Ti} is calculated through $\hat{\pi}_{Ti} = (\Psi\pi_{Ci})/(1-\hat{\pi}_{Ci}+\hat{\Psi}\hat{\pi}_{Ti})$. This manipulation produces positive estimates of π_{Ti} and π_{Ci} even for zero total event studies, allowing the inclusion of these studies without any continuity correction. When $x_{Ti} = 0$ for all i, limiting weights are calculated as follows

$$\lim_{\hat{\Psi}\to 0} \left(w_i / \sum_{i=1}^{k} w_i \right)^2 = \frac{n_{Ci}x_{Ci}/(1-x_{Ci})}{\sum_{i=1}^{k} n_{Ci}x_{Ci}/(1-x_{Ci})}.$$

The case where $x_{Ci} = 0$ for all i is handled similarly.

5.2. Risk difference

Tian *et al.* (2009) proposed a simple procedure to construct a $100(1-\alpha)$ 1-sided confidence interval (CI) of the type (a, ∞) for a common risk difference parameter Δ, based on all data from k independent studies without any artificial continuity correction. Suppose that n sets of k study-specific 1-sided CIs of any arbitrary level η can be constructed for Δ. Let $J_{ij} = (a_{ij}, \infty)$ be the η_j-level 1-sided CI obtained from the ith study, for $i = 1, \ldots, k$, and $j = 1, \ldots, n$; such that $0 < \eta_1 < \eta_2 < \ldots < \eta_n < 1$, and $a_{i1} > a_{i2} > \ldots > a_{in}$. The final combined interval for δ is (see Tian *et al.*, 2009)

$$\sum_{i=1}^{k} w_i \sum_{j=1}^{n} \tilde{w}_j \{(I(\Delta > a_{ij}) - \eta_j\} \geq c, \tag{39}$$

where $I(\cdot)$ is the indicator function, w_i is a study-specific weight, $\tilde{w}_j$ is a positive weight for η_j-level intervals, and the critical value c is chosen such that

$$Pr\left[\sum_{i=1}^{k} w_i \sum_{j=1}^{n} \tilde{w}_j (B_{ij} - \eta_j) < c\right] \leq \alpha. \tag{40}$$

In equation (40), $(B_{i1}, \ldots, B_{ik})$ are n independent random vectors whose components are correlated Bernoulli variables such that $B_{i1} \leq B_{i2} \leq \ldots \leq B_{ik}$ and $pr(B_{ij} = 1) = \eta_j$.

Tian *et al.* (2009) suggested to use $w_i = 1/(n_{Ti}+n_{Ci})$, and $\tilde{w}_j = \{\eta_j(1-\eta_j)\}^{-1}$ for the weights. Yang *et al.* (2012) showed this procedure to be a special case under the CD framework, where $F_0^{-1}(u)$ is chosen to be $\sum_{j=1}^{n} \tilde{w}_j \{I(u > 1-\eta_j) - \eta_j\}$. Then,

$$H^c(\Delta) = G_c \left\{ \sum_{i=1}^{k} w_i \sum_{j=1}^{n} \tilde{w}_j \{I(H_i(\delta) > 1-\eta_j) - \eta_j\} \right\}. \tag{41}$$

For the detailed derivation of the proof, see Yang *et al.* (2012), where alternative equivalent expression for $H^c(\Delta)$ is also obtained by using the logistic function as a transformation function.

6. Illustration

In their highly influential meta-analysis article, Nissen and Wolski (2007) concluded that rosiglitazone was associated with a significant risk of myocardial infarction [odds ratio (OR) 1.43, 95 % CI (1.03,1.98), P = 0.03] and an increase in the risk of death from cardiovascular causes, which had borderline significance [OR 1.64, 95 % CI (0.98 2.74); P = 0.06]. These conclusions were based on a fixed-effect meta-analysis using the Peto method. Soon after the release of these results, a series of reanalysis of the same data was published by others using different methods. Diamond *et al.* (2007) has conducted the meta-analysis using three conventional fixed-effect methods with two continuity corrections and including/excluding zero total event studies. Stoto (2015) reported some results based on a Localio *et al.* (2008) wide variety of statistical methods. Tian *et al.* (2009), Chu *et al.* (2012), and Liu *et al.* (2014) have used a few relatively new approaches to analyze the rosiglitazone data. They included all studies (including zero total event studies) without any continuity correction. Chu *et al.* (2012) used the beta-binomial model, whereas Tian *et al.* (2009), and Liu *et al.* (2014) used the confidence distribution methods. Estimates of various effect measures from these articles are summarized in Table 3.

7. Bayesian methods

The Bayesian methodology in meta-analysis offers flexible modeling with hierarchical structures, integrating prior information and accommodating non-normal distributions of random effects. Computational intensity has decreased with advancements in Monte Carlo techniques and computing power, supporting complex analyses without major barriers. Bayesian methodology is an alternative to the traditional meta-analysis methods. It provides a broad range of modeling alternatives with multiple levels of hierarchy and naturally integrates prior information on parameters of interest from other trials or studies. The emphasis on hierarchical modeling accounts for uncertainty in all parameters including the between-study heterogeneity. The flexibility of the Bayesian approach allows for rigorous sensitivity analysis, which is particularly important for meta-analysis of rare events. Furthermore, the Bayesian framework can be easily extended to non-normal distributions of random effects. The computational complexity of the Bayesian approach is substantially intensive compared to the traditional methods. Fortunately, software is readily available that incorporates rapidly developing Monte Carlo techniques. Due to the unprecedented rise in computational power of modern personal computers, complex computation in Bayesian analysis is no longer a major barrier.

Table 3: Various estimates of effect measures for rosiglitazone MI data

	Method	CC	OR (95% CI)	RR (SE)	RD (95% CI)
Nissen	Peto	0.5	1.43 (1.03, 1.98)		
Diamond	Fixed, IV	TAC	1.34 (.097, 1.84)		
	Fixed, IV	CC	1.29 (0.94, 1.76)		.0015 (0, .0031)
	Fixed, MH	TAC	1.36 (1.00, 1.84)		
	Fixed, MH	CC	1.28 (0.95, 1.72)		.0020 (0, .0041)
	Fixed, MH	TAC+	1.35 (1.00, 1.82)		
	Fixed, MH	CC+	1.26 (0.93, 1.69)		
Localio	Random [DL]	NA	1.31 (0.91, 1.89)		
	Random [DL]	0.5	1.31 (0.95, 1.79)		
	Random [DL]	TAC	1.33 (0.93, 1.91)		
	Conditional logistic	NA	1.45 (1.05, 2.01)		
	Exact stratified	NA	1.45 (1.03, 2.04)		
	Random intercept/slope	NA	1.37 (0.99, 1.90)		
Chu	Bivariate beta-binomial	NA		1.291 (0.382)	0.0011 (SE=0.0013)
Tian	Exact CD	NA			0.0018 (-0.008, 0.004)
Liu	Exact CD	NA	(.972, 2.00)		
	Adjusted Exact CD		(1.04, 2.01)		

CC: Constant (0.5) correction for continuity, CC+: constant correction for continuity that includes all zero total event studies, IV: inverse variance, MH: Mantel-Haenszel, TAC: treatment arm correction for continuity, TAC: treatment arm correction for continuity that includes all zero-total-event studies

The key elements of a generic Bayesian meta-analysis model are the prior distributions on both the effect and the heterogeneity parameters. The simplest form of Bayesian random-effect meta-analysis is as follows: (see Sutton and Abrams, 2001):

$$\begin{aligned} \hat{\theta}_i &\sim f(\hat{\theta}_i|\theta_i, \sigma_i^2) \\ \theta_i &\sim \pi(\theta_i|\theta, \tau^2) \\ \theta &\sim h(\theta) \\ \tau^2 &\sim h(\tau^2), \end{aligned} \tag{42}$$

where $h(\theta)$ and $h(\tau^2)$ are the prior distributions of effect parameter θ, and between-study heterogeneity parameter τ^2 respectively. The resulting posterior distribution does have the following form:

$$p(\theta, \tau, \theta_i|\hat{\boldsymbol{\theta}}_i) \propto h(\theta)h(\tau^2) \prod_{i=1}^{K} \pi(\theta_i|\theta, \tau^2) \prod_{i=1}^{K} f(\hat{\theta}_i|\theta_i, \sigma_i^2). \tag{43}$$

Inferences on parameters of interest are made from the mode of the posterior distribution (43). Except for some cases of conjugate prior distributions, the posterior mode is usually

not available in its closed form. Instead, Monte Carlo methods such as Gibbs sampling are used to numerically approximate the mode of the posterior distribution. An example of Gibbs sampling is found in a meta-analysis of randomized controlled trials comparing sodium monobuorophosphate (SMFP) to sodium Buoride (NaF) dentifrices (toothpaste) in the prevention of caries development Abrams and Sanso (1998). A complex example of a Bayesian hierarchical model that incorporates a study-level component of variability and facilitates extensive sensitivity analysis is found in Kaizar *et al.* (2006).

7.1. Strong prior

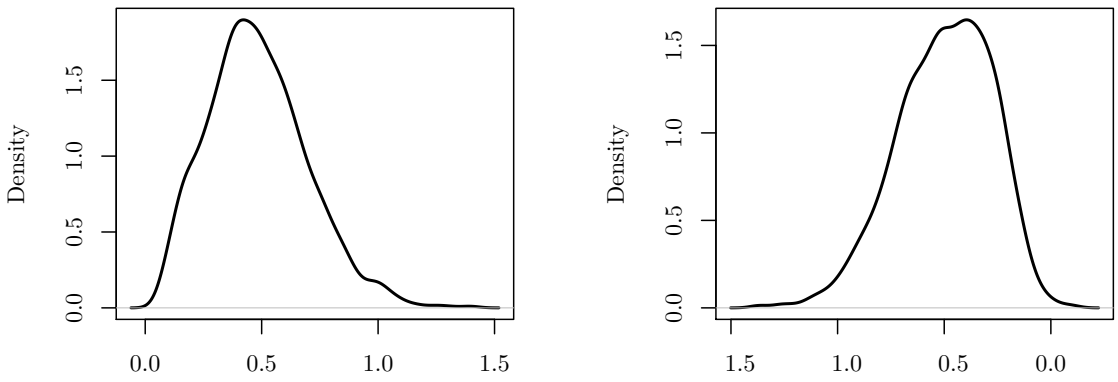
A strong prior in Bayesian analysis is one that conveys substantial prior belief or information about the parameters of interest, such as the probability of events in the control arm, treatment effect, or between-study heterogeneity. It can significantly influence the estimated outcomes of a meta-analysis by anchoring the inference towards specific values based on empirical data or subjective judgment. The prior distribution is not only a key part of Bayesian analysis, but also it is one of the most difficult and controversial aspects of the analysis. A non-informative prior is specified to express vague or general information of a parameter and to minimize a perceived subjective bias. On the other hand, an alternative prior distribution can be specified to integrate prior belief or substantiated information relevant to the estimation of the parameter of interest. Such informative priors may be formulated by considering the plausible range of the parameters, based on observed distributions from empirical studies, or based purely on subjective clinical judgment Warn *et al.* (2002). These informative priors may influence the conclusion of the meta-analysis. When binary events are of concern, prior distribution needs to be specified for the following three parameters:(1) probability of events in the control arm, (2) treatment effect, and (3) between-study heterogeneity. Strong priors on some of these parameters may have a substantial impact on the estimated overall treatment effect. The following Bayesian meta-analysis of rositaglitazone data illustrates the impact of strong priors.

Let n_{ji} be the number of the participants in the ith trial who received the jth treatment. Suppose that the probability of experiencing MI is p_{ji}. The observed MI incidences x_{ji} may be modeled under the Bayesian framework as follows:

$$\begin{aligned} X_{ji} &\sim binomial(p_{ji}, n_{ji}), \quad \text{for } j \in \{T, C\}, \text{ and } i = 1, 2, \ldots, k \\ \theta_i &\sim N(\theta, \tau^2), \quad \mu_i = logit(p_{Ci}) \\ logit(p_{Ti}) &= \mu_i + \theta_i. \end{aligned}$$

The model presented above can be easily implemented in the WinBugs program (see Warn *et al.*, 2002). To represent a plausible range of θ and τ, prior distributions $N(0, 10)$, and $U(0, 2)$ are specified, respectively. Based on the specified priors and the observed data, the WinBugs program computes posterior distributions of parameters using MCMC methods. The posterior modes of θ and τ are estimated from these posterior distributions. A graphical representation of the posterior distribution for this example data is displayed in Figure (2).

The posterior estimate of the combined odds ratio and heterogeneity from the above modeling are given in Table 4, where noninformative independent prior $U(0, 1)$ is specified for P_{Ci}. This model specification assumes a fixed background MI incidence rate and heterogeneous between-study treatment effects. The results in Table 4 show a posterior estimate



(a) Posterior Distribution of τ. **(b) Posterior Distribution of θ.**

Figure 2: Posterior Distributions of τ and θ.

Table 4: Bayesian Meta-Analysis of 42 Rosiglitazone Trials

	mean	2.5%	25%	50%	75%	97.5%
OR	0.608	0.379	0.522	0.617	0.720	0.883
τ^2	0.234	0.015	0.111	0.217	0.379	0.932

of the odds ratio of 0.608 with 95% credible interval of (0.379, 0.883), which is markedly different from the moment based estimators in Table 3. It is noteworthy that the current estimate is close to the estimate obtained from MML with random effects restricted only to the treatment effect. A priori belief regarding the incidence of MI rate among type II diabetes patients can be integrated by changing the parameters of the prior distribution of P_{Ci}. Figure 3 displays the impact of different values of parameters of the prior distribution of P_{Ci}. The posterior odds ratio remains below 1.0 as the prior becomes closer to the vague, the same as the results in Table 4. However, a strong prior of uniform(0, 0.01) provides a positive log odds ratio that is close to the moment-based results. Figure 3 essentially shows that if one is willing (or has reason) to believe *a priori* that the prevalence of MI is extremely rare, *e.g.*, less than 6/1000, in a diabetic population, then the observed data supports an elevated risk of MI among rosiglitazone users. In the absence of such prior information, the model does not support the conclusion derived from the moment-based analyses.

This example clearly demonstrates the effect of prior distributions on the conclusion of meta-analysis. A similar but less dramatic effect on the estimate of the log odds ratio is also observed for different informative prior specifications of τ. However, when only a small number of studies are available, a strong prior distribution on τ can significantly influence the results of the analysis Sutton and Abrams (2001). The hierarchical Bayesian approach is used to introduce a reasonable amount of uncertainty in the prior belief regarding distributions of the model parameters. In the rosiglitazone example, a $beta(a, b)$ prior distribution may be used to model p_c, and a $gamma(s, r)$ hyper-prior may be placed on the parameters of the

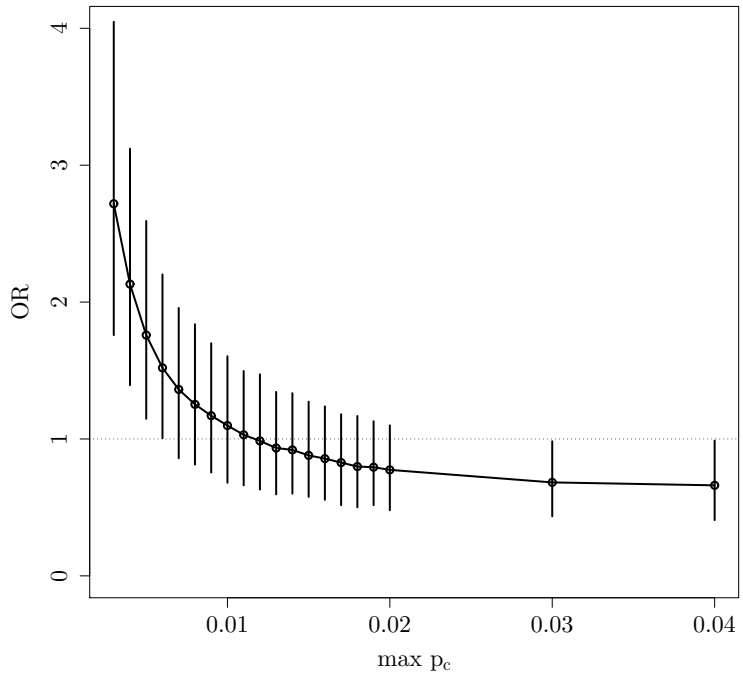


Figure 3: Posterior mean and 95% credible interval of odds ratio for varying maximum values (max p_c) of the prior distribution of $p_c \sim unif(0, \max p_c)$

beta distribution. The resulting full Bayesian model is

$$\begin{aligned} &r_{Ti} \sim binomial(p_{Ti}, n_{Ti}),\ r_{Ci} \sim binomial(p_{Ci}, n_{Ci}) \\ &\mu_i = logit(p_{Ci}),\ logit(p_{Ti}) = \mu_i + \theta_i, \\ &p_{Ci} \sim beta(a, b),\ \theta_i \sim N(\theta, \tau^2) \\ &a \sim gamma(r_a, s_a),\ b \sim gamma(r_b, s_b); \end{aligned} \tag{44}$$

where, gamma $(r, s) = \frac{s^r}{\Gamma(r)} x^{r-1} e^{-sx}$ for $x > 0, r > 0$ and $s > 0$. Table 5 presents the estimates obtained from this model for different combinations of parameters of the gamma hyper-prior distribution. For this illustration, r_a and r_b were varied while holding s_a and s_b values fixed at 0.25 and 1.5, respectively. The posterior means of the log odds ratio are clearly more consistent between 0.11 (OR=1.12) and 0.17 (OR=1.19) over different specifications of hyper-prior distributions. These estimates of log odds ratios are closer to the estimates obtained from the moment-based methods.

Table 5: Hierarchical Bayesian Analysis of 42 Rosiglitazone Trials using Different Gamma Hyper-prior Distributions.

		α	β	log odds ratio (θ)			τ^2		
r_a	r_b	mean	mean	mean	2.5%	97.5%	mean	2.5%	97.5%
0.0004	12.052	0.305	14.537	0.152	-0.252	0.513	0.069	0.000	0.651
0.0018	15.801	0.322	16.970	0.135	-0.263	0.530	0.077	0.000	0.709
0.0031	4.170	0.263	8.861	0.144	-0.206	0.529	0.062	0.000	0.600
0.0039	15.551	0.320	16.742	0.152	-0.240	0.524	0.071	0.000	0.612
0.0042	19.110	0.336	19.199	0.153	-0.251	0.551	0.065	0.000	0.622
0.0047	17.088	0.323	17.972	0.159	-0.237	0.512	0.066	0.000	0.621
0.0057	17.625	0.327	18.191	0.143	-0.223	0.549	0.061	0.000	0.627
0.0066	8.534	0.285	11.883	0.163	-0.221	0.538	0.074	0.000	0.660
0.0088	11.785	0.298	12.304	0.150	0.025	0.237	0.002	0.000	0.008
0.5884	16.796	0.332	17.959	0.146	-0.214	0.531	0.069	0.000	0.602
1.1903	7.241	0.291	11.447	0.139	-0.267	0.509	0.075	0.000	0.670
1.2831	13.824	0.317	15.963	0.112	-0.249	0.465	0.107	0.004	0.719
1.9515	1.021	0.259	7.270	0.173	-0.218	0.552	0.060	0.000	0.626
2.5283	14.976	0.342	16.910	0.137	-0.251	0.524	0.075	0.000	0.638
2.7945	16.991	0.360	18.552	0.120	-0.298	0.489	0.075	0.000	0.598
3.5064	3.797	0.292	9.580	0.161	-0.248	0.587	0.060	0.000	0.584
3.8382	15.027	0.357	17.441	0.131	-0.252	0.503	0.063	0.000	0.516
4.0664	8.528	0.326	12.920	0.123	-0.271	0.515	0.071	0.000	0.653
4.0838	7.780	0.323	12.488	0.140	-0.266	0.494	0.067	0.000	0.652

8. Discussion

Meta-analysis of safety data, particularly for rare events, poses challenges due to low event rates in randomized controlled trials (RCTs) designed primarily for efficacy. These issues include inadequate power to detect true risks and complexities arising from biases and study design differences in observational studies. Analytical methods vary in handling heterogeneity, influencing conclusions on drug safety, as seen in meta-analyses of Rosiglitazone's association with myocardial infarction, highlighting the need for cautious interpretation and sensitivity analysis. Meta-analysis of rare events data in general, and safety data in particular is a complex statistical problem with immense practical importance. Randomized control trials (RCT) are generally not designed to study safety issues related to a treatment. Therefore, individual trials may not provide adequate power to detect the true risk of adverse events, particularly when the adverse event is rare. Post-marketing safety studies are usually conducted using large observational studies. A meta-analysis from a series of large observational studies can provide a spurious degree of statistical precision, leading to acceptance of low-level associations resulting from residual confounding Henry and Hill (1999). Inherent biases and differences in study designs add further complexities to the meta-analysis of observational studies. Consequently, the assessment of drug safety partly relies on the meta-analysis of RCTs and other published literature. Although such reliance on meta-analysis holds promises of synthesizing all available evidence, it is not without serious pitfalls. Stoto (2015) discussed these issues using three high-profile examples. Stoto (2015) concluded that the precision of the results of one meta-analysis can be deceptively

low due to some typical characteristics of safety data extracted from efficacy studies. Those characteristics include low adverse event rates, untestable clinical and methodological heterogeneity, and incomplete and inconsistent reporting of adverse effects. Consequently, different syntheses can provide qualitatively different conclusions. For example, analytical methods that avoid or deal with heterogeneity in different ways may lead to different conclusions related to the risk of adverse events. A careful consideration is particularly important for safety studies, where the standard Cochrans' Q-test for detecting heterogeneity is known to be significantly underpowered (see Figure 1). These studies often possess substantive heterogeneity of the populations under study, comparison groups, and length of follow-up. The rationale for using the Peto method in such situations often points to its greater statistical power which is considered to be more important in safety analysis than the consideration of heterogeneity. However, one must not overlook a high type I error rate associated with such methods in the presence of heterogeneity. Discrepancies originating from the use of various methods are evident in the comparison of meta-analytical investigations of MI associated with rosiglitazone. A decision to place severe restrictions on the utilization of the drug was highly influenced by the results of the Peto method-based meta-analysis performed by Nissen and Wolski (2007). That analysis yielded a 95% confidence interval of (1.031, 1.979) and a p-value of 0.032 for testing that the odds ratio is 1, and thus concluded that rosiglitazone was significantly associated with myocardial infarction. The subsequent meta-analyses by others using different methods produced results that did not agree with Nissen and Wolski (2007). The varying conclusions depended on the inclusion or exclusion of zero total event studies Liu *et al.* (2014), continuity correction strategies Diamond *et al.* (2007), and effect measure (RR vs. OR) and statistical method used for analysis Stoto (2015). Furthermore, meta-analysis is itself an observational study of studies. When only a small number of adverse events are observed, meta-analysis may not be able to disentangle confounding by the indication and drug type. Over-reliance on a single analysis is not recommended when analyzing safety data. Fortunately, there are several commercial (SAS, STATA, StatXact) and freely available software (RevMan, and Rgmeta, meta, exactmeta) to facilitate an extensive sensitivity analysis when analyzing safety data involving adverse events that might occur in one per thousand patients or fewer.

References

Abrams, K. and Sanso, B. (1998). Approximate Bayesian inference for random effects meta-analysis. *Statistics in Medicine*, **17**, 201–218.

Amatya, A., Bhaumik, D. K., Normand, S.-L., Greenhouse, J., Kaizar, E., Neelon, B., and Gibbons, R. D. (2015). Likelihood-based random effect meta-analysis of binary events. *Journal of Biopharmaceutical Statistics*, **25**, 984–1004.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society*, **32**, 283–301.

Bhaumik, D. K., Amatya, A., Normand, S. T., Greenhouse, J., Kaizar, E., Neelon, B., and Gibbons, R. D. (2012). Meta-analysis of rare binary adverse event data. *Journal of the American Statistical Association*, **107**, 555–567.

Bradburn, M. J., Deeks, J. J., Berlin, J. A., and Russell L., A. (2007). Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine*, **26**, 53–77.

Chu, H., Nie, L., Chen, Y., Huang, Y., and Sun, W. (2012). Bivariate random effects models for meta-analysis of comparative studies with binary outcomes: methods for the absolute risk difference and relative risk. *Statistical Methods in Medical Research*, **21**, 621–633. Epub 2010 Dec 21.

Cox, D. (1970). The continuity correction. *Biometrika*, **57**, 217–219.

Diamond, G. A., Bax, L., and Kaul, S. (2007). Uncertain effects of rosiglitazone on the risk for myocardial infarction and cardiovascular death. *Annals of Internal Medicine*, **147**, 578–581.

Friedrich, J. O., Adhikari, N., and Beyene, J. (2007). Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data. *BMC Medical Research Methodology*, **7**.

Greenland, S. and Salvan, A. (1990). Bias in the one-step method for pooling study results. *Statistics in Medicine*, **9**, 247–252.

Henry, D. and Hill, S. (1999). Meta-analysis: its role in assessing drug safety. *Pharmacoepidemiology and Drug Safety*, **8**, 167–168.

Kaizar, E., Greenhouse, J., Seltman, H., and Kelleher, K. (2006). Do antidepressants cause suicidality in children? a Bayesian meta-analysis. *Clinical Trials*, **3**, 73–98.

Kuss, O. (2014). Statistical methods for meta-analyses including information from studies without any eventsadd nothing to nothing-and succeed nevertheless. *Statistics in Medicine*, **34**, 1097–1116.

Liu, D. (2012). *Combining Information for Heterogeneous Studies and Rare Event Studies: A Confidence Distribution Approach*. PhD thesis, Rutgers University.

Liu, D., Liu, R. Y., and Xie, M. (2014). Exact meta-analysis approach for discrete data and its application to 2× 2 tables with rare events. *Journal of the American Statistical Association*, **109**, 1450–1465.

Localio, R., Cornell, J., and Mulrow, C. (2008). Much ado about avandia: the meta-analysis of rare events in the service of health policy. In: 7th International Conference on Health Policy Statistics.

Luo, S., Chen, Y., Su, X., and Chu, H. (2014). mmeta: An R package for multivariate meta-analysis. *Journal of Statistical Software*, **56**, 11.

Nissen, S. E. and Wolski, K. (2007). Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *New England Journal of Medicine*, **356**, 2457–2471.

Rücker, G., Schwarzer, G., Carpenter, J., and Olkin, I. (2009). Why add anything to nothing? the arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine*, **28**, 721–738.

Schmid, C. H., Lau, J., McIntosh, M. W., and Cappelleri, J. C. (1998). An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Statistics in Medicine*, **17**, 1923–1942.

Shuster, J. J. (2010). Empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine*, **29**, 1259–1265.

Stoto, M. A. (2015). Drug safety meta-analysis: Promises and pitfalls. *Drug Safety*, **38**, 233–243.

Sutton, A. J. and Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, **10**, 277–303.

Sweeting, J. M., Sutton, A. J., and Lambert, P. C. (2004). What to add to nothing? use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*, **23**, 1351–1375.

Tian, L., Cai, T., Pfeffer, M. A., Piankov, N., Cremieux, P., and Wei, L. J. (2009). Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2× 2 tables with all available data but without artificial continuity correction. *Biostatistics*, **10**, 275–281.

Warn, D. E., Thompson, S. G., and Spiegelhalter, D. J. (2002). Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. *Statistics in Medicine*, **21**, 1601–1623.

Whitehead, A. and Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine*, **10**, 1665–1677.

Xie, M., Kolassa, J., Liu, R., and Liu, D. (2014). Does a zero-total-event study contain information for inference of odds ratio in meta-analysis? Technical report, Rutgers University.

Xie, M., Singh, K., and Strawderman, W. E. (2011). Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association*, **106**.

Yang, G., Liu, D., and Xie, M. (2012). Tian's exact meta-analysis method as a special example under the general framework of combining cds. Research note.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 285–309
http://www.ssca.org.in/journal

Mixtures of Linear Regressions with Measurement Error in the Response, with an Application to Gamma-Ray Burst Data

Xiaoqiong Fang[1], Andy W. Chen[2] and Derek S. Young[3]
[1]*Corporate & Investment Bank, J.P. Morgan*
Brooklyn, New York, USA
[2]*School of Business, Government, and Economics, Seattle Pacific University*
Seattle, Washington, USA
[3]*Dr. Bing Zhang Department of Statistics, University of Kentucky*
Lexington, Kentucky, USA

Received: 29 April 2024; Revised: 23 July 2024; Accepted: 30 August 2024

Abstract

Gamma-ray bursts are intense, energetic explosions of gamma rays that are usually accompanied by an afterglow, which is a longer-lived emission that is detected at longer wavelengths, like X-ray, infrared, and radio. Classic gamma-ray burst data is often analyzed using some sort of regression model (*e.g.*, linear, piecewise linear, or a broken-power law model) to relate the flux of the burst to the time since the event. While these models may provide good fits, there is also often a "flaring" phenomena that tends to noticeably deviate from the fitted model. One way we can characterize such a phenomena relative to the underlying general trend is through a mixture-of-regressions model. Some applications in astronomy, like color-luminosity relations for field galaxies, are known to have the variables in the models prone to both intrinsic scatter and measurement error. This assumption is also tenable for gamma-ray burst data where the variance of heteroscedastic measurement errors can be reasonably known. Thus, we introduce a mixture-of-linear-regressions model where the variance of the measurement error is roughly known. Estimation is accomplished using an expectation-maximization (EM) algorithm framework with a weighted least squares estimator that was developed for the non-mixture setting. The finite-sampling behavior of our proposed model's estimates is examined by a simulation study. We also demonstrate the efficacy of this approach on a dataset involving the flux measurements of gamma-ray bursts, where the variance of the measurement error for the flux measurements (the response) are known. Our results for this data problem are compared with estimates obtained using other traditional models, including the linear regression model and the mixture-of-linear-regressions model.

Key words: Astrostatistics; Bootstrap; EM algorithm; Finite mixture model; Intrinsic scatter; Weighted least squares.

AMS Subject Classifications: 62J05, 62P99

Corresponding Author: Derek S. Young
Email: derek.young@uky.edu

1. Introduction

Variability is an inherent part of the results of measurements and of the measurement process. *Measurement error models*, also called *errors-in-variables models*, account for the difference between a measured value of a quantity and its true value. The effect of such measurement error and how to incorporate it into a statistical model has been long investigated, with authoritative texts devoted to this topic, including Fuller (1987), Carroll *et al.* (2006b), and Buonaccorsi (2010). Some issues that arise due to the presence of measurement error include bias in parameter estimation for statistical models, loss of power, and masking the features of the data, thus making graphical model analysis difficult. Specifically, the text by Carroll *et al.* (2006b) covers measurement error in nonlinear models, with a special focus on bias reduction, also called *approximate consistency*. For linear regression models with measurement error in the predictors, it can cause an underestimate of the slope coefficients, known as *attenuation bias*. In nonlinear models, the direction of the bias is likely to be more complicated as treated in Carroll *et al.* (2006b). Such biases can of course lead to a loss of power as well as mask certain important features of the data.

The statistical analysis of data with measurement error has a long history, especially in econometrics, with Frisch (1935) being one of the earliest references. Measurement error models are also employed in other diverse research areas, including nutrition (Carroll *et al.*, 2006a; Murillo *et al.*, 2019), finance (Carmichael and Coën, 2008; Maddala and Nimalendran, 1996), and astrostatistics (Kelly, 2007, 2012). With respect to astronomical research, measurement error problems are widely employed due to the presence of *intrinsic scatter*, a type of measurement error regarding variations in the physical properties of astronomical sources that are not completely captured by the variables included in the (regression) model. Feigelson and Babu (1992) provided an early introduction to measurement error models for use in astronomical regressions. Morrison *et al.* (2000) studied galaxy formation with a large survey of stars in the Milky Way using star velocities, which contained heteroscedastic measurement errors. To verify galaxy formation theories, one can estimate the density function from contaminated data that are effective in unveiling the numbers of bumps or components. Kelly (2007) described a Bayesian method to account for measurement errors in linear regression of astronomical data. In another study, Andrae (2010) presented an overview of different methods for error estimation that are applicable to both model-based and model-independent parameter estimates in astronomy.

The focus of the present work will be on developing a model for gamma-ray bursts (GRBs), where we relate the flux of the burst to the time since the event. The flux measurement is prone to both intrinsic scatter and measurement error, where the variance of the measurement errors are available. Moreover, there is a "flaring" phenomena that tends to noticeably deviate from traditional models that are fit to the data; *e.g.*, linear regression models. We propose a novel mixture-of-linear-regressions model with measurement error in the response variable to characterize both the flaring phenomena relative and the underlying general trend, as well as incorporate the measurement error in the flux measurement.

In the non-mixture setting, many methods have been proposed for performing linear regression when intrinsic scatter and/or measurement error is present. Clutton-Brock (1967) proposed an *effective variance* method. Press *et al.* (1992) proposed a procedure for minimizing an *effective* χ^2*-statistic*. Stephens and Dellaportas (1992), Richardson and Gilks

(1993), Dellaportas and Stephens (1995), and Gustafson (2004) each developed Bayesian approaches for estimating measurement error models. Some methods specifically developed for and applied in astronomical research are the *bivariate correlated errors and intrinsic scatter* (BCES) estimator (Akritas and Bershady, 1996) and the FITEXY estimator (Press *et al.*, 1992).

Finite mixture models are used to characterize the presence of unobserved subpopulations (or latent classes) within an overall population. The theoretical, methodological, and computational developments concerning finite mixture models is expansive, and the application of such models have provided critical insights into problems spanning virtually every research discipline. We refer to the texts by Titterington *et al.* (1985), Lindsay (1995), McLachlan and Peel (2000), Frühwirth-Schnatter (2006), and Mengersen *et al.* (2011), as well as the numerous references therein. Mixture models have enjoyed a strong presence in a wide range of fields, spanning the biological, physical, and social sciences. In particular, they have been successfully used in agriculture, astrostatistics, bioinformatics, economics, engineering, marketing, healthcare, neuroscience, and psychology (McLachlan *et al.*, 2019). Some of the applications in astronomical research that use mixture models include classification of astronomical bodies, identification of contaminants in astronomical images, and clustering overlapping population of stars (Kuhn and Feigelson, 2019). These tasks are essential for the study of stars and planet formation as well as analyzing multi-band astronomical images (Feigelson *et al.*, 2021). There are also precedents with using mixture models in the analysis of GRBs. Tarnopolski (2019) analyzed different properties of GRBs from the Burst and Transient Source Experiment (BATSE) using mixtures of multivariate skewed distributions.

Research at the intersection of (finite) mixture models and measurement errors is fairly limited. Lindsay (1995) highlights examples where the joint distribution of observable variables (including the observed *surrogate variables*, which are the variables whose true values are subject to measurement error) has a mixture form. Richardson *et al.* (2002) provides a Bayesian treatment of mixture models in measurement error problems. For mixtures-of-linear-regressions models, measurement error has only been studied in the predictors. This model was introduced by Yao and Song (2015), who developed a deconvolution method to estimate the observed surrogates and employed a generalized expectation-maximization (GEM) algorithm (Dempster *et al.*, 1977) for performing maximum likelihood estimation. An extension of that work for the setting of mixtures of polynomial regressions was presented in Fang *et al.* (2023). The distinction with the contributions in the present paper is that we address the issue of measurement error in the response variable through a mixture structure.

This paper is organized as follows. In Section 2, we define the particular mixture model used in this study. The challenges with this model mostly concern estimation and inference, which are presented in Section 3. In particular, we extend the weighted least squares (WLS) estimator developed by Akritas and Bershady (1996), but in the context of our mixture model. In Section 4, we conduct a simulation study using our proposed algorithm. In Section 5, we perform a thorough analysis of a GRB dataset using our mixture model. We end with some concluding remarks in Section 6.

2. The model

We first consider the setup for the classic mixture-of-linear-regressions model. Suppose we have a random sample of response variables, $Y_1, \ldots, Y_n$, that are each measured with a vector of predictors, $\mathbf{X}_i = (1, X_{i,1}, \ldots, X_{i,p-1})^{\mathrm{T}}$, $p < n$, for $i = 1, \ldots, n$, such that the first entry is a 1 to accommodate an intercept. Let $\mathcal{Z}_i$ be a latent class variable with $\mathrm{P}(\mathcal{Z}_i = j|\mathbf{X}_i) = \lambda_j$ for $j = 1, \ldots, k$, where $\lambda_j > 0$ and $\sum_{j=1}^{k} \lambda_j = 1$. Given $\mathcal{Z}_i = j$, the relationship between a univariate observation Y_i and $\mathbf{X}_i$ is the linear regression model

$$Y_i = \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_j + \epsilon_j. \tag{1}$$

Here, $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$, where σ_j^2 is the error variance for class (component) j, and $\boldsymbol{\beta}_j = (\beta_{0,j}, \ldots, \beta_{p-1,j})^{\mathrm{T}}$ is the p-dimensional vector of regression coefficients. Therefore, unconditional on $\mathcal{Z}_i$, but conditional on $\mathbf{X}_i$, the Y_is follow the mixture distribution

$$Y_i \mid \mathbf{X}_i \sim \sum_{j=1}^{k} \lambda_j \mathcal{N}(\mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_j, \sigma_j^2). \tag{2}$$

Maximum likelihood estimation of mixtures of linear regressions is straightforward, and typically performed using an EM algorithm. Bayesian inference can easily be performed via classic MCMC algorithms. We refer to De Veaux (1989), Viele and Tong (2002), and Hurn *et al.* (2003) for sound treatments of both approaches, which can be implemented using, for example, the `R` package `mixtools` (Benaglia *et al.*, 2009).

Suppose now that we have additive measurement error in the response variable, which we can write using the following (additive) measurement error model:

$$Y_i^* = Y_i + \delta_i. \tag{3}$$

In the above, Y_i is the true response value, Y_i^* is the observed response variable (*i.e.*, the surrogate variable), and δ_i is the measurement error. The measurement error is assumed to be independent of the Y_i as well as to have zero mean and finite variance η_i^2. In classic measurement error models, including regression models where the measurement error occurs in the predictor, a stronger assumption of normality is usually imposed on the distribution of the δ_is. Regardless, the classic measurement error setting will seek out estimation of the variance, with such methods discussed in Carroll *et al.* (2006b). One may, however, have a known value of η_i^2s or be able to posit a good estimate. In the GRB data discussed, we can reasonably make this assumption through the reported errors in the flux measurement. Therefore, we consider the setting where we observe the following for the ith observation in the dataset:

$$(\mathbf{X}_i^{\mathrm{T}}, Y_i^*, \eta_i^2), \tag{4}$$

where the true response is assumed to arise from the mixture structure discussed above in (1) and (2).

In the non-mixture (*i.e.*, classic multiple linear regression) setting, we know that the ordinary least squares (OLS) estimator for $\boldsymbol{\beta}$ minimizes the residual sum of squares $\|\mathbf{Y} - \boldsymbol{\mathcal{X}}\boldsymbol{\beta}\|^2$, where $\mathbf{Y}$ is an n-dimensional vector consisting of the Y_is and $\boldsymbol{\mathcal{X}}$ is an $n \times p$ full-rank design matrix with ith row $\mathbf{X}_i^{\mathrm{T}}$. The OLS estimator is, thus, $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \left(\boldsymbol{\mathcal{X}}^{\mathrm{T}}\boldsymbol{\mathcal{X}}\right)^{-1} \boldsymbol{\mathcal{X}}^{\mathrm{T}}\mathbf{Y}$,

which is also equal to the maximum likelihood estimator (MLE) in this setting. In the mixture setting, when performing maximum likelihood estimation via an EM algorithm, the MLE for the jth component's regression coefficient is calculated in the M-step at the tth iteration of the algorithm as $\hat{\boldsymbol{\beta}}_j^{(t+1)} = \left(\boldsymbol{\mathcal{X}}^{\mathrm{T}}\mathbf{W}_j^{(t)}\boldsymbol{\mathcal{X}}\right)^{-1}\boldsymbol{\mathcal{X}}^{\mathrm{T}}\mathbf{W}_j^{(t)}\mathbf{Y}$. In this expression, $\mathbf{W}_j^{(t)}$ is an $n \times n$ diagonal matrix with ith entry equal to the posterior membership probability of the ith observation belonging to component j, which is determined through an application of Bayes' rule in the E-step. Note that the form of the $\hat{\boldsymbol{\beta}}_j^{(t+1)}$ is that of a WLS estimator with weighting matrix $\mathbf{W}_j^{(t)}$. If there is measurement error in the predictors, as in the setting considered by Yao and Song (2015), or in the response, as in the present consideration, then the MLE just discussed will be biased. In our measurement error setting, we can modify the WLS estimator above to reflect the WLS approach developed in Akritas and Bershady (1996) for the non-mixture setting. This is the approach developed in the next section.

3. Estimating method

3.1. A WLS-based estimate

The model presented in the previous section has non-constant error variance (heteroscedasticity) for each observation. Though WLS was employed in the previous section during estimation of the mixture-of-regression coefficients, WLS is a classic framework for addressing heteroscedasticity. By design, WLS allows one to assign individual weights to the observations, thus removing, or at least improving, the effects of heteroscedasticity. WLS is an example of the broader class of generalized least squares estimators (Aitken, 1935). The general idea of WLS is that less weight is given to those observations with a larger error variance, which forces the variance of the residuals to be constant.

Akritas and Bershady (1996) note that the optimal weight for each observation comprises both the corresponding random error variance and the intrinsic scatter (measurement error) variance. However, in a mixture-of-regressions setting, we also need to account for the uncertainty of component membership, so we incorporate the unobserved $\mathcal{Z}_{ij}$s into our method. Conditional on component membership k_i, we have

$$\begin{aligned} Y_i^* &= Y_i + \delta_i \\ &= \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_{k_i} + \epsilon_{i,k_i} + \delta_i \\ &= \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_{k_i} + \epsilon_{i,k_i}^*, \end{aligned}$$

where $\epsilon_{i,k_i} \sim \mathcal{N}(0, \sigma_{k_i}^2)$. With this setting, we may develop a WLS-type approach while working under the assumption that the variance of ϵ_{i,k_i}^* is independent of Y_i^*; see Akritas and Bershady (1996). However, we need estimates of the variance of ϵ_{i,k_i}^*. Under our assumptions, we have

$$\mathrm{Var}(\epsilon_{i,k_i}^*) = \mathrm{Var}(\epsilon_{\cdot,k_i}) + \eta_i^2. \tag{5}$$

Since $\mathrm{Var}(\epsilon_{\cdot,k_i})$ is unknown, $\mathrm{Var}(\epsilon_{i,k_i}^*)$ is also unknown. We can extend the algorithm of Akritas and Bershady (1996) combined with estimates obtained via an EM algorithm to estimate $\mathrm{Var}(\epsilon_{\cdot,1}), \ldots, \mathrm{Var}(\epsilon_{\cdot,k})$; see Algorithm 1.

As shown in Algorithm 1, an EM algorithm is employed in Step (1), and then WLS is used to adjust the regression coefficients in Step (5). The difference between the WLS-

Algorithm 1 WLS-based Algorithm

(1) Given the observed data $\left\{(\mathbf{x}_1^{\mathrm{T}}, y_1^*), \ldots, (\mathbf{x}_n^{\mathrm{T}}, y_n^*)\right\}$ and $\eta_1^2, \ldots, \eta_n^2$, obtain the mixture-of-regressions coefficient estimates $(\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \ldots, \hat{\boldsymbol{\beta}}_k^{\mathrm{T}})^{\mathrm{T}}$ using an EM algorithm.
(2) Calculate the residuals $R_{ij} = y_i^* - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_j$, for $i = 1, \ldots, n$ and $j = 1, \ldots, k$.
(3) Calculate the weighted mean of the residuals for each component membership

$$\bar{R}_{\cdot j} = \frac{\sum_{i=1}^n \hat{p}_{ij} R_{ij}}{\sum_{i=1}^n \hat{p}_{ij}},$$

where $\hat{p}_{ij}$ are the final posterior membership probabilities from the EM algorithm in Step (1).
(4) Obtain the estimates of $\mathrm{Var}(\epsilon_{\cdot,1}), \ldots, \mathrm{Var}(\epsilon_{\cdot,k})$ from

$$\widehat{\mathrm{Var}}(\epsilon_{\cdot,j}) = \frac{\sum_{i=1}^n \hat{p}_{ij}\left[\left(R_{ij} - \bar{R}_{\cdot j}\right)^2 - \eta_i^2\right]_+}{\sum_{i=1}^n \hat{p}_{ij}}.$$

(5) Set $\widehat{\mathrm{Var}}(\epsilon_{i,j}^*) = \hat{\sigma}_{ij}^{*2} = \widehat{\mathrm{Var}}(\epsilon_{\cdot,j}) + \eta_i^2$ and define $\boldsymbol{A}_j = \mathrm{diag}(\hat{\sigma}_{1j}^{*-2}\hat{p}_{1j}, \ldots, \hat{\sigma}_{nj}^{*-2}\hat{p}_{nj})$. Then, the WLS estimator based on the further weighting from the intrinsic scatter is

$$\widetilde{\boldsymbol{\beta}}_j = (\mathbf{X}^{\mathrm{T}}\boldsymbol{A}_j\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{A}_j\mathbf{Y}^*,$$

for $j = 1, \ldots, k$, where $\mathbf{Y}^* = (Y_1^*, \ldots, Y_n^*)^{\mathrm{T}}$ is the vector of observed response variables Y_i^*s.

based estimators, $\widetilde{\boldsymbol{\beta}}_1, \ldots, \widetilde{\boldsymbol{\beta}}_k$, and the MLEs from the mixture-of-regressions EM algorithm, $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_k$, will typically not be very large. The variance estimators from the classic mixture-of-regressions model will naturally be smaller than our corrected estimator, since the former excludes the variances from the response variable's measurement error. Notice in Step (3) that the weighted estimators of variances are obtained by subtracting the deviation of measurement error from the overall deviation. Thus, the value of $\left(R_{ij} - \bar{R}_{\cdot j}\right)^2 - \eta_i^2$ can be negative for some i or j, so we employ the usage of the hinge function for this difference; *i.e.*, $\left[\left(R_{ij} - \bar{R}_{\cdot j}\right)^2 - \eta_i^2\right]_+ = \left\{\left(R_{ij} - \bar{R}_{\cdot j}\right)^2 - \eta_i^2\right\} \vee 0$.

3.2. Asymptotic variance

Let $\boldsymbol{\psi}$ denote the vector of true unknown parameter values,

$$\boldsymbol{\psi} = \left(\lambda_1, \ldots, \lambda_{k-1}, \boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_k^{\mathrm{T}}, \sigma_1^2, \ldots, \sigma_k^2\right)^{\mathrm{T}}.$$

The asymptotic variance of the MLEs obtained via an EM algorithm in Step (1) of Algorithm 1 can be obtained by the inverse of the information matrix $\mathcal{I}(\boldsymbol{\psi})$ that appears in the asymptotic result

$$\sqrt{n}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \mathcal{I}^{-1}(\boldsymbol{\psi})\right).$$

However, likelihood functions for mixture models are often complicated, which translates to difficult calculations for the second derivatives of the likelihood function that comprise

$\mathcal{I}(\boldsymbol{\psi})$. Thus, other approaches are necessary (see Chapter 14 of Lange, 2010). For example, Efron and Hinkley (1978) suggested to use the observed Fisher information matrix instead. Later, Louis (1982) introduced a technique for computing the observed information by using calculations only done on the complete information when an EM algorithm is used.

The density for the k-component mixture-of-regressions model is

$$g(y_i \mid \mathbf{x}, \boldsymbol{\psi}) = \sum_{j=1}^{k} \lambda_j f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j),$$

where

$$f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j) = \frac{1}{\sigma_j} \phi\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j}{\sigma_j}\right)$$

is the probability density of the ith observation belonging to the jth component. Here, $\boldsymbol{\theta}_j = \left(\boldsymbol{\beta}_j^{\mathrm{T}}, \sigma_j\right)^{\mathrm{T}}$ is the vector of parameters of the jth component and $\phi(\cdot)$ is the density of the standard normal distribution. We can, thus, write out the observed data loglikelihood as

$$\ell_{\mathrm{O}}(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} \lambda_j f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j) \right\},$$

which can be augmented with the vector of each observation's unobserved component membership – $\mathbf{z}_i = (z_{i1}, \ldots, z_{ik})^{\mathrm{T}}$ such that $z_{ij} = \mathtt{I}\{\text{observation } i \text{ belongs to component } j\}$ – to construct the complete data loglikelihood

$$\ell_{\mathrm{C}}(\boldsymbol{\psi}) = \sum_{i=1}^{n} \sum_{j=1}^{k} \mathbf{z}_{ij} \log \left\{ \lambda_j f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j) \right\}.$$

The complete data is characterized through $\mathbf{s} = \{(\mathbf{x}_i^{\mathrm{T}}, y_i, \mathbf{z}_i^{\mathrm{T}}), i = 1, \ldots, n\}$. Since the $\mathbf{z}_i$ is unobserved, and hence "missing," use of an EM algorithm is appropriate. We forego stating the explicit E-step an M-step for this setting as it is quite standard in the mixture-of-regressions literature; see, for example, Benaglia *et al.* (2009).

To compute the observed information in the EM algorithm, let $S(\mathbf{s} \mid \boldsymbol{\psi})$ and $S((\mathbf{x}_i^{\mathrm{T}}, y_i) \mid \boldsymbol{\psi})$ be the complete data score function and observed data score function, respectively. Moreover, let $\mathcal{I}_{\mathbf{s}}(\boldsymbol{\psi})$ be the complete data information matrix; *i.e.*, the expected value of the negative of the Hessian of the complete data loglikelihood. Then, by differentiation, the observed data information matrix can be written as

$$\mathcal{I}(\hat{\boldsymbol{\psi}}) = \mathcal{I}_{\mathbf{s}}(\hat{\boldsymbol{\psi}}) - \left[\mathbb{E}_{\psi} \left\{ S(\mathbf{s} \mid \boldsymbol{\psi}) S^{\mathrm{T}}(\mathbf{s} \mid \boldsymbol{\psi}) \right\} + S\left\{ (\mathbf{x}_i^T, y_i) \mid \boldsymbol{\psi} \right\} S^{\mathrm{T}} \left\{ (\mathbf{x}_i^T, y_i) \mid \boldsymbol{\psi} \right\} \right] \Bigg|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}.$$

Thus, the asymptotic variance-covariance of the estimator $\hat{\boldsymbol{\psi}}$ can be calculated based on $\mathrm{Var}(\hat{\boldsymbol{\psi}}) = \mathcal{I}(\hat{\boldsymbol{\psi}})^{-1}$, and the estimated standard errors of the parameter estimates in $\hat{\boldsymbol{\psi}}$ are the square root of the diagonal entries of this matrix. Note that in the present setting, we are using the y_i^* in the role of the y_i that appear in the preceding formulas. Moreover, the MLE $\hat{\boldsymbol{\psi}}$ is actually based on the WLS estimators $\tilde{\boldsymbol{\beta}}_j$, $j = 1, \ldots, k$ in Step (5) of Algorithm 1, and not the $\hat{\boldsymbol{\beta}}_j$ calculated in Step (1); *i.e.*,

$$\hat{\boldsymbol{\psi}} = \left(\hat{\lambda}_1, \ldots, \hat{\lambda}_{k-1}, \tilde{\boldsymbol{\beta}}_1^{\mathrm{T}}, \ldots, \tilde{\boldsymbol{\beta}}_k^{\mathrm{T}}, \hat{\sigma}_1^2, \ldots, \hat{\sigma}_k^2 \right)^{\mathrm{T}}.$$

3.3. Bootstrap estimator for the standard errors

Even when estimation of $\boldsymbol{\psi}$ is trivial, estimation of standard errors (SEs) can be computationally burdensome, especially when measurement error is involved. One alternative strategy is to use the parametric bootstrap (Efron and Tibshirani, 1993; Davison and Hinkley, 1997), which theoretically should provide similar estimates to the standard errors compared to the method involving the information matrix. This has become especially useful for standard error estimation in mixture settings, as noted in Chapter 2 of McLachlan and Peel (2000).

Algorithm 2 Parametric Bootstrap for Standard Errors

(1) Find $\hat{\boldsymbol{\psi}}$ by implementing Algorithm 1 using the observed data $\{(\mathbf{x}_1, y_1^*), \ldots, (\mathbf{x}_n, y_n^*)\}$.
(2) Generate a bootstrap sample $\{(\mathbf{x}_1, y_1^{**}), \ldots, (\mathbf{x}_n, y_n^{**})\}$, where each y_i^{**} is a realization from the (conditional) mixture distribution $\sum_{j=1}^{k} \hat{\lambda}_j \mathcal{N}\left(\mathbf{x}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2\right)$.
(3) For each of y_i^{**}, generate the "observed" response by

$$y_i^{***} \quad = \quad y_i^{**} + \delta_i,$$

where $\delta_i \sim \mathcal{N}(0, \eta_i^2)$ is generated using the known variabilities $\eta_1^2, \ldots, \eta_n^2$.
(4) Find the estimate $\widetilde{\boldsymbol{\psi}}$ by implementing Algorithm 1 on $\{(\mathbf{x}_1, y_1^{***}), \ldots, (\mathbf{x}_n, y_n^{***})\}$.
(5) Repeat Steps (2) - (4) B times to generate the bootstrap sampling distribution $\widetilde{\boldsymbol{\psi}}^{(1)}, \widetilde{\boldsymbol{\psi}}^{(2)}, \ldots, \widetilde{\boldsymbol{\psi}}^{(B)}$.

Algorithm 2 outlines a parametric bootstrap to estimate standard errors in our mixture-of-regressions model when specifying measurement error in the response. After implementing Algorithm 2, the bootstrap variance-covariance matrix is easily computed as the sample variance-covariance matrix of the generated values $\widetilde{\boldsymbol{\psi}}^{(1)}$, $\widetilde{\boldsymbol{\psi}}^{(2)}, \ldots, \widetilde{\boldsymbol{\psi}}^{(B)}$. Thus, bootstrap standard errors are readily available. When performing a bootstrapping procedure in the mixture setting, one must be cognizant of the label switching problem, that is, we want to enforce a meaningful identifiability constraint for a particular analysis. For example, one could set $\beta_{11} < \ldots < \beta_{k1}$ (*i.e.*, a constraint on the slope for the first predictor in the model) or $\sigma_1 < \ldots < \sigma_k$. We will state the identifiability constraints used for our numerical work in the next section.

4. Numerical studies

We now study the finite sampling behavior of the proposed estimators for our mixture-of-regressions model with measurement error in the response. Our study considers mixtures of regressions with one or two predictors, as well as two or three components. The basic setting for our models involves *iid* data $(\mathbf{x}_i^{\mathrm{T}}, y_i, \eta_i)$, $i = 1, \ldots, n$ such that the response variable Y_i is drawn from the model

$$Y_i | \mathbf{X}_i = \mathbf{x}_i \sim \sum_{j=1}^{k} \lambda_j \mathcal{N}\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j, \sigma_j^2\right),$$

$$Y_i^* = Y_i + \delta_i,$$

where $\delta_i \sim \mathcal{N}(0, \eta_i^2)$ is the simulated measurement error in the response. To study the effect of the measurement error on the proposed estimator for mixtures of both simple and multiple linear regressions with different number of components, we consider the three component structures: well-separated (WS), moderately-separated (MS), and overlapping (OL). These three categorizations of separability were determined by considering component mean structures and error variances that yield varying degrees of overlap with the generated data. An explicit quantitative threshold was not employed to characterize if components are WS, MS, or OL, but rather a visual check on simulated datasets was employed to ascertain the appropriateness of the stated component structure. The 12 data-generating processes used to characterize these different structures are summarized in Table 1.

Table 1: The 12 models used for the simulation study

Model	Structure	$\boldsymbol{\beta}_1^{\mathrm{T}}$	$\boldsymbol{\beta}_2^{\mathrm{T}}$	$\boldsymbol{\beta}_3^{\mathrm{T}}$	σ_1^2	σ_2^2	σ_3^2	λ_1	λ_2
Mixtures of Simple Linear Regressions									
$M1$	WS	$(-10, 6)$	$(10, 2)$	—	4	1	—	1/2	—
$M2$	MS	$(5, 15)$	$(25, -15)$	—	4	1	—	1/2	—
$M3$	OL	$(5, 5)$	$(15, -5)$	—	4	1	—	1/2	—
$M4$	WS	$(-10, 6)$	$(10, 2)$	$(30, -5)$	4	1	9	1/3	1/3
$M5$	MS	$(5, 15)$	$(20, 20)$	$(25, -15)$	4	1	9	1/3	1/3
$M6$	OL	$(-10, 20)$	$(5, 5)$	$(15, -5)$	4	1	9	1/3	1/3
Mixtures of Multiple Linear Regressions									
$M7$	WS	$(-10, 6, 4)$	$(10, 2, 7)$	—	4	1	—	1/2	—
$M8$	MS	$(5, 15, 10)$	$(25, -15, -10)$	—	4	1	—	1/2	—
$M9$	OL	$(5, 5, 9)$	$(15, -5, 3)$	—	4	1	—	1/2	—
$M10$	WS	$(-10, 6, 4)$	$(10, 2, 7)$	$(30, -5, 10)$	4	1	9	1/3	1/3
$M11$	MS	$(5, 15, 10)$	$(20, 20, 5)$	$(25, -15, -10)$	4	1	9	1/3	1/3
$M12$	OL	$(5, 5, 9)$	$(15, -5, 3)$	$(-10, 20, 15)$	4	1	9	1/3	1/3

For each simulation condition, we randomly generated $B = 1000$ datasets for the sample sizes $n \in \{100, 250\}$. For each sample size, we generated the predictor variables as $X_{ij} \sim \mathcal{U}(0, 1)$, while different measurement errors for the response were considered for each mixture-of-regressions setting. The Monte Carlo samples for the 2-component mixtures of regressions were generated under the two conditions of $\eta_i^2 \sim \mathcal{U}(0, 0.1)$ and $\eta_i^2 \sim \mathcal{U}(2, 6)$. The Monte Carlo samples for the 3-component mixtures of regressions were generated under the two conditions of $\eta_i^2 \sim \mathcal{U}(0, 0.5)$ and $\eta_i^2 \sim \mathcal{U}(5, 10)$.

For each simulated dataset, we estimate the parameters $(\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_k^{\mathrm{T}}, \sigma_1^2, \ldots, \sigma_k^2)$ using Algorithm 1, and compare them with the estimates obtained via the "naïve" method, which simply ignores the measurement error; *i.e.*, estimation of the classic mixtures-of-regressions model without measurement error in the response. The performance of the proposed method under different conditions is assessed by calculating the mean squared error (MSE),

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}) = \frac{1}{B}\sum_{t=1}^{B}(\hat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta})^2,$$

where $\hat{\boldsymbol{\theta}}^{(t)}$ is the estimate of the parameter $\boldsymbol{\theta}$ based on the tth Monte Carlo sample and $\boldsymbol{\theta}$ is

the true value. The relative efficiencies based on the MSEs for the naïve method versus the proposed method are also calculated for all of the parameters.

4.1. Results for mixtures of simple linear regressions

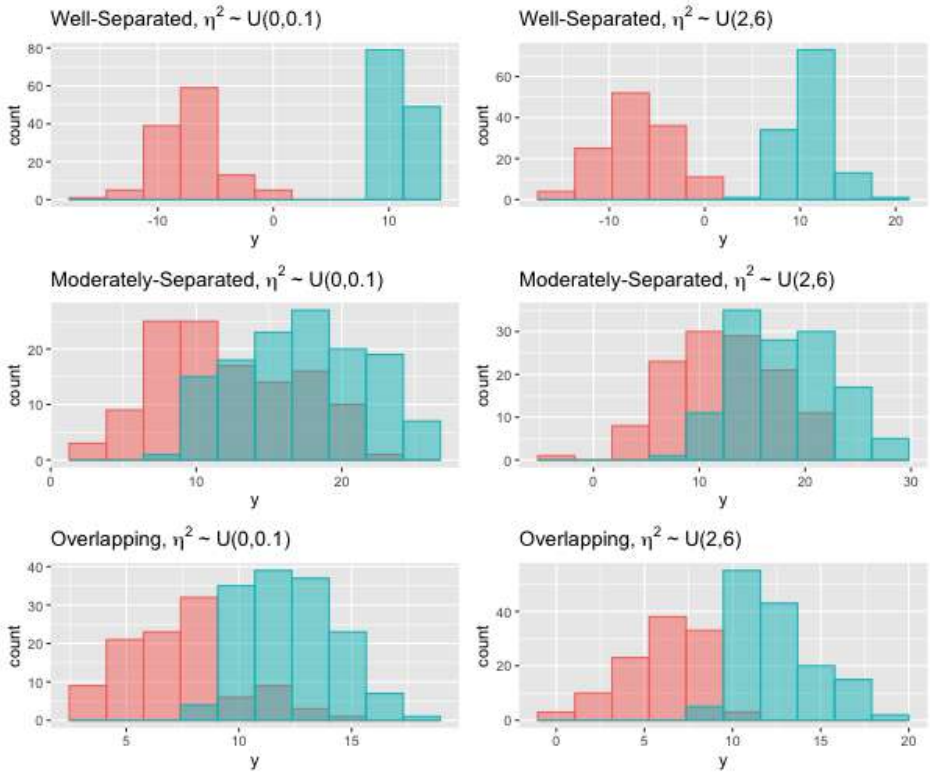


Figure 1: Histograms of observed response variables for 2-component mixtures of simple regression under different settings, with sample size $n = 250$

We first discuss the numerical results obtained for the 2-component mixtures where we have a single predictor. In particular, we first focus on models $M1$, $M2$, and $M3$ in Table 1. Figure 1 shows the histograms of observed responses y^* under different circumstances. Even though these are histograms of the unconditional distribution of the response with measurement error, it still gives an indication about the degree of separability that was incorporated in the mixtures-of-regressions structure. In the WS setting, there are two distinct regression relationships corresponding to the two different components. For the MS and OL settings, the two components have a greater degree of mixing, thus it is harder to identify to which component a certain data point belongs. Regardless, increasing the variance of the measurement errors forces the two components to be closer to each other, which compounds the ability to identify the distinct components.

Table 2 gives the MSEs and relative efficiencies (in parentheses) for the simulated datasets from models $M1$, $M2$, and $M3$. The values in the parentheses represent the relative efficiencies of MSEs for the naïve versus the proposed estimators. For example, the boldface value of 1.0552 means the MSE when estimating β_{21} using the naïve method is 1.0552 times the MSE when estimating the parameter using our proposed method. If the relative efficiency

Table 2: The MSEs and relative efficiencies (in parentheses) of the naïve estimators versus the proposed estimators for 2-component mixtures of simple linear regressions; models $M1$, $M2$, and $M3$

n	η_i^2	β_{10}	β_{11}	β_{20}	β_{21}	σ_1^2	σ_2^2
Well-Separated Components							
100	$\mathcal{U}(0, 0.1)$	0.3531	1.0550	0.0801	0.2461	0.6722	0.0425
		(1.0002)	(1.0001)	(1.0019)	(1.0008)	(0.9843)	(1.0235)
250		0.1359	0.4356	0.0338	0.1000	0.2551	0.0177
		(1.0004)	(1.0003)	(1.0016)	(1.0025)	(0.9850)	(1.0895)
100	$\mathcal{U}(2, 6)$	0.6419	2.0757	0.3878	1.2180	8.2657	11.1670
		(1.0099)	(1.0121)	(1.0580)	(1.0551)	(1.8492)	(1.2782)
250		0.2442	0.7692	0.1616	0.4966	8.5673	12.1929
		(1.0171)	(1.0192)	(1.0499)	(1.0413)	(1.8948)	(1.2908)
Moderately-Separated Components							
100	$\mathcal{U}(0, 0.1)$	0.3684	1.1907	0.0943	0.3086	0.8366	0.0553
		(0.9994)	(0.9992)	(1.0020)	(1.0017)	(1.0389)	(1.0412)
250		0.1376	0.4311	0.0345	0.1184	0.3136	0.0234
		(1.0004)	(1.0022)	(1.0016)	(1.0032)	(1.8558)	(1.0260)
100	$\mathcal{U}(2, 6)$	0.8202	3.1092	0.4664	1.7427	7.7301	10.2705
		(1.0303)	(1.023)	(1.0611)	(1.0492)	(2.0686)	(1.2932)
250		0.2920	0.9428	0.1760	0.6098	7.9266	12.2029
		(1.0598)	(1.0514)	(1.0523)	**(1.0552)**	(2.1659)	(1.3049)
Overlapping Components							
100	$\mathcal{U}(0, 0.1)$	0.3920	1.3037	0.0988	0.4589	1.0774	0.0820
		(0.9990)	(0.9997)	(1.0027)	(1.0004)	(0.9799)	(0.9861)
250		0.1587	0.5338	0.0446	0.1836	0.3580	0.0319
		(0.9927)	(1.0026)	(0.9985)	(0.9916)	(0.9582)	(1.0240)
100	$\mathcal{U}(2, 6)$	1.3720	4.5647	0.8550	3.3583	7.0853	9.1205
		(1.6076)	(1.1515)	(1.4303)	(1.1468)	(2.9174)	(1.0341)
250		0.4532	1.8502	0.3732	1.6403	4.7926	11.0519
		(1.3647)	(0.9572)	(1.0541)	(0.8900)	(3.5687)	(1.3208)

is greater than 1, it means the MSE of proposed method is smaller, which leads to greater precision of the estimator. We note that label switching did not appear to be present since a check on the estimates of β_{10} and β_{20} showed that $\hat{\beta}_{10} < \hat{\beta}_{20}$ was met for each sample. Thus, no identifiability constraint had to be enforced for this set of simulations.

Overall, the proposed method appears to behave better than the naïve method with respect to their relative efficiencies since they are greater than 1. For estimating the variances $\text{Var}(\epsilon_{\cdot,j})$ when a larger value is used (*i.e.*, when $\sigma_1 = 2$ rather than $\sigma_2 = 1$), the average relative efficiency for the settings with measurement error $\mathcal{U}(2, 6)$ is greater than 2. When the measurement error is trivial, this translates to the behaviors of both methods being nearly the same. Thus, we can conclude that our proposed method behaves better when the measurement error is larger, which accounting for measurement error in such a circumstance is likely of greater importance. Note that because our proposed method only accommodates measurement error in the response after obtaining the maximum likelihood estimates via an

EM algorithm, there is no adjustment to the mixing proportion estimates; *i.e.*, $\hat{\lambda}$ is the same under both methods and, thus, the relative efficiency is necessarily 1.

When the sample size increases from 100 to 250, the MSEs decrease. Moreover, our proposed method shows improvement over the naïve method. If we expand the values of measurement error in the response, the MSEs become larger, however, the performance of the proposed method according to the relative efficiencies is better for the same sample size. It is reasonable to infer that, if we increase the measurement error, the estimators using our proposed method will not represent our true parameters as accurately as those with smaller measurement errors, but the performance of it will be much better than the naïve method, which simply ignores the measurement error term.

Table 3: The MSEs and relative efficiencies (in parentheses) of the naïve estimators versus the proposed estimators for 3-component mixtures of simple linear regressions; models $M4$, $M5$, and $M6$

n	η_i^2	β_{10}	β_{11}	β_{20}	β_{21}	β_{30}	β_{31}	σ_1^2	σ_2^2	σ_3^2
Well-Separated Components										
100	$\mathcal{U}(0,0.5)$	0.5330	1.5660	0.1870	0.4602	1.1617	3.5029	1.0515	6.1266	6.2885
		(1.0025)	(1.0012)	(1.0158)	(1.0089)	(0.9996)	(0.9982)	(0.9757)	(1.0225)	(0.9800)
250		0.2262	0.6790	0.0617	0.1904	0.4619	1.3618	0.5769	1.3600	3.0806
		(1.0030)	(1.0025)	(1.0071)	(1.0111)	(0.9987)	(0.9992)	(1.0280)	(1.0891)	(0.9848)
100	$\mathcal{U}(5,10)$	2.2853	7.9456	2.2967	5.6084	2.8218	8.8450	41.2184	119.5947	49.2994
		(1.0224)	(1.0170)	(1.0354)	(1.0261)	(1.0474)	(1.0461)	(1.5465)	(1.2127)	(1.8582)
250		0.5122	1.6757	0.4544	1.4282	0.8378	2.7573	33.7626	53.1254	25.0650
		(1.0230)	(1.0188)	(1.0258)	(1.0275)	(1.0260)	(1.0323)	(1.5797)	(1.2139)	(2.2608)
Moderately-Separated Components										
100	$\mathcal{U}(0,0.5)$	0.6619	2.5107	1.8705	4.6683	0.7329	2.0314	1.9033	61.7355	59.8475
		(0.9995)	(0.9969)	(1.0019)	(1.0037)	(0.9983)	(0.9998)	(0.9599)	(0.9631)	(1.0482)
250		0.2350	0.7756	0.5871	1.7277	0.1041	0.2834	0.8868	61.5826	64.6231
		(1.0031)	(1.0010)	(1.0009)	(0.9993)	(1.0072)	(1.0119)	(1.0031)	(0.9576)	(1.0485)
100	$\mathcal{U}(5,10)$	6.1955	40.8465	7.4054	18.3020	11.4807	42.4403	51.5176	14.0030	167.5460
		(1.0728)	(1.0526)	(1.0209)	(1.0033)	(1.0613)	(1.0391)	(1.5418)	(2.2821)	(1.4550)
250		0.9832	5.4059	1.9183	4.3903	2.0748	5.7883	32.2413	5.4198	151.2731
		(1.0403)	(1.0278)	(0.9849)	(0.9899)	(1.0139)	(1.0287)	(1.6778)	(1.8687)	(1.4886)
Overlapping Components										
100	$\mathcal{U}(0,0.5)$	2.0540	6.7647	1.8261	5.7137	0.2518	1.1633	12.227	6.7275	0.9974
		(0.9966)	(0.9952)	(0.9980)	(0.9902)	(1.0026)	(1.0309)	(0.9672)	(0.9896)	(1.1254)
250		0.5923	2.2360	0.3429	1.7953	0.0773	0.3423	3.8101	1.9859	0.6644
		(0.9976)	(0.9932)	(0.9970)	(0.9876)	(1.0037)	(0.9989)	(0.9477)	(0.9813)	(1.2213)
100	$\mathcal{U}(5,10)$	10.0582	35.1593	24.5870	38.8456	7.3339	16.6268	49.3850	42.0632	71.0176
		(1.0882)	(1.0617)	(1.0170)	(1.0321)	(1.1401)	(1.1119)	(2.0085)	(1.6594)	(1.2376)
250		4.6846	10.0172	10.7153	18.6601	3.3252	6.3234	31.3635	36.5494	60.9078
		(1.0657)	(1.0444)	(1.0185)	(1.0413)	(1.1256)	(1.1043)	(2.2489)	(1.7373)	(1.2545)

In Table 3 we report the MSEs and relative efficiencies (in parentheses) for our simulated datasets from the 3-component setting. The models for this part of our discussion are $M4$, $M5$, and $M6$ in Table 1. Label switching was present when comparing the bootstrap samples for the moderately-separated cases. This was diagnosed by first noting that the MSEs appeared to be fairly large for some parameters when the measurement error is large. For example, the MSE of β_{21} for the moderately-separated setting with $\eta_i^2 \sim \mathcal{U}(5,10)$ and sample size $n = 100$ was first found to be 133.1943, a value much larger than expected. Since the values of β_{20} and β_{30} are close to each other, simply using the identifiability constraint

$\beta_{10} < \beta_{20} < \beta_{30}$ is not enough. To make the components distinct with each other and correct the label switching in the simulation, we imposed the identifiability constraint of β_{10} being the smallest estimated intercept of the three components and $\beta_{21} > \beta_{31}$.

When the number of components increase, the MSEs become noticeably larger since the model is growing in complexity. With a heavier-parameterized model, the estimation becomes more challenging. When we increase the sample size and decrease the variance of the measurement errors in the response, the MSEs of the unknown parameters becomes smaller. Similarly, the relative efficiencies show that for the case with larger sample size and bigger measurement error, our proposed method performs better than naïve method. For overlapping and moderately-separated cases, the MSEs are fairly large for certain parameters with large measurement error (*e.g.*, with variances $\eta_i^2 \sim \mathcal{U}(5, 10)$), since the three components are subject to heavy mixing and it becomes difficult to consistently distinguish different components, thus leading to greater uncertainty in the estimators.

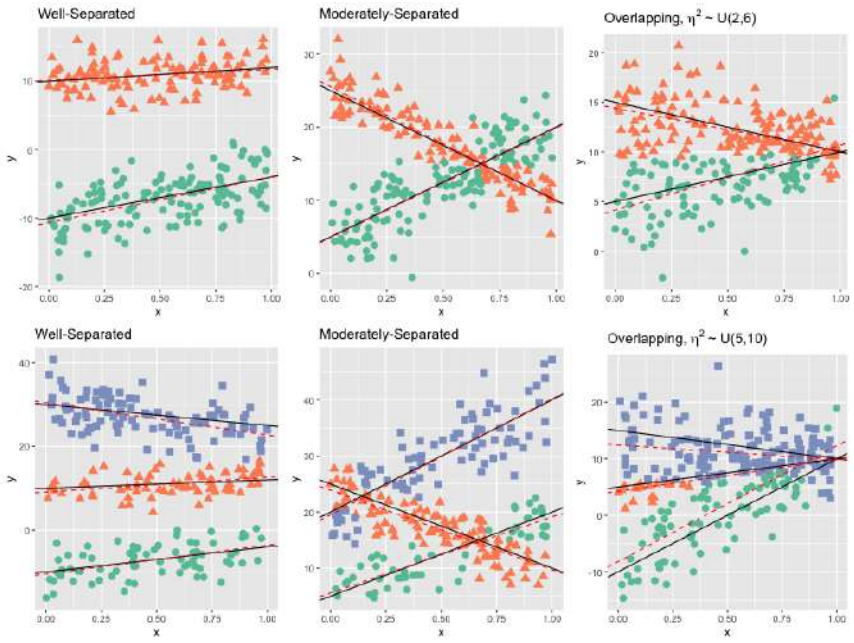


Figure 2: Scatter plots for datasets generated from each of the models $M1 - M6$, inclusive (sample size $n = 250$), where dashed red lines are the estimates obtained using Algorithm 1 and solid black lines are the lines based on the true parameters

Figure 2 shows scatterplots of datasets generated from each of the six settings (models $M1 - M6$) for mixtures of simple linear regressions with measurement error in the response. Different colors and shapes indicate from which component each observation was generated. The dashed red lines are the estimates obtained from our proposed method outlined in Algorithm 1. The solid black lines are the lines based on the true parameters. According to the scatterplots, the proposed method fits well in all settings as the dashed red lines (estimates) are similar to the solid black lines (truth). Moreover, based on the relative efficiencies reported earlier, it improves the performance of estimating parameters when compared to the naïve method. Overall, these results are consistent with demonstrating the

efficacy of our proposed method as a way to incorporate measurement error in the response when the underlying data come from a mixture-of-regressions setting.

4.2. Results for mixtures of multiple linear regressions

We next consider the 2-component mixtures of multiple linear regressions with measurement errors, which correspond to the models $M7$, $M8$, and $M9$ in Table 1. Figure 3 shows 3d scatterplots of data simulated from each of these models, where different colors represent to which component each data point belongs. In the well-separated case, the two components are very well-separated, thus making it very easy to distinguish to which component each point belongs. For the moderately-separated and overlapping cases, there are some areas where the two components are mixing, which is where we would expect to have the greatest uncertainty as to how to classify those observations if we were estimating the underlying model.

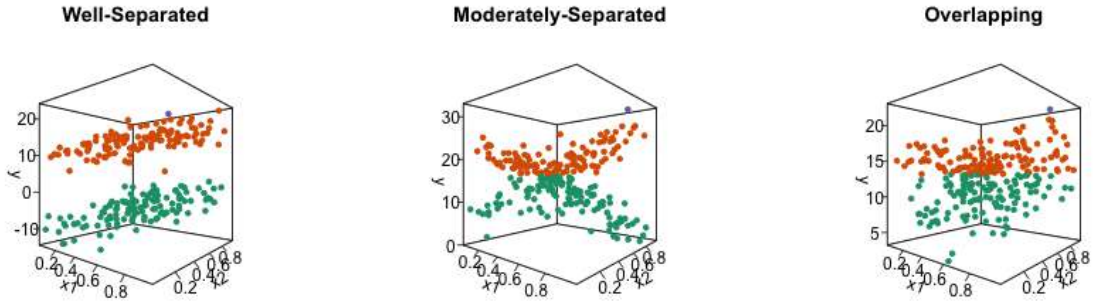


Figure 3: 3d scatterplots of the three different component structures for the 2-component mixtures of multiple linear regressions with sample size $n = 250$ and measurement error $\eta_i^2 \sim \mathcal{U}(2, 6)$ for the response

In Table 4, we report the MSEs and relative efficiencies (in parentheses) for our simulated datasets from the models $M7$, $M8$, and $M9$. Label switching did not appear to be present since the identifiability constraint $\beta_{10} < \beta_{20}$ is satisfied for all bootstrap estimates. The overall behavior of these three 2-component mixtures of multiple linear regressions are similar to those of the 2-component mixtures of simple linear regressions. When we increase the sample size from 100 to 250, the MSEs become smaller and the relative efficiencies improve. Meanwhile, because we add the predictor X_{i2}, the models are more parameterized than when the components are simple linear regressions, thus making the estimation more challenging, especially when the components are overlapping. For example, with an overlapping component, with large measurement errors (variances $\eta_i^2 \sim \mathcal{U}(2, 6)$), and with a sample size of $n = 100$, the boldface value in Table 4 is the MSE of the slope parameter for X_{i2}, β_{12}. This value of 19.2855 is a value much larger than the corresponding setting with simple linear regression components. Naturally, when increasing the number of predictor variables in settings with overlapping components, the increase in the MSEs reflect the greater difficulty in being able to estimate the true parameters.

Table 4: The MSEs and relative efficiencies (in parentheses) of the naïve estimators versus the proposed estimators for 2-component mixtures of multiple linear regressions; models $M7$, $M8$, and $M9$

n	η_i^2	β_{10}	β_{11}	β_{12}	β_{20}	β_{21}	β_{22}	σ_1^2	σ_2^2
Well-Separated Components									
100	$\mathcal{U}(0, 0.1)$	0.5943	1.0542	0.9975	0.1654	0.2692	0.2721	0.6641	0.0429
		(0.9997)	(0.9998)	(0.9994)	(1.0005)	(0.9998)	(1.0009)	(0.9711)	(0.9570)
250		0.2344	0.3588	0.4091	0.0571	0.1029	0.1001	0.2772	0.0181
		(1.0000)	(0.9999)	(1.0000)	(1.0011)	(1.0025)	(0.9999)	(0.9924)	(1.0444)
100	$\mathcal{U}(2, 6)$	1.1410	1.8997	1.9631	0.7192	1.2127	1.2058	7.8854	11.2173
		(1.0242)	(1.0242)	(1.0200)	(1.0356)	(1.0453)	(1.0334)	(1.8798)	(1.2486)
250		0.4703	0.7942	0.7993	0.2633	0.4658	0.4882	8.5387	12.1649
		(1.0264)	(1.0361)	(1.0163)	(1.0345)	(1.0322)	(1.0419)	(1.8905)	(1.2733)
Moderately-Separated Components									
100	$\mathcal{U}(0, 0.1)$	0.6763	1.2041	1.2587	0.1686	0.3052	0.3084	0.8869	0.0652
		(1.0005)	(0.9991)	(0.9999)	(1.0002)	(0.9971)	(1.0026)	(0.9788)	(0.9522)
250		0.2414	0.4074	0.4098	0.0721	0.1136	0.1233	0.3040	0.0223
		(1.0003)	(1.0008)	(0.9994)	(0.9985)	(0.9973)	(1.0015)	(0.9714)	(0.9977)
100	$\mathcal{U}(2, 6)$	1.5240	2.9314	2.8395	0.9511	2.1858	1.6698	6.8091	10.6683
		(1.0258)	(1.0379)	(1.0185)	(1.0542)	(1.0472)	(1.0416)	(2.1127)	(1.2768)
250		0.5835	0.9993	0.9861	0.3567	0.5889	0.6688	7.0279	11.6471
		(1.0181)	(1.0142)	(1.0195)	(1.0337)	(1.0452)	(1.0421)	(2.1744)	(1.2959)
Overlapping Components									
100	$\mathcal{U}(0, 0.1)$	1.2866	2.3647	1.8994	0.4989	1.0341	0.7241	1.2633	0.2225
		(1.0030)	(1.0012)	(1.0024)	(1.0071)	(1.0004)	(1.0027)	(0.9695)	(0.9831)
250		0.3486	0.6162	0.5630	0.0847	0.1826	0.1721	0.3895	0.0461
		(1.0041)	(1.00021)	(1.0033)	(1.0082)	(1.0007)	(1.0029)	(0.9744)	(0.9672)
100	$\mathcal{U}(2, 6)$	10.2329	18.2687	**19.2855**	6.5878	12.7481	7.5360	6.6059	16.4143
		(1.0901)	(1.0874)	(1.1339)	(1.1815)	(1.1073)	(1.1758)	(2.4594)	(1.1897)
250		3.0658	4.1279	3.3197	1.9051	2.8471	1.9667	6.3793	12.4284
		(1.0561)	(1.0346)	(1.0758)	(1.0923)	(1.0537)	(1.0557)	(2.2934)	(1.2622)

Finally, in Table 5, we report the MSEs and relative efficiencies (in parentheses) for our simulated datasets from the models $M10$, $M11$, and $M12$. The overall behavior of these three 3-component mixtures of multiple linear regressions are similar to those of the 3-component mixtures of simple linear regressions. When we increase the sample size from 100 to 250, the MSEs become markedly smaller and the relative efficiencies improve. Meanwhile, adding the predictor X_{i2} creates heavier-parameterized model than when the components are simple linear regressions, thus making the estimation more challenging. This, again, is especially the case when the components are overlapping. Naturally, when increasing the number of predictor variables in settings with overlapping components, the increase in the MSEs reflect the greater difficulty in being able to precisely estimate the true parameters.

4.3. Summary of simulation results

The combination of simulation conditions we considered in this section is fairly broad in ascertaining the applicability and robustness of our method. The conditions considered are more extensive relative to the most closely-related works of Yao and Song (2015) and Fang *et al.* (2023). The former only considered a two-component mixture structure ($k = 2$) in a

Table 5: The MSEs and relative efficiencies (in parentheses) of the naïve estimators versus the proposed estimators for 3-component mixtures of multiple linear regressions; models $M10$, $M11$, and $M12$

n	η_i^2	β_{10}	β_{11}	β_{12}	β_{20}	β_{21}	β_{22}	β_{30}	β_{31}	β_{32}	σ_1^2	σ_2^2	σ_3^2
Well-Separated Components													
100	$\mathcal{U}(0,0.5)$	1.2136	1.9885	3.9334	0.3336	0.5340	0.5208	2.2203	3.8362	3.7758	1.0387	12.3354	5.5941
		(1.0076)	(1.0131)	(0.9976)	(1.0177)	(1.0233)	(1.0107)	(0.9989)	(0.9997)	(0.9989)	(0.9331)	(0.9881)	(0.9505)
250		0.3811	0.6459	0.6263	0.1039	0.1737	0.1823	0.8305	1.4460	1.3632	0.4005	0.0372	2.1773
		(1.0026)	(1.0021)	(1.0029)	(0.9925)	(0.9926)	(1.0119)	(1.0002)	(1.0000)	(0.9989)	(1.0085)	(1.8628)	(0.9591)
100	$\mathcal{U}(5,10)$	3.2963	5.2986	5.0973	4.5584	8.1695	8.1475	5.7028	8.8767	12.3215	85.2660	158.6470	74.7738
		(1.0482)	(1.0178)	(1.0333)	(1.0294)	(1.0043)	(1.0053)	(1.0416)	(1.0286)	(1.0205)	(1.3193)	(1.2135)	(1.5328)
250		0.9410	1.7008	1.6351	0.7914	1.3628	1.3607	1.5043	2.6883	2.6383	34.6107	45.3006	24.9767
		(1.0164)	(1.0207)	(1.0115)	(1.0113)	(1.0046)	(1.0110)	(1.0186)	(1.0253)	(1.0261)	(1.5534)	(1.1457)	(2.2178)
Moderately-Separated Components													
100	$\mathcal{U}(0,0.5)$	1.9663	4.8574	4.1719	1.2241	2.8285	2.2239	4.9887	7.7760	6.6005	7.5266	9.7945	12.0505
		(1.0006)	(1.0011)	(0.9981)	(1.0009)	(1.0036)	(1.0086)	(1.0005)	(0.9981)	(0.9991)	(0.9960)	(1.0163)	(0.9652)
250		0.4809	1.1818	0.9333	0.1374	0.2160	0.2007	1.4692	2.5039	2.1921	0.6890	0.0400	3.3639
		(0.9995)	(0.9986)	(0.9982)	(1.0164)	(1.0111)	(1.0111)	(1.0011)	(1.0003)	(0.9995)	(0.9518)	(1.8602)	(0.9606)
100	$\mathcal{U}(5,10)$	12.9275	33.8055	22.2632	5.2212	15.3337	8.8258	18.1433	37.4492	25.1159	112.7497	70.9902	50.3589
		(1.0199)	(1.0141)	(1.0321)	(1.0872)	(1.0573)	(1.0569)	(1.0159)	(1.0131)	(1.0092)	(1.4687)	(1.2221)	(1.8285)
250		2.0909	4.3709	3.5039	1.2803	1.8202	1.7139	3.6181	6.3859	4.9530	37.6301	47.8919	23.1817
		(1.0224)	(1.0296)	(1.0131)	(1.0179)	(1.0284)	(1.0160)	(0.9911)	(0.9735)	(0.9864)	(1.6905)	(1.1922)	(2.4719)
Overlapping Components													
100	$\mathcal{U}(0,0.5)$	10.3035	20.6498	15.4182	16.7390	21.5233	33.0813	3.3270	6.4835	4.3271	20.3703	8.4015	1.2845
		(1.0063)	(1.0067)	(0.9903)	(1.0017)	(0.9917)	(1.0050)	(0.9996)	(1.0079)	(1.0006)	(0.9868)	(1.0189)	(1.1515)
250		2.0177	3.6305	2.8213	1.6731	2.9781	2.3034	0.2443	0.5121	0.4392	5.4233	2.8357	0.1046
		(0.9972)	(1.0178)	(1.0030)	(0.9998)	(0.9955)	(0.9979)	(1.0065)	(1.0029)	(1.0073)	(0.9485)	(0.9773)	(1.4291)
100	$\mathcal{U}(5,10)$	21.8372	38.7232	31.4980	40.1613	50.7183	46.2146	12.5149	26.3528	18.1346	29.0741	26.5541	47.2962
		(1.1114)	(1.0869)	(1.0810)	(1.0269)	(1.0467)	(1.0616)	(1.1389)	(1.1391)	(1.1859)	(2.4170)	(1.6763)	(0.8082)
250		11.8025	17.7110	15.0034	36.2944	43.8553	25.0780	9.3447	17.8059	10.7296	24.4411	31.1152	51.5546
		(1.0978)	(1.0866)	(1.0974)	(0.9999)	(1.0217)	(1.0725)	(1.1619)	(1.1165)	(1.1073)	(2.6009)	(1.7296)	(0.9340)

single predictor. The latter considered two-component mixture structures ($k = 2$), but where the components could be linear, quadratic, or cubic functions of a single predictor. In our simulation work, we considered two-component and three-component mixtures ($k = 2, 3$), each with one or two predictors. The parameters for the underlying regression components are then selected to be well-separated, moderately-separated, or overlapping, yielding the 12 models in Table 1. Moreover, we considered two measurement error structures and two sample sizes, further demonstrating the performance of our methods on a variety of models.

In general, the results reported in this section are consistent with results typically seen in simulations involving mixtures. When the components are well-separated, the results tend to be more stable compared to moderately-separated and overlapping settings. This, of course, follows from the variables in both moderately-separated and overlapping component models being harder to identify. Meanwhile, for the same model with the same component setting (*i.e.*, well-separated, moderately-separated, or overlapping), an increase in the sample size yields a decrease in the MSE, while an increase in the the variances of the measurement error increases the MSE.

Generally speaking, the MSEs of well-separated components are the smallest among the three different types of component settings. When we assumed a smaller measurement error, the MSEs are almost unanimously smaller, which makes sense due to smaller measurement error infusing smaller variability in the response. Overall, 2-component models had better results than the three-component models. For example, for a 3-component heavily-overlapping mixture model with measurement error $\mathcal{U}(5, 10)$ and sample size of 100, the MSEs of $\boldsymbol{\beta}_2^{\mathrm{T}} = (15, -5, 3)$ are $(40.1613, 50.7183, 46.2146)$ (see Table 4), while the 2-component heavily overlapping mixture model with measurement error $\mathcal{U}(2, 6)$ and sample

size 100 for the same $\boldsymbol{\beta}_2^{\mathrm{T}}$ has MSEs of $(6.5878, 12.7481, 7.5360)$ (see Table 5).

Our routine developed to estimate the mixture models under consideration do occasionally encounter some numerical issues, especially for the 3-component overlapping models. Sometimes, bad solutions (*i.e.*, estimates that are clearly far away from the true parameter values) were obtained. This would occasionally occur even after starting the algorithm from multiple random starting values. For practical purpose, in the 3-component simulated datasets with $B = 1000$, we trimmed $40(\approx 4\%)$ of the datasets that yield the largest deviations from the true parameter value for any single estimates from $\boldsymbol{\beta}$ vectors. After omitting those results, the MSEs were much more consistent with what was observed under the other conditions. This strategy has been employed for other simulations involving mixtures with complex structures; see, for example, Young (2014).

5. Example: Gamma-ray burst data

GRBs are key observations in gamma-ray astronomy, as they are extremely energetic explosions that occur at random times in distant galaxies. Since the Big Bang, they are considered the brightest electromagnetic events known to occur in the universe. The bursts can last from ten milliseconds to several hours. These phenomena are still the subject of intense research, but some theories suggest they arise during the birth of black holes or a massive super-giant's collapse. See the review article by Piran *et al.* (2013).

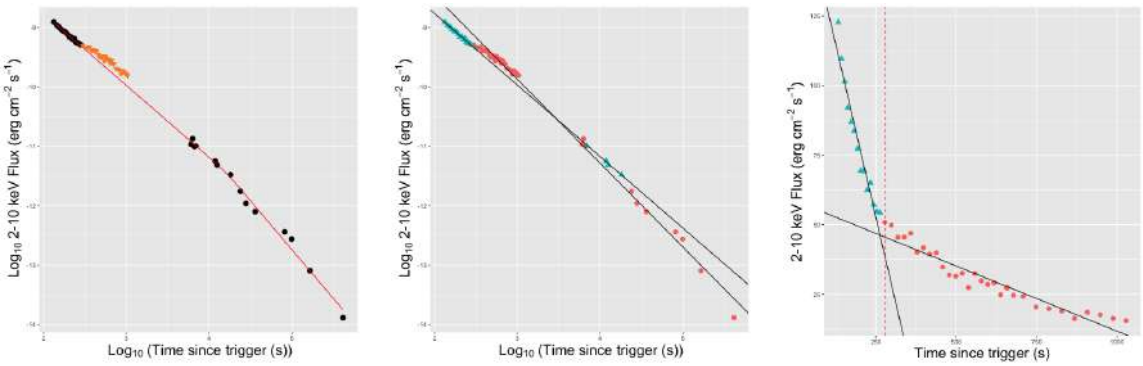
The launch of the Swift observatory (Gehrels *et al.*, 2004) modernized how we observe GRBs. The Swift observatory, which has collected and made available copious amounts of GRB data, provides rapid notification of GRB triggers to the ground using a highly-sensitive Burst Alert Telescope (BAT; Barthelmy, 2004). It also makes panchromatic observations of the burst and its afterglow. On May 25th, 2005, the Swift BAT was triggered and located GRB050525a[1] (Blustin *et al.*, 2006), the significance being that this was the first bright, low-redshift burst to have been observed using the observatory. The X-ray decay 'light curve' of GRB050525a that was obtained includes both *photo-diode* (PD) mode ($T < 2000$s) and *photon-counting* (PC) mode ($T > 2000$s) data. The data are plotted in Figure (4(a)), and like many astronomical datasets, the GRB observations suffer from measurement error due to the detection technique used.

The GRB050525a dataset consists of $n = 63$ brightness measurements in the 0.4 – 4.5 keV spectral band at times ranging from 2 minutes to 5 days after the burst. During this period, the brightness faded by a factor of 100,000. Due to the wide range in times and brightness, most analysis is done using logarithmic variables. The observations in the dataset are: time since trigger (in seconds), X-ray flux (in units of 10^{-11} erg/cm^2/s, $2-10$ keV), and the variability of the measurement error of the flux based on detector signal-to-noise values.

Blustin *et al.* (2006) fit the data with a power-law model; *i.e.*, a linear regression model. However, they note systematic deviations of the residuals at certain time points, which they attempt to capture using temporal breaks, resulting in what they call a broken power-law model; *i.e.*, a piecewise linear regression model. The data and best-fit line using a single breakpoint are shown in Figure 4(a). Blustin *et al.* (2006) note that the power-law fit

[1]The naming convention for GRBs is "GRByymmdd", where a subsequent letter (*i.e.*, a, b, c, *etc.*) denotes the observation on a day when multiple GRBs occurred.

of the pre-brightening PD mode data ($T < 280$s) extrapolates well to the pre-break PC mode data. They concluded that the brightening at about 280s in the PD mode data represents a flare in the X-ray flux, possibly similar to the sometimes much larger flares that are seen at early times in other bursts. The authors further note that the flux returns to the pre-flare decay curve prior to the start of the PC data.



(a) (b) (c)

Figure 4: Scatterplots of the GRB050525a data with (a) the best-fit line from a broken power-law model, (b) the estimated 2-component measurement error model fit, and (c) the estimated 2-component measurement error model fit on the PD mode data

Blustin *et al.* (2006) do not directly model the flaring points in their modeling. The flaring points are denoted by orange dots in Figure 4(a). In order to also capture the characteristic of the flaring part of this phenomena, we fit the data with a mixture-of-linear-regressions model, which can potentially identify separate regression models for the initial burst. Moreover, we can incorporate the reported variability of the measurement error of the flux through the model we developed in Section 2.

While we hypothesize that separate regression models could be appropriate for the initial burst and the remaining flux measurements, we will proceed to assess the number of components for the proposed mixture-of-linear-regressions model. We consider $k = 1, 2, 3, 4$ and select the best model according to results using the following model selection criteria: Akaike's information criterion (AIC; Akaike, 1973), the Bayesian information criterion (BIC; Schwarz, 1978), the Integrated Completed Likelihood criterion (ICL; Biernacki *et al.*, 2000), and the consistent AIC (cAIC; Bozdogan, 1987). The number of components is chosen based on the smallest respective model selection value. This was repeated with $N = 100$ random starts, where the scores from the best start are given in Table 6. Among the model selection criteria, AIC typically overestimates while BIC, ICL, and cAIC are good indicators for the fit of a mixture model (Wedel and DeSarbo, 1995; McLachlan and Peel, 2000). In this case, BIC, ICL, and cAIC all select $k = 2$ while AIC appears to overestimate by selecting $k = 4$. We also compare the model selection results (AIC, BIC, and cAIC) to the simple linear regression (SLR) fit[2] with no measurement error. Each of these is just slightly larger than the $k = 1$ fit, indicating that including the measurement error in the estimation provides

[2]Note that ICL, which is a penalized form of BIC, is not calculated for the SLR or the $k = 1$ fit. ICL and its variants are designed to identify the number of components in a model-based clustering framework,

a slight improvement over the traditional SLR fit. Regardless, based on these results we proceed to use the fit for the 2-component model with measurement error in the response.

Table 6: Model selection criteria for determining the number of components for the GRB dataset, where bold values indicate the number of components chosen under that criterion

k	AIC	BIC	cAIC	ICL
1	−84.935	−80.649	−78.649	—
2	−156.654	**−143.796**	**−137.796**	**−145.016**
3	−130.872	−109.440	−99.440	−111.137
4	**−158.57**	−128.568	−114.568	−131.251
SLR	−82.944	−76.515	−73.515	—

The model with known measurement errors in the responses that we fit is written as

$$y_i \sim \begin{cases} \mathbf{x}_i^T \boldsymbol{\beta}_1 + \epsilon_{i1}, & \text{with probability } \lambda \\ \mathbf{x}_i^T \boldsymbol{\beta}_2 + \epsilon_{i2}, & \text{with probability } 1-\lambda, \end{cases} \tag{6}$$
$$y_i^* = y_i + \delta_i,$$

where $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$ are independent, $i = 1, \ldots, 63$, and $j = 1, 2$, $\mathbf{x}_i = (1, \log_{10}(t_i))$, t_i is the ith observation time since trigger (in seconds), y_i^* is the logarithm (base 10) of the X-ray flux from the ith measurement, $\delta_i \sim \mathcal{N}(0, \eta_i^2)$, $\eta_i^2 = \log_{10}^2(s_i)$, s_i is the reported variability for the measurement error of the flux for the ith observation, and δ_i is independent of ϵ_{ij}.

Table 7: Parameter estimates, estimated SEs from the parametric bootstrap, and the estimated SEs using the observed information matrix

Parameter	Estimates	Bootstrap SEs	Theoretical SEs
β_{10}	−6.782	2.438	0.209
β_{11}	−1.007	0.912	0.049
β_{20}	−5.286	3.561	0.147
β_{21}	−1.552	1.178	0.022
σ_1	0.792	0.112	0.057
σ_2	1.470	0.600	0.413
λ	0.601	0.197	0.249

For the WLS estimate $\widetilde{\boldsymbol{\beta}}_j$ in our mixture-of-regressions setting, we obtain standard errors for the parameters using a parametric bootstrap with $B = 1000$. We then compare the result with variance estimates for the WLS estimators using the inverse of the observed information matrix (see Table 7). Based on the output, the standard errors from the parametric bootstrap are much larger than the inverse of observed information, especially for the

which is achieved through the estimated mean entropy that is used as the penalty term (Biernacki *et al.*, 2000; Baudry *et al.*, 2010; Bertoletti *et al.*, 2015). As noted in Bertoletti *et al.* (2015), "the ICL tends to be less prone to discriminate overlapping groups, essentially becoming an efficient model-based criterion that can be used to outline the clustering structure in the data."

intercepts. However, the standard errors for the variances, σ_1 and σ_2, and mixing proportion λ are reasonable, as well as the intercepts β_{11} and β_{21}.

The lines from the estimated model are shown in Figure 4(b), where each color represents the component based on the largest posterior membership probability. Based on this figure there are clearly two distinct components: one with time $T < 2000$s and the other with time $T > 2000$s. The result agrees with astronomers' assessment about PD mode and PC mode.

It is also worth investigating data within PD mode using our mixture model since it involves the flaring points as well as regular data points. The data within PD mode consists of the first $n = 49$ data points. We fit the non-log-transformed data (time since trigger as predictor variable x_i and X-ray flux as observed response variable y_i^*) with a 2-component mixture model using our proposed method. The fit for the model in (6) is

$$y_i \sim \begin{cases} 59.023 - 0.047x_i + \epsilon_{i1}, & \text{with probability } 0.742 \\ 179.195 - 0.510x_i + \epsilon_{i2}, & \text{with probability } 0.258, \end{cases}$$

where $\epsilon_{i1} \sim \mathcal{N}(0, 2.93^2)$ and $\epsilon_{i2} \sim \mathcal{N}(0, 4.41^2)$ for $i = 1, \ldots, 49$. The estimated regression lines from this fit are overlaid on the scatterplot of the PD mode data in Figure 4(c). Based on the calculated posterior membership probabilities, the blue triangles are those observations assigned to the first component and the red bullets are those observations assigned to the second component. While our fit identified two clear components, the clusterings are clearly affected by the time since trigger variable. Such a clustering affected by the predictor variable is called *assignment dependence*, and is treated extensively by Hennig (2000). Such a feature can be incorporated via the use of cluster-weighted models (see Gershenfeld, 1997; Ingrassia *et al.*, 2012, 2014). While our model is not a cluster-weighted model, we do note what it is identifying in this particular part of our analysis. Referring again to Figure 4(c), the red vertical dashed line is the break line of time before and after $T = 280$s. As discussed, data points with $T > 280$s are considered as flaring points, and those points classified to the second component give strong evidence in favor of this flaring assumption as they have a noticeably different linear structure than those datasets before 280s. Thus, the fit from our proposed mixture model gives evidence to the presence of a structural changepoint at this time of $T = 280$s.

6. Conclusion

Measurement error in a response variable is considered as intrinsic scatter when incorporated as part of astronomical regression models. In this paper, we discussed a mixture-of-regressions model where measurement error is treated in the response. We extended the WLS method proposed by Akritas and Bershady (1996) to the mixture setting, and used likelihood methods to compute the estimates of the parameters. Our proposed model differs from the mixture-of-regressions model introduced by Yao and Song (2015), who modeled measurement error in the predictors.

We conducted extensive simulation studies to characterize the performance of our WLS-based algorithm to reflect weighting from the intrinsic scatter. The simulation study included combinations of 2-component and 3-component models having either one or two predictors, various degrees of separability between the components, and difference amounts

of variability assumed for the measurement error. The overall results show that our method can improve the performance of estimates, especially when the measurement error is not too large. It is often the case that proposed numerical procedures for measurement error models perform best when the measurement error is not too large. Moreover, mixture models with well-separated components tend to do better in terms of their MSE and relative efficiencies when compared to the naïve estimators that do not reflect the measurement error. Again, numerical procedures for finite mixture models tend to do better under model settings with well-separated components.

Our model was motivated as a way to analyze GRB data, for which we do have a reliable estimate for the variability of the measurement error in the response variable. In particular, a 2-component mixture-of-regressions model is tenable since it can be used to characterize those flux measurements that are likely to be occurring during the flaring portion of the GRB's X-ray decay. Our model was able to make use of all of the reported data, and provided a more nuanced view of these GRB data.

There are various considerations for future research to expand on the work presented in this paper. For example, a more formal inference framework could be implemented for determining the number of components. While we just applied model selection criteria in our paper, one could proceed to perform (nonparametric) bootstrapping (McLachlan, 1987). Moreover, one could investigate bootstrapping for developing certain goodness-of-fit tests of our proposed model, some of which have appealing asymptotic properties (Babu and Rao, 2004).

Another possibility is to consider more flexibility to our general model. For example, one might assume something other than Gaussian components for the mixture structure used in this paper to achieve greater flexibility in the modeling process. Moreover, modifications to Algorithm 1 could be investigated to handle different assumptions on the measurement error δ_i. For example, one obvious setting is where the η_i are unknown, which is likely to be the more common situation encountered in practice. Another possibility is that the measurement error could also be conditioned on component membership k_i, resulting in η_i being replaced by η_{k_i} in the variance in (5). However, such an assumption surely has added identifiability issues that would require further constraints in order to perform estimation.

Another direction is how clustering can be affected by the predictor variable, which is a limitation with our work that we briefly mentioned at the end of Section 5. In the analysis of the PD mode data of the GRB, the predictor would be time since trigger. Expanding our proposed mixture-of-regressions model to also incorporate such assignment dependence would be a more flexible generalization. A cluster-weighted model could be a viable extension to our approach as it could provide a reasonable mechanism to handle measurement errors in both the response and predictor variables.

Acknowledgements

The authors would first like to thank an anonymous reviewer for helpful comments and suggestions about this manuscript. The authors are also grateful for the invitation to contribute to this special issue in memory of Prof. C. R. Rao. While a PhD student at Penn State University, the third author of this paper recalls many memories of Prof. Rao in the Department of Statistics, especially with how he engaged with graduate students. One

personal memory was when the author presented their dissertation work during the event held for awarding the C. R. and Bhargavi Rao Prize. After the talk, Prof. Rao approached them, gave them praise for their research, and asked for a copy of their dissertation. Such a gesture was quite impactful for a statistician-in-training.

References

Aitken, A. C. (1935). On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, **55**, 42–48.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akademiai Kiado, Budapest.

Akritas, M. G. and Bershady, M. A. (1996). Linear regression for astronomical data with measurement errors and intrinsic scatter. *The Astrophysical Journal*, **470**, 706–714.

Andrae, R. (2010). Error estimation in astronomy: A guide. `https://doi.org/10.48550/arXiv.1009.2755`. 1–23.

Babu, G. J. and Rao, C. R. (2004). Goodness-of-fit tests when parameters are estimated. *Sankhyā: The Indian Journal of Statistics*, **66**, 63–74.

Barthelmy, S. D. (2004). Burst Alert Telescope (BAT) on the Swift MIDEX mission. In *Proc. SPIE 5165, X-Ray and Gamma-Ray Instrumentation for Astronomy XIII*, pages 1–15.

Baudry, J.-P., Raftery, A. E., Celeux, G., Lo, K., and Gottardo, R. (2010). Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, **9**, 332–353.

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. S. (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, **32**, 1–29.

Bertoletti, M., Friel, N., and Rastelli, R. (2015). Choosing the number of clusters in a finite mixture model using an exact integrated completed likelihood criterion. *Metron*, **73**, 177–199.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 719–725.

Blustin, A. J., Band, D., Barthelmy, S., Boyd, P., Capalbi, M., Holland, S. T., Marshall, F. E., Mason, K. O., Perri, M., Poole, T., and 55 others (2006). Swift panchromatic observations of the bright gamma-ray burst GRB 050525a. *The Astrophysical Journal*, **637**, 901–913.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications*. Chapman and Hall/CRC, New York, NY.

Carmichael, B. and Coën, A. (2008). Asset pricing with errors-in-variables˙ *Journal of Empirical Finance*, **15**, 778–788.

Carroll, R. J., Midthune, D., Freedman, L. S., and Kipnis, V. (2006a). Seemingly unrelated measurement error models, with application to nutritional epidemiology. *Biometrics*, **62**, 75–84.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. (2006b). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC, New York, NY, 2nd edition.

Clutton-Brock, M. (1967). Likelihood distributions for estimating functions when both variables are subject to error. *Technometrics*, **9**, 261–269.

Davison, A. C. and Hinkley, D. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, New York, NY.

De Veaux, R. D. (1989). Mixtures of linear regressions. *Computational Statistics and Data Analysis*, **8**, 227–245.

Dellaportas, P. and Stephens, D. A. (1995). Bayesian analysis of errors-in-variables regression models. *Biometrics*, **51**, 1085–1095.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **39**, 1–38.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, **65**, 457–482.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction To The Bootstrap*. Chapman and Hall/CRC, New York, NY.

Fang, X., Chen, A. W., and Young, D. S. (2023). Predictors with measurement error in mixtures of polynomial regressions. *Computational Statistics*, **38**, 373–401.

Feigelson, E. D. and Babu, G. J. (1992). Linear regression in astronomy. II. *The Astrophysical Journal*, **397**, 55–67.

Feigelson, E. D., de Souza, R. S., Ishida, E. E. O., and Babu, G. J. (2021). 21st century statistical and computational challenges in astrophysics. *Annual Review of Statistics and Its Application*, **8**, 493–517.

Frisch, R. (1935). Statistical confluence analysis by means of complete regression systems. *The Economic Journal*, **45**, 741–742.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer, New York, NY.

Fuller, W. A. (1987). *Measurement Error Models*. John Wiley & Sons, Inc., New York, NY.

Gehrels, N., Chincarini, G., Giommi, P., Mason, K., Nousek, J., Wells, A. A., White, N. E., Barthelmy, S. D., Burrows, D. N., Cominsky, L. R., and Hurley, K. C. (2004). The Swift gamma-ray burst mission. *The Astrophysical Journal*, **611**, 1005–1020.

Gershenfeld, N. (1997). Nonliner inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences*, **808**, 18–24.

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Chapman and Hall/CRC.

Hennig, C. (2000). Identifiablity of models for clusterwise linear regression. *Journal of Classification*, **17**, 273–296.

Hurn, M., Justel, A., and Robert, C. P. (2003). Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, **12**, 55–79.

Ingrassia, S., Minotti, S. C., and Punzo, A. (2014). Model-based clustering via linear cluster-weighted models. *Computational Statistics and Data Analysis*, **71**, 159–182.

Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012). Local statistical modeling via a cluster-weighted approach with elliptical distributions. *Journal of Classification*, **29**, 363–401.

Kelly, B. C. (2007). Some aspects of measurement error in linear regression of astronomical data. *The Astrophysical Journal*, **665**, 1489–1506.

Kelly, B. C. (2012). Measurement error models in astronomy. In Feigelson, E. D. and Babu, G. J., editors, *Statistical Challenges in Modern Astronomy V*, pages 147–162. Springer-Verlag, New York, NY.

Kuhn, M. A. and Feigelson, E. D. (2019). Mixture models in astronomy. In Fruhwirth-Schnatter, S., Celeux, G., and Robert, C. P., editors, *Handbook of Mixture Analysis*, chapter 19, pages 463–483. CRC Press.

Lange, K. (2010). *Numerical Analysis for Statisticians*. Springer, New York, NY, 2nd edition.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications*, volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics and the American Statistical Association.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **44**, 226–233.

Maddala, G. S. and Nimalendran, M. (1996). 17 errors-in-variables problems in financial models. In Maddala, G. S. and Rao, C. R., editors, *Handbook of Statistics, Volume 14: Statistical Methods in Finance*, pages 507–528. North Holland - Elsevier, Amsterdam, Netherlands.

McLachlan, G. J. (1987). On Bootstrapping the Likelihood Ratio Test Stastistic for the Number of Components in a Normal Mixture. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **36**, 318–324.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and Its Application*, **6**, 355–378.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York, NY.

Mengersen, K. L., Robert, C. P., and Titterington, D. M., editors (2011). *Mixtures: Estimation and Applications*, West Sussex, England. Wiley.

Morrison, H. L., Olszewski, E. W., Mateo, M., Norris, J. E., Harding, P., Dohm-Palmer, R. C., and Freeman, K. C. (2000). Mapping the galactic halo. IV. Finding distant giants reliably with the Washington System. *The American Astronomical Society*, **121**, 37–40.

Murillo, A. L., Affuso, O., Peterson, C. M., Li, P., Wiener, H. W., Tekwe, C. D., and Allison, D. B. (2019). Illustration of measurement error models for reducing bias in nutrition and obesity research using 2-d body composition data. *Obesity*, **27**, 489–495.

Piran, T., Bromberg, O., Nakar, E., and Sari, R. (2013). The long, the short and the weak: The origin of gamma-ray bursts. *Philosophical Transactions of the Royal Society A*, **371**, 1–10.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in FORTRAN: The Art of Scientific Computing*. Cambridge University Press.

Richardson, S. and Gilks, W. R. (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology*, **138**, 430–442.

Richardson, S., Leblond, L., Jaussent, I., and Green, P. J. (2002). Mixture Models in Measurement Error Problems, with Reference to Epidemiological Studies. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **165**, 549–566.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Stephens, D. A. and Dellaportas, P. (1992). Bayesian inference of generalized linear models with covariate measurement errors. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 4*, pages 813–820. Oxford University Press, Oxford, UK.

Tarnopolski, M. (2019). Multivariate analysis of BATSE gamma-ray burst properties using skewed distributions. *The Astrophysical Journal*, **887**, 1–9.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York, NY.

Viele, K. and Tong, B. (2002). Modeling with mixtures of linear regressions. *Statistics and Computing*, **12**, 315–330.

Wedel, M. and DeSarbo, W. S. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification*, **12**, 21–55.

Yao, W. and Song, W. (2015). Mixtures of linear regression with measurement errors. *Communications in Statistics - Theory and Methods*, **44**, 1602–1614.

Young, D. S. (2014). Mixtures of regressions with changepoints. *Statistics and Computing*, **24**, 265–281.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 311–326
http://www.ssca.org.in/journal

Mixed Model Selection with Applications to Small Area Estimation

J. Sunil Rao[1] and J. N. K. Rao[2]
[1]*Division of Biostatistics, University of Minnesota, Twin Cities, USA*
[2]*Department of Mathematics and Statistics, Carleton University, Ottawa, Canada*

Received: 30 July 2024; Revised: 28 August 2024; Accepted: 31 August 2024

Abstract

Mixed models have widespread appeal in many areas of statistical modeling including small area estimation. Here we review a variety of different approaches for linear mixed model selection eventually arriving at the specific problem of selecting variables in small area models ranging from parametric and non-parametric area and unit level models to subarea small area models.

Key words: Subareas

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Many model search strategies involve trading off model fit with model complexity in a penalized goodness of fit measure. Asymptotic properties for these types of procedures in settings like linear regression and ARMA time series have been studied. Yet, these strategies do not generalize naturally to more complex models, such as those for modeling correlated data or those that involve adaptive estimation. In these cases, penalties and model complexity may not be naturally defined.

Since the introduction of Akaike's information criterion (AIC, Akaike 1973, 1974), a number of similar criteria have been proposed, including the Bayesian information criterion (BIC; Schwarz 1978), a criterion due to Hannan and Quinn (HQ; 1979), and the generalized information criterion (GIC; Nishii 1984, Shibata 1984). These procedures essentially amount to minimize a criterion function, which may be expressed as

$$\hat{D}_M + \lambda_n |M|, \tag{1}$$

where M represents a candidate model, $\hat{D}_M$ is a measure of lack of fit by M, and $|M|$ denotes the dimension of M, usually in terms of the number of estimated parameters under M. The difference is made by λ_n, where n is the sample size. This is called a "penalizer", although some authors refer $\lambda_n|M|$ as the penalizer. For example, $\lambda_n = 2$ for AIC; $\lambda_n = \log(n)$ for

Corresponding Author: J. Sunil Rao
Email: js-rao@umn.edu

BIC; and $\lambda_n = c\log\{\log(n)\}$ for HQ, where c is a constant greater than 2 (Bozdogan 1987, pp. 359).

1.1. Contributions of C. R. Rao

It should be no surprise that Professor C.R. Rao has made important contributions to model selection (in addition to the many other fundamental results he has given the field). Two specific examples are found in Rao and Wu (1989) and Bai *et al.* (1999). In both cases, the problem under study was that of linear model seletion. Specifically, the authors considered the (possibly overfit) linear model,

$$y = X\beta + \epsilon, \tag{2}$$

where y is a n-vector of response observations, X is a known design matrix, β is a p-vector of unknown regression parameters, and ϵ is a random error n-vector. Certain components of β may or may not be zero. There are thus 2^p total submodels, one of which is assumed to be the true model generating the responses.

Rao and Wu (1989) developed a criterion in the family of (1) with a flexible penalty function and proved strong consistency of model selection (that is, finding the true model). Their method allowed a wider range of penalty functions thus leading to improved small sample performance by adaptively choosing the best penalty function from the collection of candidate ones. Specifically, they entertained $\lambda_n = \alpha n^\gamma$ where $\gamma < 1$. They called a combination (α, γ) of interest if all of the models in a collection of new perturbed models built off the fitted full model are correctly selected. There may in fact be more than one combination of (α, γ) that share this property and thus Rao and Wu (1989) suggested that additional work is warranted in choosing among them. In Bai *et al.* (1999), they derived a particular choice of λ_n based on observed data, which makes it random. They then proved that the consistency property can still hold.

2. Mixed model selection

Consider the following mixed linear model:

$$y \;=\; X\beta + Z\alpha + \epsilon\,, \tag{3}$$

where $y = (y_i)_{1\le i\le N}$ is a vector of observations; $\beta = (\beta_j)_{1\le j\le p}$ is a vector of unknown regression coefficients (the fixed effects); $\alpha = (\alpha_j)_{1\le j\le m}$ is a vector of unobservable random variables (the random effects); $\epsilon = (\epsilon_i)_{1\le i\le N}$ is a vector of errors; and X, Z are known matrices. We assume that $E(\alpha) = 0$, $\mathrm{Var}(\alpha) = G$; $E(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = R$, where G and R may involve some unknown parameters such as variance components; and α and ϵ are uncorrelated.

2.1. Random factors not subject to selection

In this section, we consider the model selection problem when the random part of the model, *i.e.*, $Z\alpha$, is not subject to selection. Let $\zeta = Z\alpha + \epsilon$. Then, the problem is closely related to a regression model selection problem with correlated errors. Consider the

following general linear model:

$$y \;=\; X\beta + \zeta\ , \tag{4}$$

where ζ is a vector of correlated errors, and everything else is as in (3). We assume that there are a number of candidate vectors of covariates, $X_1, \ldots, X_q$, from which the columns of X are to be selected. Let $K = \{1, \ldots, q\}$. Then, the set of all possible models can be expressed as $\mathcal{B} = \{k : k \subseteq K\}$, and there are 2^q possible models. Let $\mathcal{A}$ be a subset of $\mathcal{B}$ that is known to contain the true model, so the selection will be within $\mathcal{A}$. In an extreme case, $\mathcal{A}$ may be $\mathcal{B}$ itself. For any matrix M, let $\mathcal{L}(M)$ be the linear space spanned by the columns of M; P_M the projection onto $\mathcal{L}(M)$: $P_M = M(M^TM)^-M^T$; and $P_M^\perp$ the orthogonal projection: $P_M^\perp = I - P_M$. For any $k \in \mathcal{B}$, let $X(k)$ be the matrix whose columns are X_j, $j \in k$, if $k \neq \emptyset$; and $X(k) = 0$ if $k = \emptyset$. We consider the following criterion for model selection:

$$C_N(k) \;=\; |y - X(k)\hat{\beta}(k)|^2 + \lambda_N|k| \;=\; |P^\perp_{X(k)}y|^2 + \lambda_N|k|\ , \tag{5}$$

$k \in \mathcal{A}$, where $|k|$ represents the cardinality of k; $\hat{\beta}(k)$ is the ordinary least squares (OLS) estimator of $\beta(k)$ for the model $y = X(k)\beta(k) + \zeta$, *i.e.*,

$$\hat{\beta}(k) \;=\; [X(k)^TX(k)]^-X(k)^Ty$$

and λ_N is a positive number satisfying certain conditions specified below. Note that $P_{X(k)}$ is understood as 0 if $k = \emptyset$. Denote the true model by k_0. If $k_0 \neq \emptyset$, we denote the corresponding X and β by X and $\beta = (\beta_j)_{1\le j\le p}$ ($p = |k_0|$), and assume that $\beta_j \neq 0$, $1 \le j \le p$. This is, of course, reasonable because otherwise the model can be further simplified. If $k_0 = \emptyset$, X, β, and p are understood as 0. For $1 \le j \le q$, Let $\{j\}^c$ represent the set $K \setminus \{j\}$. We define the following sequences: $\omega_N = \min_{1\le j\le q} |P^\perp_{X(\{j\}^c)}X_j|^2$, $\nu_N = \max_{1\le j\le q} |X_j|^2$, and $\rho_N = \lambda_{\max}(ZGZ^T) + \lambda_{\max}(R)$, where $\lambda_{\max}$ means largest eigenvalue. Let $\hat{k}$ be the minimizer of (5) over $k \in \mathcal{A}$, which will be our selection of the model. The following theorem gives sufficient conditions under which the selection is consistent in the sense that

$$P(\hat{k} \neq k_0) \;\longrightarrow\; 0\ . \tag{6}$$

Theorem 1. (Jiang and Rao 2003) Suppose that $\nu_N > 0$ for large N,

$$\rho_N/\nu_N \;\longrightarrow\; 0\ , \quad \text{while} \quad \liminf(\omega_N/\nu_N) \;>\; 0\ . \tag{7}$$

Then, (4) holds for any λ_N such that

$$\lambda_N/\nu_N \;\longrightarrow\; 0 \;\text{ and }\; \rho_N/\lambda_N \;\longrightarrow\; 0\ . \tag{8}$$

The above procedure requires selecting $\hat{k}$ from all subset of $\mathcal{A}$. Note that $\mathcal{A}$ may contain as many as 2^q subsets. When q is relatively large, alternative procedures have been proposed, in the (fixed effects) linear model context, which require less computation [*e.g.*, Zheng and Loh (1995)]. In the following, we consider an approach which is similar, in spirit, to Rao and Wu (1989). First, note that one can always express $X\beta$ in (4) as

$$X\beta \;=\; \sum_{j=1}^{q} \beta_j X_j \tag{9}$$

with the understanding that some of the coefficients β_j may be zero. It follows that $k_0 = \{1 \le j \le q : \beta_j \neq 0\}$. Let $X_{-j} = (X_u)_{1 \le u \le q, u \neq j}$, $1 \le j \le q$, $\eta_N = \min_{1 \le j \le q} |P^{\perp}_{X_{-j}} X_j|^2$, and δ_N be a sequence of positive numbers satisfying conditions specified below. Let $\hat{k}$ be the subset of K such that

$$(|P^{\perp}_{X_{-j}} y|^2 - |P^{\perp}_{X} y|^2)/(|P^{\perp}_{X_{-j}} X_j|^2 \delta_N) \;\; > \;\; 1 \, . \tag{10}$$

The following theorem states that, under suitable conditions, $\hat{k}$ is a consistent selection. Recall that ρ_N is defined above Theorem 1.

Theorem 2. (Jiang and Rao 2003) Suppose that $\eta_N > 0$ for large N, and

$$\rho_N/\eta_N \;\; \longrightarrow \;\; 0 \, . \tag{11}$$

Then, (6) holds for any δ_N such that

$$\delta_N \;\; \longrightarrow \;\; 0 \;\; \text{and} \;\; \rho_N/(\eta_N \delta_N) \;\; \longrightarrow \;\; 0 \, . \tag{12}$$

2.2. Selection of random factors

We now assume that $Z\alpha$ in (3) can be expressed as

$$Z\alpha \;\; = \;\; \sum_{j=1}^{s} Z_j \alpha_j \, , \tag{13}$$

where $Z_1, \ldots, Z_s$ are known matrices; each α_j is a vector of independent random effects with mean 0 and variance σ_j^2, which is unknown, $1 \le j \le s$. Furthermore, we assume that ϵ in (3) is a vector of independent errors with mean 0 and variance $\tau^2 > 0$, and $\alpha_1, \ldots, \alpha_s, \epsilon$ are independent. Such assumptions are customary in the mixed model context (*e.g.*, Searle, Casella, and McCulloch (1992), pp 233-234), therefore (13) represents a fairly general class of mixed linear models. If $\sigma_j^2 > 0$, we say that α_j is in the model; otherwise, it is not. Therefore, the selection of random factors is equivalent to simultaneously determining which of the variance components $\sigma_1^2, \ldots, \sigma_s^2$ are positive, and which of them are zero. The true model can be expressed as

$$y \;\; = \;\; X\beta + \sum_{j \in l_0} Z_j \alpha_j + \epsilon \, , \tag{14}$$

where $X = (X_j)_{j \in k_0}$ and $k_0 \subseteq K$ (see section 2); $l_0 \subseteq L = \{1, \ldots, s\}$ such that $\sigma_j^2 > 0$, $j \in l_0$, and $\sigma_j^2 = 0$, $j \in L \setminus l_0$.

There are some important differences between selecting the fixed covariates X_j and selecting the random factors. One difference is that, in selecting the random factors, we are going to determine whether the vector α_j, not a given component of α_j, should be in the model. In other words, the components of α_j are all "in" or all "out". Another difference is that, unlike selecting the fixed covariates, where it is reasonable to assume that the X_js are linearly independent, in a mixed linear model it is possible to have $j \neq j^T$ but $\mathcal{L}(Z_j) \subset \mathcal{L}(Z_{j^T})$.

First, note that in section 2.1 we discussed a procedure to determine the fixed part of the model, which leads to a selection $\hat{k}$ that satisfies (6). Note that the only place that the determination of $\hat{k}$ might use knowledge about Z, and hence about l_0, is through λ_N, which depends on the order of $\lambda_{\max}(ZGZ^T)$. However, under (13), $\lambda_{\max}(ZGZ^T) \leq \sum_{j=1}^{s} \sigma_j^2 \|Z_j\|^2$, where for any matrix M, $\|M\| = [\lambda_{\max}(M^T M)]^{1/2}$. Thus, an upper bound for the order of $\lambda_{\max}(ZGZ^T)$ is $\max_{1\leq j\leq s} \|Z_j\|^2$, which does not depend on l_0. Therefore, $\hat{k}$ could be determined without knowing l_0. In any case, we may write $\hat{k} = \hat{k}(l_0)$, be it dependent on l_0 or not. Now, suppose that a selection for the random part of the model, *i.e.*, a determination of l_0, is $\hat{l}$. We then define $\hat{k} = \hat{k}(\hat{l})$. The following theorem shows that the combined procedure is consistent.

Theorem 3. (Jiang and Rao 2003) Suppose that $P(\hat{l} \neq l_0) \to 0$ and $P(\hat{k}(l_0) \neq k_0) \to 0$. Then, $P(\hat{k} = k_0 \text{ and } \hat{l} = l_0) \to 1$.

How does one actually obtain $\hat{l}$? Jiang and Rao (2003) divided the vectors $\alpha_1, \ldots, \alpha_s$, or, equivalently, the matrices $Z_1, \ldots, Z_s$ into several groups. The first group is called the "largest random factors". Roughly speaking, those are Z_j, $j \in L_1 \subseteq L$ such that $\text{rank}(Z_j)$ is of the same order as N, the sample size. We can assume that $\mathcal{L}(X, Z_u, u \in L \setminus \{j\}) \neq \mathcal{L}(X, Z_u, u \in L)$, $j \in L_1$, where $\mathcal{L}(M_1, \ldots, M_t)$ represents the linear space spanned by the columns of the matrices $M_1, \ldots, M_t$. Such an assumption is reasonable because Z_j is supposed to be "largest", and hence should have contribution to the linear space. The second group consists of Z_j, $j \in L_2 \subseteq L$ such that $\mathcal{L}(X, Z_u, u \in L \setminus L_1 \setminus \{j\}) \neq \mathcal{L}(X, Z_u, u \in L \setminus L_1)$, $j \in L_2$. The ranks of the matrices in this group are of lower order of N. Similarly, the third group consists of Z_j, $j \in L_3 \subseteq L$ such that $\mathcal{L}(X, Z_u, u \in L \setminus L_1 \setminus L_2 \setminus \{j\}) \neq \mathcal{L}(X, Z_u, u \in L \setminus L_1 \setminus L_2)$, and so on. Note that if the first group, *i.e.*, the largest random factors, does not exist, the second group becomes the first, and other groups also move on. Jiang and Rao (2003) gave a procedure that determines the indexes $j \in L_1$ for which $\sigma_j^2 > 0$; then a procedure that determines the indexes $j \in L_2$ for which $\sigma_j^2 > 0$; and so on.

3. Fence methods

Although criteria like (1) are broadly used, difficulties are often encountered, especially in some non-conventional situations. For example, consider the following linear mixed model written at the unit level,

$$y_{ij} = x_{ij}^T \beta + u_i + v_j + e_{ij}, i = 1, \ldots, m_1, j = 1, \ldots, m_2, \tag{15}$$

where x_{ij} is a vector of known covariates, β is a vector of unknown regression coefficients (the fixed effects), u_i, v_j are random effects, and e_{ij} is an additional error term. It is assumed that u_i's, v_j's and e_{ij}'s are independent, and that, for the moment, $u_i \sim N(0, \sigma_u^2)$, $v_j \sim N(0, \sigma_v^2)$, $e_{ij} \sim N(0, \sigma_e^2)$. It is well-known (*e.g.*, Harville 1977, Miller 1977) that, in this case, the effective sample size for estimating σ_u^2 and σ_v^2 is not the total sample size $m_1 \cdot m_2$, but m_1 and m_2, respectively, for σ_u^2 and σ_v^2. Now suppose that one wishes to select the fixed covariates, which are components of x_{ij}, under the assumed model structure, using BIC. Then, it is not clear what should be in place of n in (1). In fact, in cases of correlated observations, such as the example above, the definition of "sample size" is often unclear.

Furthermore, suppose that normality is not assumed in the above linear mixed model. In fact, the only distributional assumptions are that the random effects and errors are independent, have zero mean and constant variances. Now, suppose that one, again, wishes to select the fixed covariates using AIC, BIC, or HQ. It is not clear how to do this because the likelihood is unknown.

Even in the conventional case, there are still practical issues regarding these criteria. For example, BIC is known to have the tendency of overly penalizing. In such a case, one may wish to replace the penalizer by $c\log(n)$, where c is a constant less than one. Question is: What c? Asymptotically, the choice of c does not make a difference in terms of consistency so long as $c > 0$. However, practically, it does. For example, comparing BIC with HQ, the penalizer of the latter is lighter in its order ($\log\{\log(n)\}$ vs $\log(n)$), but there is a constant c involved in HQ. If $n = 100$, we have $\log(n) = 4.6$ and $\log\{\log(n)\} = 1.5$, hence, if the constant c in HQ is chosen as 3, BIC and HQ are the same.

Finally, the definition of $|M|$ in (1) can also cause difficulties. In some circumstances like ordinary linear regression, this is simply the number of parameters in M, but in other situations where nonlinear, adaptive models are fitted, this can be substantially more (*e.g.*, Hastie and Tibshirani 1990, Friedman 1991, Ye 1998).

While there is extensive literature on parameter estimation in linear and generalized linear mixed models, the other component, that is, mixed model selection, has received much less attention. Only recently have some results emerge in the area of linear mixed model selection. Datta and Lahiri (2001) discussed a model selection method based on computation of the frequentist's Bayes factor in choosing between a fixed effects model and a random effects model. They focused on the following one-way balanced random effects model for the sake of simplicity: $y_{ij} = \mu + u_i + e_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, k$, where the u_i's and e_{ij}'s are normally distributed with mean zero and variances σ_u^2 and σ_e^2, respectively. As noted by the authors, the choice between a fixed effects model and a random effects one in this case is equivalent to testing the following one-sided hypothesis H_0: $\sigma_u^2 = 0$ vs H_1: $\sigma_u^2 > 0$. In fact, hypothesis testing may be regarded as a special case of model selection, but not all model selection problems can be formulated as hypothesis testing (see further discussion in subsection 8.1). Jiang and Rao (2003) developed various generalized information criteria (GICs) suitable for linear mixed model selection. Meza and Lahiri (2005) demonstrated the limitations of Mallows' C_p statistic in selecting the fixed covariates in a nested error regression model which is a special case of the linear mixed models. The nested error regression model is defined as $y_{ij} = x_{ij}^T\beta + u_i + e_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where y_{ij} is the observation, x_{ij} is a vector of fixed covariates, β is a vector of unknown regression coefficients, and u_i's and e_{ij}'s are the same as in the model above considered by Datta and Lahiri (2001). Simulation studies carried out by Meza and Lahiri (2005) showed that the C_p method without modification does not work well in the current mixed model setting when the variance σ_u^2 is large; on the other hand, a modified C_p criterion developed by these latter authors by adjusting the intra-cluster correlations performs similarly as the C_p in regression settings. Another related paper is that of Vaida and Blanchard (2005) who proposed a conditional AIC where the penalty term in this CAIC is related to the effective degrees of freedom for a linear mixed model proposed by Hodges and Sargent (2001) which reflects an intermediate level of model complexity between a full fixed effects model and a corresponding mixed model conditional on the random effects variances.

It should be pointed out that all these studies are limited to linear mixed models, while model selection in generalized linear mixed models (GLMMs) has never been seriously addressed in the literature in a general way (there are some fully Bayesian approaches for special cases like logistic mixed effects models - see Kinney and Dunson (2007) for example). In fact, our earlier simulation results suggested that in the case of GLMM selection, a procedure like GIC is much more sensitive to the choice of λ_n than in linear mixed model selection. It is these concerns, such as the above, that motivated the development of a new principle for model selection that is potentially less subjective, and applicable to both linear mixed models and GLMMs.

Jiang, Rao *et al.* (2008) proposed a new procedure for model selection, called the fence methods. An essential part of this procedure is a measure of lack-of-fit, denoted by $Q_M = Q_M(y, \theta_M)$, where M indicates the candidate model, y is an $n \times 1$ vector of observations, θ_M represents the vector of parameters under M, such that $\mathrm{E}(Q_M)$ is minimized when M is a true model and θ_M the true parameter vector under M. Here by true model we mean that M is a correct model but not necessarily the most efficient one. In the sequel we use the terms "true model" and "correct model" interchangeably. One example of Q_M is the negative log-likelihood function under a parametric model. Another example is the residual sum of squares (RSS) under a parametric or semiparametric model. For more examples, see Jiang, Rao *et al.* (2008).

The idea involves a procedure to isolate a subgroup of what are known as correct models (of which the optimal model is a member). This is accomplished by constructing a statistical *fence*, or barrier, to carefully eliminate incorrect models. Once the fence is constructed, the optimal model is selected from amongst those within the fence according to a criterion which can be made flexible and incorporate scientific or economical concerns. The fence is built by checking the following inequality for every candidate model M,

$$\hat{Q}_M - \hat{Q}_{\tilde{M}} \;\leq\; c_n \hat{\sigma}_{M,\tilde{M}}, \tag{16}$$

where $\hat{Q}_M = \inf_{\theta_M \in \Theta_M} Q_M(\theta_M, y)$, $\hat{Q}_{\tilde{M}} = \min_{M \in \mathcal{M}} \hat{Q}_M$, and $\mathcal{M}$ represents the set of candidate models. Here $\hat{\sigma}_{M,\tilde{M}}$ is an estimate of the standard deviation of the left side of (16). Finally, c_n is a tuning constant chosen below.

The motivation of (16) is to exam the closeness of $\hat{Q}_M$ to its lower bound - when the measure of lack-of-fit is close enough to the minimum the model is considered correct. The reason for the appearance of $\hat{\sigma}_{M,\tilde{M}}$ on the right side is that, when M is correct, this is an appropriate measure of the left side. Still, the constant c_n plays an important role for the finite sample performance of fence. Therefore, Jiang, Rao *et al.* (2008) proposed the following method to choose c_n adaptively.

1. Fence procedure with fixed c_n.

1. ind $\tilde{M}$ such that $\hat{Q}_{\tilde{M}} = \min_{M \in \mathcal{M}} \hat{Q}_M$.
2. For each $M \in \mathcal{M}$ such that $|M| < |\tilde{M}|$, compute $\hat{\sigma}_{M,\tilde{M}}$, an estimator of $\sigma_{M,\tilde{M}}$. Then, M belongs to $\tilde{\mathcal{M}}_-$, the set of "true" models with $|M| < |\tilde{M}|$ if (2) holds.
3. Let $\tilde{\mathcal{M}} = \{\tilde{M}\} \cup \tilde{\mathcal{M}}_-$, $m_0 = \min_{M \in \tilde{\mathcal{M}}} |M|$, and $\mathcal{M}_0 = \{M \in \tilde{\mathcal{M}} : |M| = m_0\}$. Let M_0 be the model in $\mathcal{M}_0$ such that $\hat{Q}_{M_0} = \min_{M \in \mathcal{M}_0} \hat{Q}_M$. M_0 is the selected model.

The following outlines an effective algorithm for fence. Let $d_1 < d_2 < \cdots < d_L$ be all the different dimensions of the models $M \in \mathcal{M}$.

The fence algorithm: i) Find $\tilde{M}$. ii) Compute $\hat{\sigma}_{M,\tilde{M}}$ for all $M \in \mathcal{M}$ such that $|M| = d_1$; let $\mathcal{M}_1 = \{M \in \mathcal{M} : |M| = d_1$ and (16) holds$\}$; if $\mathcal{M}_1 \neq \emptyset$, stop. Let M_0 be the model in $\mathcal{M}_1$ such that $\hat{Q}_{M_0} = \min_{M \in \mathcal{M}_1} \hat{Q}_M$; M_0 is the selected model. iii) If $\mathcal{M}_1 = \emptyset$, compute $\hat{\sigma}_{M,\tilde{M}}$ for all $M \in \mathcal{M}$ such that $|M| = d_2$; let $\mathcal{M}_2 = \{M \in \mathcal{M} : |M| = d_2$ and (16) holds$\}$; if $\mathcal{M}_2 \neq \emptyset$, stop. Let M_0 be the model in $\mathcal{M}_2$ such that $\hat{Q}_{M_0} = \min_{M \in \mathcal{M}_2} \hat{Q}_M$; M_0 is the selected model. iv) Continue until the program stops (it will at some point).

In short, the algorithm may be described as follows: Check the candidate models, from the simplest to the most complex; once one has discovered a model that falls within the fence and checked all the other models of the same simplicity (for membership within the fence), one stops. One apparent advantage of the fence algorithm is that one needs not search the entire space of candidate models in order to find the optimal model. Here the optimality is defined in terms of minimal dimension, *i.e.*, $|M|$. However, as mentioned, the criterion of optimality is flexible.

2. Forward-backward (F-B) fence procedure. The fence algorithm searches from the simplest models and therefore may not need to search the entire model space in order to determine the optimal model. On the other hand, such a procedure may still involve a lot of evaluations when the model space is large. To make the fence procedure computationally more attractive to large and complex models, the following variation of fence was proposed for situations of complex models with many predictors.

To be more specific, we let $\tilde{M}$ be the full model. The idea is to use a forward-backward procedure to generate a sequence of candidate models, among which the optimal model is selected using the fence method. We begin with a forward procedure. Let M_1 be the model that minimizes $\hat{Q}_M$ among all models with a single parameter; if M_1 is within the fence, stop the forward procedure; otherwise, let M_2 be the model that minimizes $\hat{Q}_M$ among all models that add one more parameter to M_1; if M_2 is within the fence, stop the forward procedure; and so on. The forward procedure stops when the first model is discovered within the fence. The procedure is then followed by a backward elimination. Let M_k be the final model of the forward procedure. If no submodel of M_k with one less parameter is within the fence, M_k will be our selection; otherwise, M_k is replaced by M_{k+1} which is a submodel of M_k with one less parameter and is within the fence, and so on. This approach is called the forward-backward (F-B) fence.

3. Adaptive fence procedure. Recall that $\mathcal{M}$ denotes the set of candidate models, which includes a true model. To be more specific, we assume that there is a full model $M_{\mathrm{f}} \in \mathcal{M}$, hence $\tilde{M} = M_{\mathrm{f}}$ in (16); and that every model in $\mathcal{M} \setminus \{M_{\mathrm{f}}\}$ is a submodel of a model in $\mathcal{M}$ with one less parameter than M_{f}. Let M_* denote a model with minimum dimension among $M \in \mathcal{M}$. First note that, ideally, one wishes to select c_n that maximizes the probability of choosing the optimal model. Here for simplicity the optimal model is defined as a true model that has the minimum dimension among all true models. This means that one wishes to choose c_n that maximizes

$$P \;=\; \mathrm{P}(M_0 = M_{\mathrm{opt}}), \tag{17}$$

where M_{opt} represents the optimal model, and $M_0 = M_0(c_n)$ is the model selected by the

fence procedure with the given c_n. However, two things are unknown in (17): (i) under what distribution should the probability P be computed; and (ii) what is $M_{\rm opt}$?

To solve problem (i), note that the assumptions above on $\mathcal{M}$ imply that $M_{\rm f}$ is a true model. Therefore, it is possible to bootstrap under $M_{\rm f}$. For example, one may estimate the parameters under $M_{\rm f}$, then use a model-based bootstrap to draw samples under $M_{\rm f}$. This allows us to approximate the probability P on the right side of (17).

To solve problem (ii), we use the idea of maximum likelihood. Namely, let $p^*(M) = {\rm P}^*(M_0 = M)$, where $M \in \mathcal{M}$ and $\rm P^*$ denotes the empirical probability obtained by bootstrapping. Let $p^* = \max_{M\in\mathcal{M}} p^*(M)$. Note that p^* depends on c_n. The idea is to choose c_n that maximizes p^*. It should be kept in mind that the maximization is not without restriction. To see this, note that if $c_n = 0$ then $p^* = 1$ (because when $c_n = 0$ the procedure always chooses $M_{\rm f}$). Similarly, $p^* = 1$ for very large c_n, if M_* is unique (because when c_n is large enough the procedure always chooses M_*). Therefore, what one looks for is "the peak in the middle" of the plot of p^* against c_n. This procedure is also studied in detail in Jiang *et al.* (2008).

Jiang, Rao *et al.* (2008) established consistency of fence, F-B fence and adaptive fence methods under mild regularity conditions. Here consistency is in the sense that with probability tending to one as the sample size increases the procedure will select the optimal model.

3.1. Fence method for high dimensions and subtractive measures of fit

Computation in high dimensions (p large typically), can be a challenge. If m is large, as is typically the case, this could result in a large number of $\hat{Q}(M)$'s that have to be evaluated. Jiang *et al.* (2011) introduced the idea of a subtractive measure in their work on fence methods for gene set analysis (what they called the invisible fence). Let $1, \ldots, m$ denote the candidate elements. A measure $\hat{Q}$ is said to be *subtractive* if it can be expressed as

$$\hat{Q}(M) \;=\; s - \sum_{i\in M} s_i, \tag{18}$$

where s_i, $i = 1, \ldots, m$ are some nonnegative quantities computed from the data, M is a subset of $1, \ldots, m$, and s is some quantity computed from the data that does not depend on M. Typically we have $s = \sum_{i=1}^m s_i$, but the definition does not impose such a restriction. Also the nonnegativity constraint on the s_i's is only to ensure that $\hat{Q}(M)$ behaves like a measure of lack-of-fit, that is, larger model has smaller $\hat{Q}(M)$.

For a subtractive measure, the models that minimize $\hat{Q}(M)$ at different dimensions are found almost immediately. Let $r_1, r_2, \ldots, r_m$ be the ranking of the candidate elements in terms of decreasing s_i. Then, the model that minimizes $\hat{Q}(M)$ at dimension one is r_1; the model that minimizes $\hat{Q}(M)$ at dimension two is $\{r_1, r_2\}$; the model that minimizes $\hat{Q}(M)$ at dimension three is $\{r_1, r_2, r_3\}$, and so on. This is what Jiang *et al.* (2011) called a *fast algorithm* for implementing the fence approach.

3.2. Other approaches to mixed model selection

Muller *et al.* (2013) wrote a survey paper on linear mixed model selection and discussed some other methods not discussed above. These include the marginal AIC (Vaida and Blanchard 2005), the bootstrap biased-correct mAIC of Shang and Cavanaugh (2008), Srivastava and Kubokawa (2010), conditional AIC (Vaida and Blanchard 2005), the modified Schwarz approach of Pauler (1998), minimum description length (MDL) approaches, shrinkage methods and Bayesian methods. Interested readers are directed to that survey paper for more details.

4. Mixed model selection and small area estimation

Small area estimation (SAE) has received increasing attention in recent literature. Here the term small area typically refers to a subpopulation or domain for which reliable statistics of interest cannot be produced due to certain limitations of the available data. Examples of small areas include a geographical region (*e.g.*, a state, county, municipality, *etc.*), a demographic group (*e.g.*, a specific age × sex × race group), a demographic group within a geographic region, *etc.* In absence of adequate direct samples from the small areas, methods have been developed in order to "borrow strength". See Rao and Molina (2015) for a comprehensive account of various methods used in SAE. Statistical models, especially mixed effects models, have played important roles in SAE. See Jiang and Lahiri (2006) for an overview of mixed effects models in SAE.

While there is extensive literature on inference about small areas using mixed effects models, including estimation of small area means which is a problem of mixed model prediction, estimation of the mean squared error (MSE) of the empirical best linear unbiased predictor (EBLUP; see Rao 2003), and prediction intervals (*e.g.*, Chatterjee, Lahiri, and Li 2007), model selection in SAE has received much less attention. However, the importance of model selection in SAE has been noted by prominent researchers in this field (*e.g.*, Battese, Harter, and Fuller 1988, Ghosh and Rao 1994). Datta and Lahiri (2001) discussed a model selection method based on computation of the frequentist's Bayes factor in choosing between a fixed effects model and a random effects model. They focused on the following one-way balanced random effects model for the sake of simplicity: $y_{ij} = \mu + u_i + e_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, k$, where the u_i's and e_{ij}'s are normally distributed with mean zero and variances σ_u^2 and σ_e^2, respectively. As noted by the authors, the choice between a fixed effects model and a random effects one in this case is equivalent to testing the following one-sided hypothesis H_0: $\sigma_u^2 = 0$ vs H_1: $\sigma_u^2 > 0$. Note that, however, not all model selection problems can be formulated as hypothesis testing. Fabrizi and Lahiri (2004) developed a robust model selection method in the context of complex surveys. Meza and Lahiri (2005) demonstrated the limitations of Mallows' C_p statistic in selecting the fixed covariates in a nested error regression model (Battese, Harter, and Fuller 1988), defined as $y_{ij} = x_{ij}^T\beta + u_i + e_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where y_{ij} is the observation, x_{ij} is a vector of fixed covariates, β is a vector of unknown regression coefficients, and u_i's and e_{ij}'s are the same as in the model above considered by Datta and Lahiri (2001). Simulation studies carried out by Meza and Lahiri (2005) showed that the C_p method without modification does not work well in the current mixed model setting when the variance σ_u^2 is large; on the other hand, a modified C_p criterion developed by these latter authors by adjusting the intra-cluster correlations performs similarly as the C_p in regression settings. It should be pointed out that all these studies are

limited to linear mixed models, while model selection in SAE in a generalized linear mixed model (GLMM) setting has never been seriously addressed.

4.1. Fence methods for SAE model selection

One of the advantages of fence methods is that the criterion of optimality for selecting the models within the fence is flexible. In SAE the problem of main interest is the estimation, or prediction, of the small area means. For simplicity, consider the case of linear mixed models. Then, the small area mean is typically estimated by the best linear unbiased predictor, or BLUP. Because an important measure of the accuracy of BLUP is its MSE, it makes sense to take the latter into account. Therefore, we consider the following criterion for selecting models within the fence when linear mixed models are under consideration. Suppose that one is interested in a small-area specific mixed effect (*e.g.*, the small area mean), θ_i, which is a linear combination of fixed and random effects. Let $\tilde{\theta}_i$ be the BLUP of θ_i. Let $\theta = (\theta_i)_{1\le i\le m}$ and $\tilde{\theta} = (\tilde{\theta}_i)_{1\le i\le m}$. Then, $\mathrm{MSE}(\tilde{\theta}) = \mathrm{E}(|\tilde{\theta} - \theta|^2) = \sum_{i=1}^m \mathrm{E}(\tilde{\theta}_i - \theta_i)^2 = \sum_{i=1}^m \mathrm{MSE}(\tilde{\theta}_i)$. Furthermore, an explicit expression of $\mathrm{MSE}(\tilde{\theta}_i)$ can be obtained (*e.g.*, Rao 2003, pp. 137). Note that $\mathrm{MSE}(\tilde{\theta})$ typically depends on some unknown variance components. Let $\widehat{\mathrm{MSE}}(\tilde{\theta})$ be an estimator of $\mathrm{MSE}(\tilde{\theta})$, say, by replacing the variance components by their REML estimators (*e.g.*, Jiang 2007). A model within fence is selected if (i) it has the minimum dimension; and (ii) if there are more than one models chosen by (i), select the one that has the minimal $\widehat{\mathrm{MSE}}(\tilde{\theta})$.

An interesting example is that from Jiang *et al.* (2010) who considered model selection for non-parametric SAE models. Opsomer *et al.* (2008) proposed a spline-based nonparametric model for SAE. The idea is to approximate an unknown nonparametric small-area mean function by a penalized spline (P-spline). The authors then used a connection between P-splines and linear mixed models (Ruppert, Wand, and Carroll 2003) to formulate the approximating model as a linear mixed model, where the coefficients of the splines are treated as random effects. Consider, for simplicity, the case of univariate covariate. Then, a P-spline can be expressed

$$\tilde{f}(x) \;=\; \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \gamma_1 (x - \kappa_1)_+^p + \cdots + \gamma_q (x - \kappa_q)_+^p, \tag{19}$$

where p is the degree of the spline, q is the number of knots, κ_j, $1 \le j \le q$ are the knots, and $x_+ = x1_{(x>0)}$. Clearly, a P-spline is characterized by p, q, and also the location of the knots.

Jiang *et al.* (2010) developed a simplified version of the adaptive fence in order to choose p and q. First, since the optimal model is rarely either M_{f} or M_*, the minimal model (dimensionwise; *e.g.*, a model with only the intercept). Baseline adjustment and threshold checking are used to deal with these two cases (see Jiang *et al.* 2008). The baseline adjustment is done by generating an additional vector of covariates, say, X_{a}, so that it is unrelated to the data. Then, define the model M_{f}^* as M_{f} plus X_{a}, and replace M_{f} by M_{f}^*, but let $\mathcal{M}$ remain unchanged. This way one knows for sure that the new full model, M_{f}^*, is not an optimal model (because it is not a candidate model). The threshold checking inequality is given by $\hat{Q}_{M_*} - \hat{Q}_{M_{\mathrm{f}}^*} > d_*$, where d_* is the maximum of the left side of the threshold inequality computed under the bootstrap samples generated under M_*. In case the threshold inequality holds, we ignore the right tail of the plot of p^* against c_n that

eventually goes up and stays at one.

Jiang *et al.* (2010) also constructed a a (large sample) confidence lower bound, for example,

$$p^* - 1.96\sqrt{p^*(1-p^*)/B} \tag{20}$$

where B is the bootstrap sample size. When selecting c_n that maximize p^* we take (20) into account. More specifically, suppose that there are two peaks in the plot of p^* against c_n located at $c_{n,1}$ and $c_{n,2}$ such that $c_{n,1} < c_{n,2}$. Let p_1^* and p_2^* be the heights of the peaks corresponding to $c_{n,1}$ and $c_{n,2}$. As long as p_1^* is greater than the confidence lower bound at p_2^*, that is, (4) with $p^* = p_2^*$, we choose $c_{n,1}$ over $c_{n,2}$. Clearly, the selection is in favor of smaller c_n in order to be more conservative. (In other words, we are more concerned with underfit than overfit.)

Consistency of selection under mild regularity conditions was then proven in the following Theorem:

Theorem 4. (Jiang, Nguyen, and Rao 2010). Let M_0^* denote the model selected by the fence procedure with $c_n = c_n^*$. Also, let M_{opt} denote an optimal model defined as a true model with minimum dimension and minimum MSE($\tilde{\theta}$) among all the true models within the (same) minimum dimension. Under the regularity conditions given therein, there is c_n^* which is at least a local maximum and approximate global maximum of p^*, and the corresponding M_0^* is consistent in the sense that any $\delta, \eta > 0$, there are N, N^* such that

$$\mathrm{P}\{p^*(c_n^*) \geq 1-\delta\} \wedge \mathrm{P}(M_0^* = M_{\text{opt}}) \;\; \geq \;\; 1-\eta,$$

if $m \geq N$ and $B \geq N^*$.

4.2. Variable selection for area and subarea level SAE models

In this section, we focus on variable selection under area level models and subarea level SAE models which are extensively used in practice. A basic area level model, also called the Fay-Herriot model (FH; Fay and Herriot 1979), uses direct estimators $\hat{\theta}_i$ of area means $\theta_i (i = 1, \ldots, m)$ and associated area level covariates. Direct estimators are obtained from area-specific unit level data, taking survey design into account. Area level covariates are used to link the area means. This leads to a sampling model and a linking model given by $\hat{\theta}_i = \theta_i + e_i$ and $\theta_i = x_i^T \beta + v_i$ respectively, where e_i is the sampling error, β is the vector of model parameter, x_i is the $p \times 1$ vector of area level covariates and v_i is a random area effect. Further, e_i has mean 0 and known variance ψ_i, and the sampling errors are assumed to be independent. In practice, the sampling variances ψ_i are obtained by smoothing their direct estimators using generalized variance functions. The area effect v_i has mean 0 and variance σ_v^2, and the area effects are assumed to be independent. Combining the sampling model with the linking model leads to the FH model $\hat{\theta}_i = x_i^T \beta + v_i + e_i$ which is then used for variable selection. Note that the linking model alone cannot be used for variable selection because the area means θ_i are not known.

Because of the sampling errors in the FH model, standard methods for linear regression models, such as the AIC, BIC and Mallows' C_p used for variable selection, can lead to

biased variable selection when applied to the FH model. Han (2013) used a conditional AIC method for variable selection that accounts for the sampling errors in the FH model. This method is fairly complex, and practitioners might prefer simple modifications to standard methods that can account for sampling errors in the FH model. We give a brief description of a simple method of estimating the ideal variable selection criteria under the linking model that accounts for the sampling error (Lahiri and Suntornchost 2015). The resulting estimation error is shown to converge to 0 in probability as the number of areas increases, unlike the estimation error in the naïve criteria ignoring the sampling errors. The proposed method performed well in simulations unlike the naïve method that ignores the sampling errors in the FH model.

Let, $MSE_\theta = \frac{1}{m-p}\theta^T(I_m - P)\theta$ denote the ideal mean error sum of squares, where $\theta = (\theta_1, \ldots, \theta_m)^T$, I_m is the identity matrix of order m, and $P = X(X^TX)^{-1}X$ is the standard projection matrix based on the linking model. Then the estimator of MSE_θ is given by $mse_\theta = MSE_{\hat{\theta}} - \bar{\psi}_w$, where $MSE_{\hat{\theta}}$ is obtained by replacing θ by its direct estimator $\hat{\theta}$, and $\bar{\psi}_w = \frac{1}{m-p}\sum_{i=1}^m (1-h_{ii})\psi_i$ with $h_{ii} = x_i^T(X^TX)^{-1}x_i$. We simply replace MSE_θ by mse_θ in the ideal AIC, BIC and C_p which are functions of MSE_θ. For example, the resulting $AIC = mlog\{\frac{m-p}{m}mse_\theta\} + 2p$. In the case of small m, the estimator mse_θ could take a negative value and Lahiri and Suntornchost (2015) suggested a simple modification that leads to strictly positive estimator of MSE_θ.

Two-fold subarea models are also often used in practice to estimate subarea and area means. For example, Mohadjer et al. (2012) studied adult literacy for counties (subareas) sampled from states (areas) in the United States, using data from the 2003 U. S. National Assessment of Adult Literacy (NAAL). We have areas i and subareas j are sampled from area i. Direct estimators of subarea means $\theta_{ij}(j = 1, \ldots, n_i; i = 1, \ldots, m)$ and associated subarea level covariate vector are denoted as $\hat{\theta}_{ij}$ and x_{ij} respectively. A two-fold subarea model consists of a sampling model $\hat{\theta}_{ij} = \theta_{ij} + e_{ij}$ and a linking model $\theta_{ij} = x_{ij}^T\beta + b_{ij}$ respectively, where e_{ij} is the sampling error and $b_{ij} = v_i + u_{ij}$ is the sum of the random area effect v_i and subarea effect u_{ij}. The sampling errors e_{ij} are assumed to be independent with zero means and known variances. Further, the area effect is independent of the subarea effect, and the v_i and u_{ij} are independent and identically distributed with zero means and variances σ_v^2 and σ_u^2 respectively. Under the assumptions, the composite random effects b_{ij} are correlated for each area i with covariance matrix $\Sigma_i = \sigma_v^2 1_i 1_i^T + \sigma_u^2 I_i$ where 1_i is the unit vector of length n_i and I_i is the identity matrix of order n_i.

We cannot treat the linking model for the two-fold case as a FH-type model on the subarea means because the composite random effects b_{ij} are correlated. It is necessary to transform the covariance matrix Σ_i to a diagonal covariance matrix with equal diagonal elements across areas i, and then apply the variable selection method to the transformed linking model to get the ideal error mean sum of squares. Cai *et al.* (2020) obtained a parameter-free transformation matrix A_i of order $(n_i - 1) \times n_i$ and full rank that makes the covariance matrix of $A_i b_i$ diagonal with equal diagonal elements across $i = 1, \ldots, m$, where $b_i = (b_{i1}, \ldots, b_{im})^T$ (Li and Lahiri (2019) used a similar transformation in the context of unit level models). The transformed linking model is then used to get the ideal mean square error sum of squares MSE_{θ^*} and its estimator mse_{θ^*} along the lines of the method used for the FH linking model. Note that the transformed vector $\theta_i^* = A_i\theta_i$ has length $n_i - 1$ unlike the vector

θ_i with elements $\theta_{ij}, j = 1, \ldots, n_i$, and as a result each area loses one degree of freedom after transformation. The variable selection criteria can then be computed using mse^*_θ, as in the case of the FH model. Cai *et al.* (2020) report simulation results showing that the proposed transformation method performs well in variable selection, unlike the naive method treating the linking model as a FH-type model ignoring the correlations, especially as σ_v^2 increases.

Three-fold models linking sub-subarea means to related covariates and random effects at the area, subarea and sub-subarea levels are also used in practice to estimate sub-subarea means as well as subarea means. For example, the Program for the International Assessment of Adult Competencies (PIAAC) in the United States used a three-fold model with census divisions as areas, states within a census division as subareas and counties within a state as sub-subareas. Cai and Rao (2022) extended the two-fold model variable selection method of Cai *et al.* (2020) to variable selection to variable selection under three-fold models.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory* (B. N. Petrov and F. Csaki eds.). Akademiai Kiado, Budapest, 267-281.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716-723.

Bai, Z., Rao, C. R., and Wu, Y. (1999). Model selection with data-oriented penalty. *Journal of Statistical Planning and Inference*, **77**, 103-117.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **80**, 28-36.

Bozdogan, H. (1994). Editor's general preface, in *Engineering and Scientific Applications*, Vol. 3 (H. Bozdogan ed.), *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, pages ix-xii. Kluwer Academic Publishers, Dordrecht, Netherlands.

Cai, S. and Rao, J. N. K. (2022). Selection of auxiliary variables for three-fold linking models in small area estimation: a simple and effective method. *Stats*, **5**, 128-138.

Cai, S., Rao, J. N. K., Dumitrescu. L., and Chatrchi, G. (2020). Effective transformation -based variable selection under two-fold subarea models in small area estimation. *Statistics in Transition*, **21**, 68 – 83.

Chatterjee, S., Lahiri, P., and Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP, and related prediction intervals in linear mixed models. *Annals of Statistics*, **36**, 1221-1245.

Datta, G. S. and Lahiri, P. (2001). Discussions on a paper by Efron and Gous. *Model Selection*, IMS Lecture Notes/Monograph 38.

Fabrizi, E. and Lahiri, P. (2004). A new approximation to the Bayes information criterion in finite population sampling. Tech. Report, Department of Mathematics, University of Maryland.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269-277.

Friedman, J. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, **19**, 1-141.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, **9**, 55-93.

Han, B. (2013). Conditional Akaike information criterion in the Fay-Herriot model. *Statistical Methodology*, **11**, 53-67.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society B*, **41**, 190-195.

Harville, D. A. (1977). Maximum likelihood approaches to variance components estimation and related problems. *Journal of the American Statistical Association*, **72**, 320-340.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, New York.

Hodges, J. S. and Sargent, D. J. (2001), Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, **88**, 367-379.

Jiang, J. and Rao, J. S. (2003). Consistent procedures for mixed linear model selection. *Sankhya A*, **65**, 23-42.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation (with discussion). *Test*, **15**, 1-96.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, New York.

Jiang, J., Rao, J. S., Gu, Z., and Nguyen, T. (2008). Fence methods for mixed model selection. *Annals of Statistics*, **36**, 1669-1692.

Jiang, J., Nguyen, T., and Rao, J. S. (2010). A simplified adaptive fence procedure. *Statistics and Probability Letters*, **79**, 625-629.

Jiang, J., Rao, J. S., and Nguyen, T. (2011). Invisible fence methods and the identification of differentially expressed gene sets. *Statistics and Its Interface*, **4**, 403-415.

Kinney, S. K. and Dunson, D. B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics*, **63**, 690-698.

Kubokawa, T. (2011). Conditional and unconditional methods for selecting variables in linear mixed models. *Journal of Multivariate Analysis*, **102**, 641–660.

Lahiri, P. and Suntornchost, J. (2015). Variable selection for linear mixed models with application to small area estimation. *Sankhya B*, **77**, 312-320.

Li, Y. and Lahiri, P. (2019). A simple adaptation of variable selection software for regression models to select variables in nested error regression models. *Sankhya B*, **81**, 302 – 317.

Meza, J. and Lahiri, P. (2005). A note on the Cp statistic under the nested error regression model. *Survey Methodology*, **31**, 105-109.

Miller, J. J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of analysis of variance. *Annals of Statistics*, **5**, 746-762.

Mohadjer, L., Rao, J. N. K., Liu, B., Krenzke, T., and Van De Kerckhove, W. (2012). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *Journal of the Indian Society of Agricultural Statistics*, **66**, 55-63.

Muller, S., Scealy, J. S., and Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science*, **28**, 135-167.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics*, **12**, 758-765.

Opsomer, J. D., Breidt, F. J., Claeskens, G., Kauermann, G., and Ranalli, M. G. (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society B*, **70**, 265-286.

Pauler, D. K. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika*, **85**, 13–27.

Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika*, **76**, 369-374.

Rao, J. N. K. (2003). *Small Area Estimation*. Wiley, New York.

Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation, 2nd ed*. Wiley, New York.

Ruppert, R., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*. John Wiley and Sons, New York.

Shang, J. and Cavanaugh, J. E. (2008). Bootstrap variants of the Akaike information criterion for mixed model selection. *Computational Statistics and Data Analysis*, **52**, 2004–2021.

Shibata, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika*, **71**, 43-49.

Srivastava, M. S. and Kubokawa, T. (2010). Conditional information criteria for selecting variables in linear mixed models. *Journal of Multivariate Analysis*, **101**, 1970–1980.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed effects models. *Biometrika*, **92**, 351-370.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, **93**, 120-131.

Zheng, X., and Loh, W. Y. (1995). Consistent variable selection in linear models. *Journal of the American Statistical Association*, **90**, 151-156.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 327–360
http://www.ssca.org.in/journal

Gene-Gene and Gene-Environment Interactions in Case-Control Studies Based on Hierarchies of Dirichlet Processes

Durba Bhattacharya[1] **and Sourabh Bhattacharya**[2]
[1]*Department of Statistics, St. Xavier's College (Autonomous), Kolkata*
[2]*Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata*

Received: 29 May 2024; Revised: 30 August 2024; Accepted: 06 September 2024

Abstract

It is becoming increasingly clear that complex interactions among genes and environmental factors play crucial roles in triggering complex diseases. Thus, understanding such interactions is vital, which is possible only through statistical models that adequately account for such intricate, albeit unknown, dependence structures. In this article, we propose and develop a novel nonparametric Bayesian model for case-control genotype data using hierarchies of Dirichlet processes that offers a more realistic and nonparametric dependence structure among the genes, induced by the environmental variables. In this regard, we propose a novel and highly parallelisable MCMC algorithm that is rendered quite efficient by the combination of modern parallel computing technology, effective Gibbs sampling steps, retrospective sampling and Transformation based Markov Chain Monte Carlo (TMCMC). We devise appropriate Bayesian hypothesis testing procedures to detect the roles of genes and environment in case-control studies. Applying our ideas to 5 biologically realistic case-control genotype datasets simulated under distinct set-ups, we obtain encouraging results in each case. We finally apply our ideas to a real, myocardial infarction dataset, and obtain interesting results on gene-gene and gene environment interaction, that broadly agree with the results reported in the literature, but provide further important insights.

Key words: Case-control study; Hierarchical Dirichlet process; Gene-gene and gene-environment interaction; Myocardial Infarction; Parallel processing; Transformation based MCMC.

1. Introduction

In spite of much research on gene-gene interaction, including genome-wide association studies (GWAS), it has become increasingly clear that gene-gene interaction alone is insufficient for explaining most complex diseases. Investigating environmental factors independently of the genetic factors is not sufficient either – biomedical research points towards the importance of interactions between genes and the environment in explaining complex diseases. Indeed, according to Hunter (2005) (see also Mather and Caligary (1976)), considering only the separate contributions of genes and environment to a disease, ignoring their

Corresponding Author: Durba Bhattacharya
Email: durba@sxccal.edu

interactions, will lead to incorrect estimation of the disease proportion (the "population attributable fraction") that is explained by genes, the environment, and their joint effect. In particular, environmental exposures are expected to influence gene-gene interactions of the individuals. A comprehensive overview of gene-environment interaction with various examples is provided in Bhattacharya and Bhattacharya (2020).

Since no simple relationship exists between the genes and environment, it is clear that linear or additive models, as are mostly used so far, are inadequate for modeling gene-environment interactions. Also, the logistic model based approaches, (see for example Ahn *et al.* (2013), Wen and Stephens (2014) and Liu *et al.* (2015)) resting on Fisher's definition of interaction result in the inclusion of a large number of interaction terms even with a moderate number of genetic and environmental factors.

The fact that the genetic data may arise from a stratified population with an unknown number of subpopulations makes the problem all the more demanding. Wen and Stephens (2014), in their attempt to study the genetic association with respect to genetic data arising from multiple potentially-heterogeneous subgroups, assume the number of subgroups to be known in advance. Also, the problem of quantifying the strength of heterogeneity, as acknowledged by Wen and Stephens (2014), remains unanswered due to the above considerations and the need of an appropriate prior. The Bayesian semiparametric model proposed by Bhattacharya and Bhattacharya (2020) takes care of the above mentioned problems by proposing a model based on Dirichlet Processes (DP) and a hierarchical matrix-normal distribution, which encapsulates the mechanism of dependence among genes under environmental effects with respect to genotype data arising out of a possibly stratified population. In a somewhat similar spirit, Urbut *et al.* (2019) and Yang *et al.* (2024) propose mixture of multivariate nornal distributions with appropriate covariance matrices relevant for the phenomenon under study.

We now elaborate on a possible drawback of the dependence structure induced by the modeling strategy of Bhattacharya and Bhattacharya (2020), which motivated us to develop our present work based on Hierarchical Dirichlet Processes.

In their model, the relevant gene-gene covariance matrix for individual i is $\tilde{\sigma}_{ii}\boldsymbol{A}$, where $\boldsymbol{A}$ is the gene-gene interaction matrix common to all the individuals in the absence of environmental variables, and $\tilde{\sigma}_{ii} = \sigma_{ii} + \phi$, with σ_{ii} being the i-th diagonal element of a symmetric, positive definite matrix not associated with the environmental variable, and ϕ is a non-negative parameter, to be interpreted as the effect of the environmental variable $\boldsymbol{E}$ on gene-gene interaction. Note that Bhattacharya and Bhattacharya (2020) assumed that the covariance matrices for all the individuals are affected in the same way by the environmental variable, which seems to be a limitation of the covariance structure. The environmental variables may affect the gene-gene interactions of individuals differently depending on the extent and type of their exposure to the environmental factors.

In this article, we introduce a novel Bayesian nonparametric model for gene-gene and gene-environment interactions for case-control genotype data that solves the issues detailed above. Our model represents the individual genotype data as finite mixtures based on Dirichlet processes as before, but instead of the hierarchical matrix normal distribution, we introduce a hierarchy of Dirichlet processes that create appropriate nonparametric dependence among the genes induced by the environment, case-control dependence, and de-

pendence among the individuals. As we show, our modeling strategy satisfies all the desirable properties, bypassing the drawbacks of the matrix-normal based model of Bhattacharya and Bhattacharya (2020). The key idea of inducing such nonparametric dependence is to ensure that the minor allele frequencies associated with every sub-population, individual, gene and case/control status share a global pool of random parameters, in such a manner that only the dependence structure is influenced by the environmental variables, not the marginal distributions of the minor allele frequencies. The last point is important biologically and so, it requires care to model such intricate dependence.

Although our hierarchical Dirichlet process (HDP) model has parallels with the HDP introduced by Teh *et al.* (2006), our HDP has one more level of hierarchy compared to Teh *et al.* (2006). Moreover, the aforementioned special and intricate dependence structure has not been considered in any previous HDP application.

Exploiting conditional independence structures of our Bayesian model, we develop a novel and highly parallelisable Markov Chain Monte Carlo (MCMC) methodology that combines the efficiencies of modern parallel computing infrastructure, Gibbs steps, retrospective sampling methods, and Transformation based Markov Chain Monte Carlo (TMCMC). For the hypothesis testing procedures, we essentially adopt and extend the ideas provided in Bhattacharya and Bhattacharya (2020). Application of our model and methods to five simulation experiments for the validation purpose yielded quite encouraging results, and application to a real myocardial infarction (MI) case-control type dataset yielded results that are broadly in agreement with the results reported in the literature, but provided new and interesting insights into the mechanisms of gene-gene and gene-environment interactions.

The rest of our paper is structured as follows. We introduce our model in Section 2, and in Section 3 discuss the relevant dependence structures induced by our model. In Section 4 we extend the Bayesian hypothesis testing procedures proposed in Bhattacharya and Bhattacharya (2020) to learn about the roles of genes, environmental variables and their interactions in case-control studies. In Section 5 we briefly discuss the results of application of our model and methodologies to 5 biologically realistic simulated data sets, the details of which are provided in the Annexure, described below. In Section 6 we analyze the real MI dataset using our ideas, demonstrating quite interesting and insightful outcome. Finally, we summarize our work with concluding remarks in Section 7.

Additional details are provided in the Annexure, whose sections have the prefix "A-" when referred to in this paper.

2. A new Bayesian nonparametric model for gene-gene and gene-environme nt interactions

2.1. Case-control genotype data

For $s = 1, 2$ denoting the two chromosomes, let $y^s_{ijkr} = 1$ and $y^s_{ijkr} = 0$ indicate the presence and absence of the minor allele of the i-th individual belonging to the k-th group (either control or case), for $k = 0, 1$, with $k = 1$ denoting case; at the r-th locus of j-th gene, where $i = 1, \ldots, N_k$; $r = 1, \ldots, L_j$ and $j = 1, \ldots, J$; let $N = N_1 + N_2$. Let $\boldsymbol{E}_i$ denote a set of environmental variables associated with the i-th individual. In what follows, we model this case-control genotype and the environmental data using our Bayesian nonparametric model,

described in the next few sections.

2.2. Mixture models based on Dirichlet processes

Let $\boldsymbol{y}_{ijkr} = (y^1_{ijkr}, y^2_{ijkr})$, and if $L = \max\{L_1, \ldots, L_J\}$, let $\boldsymbol{Y}_{ijk} = (\boldsymbol{y}_{ijk1}, \boldsymbol{y}_{ijk2}, \ldots, \boldsymbol{y}_{ijkL_j})$ and $\tilde{\boldsymbol{Y}}_{ijk} = (\tilde{\boldsymbol{y}}_{ijk,L_j+1}, \ldots, \tilde{\boldsymbol{y}}_{ijkL})$, where $\tilde{\boldsymbol{Y}}_{ijk}$ are unobserved and assumed to be missing. We introduce these unobserved variables to match the number of loci for all the genes, which is required so that the vectors of minor allele frequencies come from the distribution having the same dimension. This "dimension-matching" is required for the theoretical development of our modeling ideas; see (5) and (6).

We assume that for every triplet (i, j, k), $\boldsymbol{X}_{ijk} = (\boldsymbol{x}_{ijk1}, \ldots, \boldsymbol{x}_{ijkL}) = (\boldsymbol{Y}_{ijk}, \tilde{\boldsymbol{Y}}_{ijk})$ have the mixture distribution

$$[\boldsymbol{X}_{ijk}] = \sum_{m=1}^{M} \pi_{mijk} \prod_{r=1}^{L} f\left(\boldsymbol{x}_{ijkr} | p_{mijkr}\right), \tag{1}$$

where $f\left(\cdot | p_{mijkr}\right)$ is the Bernoulli mass function given by

$$f\left(\boldsymbol{x}_{ijkr} | p_{mijkr}\right) = \left\{p_{mijkr}\right\}^{x^1_{ijkr}+x^2_{ijkr}} \left\{1 - p_{mijkr}\right\}^{2-(x^1_{ijkr}+x^2_{ijkr})}. \tag{2}$$

In the above, M denotes the *maximum* number of mixture components and p_{mijkr} stands for the minor allele frequency at the r-th locus of the j-th gene for the i-th individual of the k-th case/control group. Note that minor allele frequency is the frequency at which the second most common allele occurs in a given population.

Allocation variables z_{ijk}, with probability distribution

$$[z_{ijk} = m] = \pi_{mijk}, \tag{3}$$

for $i = 1, \ldots, N_k$ and $m = 1, \ldots, M$, allow representation of (1) as

$$[\boldsymbol{X}_{ijk} | z_{ijk}] = \prod_{r=1}^{L} f\left(\boldsymbol{x}_{ijkr} | p_{z_{ijk}ijkr}\right). \tag{4}$$

Following Majumdar *et al.* (2013), Bhattacharya and Bhattacharya (2018), we set $\pi_{mijk} = 1/M$, for $m = 1, \ldots, M$, and for all (j, k).

Letting $\boldsymbol{p}_{mijk} = (p_{mijk1}, p_{mijk2}, \ldots, p_{mijkL})$, we next assume that

$$\boldsymbol{p}_{1ijk}, \boldsymbol{p}_{2ijk}, \ldots, \boldsymbol{p}_{Mijk} \overset{iid}{\sim} \boldsymbol{G}_{ijk}; \tag{5}$$

$$\boldsymbol{G}_{ijk} \sim \mathrm{DP}\left(\alpha_{G,ik} \boldsymbol{G}_{0,jk}\right), \tag{6}$$

where $\mathrm{DP}\left(\alpha_{G,ik} \boldsymbol{G}_{0,jk}\right)$ stands for Dirichlet process with expected probability measure $\boldsymbol{G}_{0,jk}$ having precision parameter $\alpha_{G,ik}$, with

$$\log(\alpha_{G,ik}) = \mu_G + \boldsymbol{\beta}_G^T \boldsymbol{E}_{ik}, \tag{7}$$

where $\boldsymbol{E}_{ik}$ is a d-dimensional vector of continuous environmental variable for the i-th individual in the k-th group, $\boldsymbol{\beta}_G$ is a d-dimensional vector of regression coefficients, and μ_G is the intercept term. The model can be easily extended to include categorical environmental variables along with the continuous ones.

2.3. Hierarchical Dirichlet processes to induce dependence between the genes and case-control status

We further assume that for $k = 0, 1$,

$$\boldsymbol{G}_{0,jk} \overset{iid}{\sim} DP\left(\alpha_{G_0,k}\boldsymbol{H}_k\right);\ j = 1, \ldots, J, \tag{8}$$

where

$$\log(\alpha_{G_{0,k}}) = \mu_{G_0} + \boldsymbol{\beta}_{G_0}^T \bar{\boldsymbol{E}}_k, \tag{9}$$

with

$$\bar{\boldsymbol{E}}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} \boldsymbol{E}_{ik}. \tag{10}$$

We postulate the last level of hierarchy as

$$\boldsymbol{H}_k \overset{iid}{\sim} DP\left(\alpha_H \tilde{\boldsymbol{H}}\right);\ k = 0, 1, \tag{11}$$

where

$$\log(\alpha_H) = \mu_H + \boldsymbol{\beta}_H^T \bar{\bar{\boldsymbol{E}}}, \tag{12}$$

with

$$\bar{\bar{\boldsymbol{E}}} = \frac{\bar{\boldsymbol{E}}_0 + \bar{\boldsymbol{E}}_1}{2}. \tag{13}$$

We specify the base probability measure $\tilde{\boldsymbol{H}}$ as follows: for $m = 1, \ldots, M$, $i = 1, \ldots, N_k$, $k = 0, 1$, and $r = 1, \ldots, L$,

$$p_{mijkr} \overset{iid}{\sim} \text{Beta}\left(\nu_1, \nu_2\right), \tag{14}$$

under $\tilde{\boldsymbol{H}}$, where $\nu_1, \nu_2 > 0$.

This completes the specification of a hierarchy of Dirichlet processes to build dependence among the genes and the distributions of genotypes of cases-controls given data. Note that our model consists of one more level of hierarchy of Dirichlet processes than considered in the applications of Teh *et al.* (2006), who introduce hierarchical Dirichlet processes (HDP). Specifically, for given k and H_k, our hierarchy levels are comparable to that of Teh *et al.* (2006), but our extra level of hierarchy comes from (11), which creates dependence between case and control; details and reasons for insisting on such dependence structure are provided in Section 3.3.

Moreover, our likelihood based on Dirichlet processes ensuring at most M mixture components, is significantly different from those considered in the applications of Teh *et al.* (2006), which are based on the traditional DP mixture; see Mukhopadhyay *et al.* (2011), Mukhopadhyay *et al.* (2012), Mukhopadhyay and Bhattacharya (2013) for details on the conceptual, computational and asymptotic advantages of our modeling style over the traditional DP mixture.

2.4. The Chinese restaurant analogy

An extended version of the Chinese restaurant metaphor used by Teh *et al.* (2006) may be considered to illustrate our model. For $k = 0, 1$, the set of random probability measures $\{\boldsymbol{G}_{0,jk}; j = 1, \ldots, J\}$ can be associated with J restaurants. Letting τ_{ijk} denote the number of tables at the j-th restaurant associated with the i-th individual, we denote by $\boldsymbol{\phi}_{lijk}$ the dish being served at table l of the j-th restaurant for the i-th individual. Note that $\{\boldsymbol{\phi}_{lijk}; l = 1, \ldots, \tau_{ijk}; i = 1, \ldots, N_k\}$ is a set of *iid* realizations from $\boldsymbol{G}_{0,jk}$. Thus, we have different sets of realizations from $\boldsymbol{G}_{0,jk}$ for different individuals i.

For $k = 0, 1$, we also let $\boldsymbol{\Xi}_{R_k k} = \{\boldsymbol{\xi}_{1k}, \ldots, \boldsymbol{\xi}_{R_k k}\}$ denote a set of R_k *iid* realizations from $\boldsymbol{H}_k$. Then it follows that for $l = 1, \ldots, \tau_{ijk}$, $i = 1, \ldots, N_k$, and for $j = 1, \ldots, J$, $\boldsymbol{\phi}_{lijk} \in \boldsymbol{\Xi}_{R_k k}$. In other words, $\boldsymbol{\Xi}_{R_k k}$ is the set of distinct elements in the set $\{\boldsymbol{\phi}_{lijk}; l = 1, \ldots, \tau_{ijk}; i = 1, \ldots, N_k; j = 1, \ldots, J\}$, and, from the Chinese restaurant perspective, is the set of global dishes among all the restaurants, given k.

Finally, let $\boldsymbol{\zeta}_S = \{\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_S\}$ denote a set of S *iid* realizations from $\tilde{\boldsymbol{H}}$. Then it follows that $\boldsymbol{\zeta}_S$ is the set of distinct elements in $\{\boldsymbol{\Xi}_{R_k k} : k = 0, 1\}$. In other words, $\boldsymbol{\zeta}_S$ is the set of global dishes served in all the restaurants, irrespective of $k = 0$ or $k = 1$.

3. Discussion of the dependence structure induced by our HDP-based model

3.1. Dependence among individuals

It follows from the discussion in Section 2.4 that $\{\boldsymbol{\phi}_{lijk}; l = 1, \ldots, T_{mijk}; i = 1, \ldots, N_k\} \in \{\boldsymbol{\xi}_{1k}, \ldots, \boldsymbol{\xi}_{R_{mk}k}\}$, where $\boldsymbol{\xi}_{1k}, \ldots, \boldsymbol{\xi}_{R_{mk}k} \overset{iid}{\sim} \boldsymbol{H}_k$. This shows that $\{\boldsymbol{\phi}_{lijk}; l = 1, \ldots, T_{mijk}; i = 1, \ldots, N_k\}$ in (15) are shared among the individuals, thus creating dependence among the subjects.

For more precise insights regarding the dependence structure, let us first marginalize over $\boldsymbol{G}_{ijk}$ to obtain the joint distribution of $\boldsymbol{P}_{Mijk} = \{\boldsymbol{p}_{1ijk}, \ldots, \boldsymbol{p}_{Mijk}\}$ using the following Polya urn distributions: given $\boldsymbol{G}_{0,jk}$, $\boldsymbol{p}_{1ijk} \sim \boldsymbol{G}_{0,jk}$, and for $m = 2, \ldots, M$,

$$[\boldsymbol{p}_{mijk} | \boldsymbol{p}_{lijk}; l < m] = \frac{\alpha_{G,ik}}{\alpha_{G,ik} + m - 1} \boldsymbol{G}_{0,jk}(\boldsymbol{p}_{mijk}) + \frac{1}{\alpha_{G,ik} + m - 1} \sum_{t=1}^{T_{mijk}} \tilde{n}_{tmijk} \delta_{\boldsymbol{\phi}_{tijk}}(\boldsymbol{p}_{mijk}), \tag{15}$$

where $\sum_{t=1}^{T_{mijk}} \tilde{n}_{tmijk} = m - 1$. Here $\tilde{n}_{tmijk} = \#\{l < m : \boldsymbol{p}_{lijk} = \boldsymbol{\phi}_{tijk}\}$.

Since conditionally on $\boldsymbol{G}_{0,jk}$, the marginal distribution of $\boldsymbol{p}_{mijk}$, for $m = 1, \ldots, M$ and $i = 1, \ldots, N_k$, is $\boldsymbol{G}_{0,jk}$, the marginal is unaffected by the environmental variable, but the joint distribution of $\boldsymbol{P}_{Mijk}$ implied by the Polya urn distributions (15) shows that the dependence structure of $\boldsymbol{P}_{Mijk}$ is influenced by the regression on $\boldsymbol{E}_{ik}$ through $\alpha_{G,ik}$. This is a very desirable property of our modeling approach, since, in reality, the population minor allele frequencies for the case-control group are not expected to be affected by environmental variables, although environmental exposure is expected to influence dependence among individuals and gene-gene interactions in individuals. Note that marginal distributions depending upon environmental variables may be envisaged only under mutation, but since it is an extremely rare phenomenon and the type of case control type genotype data we are dealing with is not appropriate for such studies, we do not include mutational effects in our

model.

3.2. Dependence among the genes

We now show that the gene-gene interactions of the i-th individual are affected by $\boldsymbol{E}_{ik}$, but not the marginal effects of the genes.

Dependence among the genes for the i-th individual is induced by $\{\boldsymbol{\phi}_{tijk}; t = 1, \ldots, \tau_{ijk}; j = 1, \ldots, J\}$, where, for $t = 1, \ldots, \tau_{ijk}$, $\boldsymbol{\phi}_{tijk} \overset{iid}{\sim} \boldsymbol{G}_{0,jk}$, with $\boldsymbol{G}_{0,jk} \sim DP\left(\alpha_{G_0,k}\boldsymbol{H}_k\right)$. In fact, marginalizing over $\boldsymbol{G}_{0,jk}$ yields the following Polya urn scheme for $\{\boldsymbol{\phi}_{tijk}; t = 1, \ldots, \tau_{ijk}\}$:

$$[\boldsymbol{\phi}_{tijk}|\boldsymbol{\phi}_{lijk}; l < t] = \frac{\alpha_{G_0,k}}{\alpha_{G_0,k} + t - 1}\boldsymbol{H}_k\left(\boldsymbol{\phi}_{tijk}\right) + \frac{1}{\alpha_{G_0,k} + t - 1}\sum_{l=1}^{R_{tk}} \bar{n}_{ltik}\delta_{\boldsymbol{\xi}_{lk}}\left(\boldsymbol{\phi}_{tijk}\right), \tag{16}$$

where $\bar{n}_{ltik} = \#\left\{\ell < t : \boldsymbol{\phi}_{\ell ijk} = \boldsymbol{\xi}_{lk}\right\}$. Note that $\sum_{l=1}^{R_{tk}} \bar{n}_{ltik} = t - 1$.

It is clear from (16) that $\{\boldsymbol{\phi}_{tijk}; j = 1, \ldots, J\}$ share $\{\boldsymbol{\xi}_{lk}; l = 1, \ldots, R_k\}$, so that the latter set creates dependence among the genes. Moreover, it is also clear from (16) that the dependence structure does not depend directly upon $\boldsymbol{E}_{ik}$, but upon $\bar{\boldsymbol{E}}_k$, through the regression of $\log(\alpha_{G_0,k})$ on $\bar{\boldsymbol{E}}_k$; see (9). In other words, the gene-gene dependence structure of any individual is not directly influenced by the corresponding environmental variable. However, the dependence structure is also influenced by $\bar{n}_{ltik}$, which depends upon the i-th individual in the k-th case-control group through τ_{ijk}, which is directly influenced by $\boldsymbol{E}_{ik}$ through $\alpha_{G,ik}$. Thus, as is desirable, our modeling style induces gene-gene interactions that are specific to the individuals and are influenced by the corresponding environmental variables and the averages of the environmental variables within the case-control groups that the individuals belong to.

It is also interesting to observe that in spite of the individual-specific gene-gene interactions, the marginal distributions of $\boldsymbol{\phi}_{tijk}$ remains $\boldsymbol{G}_{0,jk}$ for the non-marginalized version and $\boldsymbol{H}_k$ for the marginalized version characterized by (16), signifying that the individual genes are not affected by $\boldsymbol{E}_{ik}$.

3.3. Case-control dependence

Finally, we note that

$$[\boldsymbol{\xi}_{sk}|\boldsymbol{\xi}_{lk}; l < s] = \frac{\alpha_H}{\alpha_H + s - 1}\tilde{\boldsymbol{H}}\left(\boldsymbol{\xi}_{sk}\right) + \frac{1}{\alpha_H + s - 1}\sum_{l=1}^{S_{sk}} \breve{n}_{lsk}\delta_{\boldsymbol{\zeta}_l}\left(\boldsymbol{\xi}_{sk}\right), \tag{17}$$

where $\breve{n}_{lsk} = \#\left\{\ell < s : \boldsymbol{\xi}_{\ell k} = \boldsymbol{\zeta}_l\right\}$ and $\sum_{l=1}^{S_{sk}} \breve{n}_{lsk} = s - 1$. So, $\{\boldsymbol{\xi}_{sk}; s = 1, \ldots, R_k; k = 0, 1\}$ share $\{\boldsymbol{\zeta}_l; l = 1, \ldots, S\}$, creating dependence between case and control status. Dependence between case and control status are likely to be caused by various implicit factors and environmental variables that are not accounted for in the study. These factors and environmental variables may be insignificant individually, but together may exert non-negligible influence on cases and controls.

A schematic diagram of our HDP-based model and the dependence structure is depicted in Figure 1. We remark that in a much simpler set-up, the original HDP proposed

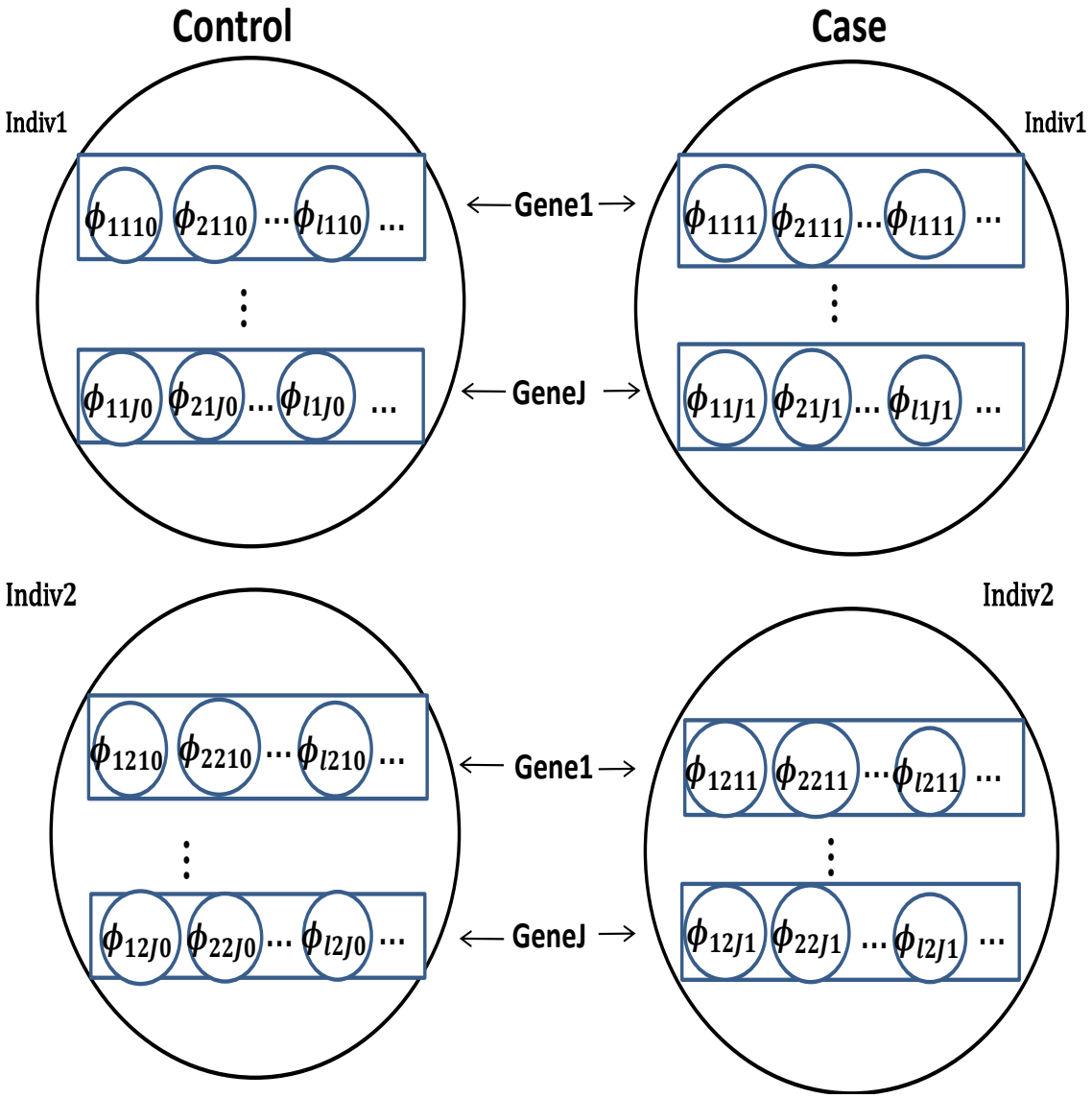


Figure 1: Schematic diagram of our HDP-based Bayesian model.

in Teh *et al.* (2006) has also been used by De Iorio *et al.* (2015) for inferring population admixture, allowing for correlations between loci due to linkage disequilibrium.

In Section A-1 we propose an MCMC procedure for the inferential purpose, and in Section A-2 we provide a parallel algorithm for implementing the MCMC method.

4. Detection of the roles of environment, genes and their interactions with respect to our HDP based model

4.1. Formulation of the tests and interpretation of their results

4.1.1. Bayesian test for the impact of the genes on case-control

To test if genes have any effect on case-control, we formulate as in Bhattacharya and Bhattacharya (2018) and Bhattacharya and Bhattacharya (2020), the following hypotheses:

$$H_{01} : h_{0j} = h_{1j};\ j = 1, \ldots, J, \tag{18}$$

versus

$$H_{11} : \text{not } H_0, \tag{19}$$

where

$$h_{0j}(\cdot) = \frac{1}{M} \sum_{m=1}^{M} \prod_{r=1}^{L_j} f\left(\cdot | p^r_{mi_0jk=0}\right); \tag{20}$$

$$h_{1j}(\cdot) = \frac{1}{M} \sum_{m=1}^{M} \prod_{r=1}^{L_j} f\left(\cdot | p^r_{mi_1jk=1}\right). \tag{21}$$

In the above, for $k = 0, 1$, i_k is the index such that $\boldsymbol{P}_{Mi_kjk} = \{\boldsymbol{p}_{1i_kjk}, \boldsymbol{p}_{2i_kjk}, \ldots, \boldsymbol{p}_{Mi_kjk}\}$ is some measure of central tendency of $\{\boldsymbol{P}_{Mijk} = \{\boldsymbol{p}_{1ijk}, \boldsymbol{p}_{2ijk}, \ldots, \boldsymbol{p}_{Mijk}\}; i = 1, \ldots, N_k\}$. Appropriate measures of central tendency, based on clusterings, is discussed in Section 4.2.1.

4.1.2. Bayesian test for significance of the environmental variables

To check if the environmental variables are significant, we shall test the following: for $\ell = 1, \ldots, d$,

$$H_{02\ell} : \beta_{G,\ell} = 0 \text{ versus } H_{12\ell} : \beta_{G,\ell} \neq 0, \tag{22}$$

$$H_{03\ell} : \beta_{G_0,\ell} = 0 \text{ versus } H_{13\ell} : \beta_{G_0,\ell} \neq 0, \tag{23}$$

and

$$H_{04\ell} : \beta_{H,\ell} = 0 \text{ versus } H_{14\ell} : \beta_{H,\ell} \neq 0. \tag{24}$$

4.1.3. Bayesian test for significance of gene-gene interaction

In our HDP based nonparametric model there is no readily available quantification of gene-gene interaction unlike the models of Bhattacharya and Bhattacharya (2018) and Bhattacharya and Bhattacharya (2020). Thus, in order to test for gene-gene interaction, it is necessary to first reasonably define such a measurement.

A measure of gene-gene interaction influenced by environmental variables

For our purpose, we first define

$$\bar{p}_{mijk} = \frac{\sum_{r=1}^{L_j} p_{mijkr}}{L_j}. \tag{25}$$

With the above definition, for subject i belonging to case-control group k, we consider the following covariance

$$C(i, j_1, j_2, k) = cov\left(\text{logit}(\bar{p}_{z_{ij_1k}ij_1k}), \text{logit}(\bar{p}_{z_{ij_2k}ij_2k})\right), \tag{26}$$

as quantification of subject-wise gene-gene dependence that accounts for population memberships of subject i with respect to genes j_1 and j_2, through z_{ij_1k} and z_{ij_2k}, where for any $p \in (0,1)$, $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$. Thus, gene-gene interaction associated with our model is subject-specific.

While implementing our model using our parallelised MCMC methodology, we simulate $C(i, j_1, j_2, k)$ at each iteration by generating $\{p_{mijkr} : r = 1, \ldots, L_j\}$ as many times as required from the respective full conditionals holding the remaining parameters fixed, and then compute the empirical covariance corresponding to (26) using the generated *iid* samples conditionally on the remaining parameters to approximate (26).

Formulation of the Bayesian tests for gene-gene interactions

To test for subject-wise gene-gene interaction, we consider the following tests: for $i = 1, \ldots, N_k$, $k = 0, 1$, and for $j_1, j_2 \in \{1, \ldots, J\}$,

$$H_{05ij_1j_2k} : C(i, j_1, j_2, k) = 0 \text{ versus } H_{15ij_1j_2k} : C(i, j_1, j_2, k) \neq 0. \tag{27}$$

4.1.4. Interpretations of the results of the above tests

The cases that can possibly arise and the respective conclusions are the following:

- For some appropriate divergence measure d between two distributions, if $\max_{1 \leq j \leq J} d(h_{0j}, h_{1j})$, is significantly small with high posterior probability, then H_{01} is to be accepted. If h_{0j} and h_{1j} are not significantly different, then it is plausible to conclude that the j-th gene is not marginally significant in the case-control study.
- Suppose that H_{01} is accepted (so that genes have no significant role) and that at least one of $\beta_{G,\ell}$ or $\beta_{G_0,\ell}$ or $\beta_{H,\ell}$ is significant, at least for some ℓ. This may be interpreted as the environmental variable $\boldsymbol{E}$ having some altering effect on all the genes in a way that doesn't affect the disease status. If $C(i, j_1, j_2, k)$ turns out to be significant, then this would additionally imply that the environmental variable $\boldsymbol{E}$ influences interaction between genes j_1 and j_2 for the i-th individual, but not in a way that is responsible for the case/control status.
- If H_{01} is rejected, indicating that the genes are significant, but none of the $\beta_{G,\ell}$, $\beta_{G_0,\ell}$, $\beta_{H,\ell}$ or $C(i, j_1, j_2, k)$ are significant, then only the genes, not $\boldsymbol{E}$, are responsible for the disease. In that case, one may conclude that the disease is of purely genetic nature.

- Suppose that H_{01} is rejected, none of $\beta_{G,\ell}$, $\beta_{G_0,\ell}$, $\beta_{H,\ell}$ is significant, but $C(i, j_1, j_2, k)$ is significant for at least some i, j_1, j_2, k. Then the environmental variable is not significant, and the case/control status of the individuals associated with significant gene-gene interactions can be attributed to purely genetic causes triggered by gene-gene interactions of the individuals.

- Now suppose that H_{01} is rejected, and at least one of $\beta_{G,\ell}$, $\beta_{G_0,\ell}$, $\beta_{H,\ell}$ is significant, but none of the subject-wise gene-gene interactions is significant. Then the environmental variable $\boldsymbol{E}$ does not significantly affect the interactions to determine the case/control status, and marginal effects of the individual genes are responsible for the case/control status of an individual.

- If, on the other hand, H_{01} is rejected, at least one of $\beta_{G,\ell}$, $\beta_{G_0,\ell}$, $\beta_{H,\ell}$ is significant, and $C(i, j_1, j_2, k)$ is significant for at least some i, j_1, j_2, k, then the environmental variable is significant and is responsible for influencing gene-gene interactions within the individuals with significant $C(i, j_1, j_2, k)$, which, in turn, affects the case/control status of the individuals.

4.2. Methodologies for implementing the Bayesian tests

4.2.1. Hypothesis testing based on clustering modes

As in Bhattacharya and Bhattacharya (2018) and Bhattacharya and Bhattacharya (2020), here we exploit the concept of "central" clustering introduced by Mukhopadhyay *et al.* (2011). Briefly, central clustering may be interpreted as a suitable measure of central tendency of a set of clusterings. Mukhopadhyay *et al.* (2011) particularly consider the mode(s) of the set of clusterings, and provide methods for appropriately obtaining the mode(s) using a suitable metric that they propose to quantify distances between any two clusterings. Their proposed metric is also computationally inexpensive, which makes the concept based on central clusterings extremely useful in practice.

For $k = 0, 1$, let i_k denote the index of the central clusterings of $\boldsymbol{P}_{Mijk} = \{\boldsymbol{p}_{1ijk}, \boldsymbol{p}_{2ijk}, \ldots, \boldsymbol{p}_{Mijk}\}$, $i = 1, \ldots, N_k$. We then study the divergence between the two clusterings of

$$\boldsymbol{P}_{Mi_0jk=0} = \{\boldsymbol{p}_{1i_0jk=0}, \boldsymbol{p}_{2i_0jk=0}, \ldots, \boldsymbol{p}_{Mi_0jk=0}\}$$

and

$$\boldsymbol{P}_{Mi_1jk=1} = \{\boldsymbol{p}_{1i_1jk=1}, \boldsymbol{p}_{2i_1jk=1}, \ldots, \boldsymbol{p}_{Mi_1jk=1}\},$$

for $j = 1, \ldots, J$. A schematic diagram illustrating the idea can be found in Bhattacharya and Bhattacharya (2020).

Significantly large divergence between the two clusterings clearly indicates that the j-th gene is marginally significant.

4.2.2. Enhancement of clustering metric based inference using Euclidean distance

As argued in Bhattacharya and Bhattacharya (2018), significantly large clustering distance between $\boldsymbol{P}_{Mjk=0}$ and $\boldsymbol{P}_{Mjk=1}$ indicates rejection of H_0, but insignificant clustering distance does not necessarily provide strong evidence in favour of the null. In this regard, Bhattacharya and Bhattacharya (2018) (see also Bhattacharya and Bhattacharya (2020)) argue

that the Euclidean distance is an appropriate candidate to be tested for significance before arriving at the final conclusion. Briefly, we first compute the averages $\bar{p}_{mijk} = \sum_{r=1}^{L_j} p_{m,ijkr}/L_j$, then consider their logit transformations $\text{logit}\,(\bar{p}_{mijk}) = \log\{\bar{p}_{mijk}/(1-\bar{p}_{mijk})\}$. Then, we compute the Euclidean distance between the vectors

$$\text{logit}\left(\bar{\boldsymbol{P}}_{Mi_0jk=0}\right) = \{\text{logit}\,(\bar{p}_{1i_0jk=0}), \text{logit}\,(\bar{p}_{2i_0jk=0}), \ldots, \text{logit}\,(\bar{p}_{Mi_0jk=0})\}$$

and

$$\text{logit}\left(\bar{\boldsymbol{P}}_{Mi_1jk=1}\right) = \{\text{logit}\,(\bar{p}_{1i_1jk=1}), \text{logit}\,(\bar{p}_{2i_1jk=1}), \ldots, \text{logit}\,(\bar{p}_{Mi_1jk=1})\}.$$

We denote the Euclidean distance associated with the j-th gene by

$$d_{E,j} = d_{E,j}\left(\text{logit}\left(\bar{\boldsymbol{P}}_{Mi_0jk=0}\right), \text{logit}\left(\bar{\boldsymbol{P}}_{Mi_1jk=1}\right)\right),$$

and denote $\max\limits_{1\le j\le J} d_{E,j}$ by d_E^*.

4.2.3. Formal Bayesian hypothesis testing procedure integrating the above developments

In our problem, we need to test the following for reasonably small choices of ε's:

$$H_{0,d^*}: d^* < \varepsilon_{d^*} \text{ versus } H_{1,d^*}: d^* \ge \varepsilon_{d^*}; \tag{28}$$

$$H_{0,d_E^*}: d_E^* < \varepsilon_{d_E^*} \text{ versus } H_{1,d_E^*}: d_E^* \ge \varepsilon_{d_E^*}; \tag{29}$$

for $\ell = 1, \ldots, d$,

$$H_{0,\beta_{G,\ell}}: |\beta_{G,\ell}| < \varepsilon_{G,\ell} \text{ versus } H_{1,\beta_{G,\ell}}: |\beta_{G,\ell}| \ge \varepsilon_{G,\ell}, \tag{30}$$

$$H_{0,\beta_{G_0,\ell}}: |\beta_{G_0,\ell}| < \varepsilon_{G_0,\ell} \text{ versus } H_{1,\beta_{G_0,\ell}}: |\beta_{G_0,\ell}| \ge \varepsilon_{G_0,\ell}, \tag{31}$$

$$H_{0,\beta_{H,\ell}}: |\beta_{H,\ell}| < \varepsilon_{H,\ell} \text{ versus } H_{1,\beta_{H,\ell}}: |\beta_{H,\ell}| \ge \varepsilon_{H,\ell}, \tag{32}$$

and, for $i = 1, \ldots, N_k$, $k = 0, 1$, $j_1, j_2 \in \{1, \ldots, J\}$,

$$H_{0,C_{i,j_1,j_2,k}}: |C_{i,j_1,j_2,k}| < \varepsilon_{C,ij_1j_2k} \text{ versus } H_{1,C_{i,j_1,j_2,k}}: |C_{i,j_1,j_2,k}| \ge \varepsilon_{C,ij_1j_2k}, \tag{33}$$

If H_0 is rejected in (28) or in (29), we could also test if the j-th gene is influential by testing, for $j = 1, \ldots, J$, $H_{0,\hat{d}_j}: \hat{d}_j < \varepsilon_{\hat{d}_j}$ versus $H_{1,\hat{d}_j}: \hat{d}_j \ge \varepsilon_{\hat{d}_j}$, where $\hat{d}_j = \hat{d}\,(\boldsymbol{P}_{Mi_0jk=0}, \boldsymbol{P}_{Mi_1jk=0})$; we could also test $H_{0,d_{E,j}}: d_{E,j} < \varepsilon_{d_{E,j}}$ versus $H_{1,d_{E,j}}: d_{E,j} \ge \varepsilon_{d_{E,j}}$.

4.2.4. Null model and choice of ε

To obtain the null posterior distribution, we fit our HDP-based Bayesian model to the dataset generated from the HDP-based model where the genes are independent and not influenced by the environmental variable, and where there is no difference between the probabilities associated with case and control. For the null data we chose the same number of genes, the same number of loci for each gene, and the same number of cases and controls as the non-null data. We also choose the same value M as in the non-null model, but set $\beta_G = \beta_{G_0} = \beta_H = 0$. To generate the data from the null model, we first

simulate, independently for $j = 1, \ldots, J$, the set $\{\boldsymbol{p}_{m1j0} : m = 1, \ldots, M\}$, using the Polya urn scheme involving $\tilde{\boldsymbol{H}}$ and α_H, and set $\{\boldsymbol{p}_{m1j1} : m = 1, \ldots, M\} = \{\boldsymbol{p}_{m1j0} : m = 1, \ldots, M\}$, so that there is no difference between the probabilities associated with case and control, and that the genes are independent. Since the simulation method is independent of the environmental variable, it is clear that the genes are not influenced by the environment. Given the probabilities $\{\boldsymbol{p}_{m1j1} : m = 1, \ldots, M\}$ and $\{\boldsymbol{p}_{m1j0} : m = 1, \ldots, M\}$, we then simulate the data using our Bernoulli model. To the data thus generated, we fit our full HDP-based Bayesian model, to obtain the null posterior.

As in Bhattacharya and Bhattacharya (2018) here also we specify ε's as $F^{-1}(0.55)$, where F is the distribution function of the relevant benchmark null posterior distribution. Bhattacharya and Bhattacharya (2018) showed that the choice $F^{-1}(0.55)$, rather than the median, ensures that the correct null hypothesis is accepted under the "$0-1$" loss. Note that, for the median, the posterior probability of the true null is 0.5, while under the "$0-1$" loss, the true null will be accepted if its posterior probability is greater than $1/2$.

5. Simulation studies

For simulation studies, we first generate realistic biological data for stratified population with known gene-environment interaction from the GENS2 software of Pinelli *et al.* (2012). To this data, we then apply our model and methodologies in an effort to detect gene-environment interaction effects that are present in the data. We consider simulation studies in 5 different true model set-ups: (a) presence of gene-gene and gene-environment interaction, (b) absence of genetic or gene-environmental interaction effect, (c) absence of genetic and gene-gene interaction effects but presence of environmental effect, (d) presence of genetic and gene-gene interaction effects but absence of environmental effect, and (e) independent and additive genetic and environmental effects. The details of our simulation experiments are provided in Section A-3 of the supplement. Here we briefly summarize the results of our experiments.

In case (a), we correctly obtained clear significance of the influence of genetic effects. Also, β_H turned out to be very significant, demonstrating significant overall impact of the environmental variable on the genes. Interestingly, as one may expect, there are more instances of significant gene-gene interactions in the case group compared to the control group. The posteriors of the number of sub-populations gave high probabilities to the correct number of sub-populations in all the 5 simulation experiments. Quite importantly, we demonstrate in cases (a), (d) and (e) where the genes are relevant, that our HDP model can detect disease predisposing loci (DPL) with more precision compared to the matrix-normal-inverse-Wishart model for gene-environment interactions proposed in Bhattacharya and Bhattacharya (2020). In case (b) using our ideas in conjunction with significance testing in a simple logistic regression framework, we are correctly able to conclude that the genetic or gene-environmental effects are insignificant. As in Bhattacharya and Bhattacharya (2020), the right conclusion is arrived at in case (c) by utilizing our ideas in conjunction with the Akaike Information Criterion (AIC) in the context of simple logistic regression. Using our Bayesian testing procedure along with the aid of logistic regression, we have been able to correctly obtain insignificance of the environmental variable and significance of the genes. In this experiment, we found no gene-gene interaction in the control group and found two (marginal) instances of gene-gene interaction among the cases. As regards case (e), we note as in Bhattacharya and

Bhattacharya (2020) that additivity of genetic and environmental effects is not supported even by our current HDP-based Bayesian model. In spite of this, we correctly obtained significance of the environmental variable and precisely obtained the DPLs. But the lack of the additivity criterion in our model seems to have forced gene-environment interactions. Bhattacharya and Bhattacharya (2020) report similar results, who obtained, after eventually resorting to logistic regression, the AIC-based best model consisting of the additive marginal effects of the first gene and the environmental variable, along with an additive intercept, which is broadly consistent with the data-generating mechanism.

6. Application of our HDP based ideas to a real, case-control dataset on Myocardial Infarction

We now consider application of our model and methods to a case-control dataset on early-onset of myocardial infarction (MI) from MI Gen study, obtained from the dbGaP database **http://www.ncbi.nlm.nih.gov/gap**. The same dataset has been analyzed by Bhattacharya and Bhattacharya (2018) without considering the sex variable as the covariate, and by Bhattacharya and Bhattacharya (2020), who incorporate the sex variable in their gene-environment interaction model. Although Bhattacharya and Bhattacharya (2018) obtained significant genetic and gene-gene interaction effects, their later study after considering sex as the environmental variable, revealed strong effects of the sex variable but no significant gene-gene interaction, although many of the genes turned out to be individually significant. In our current HDP based analysis, we once again obtain strong effects of the sex variable, but in contrast with Bhattacharya and Bhattacharya (2020), although we obtain significant genetic effects, none of the genes turned out to be significant individually. Moreover, the subject-wise gene-gene interactions, although of small magnitude, turned out to be significant in some cases, and interestingly (and apparently counter-intuitively) seem to be instrumental in counter-acting the disease rather than provoking it.

6.1. Data description

We recall that the MI Gen data obtained from dbGaP consists of observations on presence/absence of minor alleles at 727478 SNP markers associated with 22 autosomes and the sex chromosomes of 2967 cases of early-onset myocardial infarction, 3075 age and sex matched controls. The average age at the time of MI was 41 years among the male cases and 47 years among the female cases. The data broadly represents a mixture of four subpopulations: Caucasian, Han Chinese, Japanese and Yoruban. Using the Ensembl human genome database (**http://www.ensembl.org/**) we could categorize 446765 markers out of 727478 with respect to 37233 genes.

As in Bhattacharya and Bhattacharya (2020) we considered 32 genes covering 1251 loci, for 200 individuals. These 1251 loci include 33 SNPs that are believed to be associated with MI and also those that are believed to be associated with different cardiovascular end points like LDL cholesterol, smoking, blood pressure, body mass, etc. Other than the 33 SNPs, the remaining 1218 SNPs are not known to be associated with the disease. See Bhattacharya and Bhattacharya (2020) for the details and the relevant references.

Since the four broad sub-populations are not unlikely to admit further genetic subdivisions, it makes sense to set the maximum number of mixture components, M, to a

value much larger than 4. As before, we set $M = 30$; we also set $\nu_1 = \nu_2 = 1$, so that $\tilde{\boldsymbol{H}}$ is the uniform distribution on $[0, 1]$. As in the simulation experiments, here also the structures $\alpha_{G,ik} = 0.1 \times \exp(100 + \mu_G + \beta_G E_{ik})$, $\alpha_{G_0,k} = 0.1 \times \exp\left(100 + \mu_{G_0} + \beta_{G_0}\bar{E}_k\right)$ and $\alpha_H = 0.1 \times \exp\left(100 + \mu_H + \beta_H \bar{\bar{E}}\right)$, where $\mu_G, \mu_{G_0}, \mu_H \overset{iid}{\sim} U(0, 1)$ and $\beta_G, \beta_{G_0}, \beta_H \overset{iid}{\sim} U(-1, 1)$, ensured adequate number of sub-populations and satisfactory mixing of MCMC. For the null data and model, we follow the same procedure as discussed in Section 4.2.4.

6.2. Remarks on model implementation

Our parallel MCMC algorithm detailed in Section A-2 takes about 7 days to generate 30,000 iterations on our VMware consisting of 50 double-threaded, 64-bit physical cores, each running at 2493.990 MHz. We discard the first $10,000$ iterations as burn-in, using the subsequent 20,000 iterations for our Bayesian inference. Satisfactory mixing properties are indicated by informal convergence diagnostics such as trace plots.

6.3. Results of the real data analysis

6.3.1. Effect of the sex variable

We obtain $P(|\beta_G| < \varepsilon_{\beta_G}|\text{Data}) \approx 0$, $P(|\beta_{G_0}| < \varepsilon_{\beta_{G_0}}|\text{Data}) \approx 0$ and $P(|\beta_H| < \varepsilon_{\beta_H}|\text{Data}) \approx 1$. In other words, although $\bar{\bar{E}}$ (here E being the sex variable) is insignificant, both E_{ik} and $\bar{E}_k$ are very significant. Thus, in this study, sex seems to play an important role in influencing the genes.

6.3.2. Roles of individual genes

With the clustering metric we obtained $P(d^* < \epsilon_1|\text{Data}) \approx 0.030$ while that with the Euclidean distance we obtained $P(d_E^* < \epsilon_2|\text{Data}) \approx 0.540$. That is, the maximum of the gene-wise clustering metrics turns out to be significant, while the maximum of the gene-wise Euclidean metrics is seen to be insignificant. The same ambiguity was also obtained by Bhattacharya and Bhattacharya (2020). The tests of the marginal genes are expected to shed some light regarding this dilemma. The posterior probabilities of the null hypotheses (of no significant genetic influence) reveal that none of the individual genes are significant, for either the clustering metric or the Euclidean metric. Our result is not much different from that of Bhattacharya and Bhattacharya (2020) who also note that their marginal probabilities, at least for the clustering metric, are not significantly small to provide strong enough evidences against the nulls.

Now, at least from the clustering metric perspective, it is necessary to explain the issue that all the genes are insignificant individually but still the maximum of the gene-wise clustering metric values is significant. The key to this issue seems to be the correlations between the distances, which are induced by gene-gene interactions. We explain this phenomenon using a bivariate normal example. Let (X_1, X_2) have a bivariate normal distribution with means 0, variances 1, and correlation ρ. Figure 2 depicts the median of $\max\{X_1, X_2\}$ as a function of ρ, which is seen to be increasing as ρ decreases from 1 to -1. On the other hand, the medians of the marginal distributions of X_1 and X_2 remain zero, irrespective of the value of ρ. Thus, it seems that gene-gene interaction does have some role to play in this study.

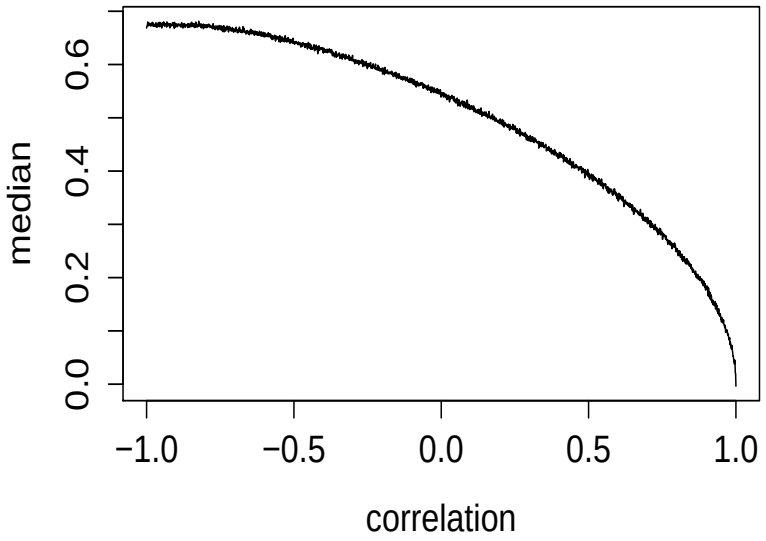


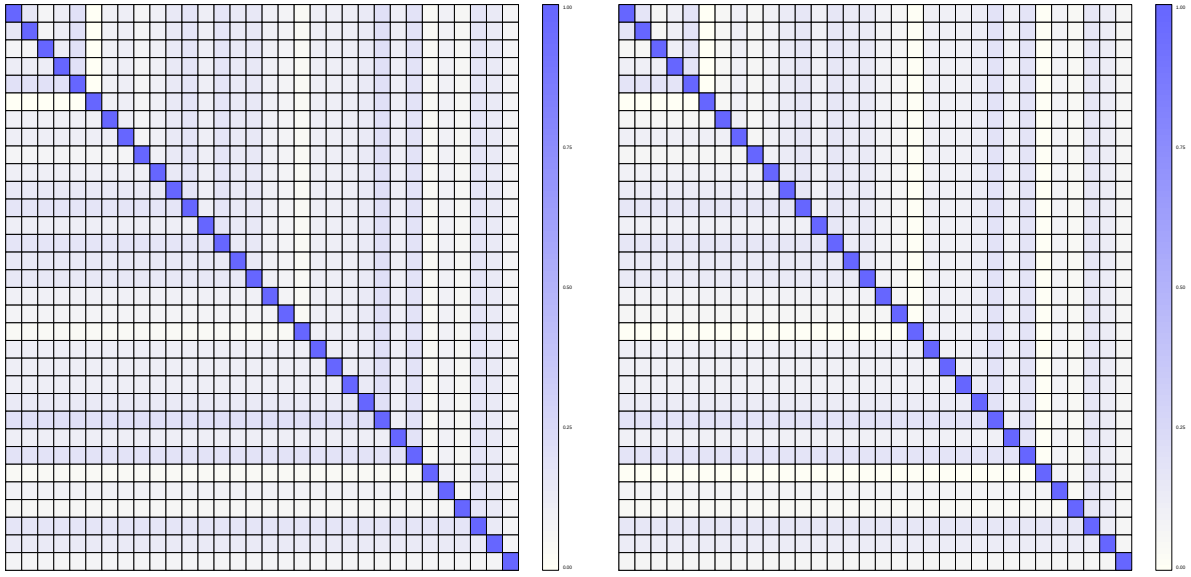
Figure 2: Bivariate normal example: Plot of the median of $\max\{X_1, X_2\}$ **with respect to the correlation** ρ**.**

6.3.3. Gene-gene interactions

Unlike Bhattacharya and Bhattacharya (2020), where there is a single gene-gene correlation structure for all the individuals, our current model has provision for subject-specific gene-gene correlations. Figures 3 and 4 show the typical gene-gene correlations representative of cases and controls in all males and females respectively. Essentially, the pictures represent the gene-gene correlation patterns for all the subjects. The color intensities correspond to the absolute values of the correlations. Although the correlations are small in all the subjects, the tests of hypotheses reveal some interesting structures. Figures 5 and 6 represent the all possible interacting patterns found in the study. Panel (a) of Figure 5 represents 9 male cases where no gene-gene interaction is significant. Panel (b) shows the genetic interaction pattern in some male cases where $AP006216.10$ and $C6orf106$, interact with all the other genes. Panel (c) shows the results of significance tests of gene-gene interactions for some male cases, for whom only $AP006216.10$ interacts with all the other genes in the study. A representative interaction pattern for the male controls shown in panel (d), indicates that only $C6orf106$ or only $AP006216.10$ interacts with every gene, but in a few subjects both $AP006216.10$ and $C6orf106$ interact with all the genes.

Even for the females, the two genes, $AP006216.10$ and $C6orf106$, play significant roles in gene-gene interactions. Indeed, in our data, unlike the 9 male cases, there is no female for whom all gene-gene interactions are insignificant. The relevant representative plots for the females, given by Figure 6, shows that for all the female cases, only $AP006216.10$ interacts with the other genes. For the female controls, either only $AP006216.10$ or only $C6orf106$ interacts with the other genes, or both $AP006216.10$ and $C6orf106$ interact significantly with the other genes included in the study.

The messages gained from our analysis seem to be somewhat counter-intuitive but perhaps quite insightful. Our tests indicate that the genes have insignificant marginal effect. Thus, some external or non-genetic factors might have some significant role to play. But for most of the subjects, at least one of the genes $AP006216.10$ and $C6orf106$ interact

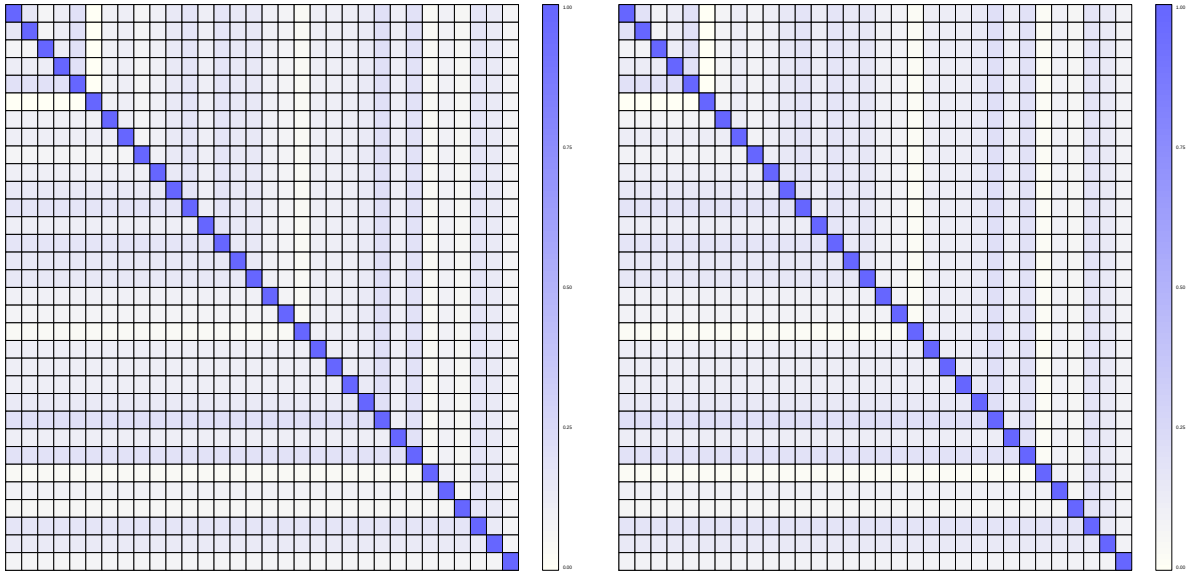


(a) Male case. (b) Male control.

Figure 3: Typical median gene-gene posterior correlation plot for male cases and male control.

with every other gene. The subjects, for whom no significant genetic interactions involving $AP006216.10$ and $C6orf106$ were detected, turned out to be male cases, indicating that the lack of genetic interaction in these males failed to get them any preventive measure against MI. On the other hand, the interactions of the genes $AP006216.10$ and $C6orf106$ with all the genes seemed to reduce the risk of the disease for the other subjects. Thus, in this study, the gene-gene interactions seem to have a beneficial effect on the subjects. It also seems that only a small proportion of males are prone to the risk of having no beneficial gene-gene interactions.

Note that our results are broadly consistent with those obtained by Bhattacharya and Bhattacharya (2020) but are more precise and informative. Indeed, they also noted relatively small impact of the individual genes and small gene-gene correlations. Our current ideas and analyses also support their conclusion that external factors (in particular, sex) are perhaps playing a bigger role in explaining case-control with respect to MI. We recall (see Bhattacharya and Bhattacharya (2020)) that with respect to the data that we used, the empirical conditional probability of a male given case is about 0.38, and that of a male given control is about 0.50, so that females seem to be more at risk, given our data. The inherent coherence of the Bayesian paradigm upholds the sex factor by attaching little importance to the individual genes. However, in contrast with Bhattacharya and Bhattacharya (2020) who found no interacting genes, here it turns out that the genes $AP006216.10$ and $C6orf106$ in interaction with other genes generally lower the risk of the individuals with respect to MI. Importantly, each of the few males having no such interactions turned out to be a case. This seems to be roughly in accordance with the popular belief that males are more susceptible to MI than females. Our Bayesian model coherently weaves together the prior and the data and brings out this information in spite of the data-driven information that females are more prone to MI than males. We also note that Lucas *et al.* (2012), who analyzed the same MI



(a) Female case. (b) Female control.

Figure 4: Typical median gene-gene posterior correlation plot for female cases and female controls.

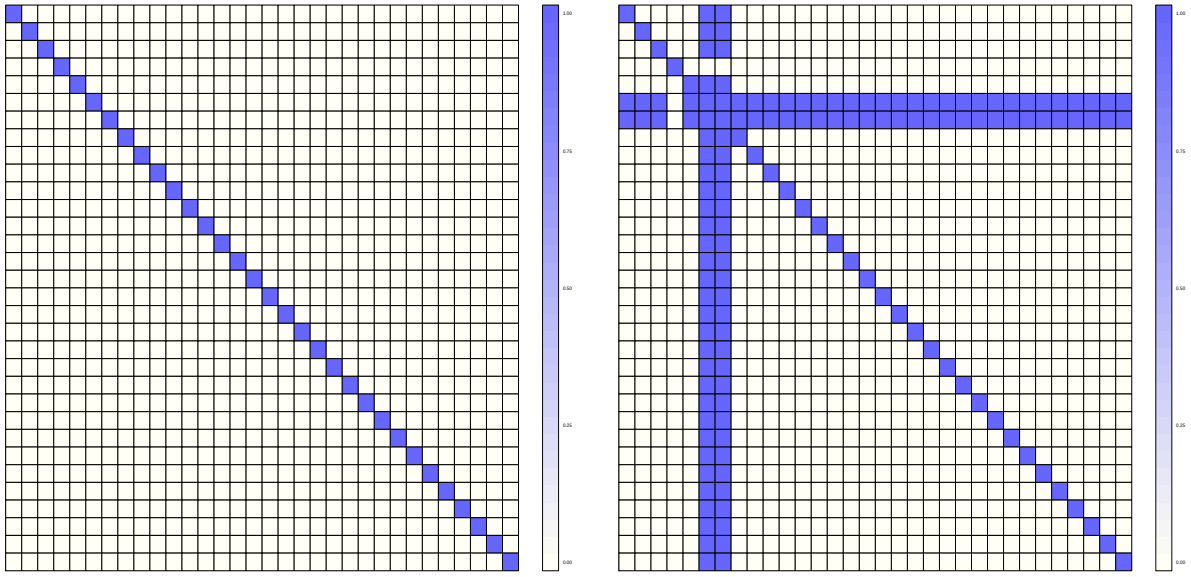
dataset using logistic regression, reached the conclusion that there is no significant gene-gene interaction. Thus, their result completely supports that of Bhattacharya and Bhattacharya (2020) and are also very much in keeping with our current results.

6.3.4. Posteriors of the number of sub-populations

The posterior distributions of the number of sub-populations for the males and females turned out to be quite similar, irrespective of case and control, with the mode at 3 and 4 components receiving the next highest probability. Thus, the 4 sub-populations, irrespective of sex, are well-supported by our model, showing that these can not be further sub-divided genetically. This is not unexpected, since the roles of the individual genes are not significant in our study. Our result broadly agrees with Bhattacharya and Bhattacharya (2020) who obtained for different genes, the modes at 5 components, with 4 components receiving the next highest posterior mass.

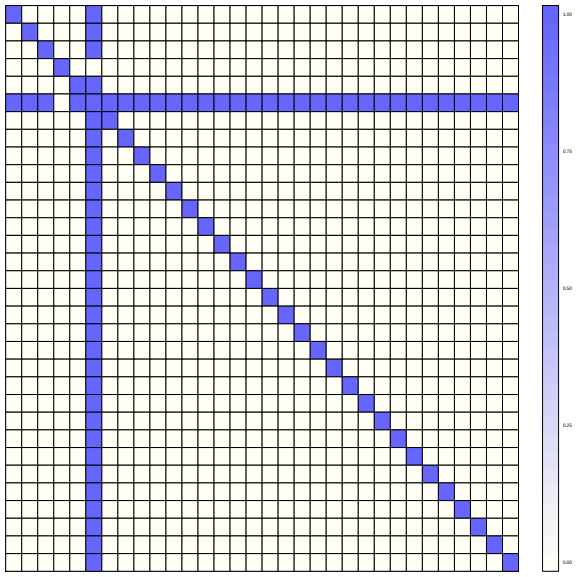
7. Summary and conclusion

In this paper, we have proposed a novel Bayesian nonparametric gene-gene and gene-environment interaction model based on hierarchies of Dirichlet processes. This model is a significant improvement over that of Bhattacharya and Bhattacharya (2020) in the sense of much clear interpretability and accounting for subject-specific gene-gene interactions. Moreover, the interactions arise as natural by-products of our nonparametric structure based on HDP, and are not based on matrix normal distributions, as in Bhattacharya and Bhattacharya (2018) and Bhattacharya and Bhattacharya (2020), and hence, are more realistic. We propose a novel parallel MCMC algorithm to implement our model, that combines powerful technology with conditionally independent structures inherent within our HDP based

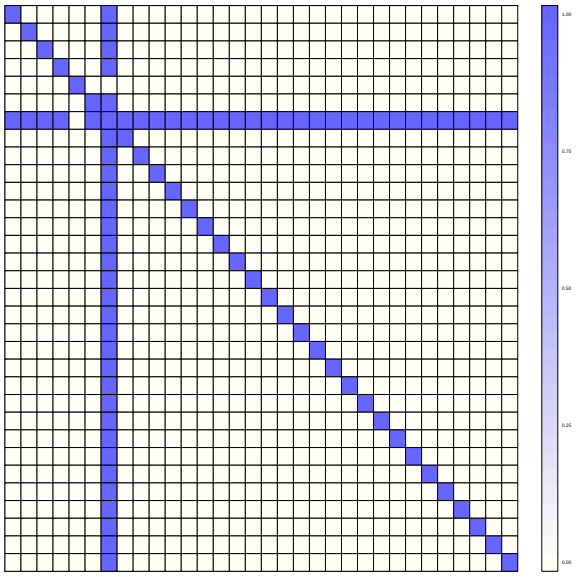


(a) Male case.

(b) Male case.

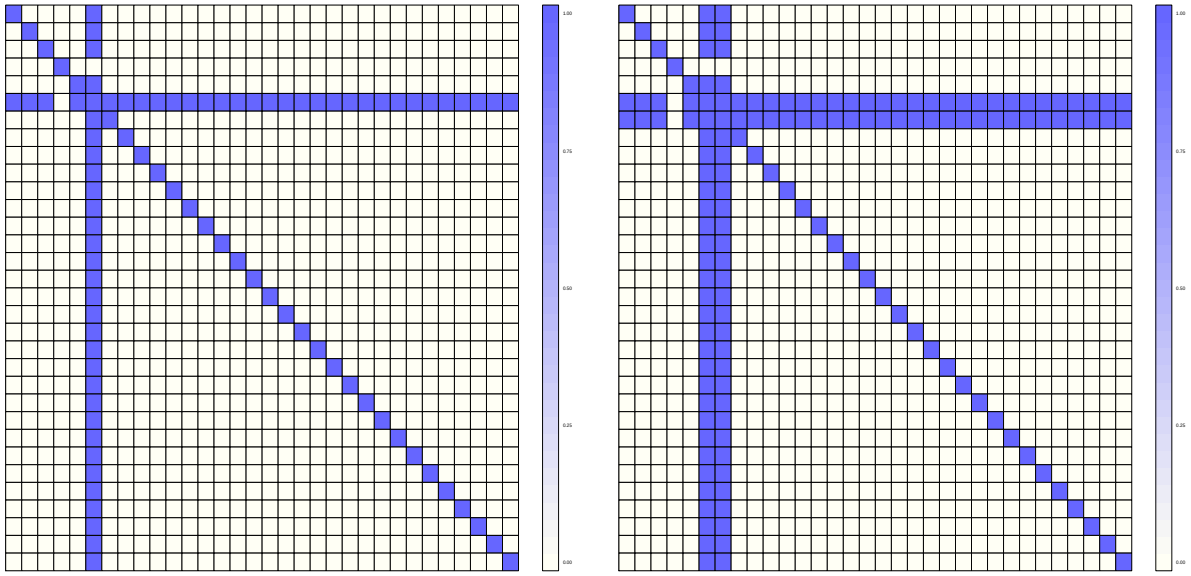


(c) Male case.



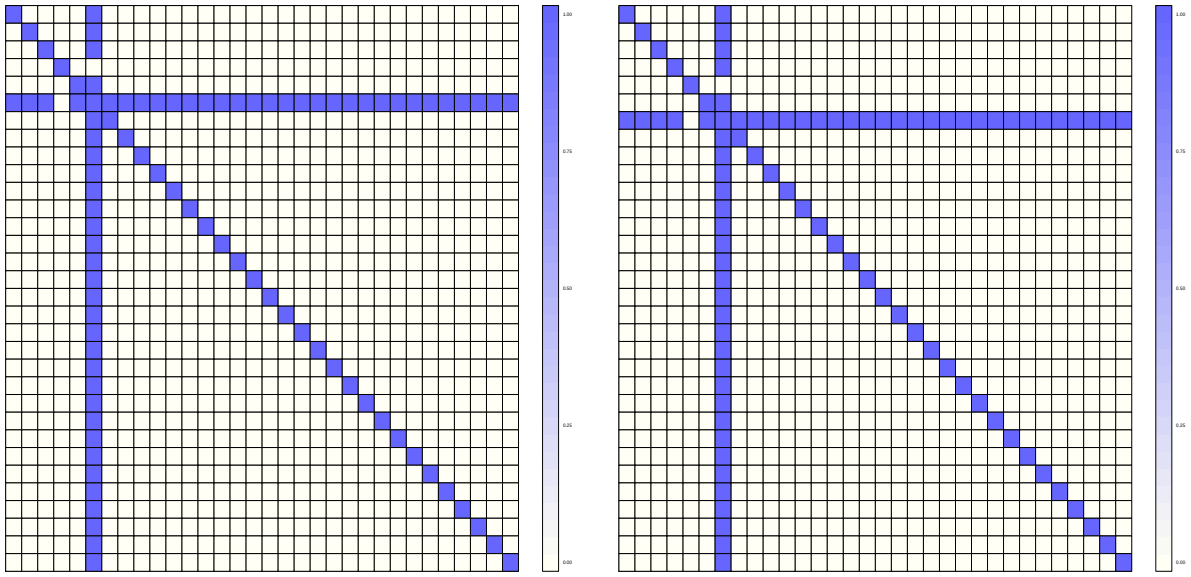
(d) Male control.

Figure 5: Presence/absence of gene-gene interactions for typical male cases and controls: Blue denotes presence and white represents absence of gene-gene interaction.



(a) Female case.

(b) Female control.



(c) Female control.

(d) Female control.

Figure 6: Presence/absence of gene-gene interactions for typical female cases and controls: Blue denotes presence and white represents absence of gene-gene interaction.

model and efficient TMCMC methods. The Bayesian tests of hypotheses that we employ in this paper are are appropriately modified versions of those proposed in Bhattacharya and Bhattacharya (2020).

Applications of our ideas to biologically realistic datasets generated under 5 different set-ups characterized by different combinations and structures associated with gene-gene and gene-environment interactions demonstrated encouraging performance of our model and methods. Our analysis of the MI dataset showed strong impact of the sex variable, which is consistent with the results of Bhattacharya and Bhattacharya (2020). Our tests showed no effect of the individual genes, which is also in keeping with Bhattacharya and Bhattacharya (2020) who obtained relatively weak marginal effects. But most interestingly, even though we obtained very weak gene-gene correlations in accordance with Bhattacharya and Bhattacharya (2020) and Lucas *et al.* (2012), our tests on gene-gene interaction showed that two genes, $AP006216.10$ and $C6orf106$, generally interact with all the other genes in a beneficial way so as to fight the disease. Moreover, the only situations where all the gene-gene interactions turned out to be insignificant, were the male cases, showing that the usual belief that males are more prone to heart attack than females may hold some value from this perspective.

Although many standard methods are commonly used in GWAS to identify the genetic and the environmental effects, there are several reasons that point towards the fact that our approach is not comparable with the existing methods.

So far, due to insufficient computational resources, we are compelled to restrict focus on a relatively small portion of the data. For improving our computing infrastructure, we have already taken the initiative of procuring supercomputing facilities from the Govt. of India, a project led, on behalf of Indian Statistical Institute, by the second author of this paper. With such a facility, we will be able to analyze the entire MI dataset with much ease.

Acknowledgment

We are sincerely grateful to the Editor-in-Chief and the anonymous reviewer, whose comments have led to significant improvement of our manuscript.

References

Ahn, J., Mukherjee, B., Gruber, S. B., and Ghosh, M. (2013). Bayesian semiparametric analysis for two-phase studies of gene-environment interaction. *The Annals of Applied Statistics*, **7**, 543–569.

Bhattacharya, D. and Bhattacharya, S. (2018). A Bayesian semiparametric approach to learning about gene-gene interactions in case-control studies. *Journal of Applied Statistics*, **45**, 1–23.

Bhattacharya, D. and Bhattacharya, S. (2020). Effects of gene-environment and gene-gene interactions in case-control studies: A Novel Bayesian Semiparametric Approach. *Brazilian Journal of Probability and Statistics*, **34**, 71–89.

De Iorio, M., Elliott, L. T., Favaro, S., Adhikari, K., and Teh, Y. W. (2015). Modeling population structure under hierarchical Dirichlet processes. Available at "https://arxiv.org/abs/1503.08278".

Dey, K. K. and Bhattacharya, S. (2017). On geometric ergodicity of additive and multiplicative transformation based Markov chain Monte Carlo in high dimensions. *Brazilian Journal of Probability and Statistics*, **31**, 569–617. Also available at "http://arxiv.org/pdf/1312.0915.pdf".

Hunter, D. J. (2005). Gene environment interactions in human diseases. *Nature Publishing Group*, **6**, 287–298.

Liu, C., Ma, J., and Amos, C. I. (2015). Bayesian variable selection for hierarchical gene-environment and gene-gene interactions. *Human Genetics*, **134**, 23–36.

Lucas, G., Lluis-Ganella, C., Subirana, I., Masameh, M. D., and Gonzalez, J. R. (2012). Hypothesis-based analysis of gene-gene interaction and risk of myocardial infraction. *Plos One*, **7**, 1–8.

Majumdar, A., Bhattacharya, S., Basu, A., and Ghosh, S. (2013). A novel Bayesian semiparametric algorithm for inferring population structure and adjusting for case-control association tests. *Biometrics*, **69**, 164–173.

Mather, K. and Caligary, P. (1976). Genotype x environmental interactions. *Heredity*, **36**, 41–48.

Mukhopadhyay, S. and Bhattacharya, S. (2013). Bayesian MISE convergence rates of mixture models based on the Polya urn model: Asymptotic comparisons and choice of prior parameters. Available at http://arxiv.org/abs/1205.5508.

Mukhopadhyay, S., Bhattacharya, S., and Dihidar, K. (2011). On Bayesian "central clustering": Application to landscape classification of Western Ghats. *Annals of Applied Statistics*, **5**, 1948–1977.

Mukhopadhyay, S., Roy, S., and Bhattacharya, S. (2012). Fast and efficient Bayesian semiparametric curve-fitting and clustering in massive data. *Sankhya. Series B*, **71**, 77–106.

Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet processes hierarchical models. *Biometrika*, **95**, 169–186.

Pinelli, M., Scala, G., Amato, R., Cocozza, S., and Miele, G. (2012). Simulating gene-gene and gene-environment interactions in complex diseases: Gene-environment interaction simulator 2. *BMC Bioinformatics*, **13**.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–650.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, **101**, 1566–1581.

Urbut, S. M., Wang, G., Carbonetto, P., and Stephens, M. (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics*, **51**, 187–195.

Wen, X. and Stephens, M. (2014). Bayesian methods for genetic association analysis with heterogenous subgroups: From meta-analyses to gene-environment iInteractions. *Annals of Applied Statistics*, **8**, 176–203.

Yang, Y., Carbonetto, P., Gerard, D., and Stephens, M. (2024). Improved methods for empirical Bayes multivariate multiple testing and effect size estimation. Available at http://arxiv.org/abs/2406.08784.

ANNEXURE

A-1. An MCMC method using Gibbs sampling and TMCMC

A-1.1. Full conditionals

Full conditional of $\boldsymbol{H}_k$

First observe that for $k = 0, 1$, the full conditional of $\boldsymbol{H}_k$ is given by

$$[\boldsymbol{H}_k|\cdots] \sim DP\left(\alpha_H + n_{\cdot k}, \frac{\alpha_H \tilde{\boldsymbol{H}} + \sum_{s=1}^{S} n_{sk}\delta_{\boldsymbol{\eta}_s}}{\alpha_H + n_{\cdot k}}\right), \tag{34}$$

where $n_{sk} = \#\{r \in \{1, \ldots, R_k\} : \boldsymbol{\xi}_{rk} = \boldsymbol{\eta}_s\}$ and $n_{\cdot k} = \sum_{s=1}^{S} n_{sk}$.

Full conditional of $\boldsymbol{G}_{0,jk}$

Similarly, the full conditional of $\boldsymbol{G}_{0,jk}$ is given, for $j = 1, \ldots, J$ and $k = 0, 1$, by

$$[\boldsymbol{G}_{0,jk}|\cdots] \sim DP\left(\alpha_{G_0,k} + n_{\cdot jk}, \frac{\alpha_{G_0,k}\boldsymbol{H}_k + \sum_{l=1}^{R_k} n_{ljk}\delta_{\boldsymbol{\xi}_{lk}}}{\alpha_{G_0,k} + n_{\cdot jk}}\right), \tag{35}$$

where $n_{ljk} = \#\{(t, i) \in \{1, \ldots, \tau_{ijk}\} \times \{1, \ldots, N_k\} : \boldsymbol{\phi}_{tijk} = \boldsymbol{\xi}_{lk}\}$ and $n_{\cdot jk} = \sum_{l=1}^{R_k} n_{ljk}$.

The full conditionals of $\boldsymbol{H}_k$ and $\boldsymbol{G}_{0,jk}$ given by (34) and (35) indicate generating the infinite-dimensional random probability measures using Sethuraman's characterization of Dirichlet processes (see Sethuraman (1994)). However, in our case, forming the infinite-dimensional Sethuraman's construction is not necessary; instead, it will be required to simulate from the random probability measures having distributions (34) and (35). Such simulations are possible using the retrospective method (see Papaspiliopoulos and Roberts (2008)) which avoids dealing with infinitely many objects.

Full conditional of $\boldsymbol{p}_{mijk}$

The associated Polya urn distribution of $\boldsymbol{p}_{mijk}$ given $\boldsymbol{P}_{Mijk}\backslash\{\boldsymbol{p}_{mijk}\}$, derived by marginalizing over $\boldsymbol{G}_{ijk}$, is the following:

$$[\boldsymbol{p}_{mijk}|\boldsymbol{P}_{Mijk}\backslash\{\boldsymbol{p}_{mijk}\}] = \frac{\alpha_{G,ik}}{\alpha_{G,ik} + M - 1}\boldsymbol{G}_{0,jk}(\boldsymbol{p}_{mijk}) + \frac{1}{\alpha_{G,ik} + M - 1}\sum_{m2\neq m=1}^{M}\delta_{\boldsymbol{p}_{m2ijk}}(\boldsymbol{p}_{mijk}) \tag{36}$$

where $M_{tijk} = \#\{m2 \in \{1, \ldots, M\}\backslash\{m\} : \boldsymbol{p}_{m2ijk} = \boldsymbol{\phi}_{tijk}\}$ and $\delta_{\boldsymbol{\phi}_{tijk}}(\cdot)$ denotes point mass at $\boldsymbol{\phi}_{tijk}$.

Given $z_{ijk} = m$, on combining the Polya urn distribution with the likelihood $\prod_{r=1}^{L} f(x_{ijkr}|p_{mijkr})$ we obtain the following full conditional of $\boldsymbol{p}_{mijk}$:

$$[\boldsymbol{p}_{mijk}|\cdots] \propto \alpha_{G,ik}\prod_{r=1}^{L} f(x_{ijkr}|p_{mijkr})\boldsymbol{G}_{0,jk}(\boldsymbol{p}_{mijk}) + \sum_{t=1}^{\tau_{ijk}} M_{tijk}\prod_{r=1}^{L} f(x_{ijkr}|\phi_{tijkr})\delta_{\boldsymbol{\phi}_{tijk}}(\boldsymbol{p}_{mijk}). \tag{37}$$

Note that in (37), $\boldsymbol{G}_{0,jk}$, drawn from (35), is not available in closed form and only admits the form dictated by Sethuraman's construction, given, almost surely, by

$$\boldsymbol{G}_{0,jk} = \sum_{l=1}^{\infty} \tilde{p}_l \delta_{\tilde{\boldsymbol{\xi}}_{ljk}}, \tag{38}$$

where $\tilde{p}_1 = V_1$, $\tilde{p}_l = V_l \prod_{s<l}(1-V_s)$, for $l \geq 2$, with $V_1, V_2, \ldots \overset{iid}{\sim} \text{Beta}\,(\alpha_{G_0,k} + n_{\cdot jk}, 1)$, and for $l = 1, 2, \ldots$, $\tilde{\boldsymbol{\xi}}_{ljk} \overset{iid}{\sim} \frac{\alpha_{G_0,k}\boldsymbol{H}_k + \sum_{l=1}^{R_k} n_{ljk}\delta_{\boldsymbol{\xi}_{lk}}}{\alpha_{G_0,k}+n_{\cdot jk}}$.

In (37), the posterior proportional to $\prod_{r=1}^{L} f(x_{ijkr}|p_{mijkr})\boldsymbol{G}_{0,jk}\,(\boldsymbol{p}_{mijk})$, which we denote by $[\boldsymbol{G}_{0,jk}|\boldsymbol{X}_{ijk}]$, is the discrete distribution that puts mass $C_{ijk}\tilde{p}_t \prod_{r=1}^{L} f(x_{ijkr}|\tilde{\xi}_{tjkr})$ to the point $\tilde{\boldsymbol{\xi}}_{tjk}$, for $t = 1, 2, \ldots$, where

$$C_{ijk} = \left(\sum_{t=1}^{\infty} \tilde{p}_t \prod_{r=1}^{L} f(x_{ijkr}|\tilde{\xi}_{tjkr})\right)^{-1} \tag{39}$$

is the normalizing constant. Combining these with (37) it follows that

$$[\boldsymbol{p}_{mijk}|\cdots] = \alpha_{G,ik}\bar{C}C_{ijk}^{-1}[\boldsymbol{G}_{0,jk}\,(\boldsymbol{p}_{mijk})\,|\boldsymbol{X}_{ijk}] + \bar{C}\sum_{t=1}^{\tau_{ijk}} M_{tijk} \prod_{r=1}^{L} f(x_{ijkr}|\phi_{tijkr})\delta_{\boldsymbol{\phi}_{tijk}}\,(\boldsymbol{p}_{mijk})\,, \tag{40}$$

where

$$\bar{C} = \left[\alpha_{G,ik}C_{jk}^{-1} + \sum_{t=1}^{\tau_{ijk}} M_{tijk} \prod_{r=1}^{L_j} f(x_{ijkr}|\phi_{tijkr})\right]^{-1}$$

is the normalizing constant of $[\boldsymbol{p}_{mijk}|\cdots]$.

A-1.2. Retrospective sampling methods

Retrospective method for simulating from $[\boldsymbol{p}_{mijk}|\cdots]$

From (40) it follows that, to draw from $[\boldsymbol{p}_{mijk}|\cdots]$, it is required to simulate from $[\boldsymbol{G}_{0,jk}\,(\boldsymbol{p}_{mijk})\,|\boldsymbol{X}_{ijk}]$ with probability proportional to C_{ijk}^{-1}. However, since C_{ijk} involves an infinite series, its calculation is infeasible. The same issue also prevents the traditional simulation methods to draw from the discrete distribution $[\boldsymbol{G}_{0,jk}|\boldsymbol{X}_{ijk}]$. In this case, the retrospective sampling method proposed in Section 3.5 of Papaspiliopoulos and Roberts (2008) is the appropriate method for our purpose. We first briefly discuss the role of such method in simulating from $[\boldsymbol{G}_{0,jk}|\boldsymbol{X}_{ijk}]$, and then argue that a by-product of the method can be used to estimate C_{ijk} arbitrarily accurately.

Retrospective method to draw from $[\boldsymbol{G}_{0,jk}\,(\boldsymbol{p}_{mijk})\,|\boldsymbol{X}_{ijk}]$

Note that the retrospective method requires $\prod_{r=1}^{L} f(x_{ijkr}|\phi_{tijkr})$ in our case to be uniformly bounded for all $\boldsymbol{\phi}_{tijk}$, which holds in our case, as $f(x_{ijkr}|\phi_{tjkr})$ represents the Bernoulli distribution, which is bounded above by 1. We briefly describe the method as follows. Let

$$c_\ell(K) = \sum_{a=1}^{K} \tilde{p}_a \prod_{r=1}^{L} f(x_{ijkr}|\tilde{\xi}_{ajkr}) \tag{41}$$

and

$$c_u(K) = c_\ell(K) + (1 - \sum_{a=1}^{K} \tilde{p}_a). \tag{42}$$

Let us also define $\breve{p}_{\ell,a}(K) = \tilde{p}_a \prod_{r=1}^{L} f(x_{ijkr}|\tilde{\xi}_{ajkr})/c_\ell(K)$ and $\breve{p}_{u,a}(K) = \tilde{p}_a \prod_{r=1}^{L} f(x_{ijkr}|\tilde{\xi}_{ajkr})/c_u(K)$. To simulate from $[\boldsymbol{G}_{0,jk}|\boldsymbol{X}_{ijk}]$ we first generate $U \sim \text{Uniform}(0,1)$, and given U, choose $\tilde{\boldsymbol{\xi}}_{tjk}$ when

$$\sum_{a=1}^{t-1} \breve{p}_{u,a}(K) \leq U \leq \sum_{a=1}^{t} \breve{p}_{\ell,a}(K). \tag{43}$$

In fact, K needs to be increased and $\tilde{p}_t$ and $\prod_{r=1}^{L} f(x_{ijkr}|\tilde{\xi}_{ajkr})$ simulated retrospectively, till (43) is satisfied for some $t \leq K$.

Retrospective method for estimating C_{ijk} arbitrarily accurately

By choosing K to be large enough, the quantities $c_\ell(K)$ and $c_u(K)$ given by (41) and (42), respectively, can be made arbitrarily close. In other words, for any $\epsilon > 0$, there exists $K_0 \geq 1$ such that $|c_\ell(K) - c_u(K)| < \epsilon$, for $K \geq K_0$. Thus, for any such $K \geq K_0$, one may approximate C_{ijk} with $[c_\ell(K)]^{-1}$. In practice, it is only required to simulate $\tilde{U} \sim \text{Uniform}(0,1)$ and simulate from $[\boldsymbol{G}_{0,jk}(\boldsymbol{p}_{mijk})|\boldsymbol{X}_{ijk}]$ if $\tilde{U} \leq \bar{C}C_{ijk}^{-1}$. For sufficiently small ϵ and for finite number of simulations, it will generally hold that $\tilde{U} \leq \bar{C}C_{ijk}^{-1}$ if and only if $\tilde{U} \leq \bar{C}_\epsilon c_\ell(K)$, for $K \geq K_0$, where

$$\bar{C}_\epsilon = \left[c_\ell^{-1}(K) + \sum_{t=1}^{\tau_{ijk}} M_{tijk} \prod_{r=1}^{L} f(x_{ijkr}|\phi_{tijkr}) \right]^{-1}.$$

Retrospective method to simulate from $\frac{\alpha_{G_0,k}\boldsymbol{H}_k + \sum_{l=1}^{R_k} n_{ljk}\delta_{\boldsymbol{\xi}_{lk}}}{\alpha_{G_0,k}+n_{\cdot jk}}$

The retrospective simulation method requires simulation of $\tilde{\boldsymbol{\xi}}_{ljk} \overset{iid}{\sim} \frac{\alpha_{G_0,k}\boldsymbol{H}_k + \sum_{l=1}^{R_k} n_{ljk}\delta_{\boldsymbol{\xi}_{lk}}}{\alpha_{G_0,k}+n_{\cdot jk}}$, for $l = 1, 2, \ldots$. This requires simulation from $\boldsymbol{H}_k$ with probability proportional to $\alpha_{G_0,k}$. For this, we first simulate $U \sim \text{Uniform}(0,1)$. We then simulate a realization from $\boldsymbol{H}_k$ after generating $\boldsymbol{H}_k$ from the Dirichlet process given by (34). Note that we do not have to generate the entire random probability measure $\boldsymbol{H}_k$ for this; we only need to generate as many realizations $\boldsymbol{\eta}_{lk}^*$'s from $\frac{\alpha_H \tilde{\boldsymbol{H}} + \sum_{s=1}^{S} n_{sk}\delta_{\boldsymbol{\eta}_s}}{\alpha_H + n_{\cdot k}}$ and as many $p_{lk}^* = V_{lk}^* \prod_{s<l}(1 - V_{lk}^*)$; $l = 1, 2 \ldots$, with $p_{1k}^* = V_{1k}^*$, with $V_{lk}^* \overset{iid}{\sim} \text{Beta}(\alpha_H + n_{\cdot k}, 1)$, as required to satisfy $\sum_{l=1}^{t-1} p_{lk}^* < U \leq \sum_{l=1}^{t} p_{lk}^*$, for some $t \geq 1$ (with $p_0^* = 0$). We then report $\tilde{\boldsymbol{\xi}}_{1jk} = \boldsymbol{\eta}_{tk}^*$ with probability proportional to $\alpha_{G_0,k}$ and $\tilde{\boldsymbol{\xi}}_{1jk} = \boldsymbol{\xi}_{\tilde{l}jk}$ with probability proportional to $n_{\tilde{l}jk}$, for $\tilde{l} \in \{1, \ldots, R_k\}$. We repeat this procedure for generating $\boldsymbol{\xi}_{ljk}$; $l \geq 2$, by sequentially augmenting the existing simulations of $\boldsymbol{\eta}_{lk}^*$'s and p_{lk}^*'s with new draws from $\tilde{\boldsymbol{H}}$ and $\text{Beta}(\alpha_H + n_{\cdot k}, 1)$, if needed. Indeed, note that for augmentation of p_{lk}^*'s, only extra V_{lk}^*'s need to be generated from $\text{Beta}(\alpha_H + n_{\cdot k}, 1)$.

A-1.3. Updating the allocation and proportion variables

Updating procedure for z_{ijk} and $\boldsymbol{p}_{mijk}$

The full conditional of z_{ijk} is given by the following:

$$[z_{ijk} = m|\cdots] \propto \pi_{mijk} \prod_{r=1}^{L_j} f\left(\boldsymbol{x}_{ijkr}|p_{mijkr}\right); \tag{44}$$

for $m = 1, \ldots, M$.

Recall that we have devised a method of simulating from the full conditional of $\boldsymbol{p}_{mijk}$ given the data and the remaining variables. For our convenience, we re-formulate the full conditional in terms of the dishes $\boldsymbol{\phi}_{tjk}$ and the indicators of the dishes, which we denote by t_{mijk}, where $t_{mijk} = t$ if and only if $p_{mijk} = \phi_{tijk}$; $t = 1, \ldots, \tau_{ijk}$.

Now let $\tau_{ijk}^{(m)}$ denote the number of elements in $\boldsymbol{P}_{Mijk} \setminus \{\boldsymbol{p}_{mijk}\}$ that arose from $[\boldsymbol{G}_{0,jk}|\boldsymbol{X}_{ijk}]$. Also let $\boldsymbol{\phi}^{m*}_{tijk} = \left\{\phi^{m*}_{tijkr};\ r = 1, \ldots, L\right\};\ t = 1, \ldots, \tau_{ijk}^{(m)}$ denote the parameter vectors arising from $[\boldsymbol{G}_{0,jk}|\boldsymbol{X}_{ijk}]$. Further, let $\boldsymbol{\phi}^{m*}_{tijk}$ occur M_{mtijk} times.

Then we update t_{mijk} using Gibbs steps, where the full conditional distribution of t_{mijk} is given by

$$[t_{mijk} = t|\cdots] \propto \begin{cases} q^*_{t,mijk} & \text{if} \quad t = 1, \ldots, \tau_{ijk}^{(m)}; \\ q_{0,mijk} & \text{if} \quad t = \tau_{ijk}^{(m)} + 1, \end{cases} \tag{45}$$

where

$$q_{0,mijk} = \alpha_{G,ik} C^{-1}_{ijk}; \tag{46}$$

$$q^*_{t,mijk} = M_{mtijk} \prod_{r=1}^{L_j} \left\{\phi^{m*}_{tijkr}\right\}^{n_{1mijkr}} \left\{1 - \phi^{m*}_{tijkr}\right\}^{n_{2mijkr}}. \tag{47}$$

In (46) and (47), n_{1mijkr} and n_{2mijkr} denote the number of "a" and "A" alleles, respectively, at the r-th locus of the j-th gene of the i-th individual, associated with the m-th mixture component. In other words, $n_{1mijkr} = x^1_{ijkr} + x^2_{ijkr}$ and $n_{2mijkr} = 2 - \left(x^1_{ijkr} + x^2_{ijkr}\right)$.

Let $n^*_{1tijkr} = \sum_{m:t_{mijk}=t} n_{1mijkr}$ and $n^*_{2tijkr} = \sum_{m:t_{mijk}=t} n_{2mijkr}$. Then, for $t = 1, \ldots, \tau_{ijk};\ r = 1, \ldots, L_j;\ j = 1, \ldots, J$ and $k = 0, 1$, update $\boldsymbol{\phi}^*_{tijk}$ by simulating from its full conditional distribution, given by

$$[\boldsymbol{\phi}^*_{tijk}|\cdots] \sim [\boldsymbol{G}_{0,jk}|\boldsymbol{X}_{ijk}]. \tag{48}$$

The above simulation from $[\boldsymbol{\phi}^*_{tijk}|\cdots]$ is to be carried out by the retrospective method as discussed above.

A-1.4. Updating the missing data

Updating $\tilde{\boldsymbol{Y}}_{ijk}$

From (4) it follows that

$$[\tilde{\boldsymbol{Y}}_{ijk}|z_{ijk}] = \prod_{r=L_j+1}^{L} f\left(\boldsymbol{y}_{ijkr}|p_{z_{ijk}ijkr}\right). \tag{49}$$

Hence, given the other unknowns, $\tilde{\boldsymbol{Y}}_{ijk}$ can be updated by simply simulating from the Bernoulli distributions given by (49).

A-1.5. Relevant factor aggregations for updating the fixed-dimensional parameters

Relevant factors for updating μ_G and $\boldsymbol{\beta}_G$

Let

$$\mathcal{L}_G(\mu_G, \boldsymbol{\beta}_G) = \prod_{k=0}^{1} \prod_{i=1}^{N_k} \prod_{j=1}^{J} \prod_{m=2}^{M} [\boldsymbol{p}_{mijk} | \boldsymbol{p}_{lijk}; l < m],$$

where $[\boldsymbol{p}_{mijk} | \boldsymbol{p}_{lijk}; l < m]$ is given by (15). Let $\pi_G(\mu_G, \boldsymbol{\beta}_G)$ denote the prior on $(\mu_G, \boldsymbol{\beta}_G)$. Note that $\pi_G(\mu_G, \boldsymbol{\beta}_G)\mathcal{L}_G(\mu_G, \boldsymbol{\beta}_G)$ is the product of the only factors in the joint model consisting of μ_G and $\boldsymbol{\beta}_G$.

Relevant factors for updating μ_{G_0} and $\boldsymbol{\beta}_{G_0}$

Now let

$$\mathcal{L}_{G_0}(\mu_{G_0}, \boldsymbol{\beta}_{G_0}) == \prod_{k=0}^{1} \prod_{i=1}^{N_k} \prod_{j=1}^{J} \prod_{t=2}^{\tau_{ijk}} [\boldsymbol{\phi}_{tijk} | \boldsymbol{\phi}_{lijk}; l < t],$$

where $[\boldsymbol{\phi}_{tijk} | \boldsymbol{\phi}_{lijk}; l < t]$ is given by (16).

Let $\pi_{G_0}(\mu_{G_0}, \boldsymbol{\beta}_{G_0})$ denote the prior on $(\mu_{G_0}, \boldsymbol{\beta}_{G_0})$. Then $\pi_{G_0}(\mu_{G_0}, \boldsymbol{\beta}_{G_0})\mathcal{L}_{G_0}(\mu_{G_0}, \boldsymbol{\beta}_{G_0})$ is the functional form associated with μ_{G_0} and $\boldsymbol{\beta}_{G_0}$.

Relevant factors for updating μ_H and $\boldsymbol{\beta}_H$

Finally, we let

$$\mathcal{L}_H(\mu_H, \boldsymbol{\beta}_H) = \prod_{k=0}^{1} \prod_{s=2}^{R_k} [\boldsymbol{\xi}_{sk} | \boldsymbol{\xi}_{lk}; l < s],$$

where $[\boldsymbol{\xi}_{sk} | \boldsymbol{\xi}_{lk}; l < s]$ is given y (17).

Let $\pi_H(\mu_H, \boldsymbol{\beta}_H)$ be the prior on $(\mu_H, \boldsymbol{\beta}_H)$. Then $\pi_H(\mu_H, \boldsymbol{\beta}_H)\mathcal{L}_H(\mu_H, \boldsymbol{\beta}_H)$ is the functional form to be considered for updating μ_H and $\boldsymbol{\beta}_H$.

A-1.6. Mixture of additive and multiplicative TMCMC for updating the fixed-dimensional parameters in a single block

We shall update all the parameters μ_G, $\boldsymbol{\beta}_G$, μ_{G_0}, $\boldsymbol{\beta}_{G_0}$, μ_H and $\boldsymbol{\beta}_H$ using a mixture of additive and multiplicative TMCMC, where all the aforementioned parameters are given either the additive move or the multiplicative move with equal probability, and where the acceptance ratio will be calculated by evaluating the functional form

$$\pi_G(\mu_G, \boldsymbol{\beta}_G)\mathcal{L}_G(\mu_G, \boldsymbol{\beta}_G) \times \pi_{G_0}(\mu_{G_0}, \boldsymbol{\beta}_{G_0})\mathcal{L}_{G_0}(\mu_{G_0}, \boldsymbol{\beta}_{G_0}) \times \pi_H(\mu_H, \boldsymbol{\beta}_H)\mathcal{L}_H(\mu_H, \boldsymbol{\beta}_H)$$

at the numerator and the denominator corresponding to the proposed and the current values of μ_G, $\boldsymbol{\beta}_G$, μ_{G_0}, $\boldsymbol{\beta}_{G_0}$, μ_H and $\boldsymbol{\beta}_H$, with all other unknowns held fixed at their current values, multiplied by an appropriate Jacobian whenever the multiplicative move is chosen. For details regarding mixture of additive and multiplicative TMCMC, see Dey and Bhattacharya (2017).

A-2. A parallel algorithm for implementing our MCMC procedure

Recall that the mixtures associated with gene $j \in \{1, \ldots, J\}$, and individual $i \in \{1, \ldots, N_k\}$ and case-control status $k \in \{0, 1\}$, are conditionally independent of each other, given the interaction parameters. This allows us to update the mixture components in separate parallel processors, conditionally on the interaction parameters. Once the mixture components are updated, we update the interaction parameters using a specialized form of TMCMC, in a single processor. Furthermore, the parameters of the HDP are also amenable to efficient parallelization. The details are as follows.

(1) (a) In processes numbered 0 and 1, simultaneously obtain the set of distinct elements $\boldsymbol{\Xi}_{R_k,k}$; $k = 0, 1$, from $\{\phi_{tijk}; t = 1, \ldots, \tau_{ijk}; i = 1, \ldots, N_k; j = 1, \ldots, J\}$; $k = 0, 1$.

(b) Communicate $\boldsymbol{\Xi}_{R_k,k}$; $k = 0, 1$, to all the processes.

(2) (a) In process 0, obtain the set of distinct elements $\boldsymbol{\zeta}_S$ from $\{\boldsymbol{\Xi}_{R_k,k}; k = 0, 1\}$.

(b) Communicate $\boldsymbol{\zeta}_S$ to all the processes.

(3) In processes numbered 0 and 1, do the following in parallel for $k = 0, 1$:

(a) Simulate, following the retrospective method. $\boldsymbol{\eta}^*_{lk} \overset{iid}{\sim} \frac{\alpha_H \tilde{\boldsymbol{H}} + \sum_{s=1}^{S} n_{sk} \delta_{\boldsymbol{\eta}_s}}{\alpha_H + n_{\cdot k}}$; $l = 1, 2, \ldots, \mathcal{L}$, for sufficiently large $\mathcal{L}$.

(b) Communicate the simulated values to all the processes.

(3) Split $\{(j, k) : \ j = 1, \ldots, J; \ k = 0, 1\}$ in the available parallel processes.

(a) For each (j, k), simulate, following the retrospective method. $\tilde{\boldsymbol{\xi}}_{ljk} \overset{iid}{\sim} \frac{\alpha_{G_0,k} \boldsymbol{H}_k + \sum_{l=1}^{R_k} n_{ljk} \delta_{\boldsymbol{\xi}_{lk}}}{\alpha_{G_0,k} + n_{\cdot jk}}$; $l = 1, 2, \ldots, \mathcal{L}$.

(b) Communicate the simulated values to all the processes.

(4) (a) Split the triplets $\{(i, j, k) : \ i = 1, \ldots, N_k; \ j = 1, \ldots, J; \ k = 0, 1\}$ in the available parallel processes sequentially into

$$\mathcal{T}_1 = \{(i, j, 0) : \ i = 1, \ldots, N_0; \ j = 1, \ldots, J\}$$

and

$$\mathcal{T}_2 = \{(i, j, 1) : \ i = 1, \ldots, N_1; \ j = 1, \ldots, J\}.$$

(b) Then parallelise updating of the mixtures associated with $\mathcal{T}_1$, followed by those of $\mathcal{T}_2$.

(c) If, for any (i, j, k), retrospective simulation from $[\boldsymbol{G}_{0,jk} | \boldsymbol{X}_{ijk}]$ requires more than $\mathcal{L}$ simulations of $\tilde{\boldsymbol{\xi}}_{ljk}$ in step (3) (a), then increase $\mathcal{L}$ to $\mathcal{L}^*$, and

(i) For $k = 0, 1$, augment the simulations of $\{\boldsymbol{\eta}^*_{lk}; l = 1, \ldots, \mathcal{L}\}$ with new simulations $\{\boldsymbol{\eta}^*_{lk}; l = \mathcal{L} + 1, \ldots, \mathcal{L}^*\}$.

(ii) For $j = 1, \ldots, J$ and for $k = 0, 1$, augment the simulations of $\left\{\tilde{\boldsymbol{\xi}}_{ljk}; l = 1, \ldots, \mathcal{L}\right\}$ with new simulations $\left\{\tilde{\boldsymbol{\xi}}_{ljk}; l = \mathcal{L} + 1, \ldots, \mathcal{L}^*\right\}$.

(iii) Repeat (4) (a) and (4) (b).

(5) During each MCMC iteration, for each (i, j, k) in each available parallel processor, update the allocation variables z_{ijk}, the proportions $\boldsymbol{p}_{mijk}$; $m = 1, \ldots, M$, and the missing data $\tilde{\boldsymbol{Y}}_{ijk}$, using the methods proposed in Sections A-1.3 and A-1.4.

(6) Communicate the results of updating in (4) and (5) to all the processes.

(7) (a) During each MCMC iteration, update the parameters μ_G, $\boldsymbol{\beta}_G$, μ_{G_0}, $\boldsymbol{\beta}_{G_0}$, μ_H and $\boldsymbol{\beta}_H$ in a single block using a mixture of additive and multiplicative TMCMC, as proposed in Section A-1.6, in process number 0.

(b) Communicate the updated results to all the processes.

A-3. Simulation studies

For simulation studies, we first generate realistic biological data for stratified population with known gene-environment interaction from the GENS2 software of Pinelli *et al.* (2012). To this data, we then apply our model and methodologies in an effort to detect gene-environment interaction effects that are present in the data. We consider simulation studies in 5 different true model set-ups: (a) presence of gene-gene and gene-environment interaction, (b) absence of genetic or gene-environmental interaction effect, (c) absence of genetic and gene-gene interaction effects but presence of environmental effect, (d) presence of genetic and gene-gene interaction effects but absence of environmental effect, and (e) independent and additive genetic and environmental effects.

As we demonstrate, our model and methodologies successfully identify the effects of the individual genes, gene-gene and gene-environment interactions, and the number of sub-populations. In all our applications, we set $M = 30$, $\nu_1 = \nu_2 = 1$, so that $\tilde{\boldsymbol{H}}$ is the uniform distribution on $[0, 1]$. We set $\alpha_{G,ik} = 0.1 \times \exp(100 + \mu_G + \beta_G E_{ik})$, $\alpha_{G_0,k} = 0.1 \times \exp\left(100 + \mu_{G_0} + \beta_{G_0}\bar{E}_k\right)$ and $\alpha_H = 0.1 \times \exp\left(100 + \mu_H + \beta_H\bar{\bar{E}}\right)$, where we assumed $\mu_G, \mu_{G_0}, \mu_H \overset{iid}{\sim} U(0, 1)$ and $\beta_G, \beta_{G_0}, \beta_H \overset{iid}{\sim} U(-1, 1)$. This structure ensured adequate number of sub-populations and satisfactory mixing of MCMC.

A-3.1.First simulation study: presence of gene-gene and gene-environment interaction

A-3.1.1.Data description

As in Bhattacharya and Bhattacharya (2020) we consider two genetic factors as allowed by GENS2 and simulated 5 data sets with gene-gene and gene-environment interaction with a one-dimensional environmental variable, associated with 5 sub-populations. One of the genes consists of 1084 SNPs and another has 1206 SNPs, with one disease pre-disposing locus (DPL) at each gene. There are 113 individuals in each of the 5 data sets, from which we selected a total of 100 individuals without replacement with probabilities assigned to the 5 data sets being $(0.1, 0.4, 0.2, 0.15, 0.15)$. Our final dataset consists of 46 cases and 54 controls. Since, in our case, the environmental variable is one-dimensional, $d = 1$.

A-3.1.2.Model implementation

We implemented our parallel MCMC algorithm on 50 cores in a 64-bit VMware with 64-bit physical cores, each running at 2793.269 MHz. Our code is written in C in conjunction with the Message Passing Interface (MPI) protocol for parallelisation.

The total time taken to implement $30,000$ MCMC iterations, where the first $10,000$ are discarded as burn-in, is approximately 20 hours. We assessed convergence informally with trace plots, which indicated adequate mixing properties of our algorithm.

A-3.1.3.Specifications of the thresholds ε's using null distributions

Following the method outlined in Section 4.2.4 and setting M to be 30, we obtain $\varepsilon_{d^*} = 0.200$, $\varepsilon_{\hat{d}_1} = 0.167$, $\varepsilon_{\hat{d}_2} = 0.167$, $\varepsilon_{d^*_E} = 0.250$, $\varepsilon_{d^*_{E,1}} = 0.185$, $\varepsilon_{d^*_{E,2}} = 0.173$, $\varepsilon_{\beta_G} = 0.874$, $\varepsilon_{\beta_{G_0}} = 0.128$, $\varepsilon_{\beta_H} = 0.219$.

A-3.1.4.Results of fitting our model

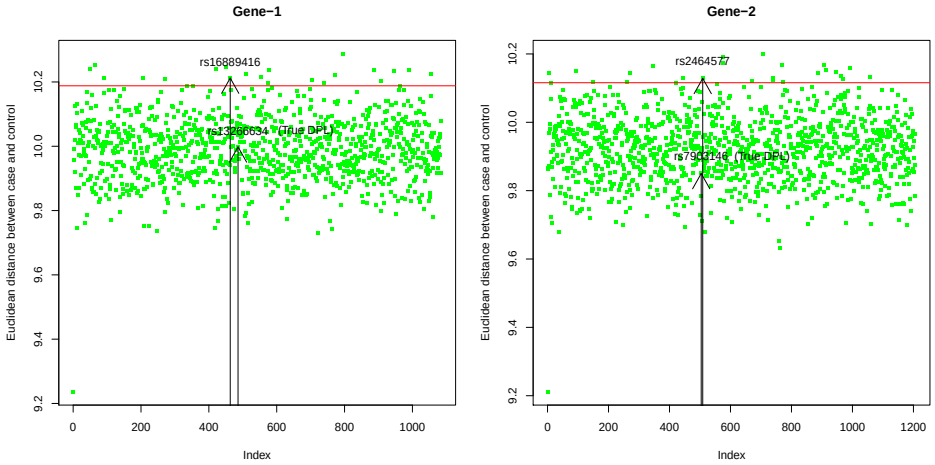
The posterior probabilities $P\left(d^* < \varepsilon_{d^*}|\text{Data}\right)$, $P\left(\hat{d}_1 < \varepsilon_{\hat{d}_1}|\text{Data}\right)$ and $P\left(\hat{d}_2 < \varepsilon_{\hat{d}_2}|\right.$ Data) empirically obtained from $20,000$ MCMC samples, turned out to be 0.378, 0.317 and 0.324, respectively. Hence, H_{0,d^*}, $H_{0,\hat{d}_1}$ and $H_{0,\hat{d}_2}$ are rejected, suggesting the influence of significant genetic effects in the case-control study.

However, $P\left(d^*_E < \varepsilon_{d^*_E}|\text{Data}\right)$, $P\left(\hat{d}_{E,1} < \varepsilon_{\hat{d}_{E,1}}|\text{Data}\right)$ and $P\left(\hat{d}_{E,2} < \varepsilon_{\hat{d}_{E,2}}|\text{Data}\right)$ are given, approximately, by 0.558, 0.561 and 0.550, respectively, which seem to contradict the results of the clustering based hypothesis tests. This can be explained as follows. Since $\boldsymbol{G}_{0,jk}$ are discrete, the parameters $\boldsymbol{p}_{mijk}$, even if generated from $\boldsymbol{G}_{0,jk}$, coincide with positive probability, so that the effective dimensionalities of logit $\left(\bar{P}_{Mi_0jk=0}\right)$ and logit $\left(\bar{P}_{Mi_1jk=1}\right)$ are drastically reduced, so that the Euclidean distance between these two vectors is substantially small. As such, the Euclidean distance fails to reject the null even if it is false. As noted in Bhattacharya and Bhattacharya (2020), even the clustering metric in this scenario is not completely satisfactory since this involves clustering distance between two empirically obtained central clusterings which may not be very accurate unless the sample sizes for case and control are very large. However, compared to the Euclidean distance, the clustering metric turns out to be far more reliable.

To check the influence of the environmental variable on the genes we compute the posterior probabilities $P\left(|\beta_G| < \varepsilon_{\beta_G}|\text{Data}\right)$, $P\left(|\beta_{G_0}| < \varepsilon_{\beta_{G_0}}|\text{Data}\right)$ and $P\left(|\beta_H| < \varepsilon_{\beta_H}|\text{Data}\right)$. The probabilities turned out to be 0.544, 0.550 and 0.191, respectively, showing that β_H is very significant. That is, the environmental variable has a significant overall effect on the genes.

The posterior probabilities of no gene-gene interactions for the controls and cases, showed the prominence of several gene-gene interactions in both control and case groups. As to be expected, in the case group, more instances of gene-gene interactions turned out to be significant compared to the control group.

Also, encouragingly, The posteriors of the number of sub-populations gave high prob-



(a) Index plot for the first gene (b) Index plot for the second gene.

Figure A-1: Presence of gene-gene and gene-environment interaction: Plots of the Euclidean distances $\left\{d_j^r\left(\textbf{logit}\left(\boldsymbol{p}_{i_0jk=0}^r\right), \textbf{logit}\left(\boldsymbol{p}_{i_1jk=1}^r\right)\right);\ r=1,\ldots,L_j\right\}$ against the indices of the loci, for $j=1$ (panel (a)) and $j=2$ (panel (b)).

abilities to 5, the true number of sub-populations.

A-3.1.5.Detection of DPL

The correct positions of the DPL, provided by GENS2, are $rs13266634$ and $rs7903146$, for the first and second gene respectively. Due to the LD effects implied by the highly correlated structure of our current HDP based model, the actual DPL are difficult to locate. Notably, our model is considerably more structured than those of Bhattacharya and Bhattacharya (2018) and Bhattacharya and Bhattacharya (2020), and any inappropriate dependence structure would render the task of DPL finding far more difficult than our previous models. Nevertheless, we demonstrate that our HDP model can detect DPLs with more precision compared to our previous matrix-normal-inverse-Wishart model for gene-environment interactions.

Following Bhattacharya and Bhattacharya (2018) and Bhattacharya and Bhattacharya (2020), and writing $\boldsymbol{p}_{ijk}^r = \{p_{mijkr} : m = 1,\ldots,M\}$, we declare the r-th locus of the j-th gene as disease pre-disposing if, for the r-th locus, the Euclidean distance $d_j^r\left(\text{logit}\left(\boldsymbol{p}_{i_0jk=0}^r\right), \text{logit}\left(\boldsymbol{p}_{i_1jk=1}^r\right)\right)$, between logit $\left(\boldsymbol{p}_{i_0jk=0}^r\right)$ and logit $\left(\boldsymbol{p}_{i_1jk=1}^r\right)$, is significantly larger than $d_j^{r2}\left(\boldsymbol{p}_{i_0jk=0}^{r2}, \boldsymbol{p}_{i_1jk=1}^{r2}\right)$, for $r2 \neq r$. We adopt the graphical method as in our previous works. The red, horizontal lines in the panels of Figure A-1 represent the cut-off value such that the points above the horizontal line are those with the highest 2% Euclidean distances. The actual DPLs of the two genes, as well as their nearest neighbours with Euclidean distances on or above the red, horizontal lines, are shown in the figures. That even such small sets of SNPs with highest 2% Euclidean distances consist of close neighbours of the true DPLs, is quite encouraging. Observe that the DPL detection is more precise for the second gene in the sense that the closest neighbour of the actual DPL above the red, horizontal line is closer to the true DPL than for the first gene.

The above results on DPL detection is also a significant improvement over Bhattacharya and Bhattacharya (2020) where highest 10% Euclidean distances were considered, suggesting that our current HDP based model is more appropriate compared to our previous matrix-normal-inverse-Wishart model for gene-environment interaction.

A-3.2.Second simulation study: no genetic or environmental effect

Here we use the same case-control genotype data set as used by Bhattacharya and Bhattacharya (2018) in their second simulation study where genetic effects are absent, consisting of 49 cases and 51 controls and 5 sub-populations with the mixing proportions $(0.1, 0.4, 0.2, 0.15, 0.15)$. We use the same environmental data set generated in our first simulation study described in Section A-3.1, which is unrelated to this genotype data.

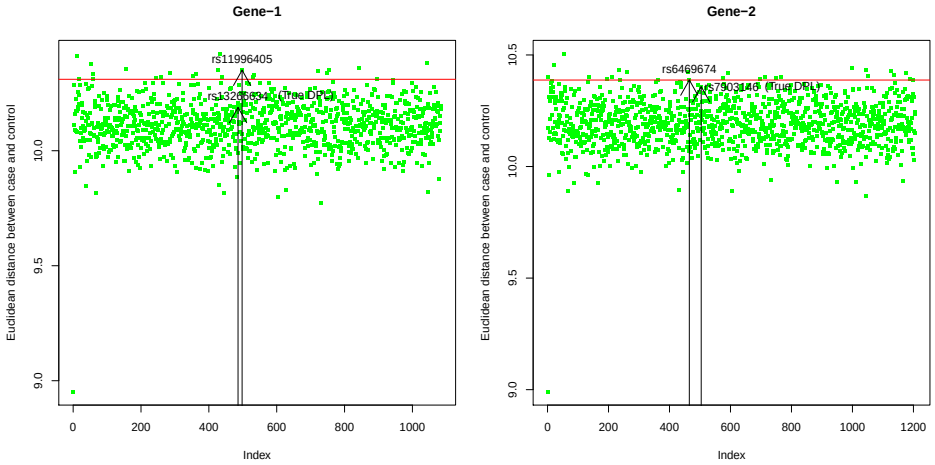
Here we obtain $P\left(d^* < \varepsilon_{d^*}|\text{Data}\right) \approx 0.407$. Although this does not cross the 0.5 benchmark, there is significant evidence in favour of the null, and falling short of 0.5 can be attributed to the slight deficiency of the distance between the two approximate central clusterings associated with case and control, as already discussed in the context of the first simulation study.

Also, in this study, $P\left(|\beta_G| < \varepsilon_{\beta_G}|\text{Data}\right)$, $P\left(|\beta_{G_0}| < \varepsilon_{\beta_{G_0}}|\text{Data}\right)$ and $P\left(|\beta_H| < \varepsilon_{\beta_H}|\right.$ Data) are given by 0.549, 0.550 and 0.649, respectively, suggesting insignificance of the effect of the environmental variable on gene-gene interaction. As noted in Bhattacharya and Bhattacharya (2020), however, it is not straightforward to test whether or not the environment is responsible for the case-control status. This is because we have modeled the genotype data conditionally on case-control instead of modeling the case-control status conditionally on the environmental variable. Bhattacharya and Bhattacharya (2020) use significance testing in a simple logistic regression framework to show insignificance of the environmental variable. As before, our model assigned high posterior probability to 5 sub-populations. Note that since there is no genetic effect in this study, the question of detecting DPLs does not arise here.

A-3.3.Third simulation study: absence of genetic and gene-gene interaction effects but presence of environmental effect

In this study we consider a case-control genotype data set simulated from GENS2 where case-control status depends only upon the environmental data. The number of cases here is 47 and the number of controls is 53. This is the same case-control genotype data set as used by Bhattacharya and Bhattacharya (2020) in their third simulation study.

In this case, we find that $P\left(d^* < \varepsilon_{d^*}|\text{Data}\right) \approx 0.400$, which provides reasonable evidence in favour of the null, even though the 0.5 benchmark is not crossed. Moreover, $P\left(|\beta_G| < \varepsilon_{\beta_G}|\text{Data}\right) \approx 0.536$, $P\left(|\beta_{G_0}| < \varepsilon_{\beta_{G_0}}|\text{Data}\right) \approx 0.518$ and $P\left(|\beta_H| < \varepsilon_{\beta_H}|\text{Data}\right) \approx 0.504$, suggesting that the environmental variable does not affect the genetic structure. Bhattacharya and Bhattacharya (2020) show by means AIC, in the context of simple logistic regression, that the best model consists of the marginal effects of the second gene and the environment. In conjunction with our HDP-based model which produces reasonable evidence in favour of accepting the hypothesis of no genetic effect, it may be possible to conclude that the environmental variable is responsible for the case-control status.



(a) Index plot for the first gene (b) Index plot for the second gene.

Figure A-2: Presence of genetic and gene-gene interaction effects but absence of environmental effect: Plots of the Euclidean distances $\left\{d_j^r\left(\textbf{logit}\left(\boldsymbol{p}_{i_0jk=0}^r\right), \textbf{logit}\left(\boldsymbol{p}_{i_1jk=1}^r\right)\right);\ r=1,\ldots,L_j\right\}$ against the indices of the loci, for $j=1$ (panel (a)) and $j=2$ (panel (b)).

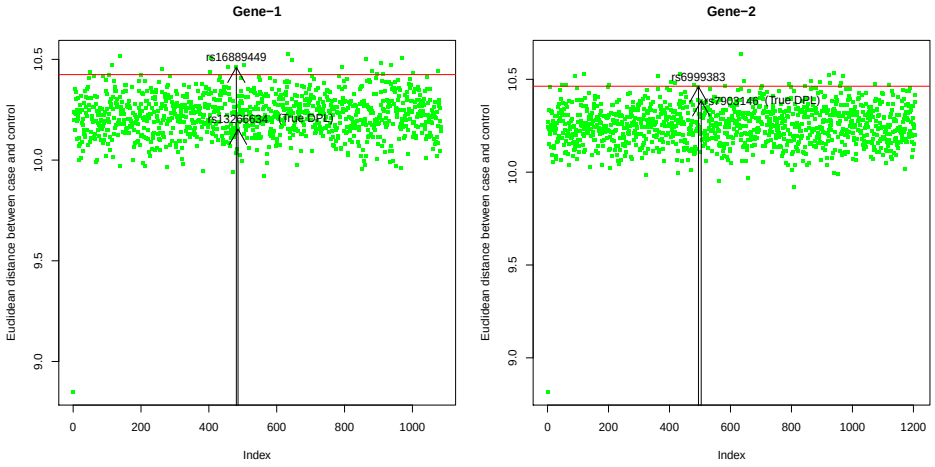
As before, 5 subpopulations get significant weight by our posterior distribution, and again, the question of DPL detection is irrelevant here since there is no genetic effect.

A-3.4.Fourth simulation study: presence of genetic and gene-gene interaction effects but absence of environmental effect

Here we use the same genotype data set as used by Bhattacharya and Bhattacharya (2018) in their first simulation study associated with genetic and gene-gene interaction effects, consisting of 41 cases and 59 controls and 5 sub-populations with the mixing proportions $(0.1, 0.4, 0.2, 0.15, 0.15)$. We use the same environmental data set generated in our first simulation study described in Section A-3.1, which is unrelated to this case-control genotype data.

Here we obtain $P\left(|\beta_G| < \varepsilon_{\beta_G}|\text{Data}\right) \approx 0.549$, $P\left(|\beta_{G_0}| < \varepsilon_{\beta_{G_0}}|\text{Data}\right) \approx 0.542$ and $P\left(|\beta_H| < \varepsilon_{\beta_H}|\text{Data}\right) \approx 0.552$, correctly suggesting insignificance of the environmental variable with respect to its effect on the genetic structure. Using logistic regression, Bhattacharya and Bhattacharya (2020) conclude that the environmental variable has no role to play in the case-control status. Furthermore, we obtain $P\left(d^* < \varepsilon_{d^*}|\text{Data}\right) \approx 0.390$, $P\left(\hat{d}_1 < \varepsilon_{\hat{d}_1}|\text{Data}\right) \approx 0.336$ $P\left(\hat{d}_2 < \varepsilon_{\hat{d}_2}|\text{Data}\right) \approx 0.324$. so that importance of genes is correctly indicated by our tests. Interestingly, study of the posterior probabilities of no gene-gene interactions for controls and cases showed no gene-gene interaction in the control group and only two (marginal) instances of gene-gene interaction among the cases.

Figure A-2 shows the plots of Euclidean distances between cases and controls for the loci of the two genes. In this case, Gene-1 has been located quite precisely, and for Gene-2 the Euclidean distance for even the true DPL is very close to the red, horizontal line, indicating encouraging performance.



(a) Index plot for the first gene (b) Index plot for the second gene.

Figure A-3: Independent and additive genetic and environmental effects: Plots of the Euclidean distances $\left\{d_j^r\left(\textbf{logit}\left(\boldsymbol{p}_{i_0jk=0}^r\right),\textbf{logit}\left(\boldsymbol{p}_{i_1jk=1}^r\right)\right);\ r=1,\ldots,L_j\right\}$ against the indices of the loci, for $j=1$ (panel (a)) and $j=2$ (panel (b)).

A-3.5.Fifth simulation study: independent and additive genetic and environmental effects

As in Bhattacharya and Bhattacharya (2020), we consider the situation where the genetic and environmental effects are independent of each other and additive; the data consists of 57 cases and 43 controls.

Note that, as in Bhattacharya and Bhattacharya (2020), in our current HDP-based Bayesian model also there is no provision for additivity of genetic and environmental effects. As such, it is not expected to capture the true data-generating mechanism accurately. Indeed, here we obtain $P\left(d^*<\varepsilon_{d^*}|\text{Data}\right)\approx 0.389$, $P\left(\hat{d}_1<\varepsilon_{\hat{d}_1}|\text{Data}\right)\approx 0.337$ and $P\left(\hat{d}_2<\varepsilon_{\hat{d}_2}|\text{Data}\right)\approx 0.331$, indicating significance of the genes. However, the test with d_E^* does not yield overwhelming evidence against the null. Our tests of gene-gene interaction indicated significant interactions for controls and particularly for cases. Also, $P\left(|\beta_G|<\varepsilon_{\beta_G}|\text{Data}\right)$, $P\left(|\beta_{G_0}|<\varepsilon_{\beta_{G_0}}|\text{Data}\right)$ and $P\left(|\beta_H|<\varepsilon_{\beta_H}|\text{Data}\right)$ are given, approximately, by 0.547, 0.550 and 0.367, the last value showing that the environmental variable does affect gene-gene interaction. The lack of the additivity provision in our model seems to have forced the gene-environment interaction in this case.

In spite of the lack of additivity of our model the Euclidean distances between cases and controls for the gene-wise SNPs are not adversely affected, and the actual DPLs are detected quite accurately; see Figure A-3. This brings forth the generality and usefulness of our nonparametric dependence structure. As before, 5 sub-populations received significant posterior probabilities.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 361–399
http://www.ssca.org.in/journal

Bayesian Predictive Inference for Nonprobability Samples with Spatial Poststratification

Dhiman Bhadra[1] **and Balgobin Nandram**[2]
[1]*Operations and Decision Sciences Area, Indian Institute of Management Ahmedabad, Gujarat - 380015, India*
[2]*Department of Mathematical Sciences, Worcester Polyechnic Institute, Worcester, Massachusetts - 01609, USA*

Received: 27 April 2024; Revised: 07 September 2024; Accepted: 10 September 2024

Abstract

Non-probability sampling involves selecting samples from a population in which the probability of selection is unknown and some population units may have zero selection probabilities. This differentiates it from probability sampling where selection is governed by a probability model and every population unit has a non-zero chance of being selected. Non-probability samples usually suffer from selection bias and hence may not represent the target population accurately. An important problem that arises in this context is the prediction of responses corresponding to non-sampled units, which should ideally have been sampled. In this article, we propose three modeling frameworks to address this issue. We use propensity scores to balance the sampled and non-sampled units and a Bayesian estimation scheme for parameter inference and prediction. We incorporate a spatial poststratification scheme to assess the predictive ability of our models on a simulated dataset. In addition, we perform model selection routines to identify the optimal model having the best predictive ability.

Key words: Beta-Bernoulli; Metropolis Hastings sampler; Non-probability samples; Propensity scores; Spatial poststratification.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

One of the most important aspects of any statistical investigation is the formulation of a realistic and objective plan for data collection. These data should ideally be derived from a sample that is a good representation of the target population in the sense that it reflects all the conspicuous categories of the population adequately. Traditionally, the selection of such samples is guided by an underlying probabilistic mechanism which ensures that each and every population unit has a positive probability of being selected. The most well known of these selection mechanisms is the so called simple random sampling, which has the property that every sample of size, say n, has the same chance of being selected. This implies that each

Corresponding Author: Dhiman Bhadra
Email: dhiman@iima.ac.in

population unit has the same chance of being selected in the final sample. As a result, this kind of sample is known as a probability sample and the corresponding plan is designated as a probability sampling plan. Commonly used sampling mechanisms such as stratified, cluster or systematic sampling (Neyman, 1934) and their combinations are all grounded in the principle of probability sampling as opposed to non-probability sampling.

However, obtaining a truly parsimonious and representative probability sample is often prohibitively difficult in a real setting due to various constraints. Even if such a sampling scheme is implemented, it is a formidable task to obtain the requisite responses from the selected sample units. In fact, response rates of major surveys have been declining rapidly, casting doubts on the validity of probability samples as a proper representation of the population. According to Pew Research Center, the response rates in typical telephone surveys dropped from 36% in 1997 to only 9% in 2012 (Kohut *et al.*, 2012). Such low response rates, coupled with the complexity of implementation of probability-based survey designs raise serious doubts as to the viability of such sampling frameworks in real-life settings.

The above considerations along with an explosion of data being generated through various channels have led to an upsurge in the usage of non-probability sampling schemes. These schemes, as the term suggests, does not involve any underlying probabilistic mechanism for implementation. As a result, such schemes are convenient to use and hence are also referred to as convenience sampling schemes. Inferences from such samples are generally model based. However, as population units "self-select" themselves, the samples so obtained, often suffer from selection bias. This often results in the sample being non-representative of the target population in the sense that the sample may fail to incorporate all the relevant segments of a target population in the correct proportion. For example, in an email survey, only those who are willing to participate respond, probably having particular demographic characteristics. As a result, the demographic characteristics of those who do not participate are under-represented in the sample. Having said that, there is a subtle difference between selection bias and undercoverage in which certain sections of the population have absolutely no representation in the sample. In other words, it can be said that undercoverage is an extreme form of selection bias where a certain section of the population have absolutely zero chance of being selected in the sample. In this context, we would like to state that the proposed modeling frameworks have been designed to account for selection bias, not necessarily undercoverage.

In order to explore the applicability of non-probability sampling schemes for sampling from finite populations, the American Association of Public Opinion Research (AAPOR) constituted two task forces, neither of which favoured their use (Baker *et al.*, 2013). It was also suggested that inferences about a population drawn from a non-probability sample is valid subject to the verification of the modeling assumptions underlying the sampling scheme, a rather difficult proposition. The report also outlined various forms of non-probability samples such as convenience, snowball, network, mall-intercept and volunteer samples. One common aspect of all these schemes is the non-probabilistic aspect of sample selection, which results in biases, as mentioned before. Techniques for controlling biases have also been proposed such as sampling match which involves selecting non-probability sample units such that their characteristics match those in the population. This leads to the reduction of selection bias specially when the distribution of covariates used for matching are similar for the non-probability sample and the target population. A modified matching principle can

be adopted for observational studies in which the non-probability sample units are matched with those in a probability sample. Each unit in the non-probability sample can then be assigned a weight as a form of quantification of its degree of matching with the probability sample (Rubin, 1979). Rosenbaum and Rubin (1983) illustrated the use of propensity scores in the context of observational studies when the distribution of covariates is different in the treatment and control groups. This technique can be adopted for non-probability sampling as well, since the covariate characteristics may differ between the sampled and non-sampled groups. An extensive overview of matching procedures for causal inference and their applicability in diverse fields have been provided by Stuart (2010).

Smith (1983) introduced the notion of non-probability sampling and discussed general approaches for making inference from such samples. The basic formulation outlined therein is to model the joint distribution of the response observations and the selection probabilities of the population units. This formulation resembles the works of Rubin (1976), Little (1982) and Little and Rubin (2002) on selection mechanisms and survey responses. Smith (1983) also introduced the concept of poststratification and discussed its application on quota sampling. In the context of the above framework, Elliott and Valliant (2017) proposed two specific approaches of inference from non-probability samples, namely quasi-randomization and super-population. The underlying idea for these two approaches is to decompose the aforementioned joint distribution into the product of a conditional distribution of the response vector given that of the selection probabilities and the distribution of the response vector given the covariates. Quasi-randomization involves modelling the first component and estimating the selection probabilities as a way of correcting for the selection bias. On the other hand, the superpopulation approach involves modeling the second component. Although both approaches involve modeling, those are fundamentally different in their character. However, both approaches are aimed at nullifying or correcting for the effect of selection bias so as to make the resulting non-probability sample a better representation of the population.

One approach is to use propensity scores to estimate the survey weights of the non-probability sample and then proceed as in a regular probability sample; see Elliott and Valliant (2017) for an informative review of quasi-randomization and the super-population approach for non-probability samples. Chen *et al.* (2020) supplemented a non-probability sample with a probability sample using only the observed covariates to estimate propensity scores via logistic regression. Another approach is to use a nonignorable selection model to remove the selection bias; see Smith (1983) for pioneering work in this direction. Xu and Nandram (2019) used this approach to obtain full Bayesian analyzes; the references therein provide a historical development of this area. It is difficult to make valid inference from a non-probability sample with considerable selection bias. After all, a probability sample is the gold standard (high quality), but a non-probability sample is likely to have low quality (large bias, large mean squared error but unrealistically small variance). The key problem of a non-probability sample is that it is very likely to lead to seriously biased estimates of finite population quantities. Therefore, the large well-documented literature on selection bias is pertinent in the study of non-probability samples; these articles are too numerous to mention here; see Xu *et al.* (2020) and Choi *et al.* (2021) for recent applications, and the references therein.

It is also possible to make inference about a finite population quantity using a sin-

gle non-probability sample only; see Rao (2021) for a discussion. Supplementing a non-probability sample with a small probability sample has recently received some attention. But it is not quite practical to run a small probability survey in parallel with a non-probability sample. Therefore, if one can make accurate finite population inference from a non-probability sample only, this can be useful and economical. After all it costs money and time to design and field even a small survey, and it is less practical that both a non-probability sample and a probability sample will be available at the same time.

Survey samplers have long been using probability samples from one or more sources in conjunction with census and administrative data to make valid and efficient inferences on finite population parameters. This topic has received a lot of attention more recently in the context of data from non-probability samples such as transaction data, web surveys and social media data. Rao (2021) reviewed various probability survey methods that are used to make valid inferences about finite population parameters. This allowed him to show how these models can be extended to non-probability samples that can lead to "valid inferences by themselves or when combined with probability samples". Beaumont (2020) also reviewed some approaches that can "reduce, or even eliminate the use of probability surveys, all while preserving a valid statistical inference framework". However, naive use of such data can lead to serious sample selection bias and without adjustment to reduce selection bias it can lead to the "big data paradox: the bigger the data, the surer we fool ourselves" (Meng, 2018). Inevitably, non-probability samples will be more widely used in the future, and we need to continue researching methods for obtaining valid (or at least acceptable) inferences from them, possibly in combination with probability samples as illustrated in several papers. Falling response rates and increasing respondent burden are often given as reasons for using non-probability samples, especially in socioeconomic surveys. Robustness to model misspecification is also important in non-probability samples; see, for example, Marella (2023) and Rafei *et al.* (2022).

It is possible to use post-stratification to make satisfactory inference from a non-probability sample only, and it is convenient to do so. It is not necessary to estimate directly the selection probabilities for the non-probability sample; see Little (1993), Wang *et al.* (2015), Wang *et al.* (2021), Nandram and Choi (2005, 2010). Propensity scores are used to stratify the population, and they are not used as survey weights. However, too many strata can lead to sparseness and some strata can be empty. Cochran and Chambers (1965) suggested an optimal number of five strata (using quintiles); while this is good for small samples, it may not be so good for large samples. For larger samples, we can use more thinning, and a larger number of strata might be more efficient, say ten strata (using deciles). We may not know the nonsampled covariates, but the minimal we can assume is that the population size and the average covariates are known, a practical scenario. It is then possible to generate surrogates of the the nonsampled covariates using a bootstrap procedure.

Here, we are not concerned with data integration nor small area estimation. But there is also an emerging area in this direction; see Nandram and Rao (2024), Nandram and Rao (2023), Nandram and Rao (2021) and Nandram *et al.* (2021) for a Bayesian approach using propensity scores to estimate the selection probabilities with assistance from a small probability sample; there are other Bayesian approaches such as Sakshaug *et al.* (2019), Wiśniowski *et al.* (2020), Salvatore *et al.* (2024) and Rafei *et al.* (2022), who used the non-probability sample to supplement the probability sample. There has been a non-Bayesian

literature also. One leading paper is Chen *et al.* (2020), who use the design approach with double robustness and asymptotic theory. However, models will be better, if in fact, inference about the finite population parameters is robust to the of assumptions of the models for both the study variable and the participation variable. Other non-Bayesian approaches are discussed by Elliot (2009), Elliott and Haviland (2007) and Robbins *et al.* (2021).

The primary objective of this article is to propose methodologies aimed towards reducing selection bias when there exists significant difference in the characteristics between sampled and non-sampled units. A related problem that will be addressed is the prediction of the response observations for the non-sampled units for given values of their covariates. In doing so, we will build on established ideas, for instance, post-stratification (Smith, 1983), superpopulation approach (Elliot and Valliant, 2017) and propensity scores matching (Rosenbaum and Rubin, 1983). The crux of our methodology will be to treat the non-sampled units vis-a-vis their response as missing data (Rubin, 1976; Little and Rubin, 2002). Propensity scores and post-stratification will be used to balance the covariate distributions between the sampled and non-sampled groups. Lastly, we will perform the analysis in a hierarchical Bayesian setup through Markov chain Monte Carlo methodology. Application of Bayesian methodology in the context of non-probability sampling is a relatively unexplored domain. A recent article in this space is by Sakshaug *et al.* (2019) which examines the exchangeability of probability and non-probability sampling schemes by supplementing small probability samples with non-probability ones in a Bayesian paradigm. Their proposed method is applied simultaneously on probability and non-probability surveys and is shown to reduce the variance and mean squared error of model based predictions corresponding to non-probability samples relative to probability-only samples. However, the novel aspect of the methodology proposed in this article is the integration of a spatial dimension in the model for binary responses which enables us to better predict the response for the non-sampled units. Having said so, we must emphasize that our target of inference and/or prediction is the finite population proportion of success in the non-sampled group. However, we believe that our framework can be effortlessly extended to estimate population means in general, arising from continuous or discrete response variables by broadening the distributional structure of the said variables.

This paper is organized as follows. In Section 2, we describe the simulation mechanism for generating test data. In Section 3, we outline a modeling framework based on the Beta-Bernoulli distribution for the purpose of prediction of responses for non-sampled units. In Section 4, we introduce a modified model that incorporates a spatial dimension to the existing modeling framework. In Section 5, we propose another spatial model that leads to more precise prediction of responses for the non-sampled units. For each of these frameworks, we discuss the mechanism of estimation and prediction in a hierarchical Bayesian setup. In Section 6, we discuss some diagnostic measures for comparing the relative predictive abilities of the aforementioned models, followed by concluding remarks and a discussion of future work in Section 7.

2. Data simulation

As mentioned before, one of the principal characteristics of non-probability sampling is that the distribution of covariates is different for the sampled and non-sampled groups. This will be the basis for our simulation exercise aimed at generating the dataset on which

our proposed methodologies will be tested later. For the purpose of simulation, let the population size be 10,000, denoted as N, while the size of the sampled group be 1000, denoted as n. We consider four (4) covariates, namely Age (X_1), Race (X_2), Gender (X_3) and Education level (X_4) and a response Y such that

$$X_{2i} = \begin{cases} 1 & \text{if } i^{th} \text{ subject is white} \\ 0 & \text{if } i^{th} \text{ subject is black;} \end{cases} \quad X_{3i} = \begin{cases} 1 & \text{if } i^{th} \text{ subject is male} \\ 0 & \text{if } i^{th} \text{ subject is female;} \end{cases}$$

$$X_{4i} = \begin{cases} 1 & \text{if } i^{th} \text{ subject's education is college or higher} \\ 0 & \text{if } i^{th} \text{ subject's education is highschool or lower;} \end{cases}$$

$$Y_i = \begin{cases} 1 & \text{if } i^{th} \text{ subject's response is Yes} \\ 0 & \text{if } i^{th} \text{ subject's response is No,} \end{cases}$$

where $i = 1, ..., N$. Finally, we assume the following distributions for the covariates

$$\begin{aligned} X_{1i} &\sim N(55, 5^2), & X_{1i} &\sim N(65, 5^2); \\ X_{2i} &\sim \text{Bernoulli}(0.3), & X_{2i} &\sim \text{Bernoulli}(0.5); \\ X_{3i} &\sim \text{Bernoulli}(0.5), & X_{3i} &\sim \text{Bernoulli}(0.4); \\ X_{4i} &\sim \text{Bernoulli}(0.5), & X_{4i} &\sim \text{Bernoulli}(0.6), \end{aligned}$$

where the distributions in the first (left) column correspond to the subjects in the sampled group ($i = 1, ..., n$) while those in the second (right) column correspond to those in the non-sampled group ($i = n+1, ..., N$). The above choice of parameters was guided by the fact that the distributions of each covariate for the sampled and non-sampled groups should not be too different. This is critical, because in the poststratification step to be implemented next, it is necessary for every stratum to have some sampled units. In other words, if the distributions of particular covariates in the sampled and non-sampled groups are very different, there may be stratum which will be devoid of any units from the sampled group. If so, it would not be possible for us to predict the response of the non-sampled units for that stratum.

Finally, we assume that $Y_i|p_i \overset{ind}{\sim} \text{Bernoulli}\ (p_i)$. Once the above covariate values are simulated, we generate the probability of success (*i.e..*, $Y_i = 1$) using the following logistic regression function

$$p_i = P(Y_i = 1) = \frac{e^{\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{4i} + \epsilon_i}}{1 + e^{\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{4i} + \epsilon_i}}, \quad i = 1, 2, ..., N,$$

where the ϵ follow a standard normal distribution *i.e..* $\epsilon_i \sim N(0, 1)$. We assume $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.1, 0.01, 4, -5, -1)$. Once the N simulated values of p_i are obtained, the corresponding values of Y_i are drawn from Bernoulli (p_i). Table 1 depicts part of the simulated data.

Here R_i is such that

$$R_i = \begin{cases} 1 & \text{if unit } i \text{ belongs to the sampled group} \\ 0 & \text{if unit } i \text{ belongs to the non-sampled group,} \quad i = 1, 2, ..., N. \end{cases}$$

It is important to note that $Y_i, (i = 1001, ..., 10,000)$ will be assumed to be unobserved since they relate to the non-sampled units. However, the covariates, X_i's are always observed.

Table 1: Simulated data set of the population

i	R_i	X_{1i}	X_{2i}	X_{3i}	X_{4i}	Y_i
1	1	48	0	0	0	1
2	1	63	1	0	1	1
3	1	50	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
999	1	54	0	0	0	0
1000	1	56	1	1	1	0
1001	0	47	1	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
9998	0	64	1	0	0	1
9999	0	66	1	0	0	1
10000	0	59	1	0	0	1

Our sole purpose will be to predict these unobserved responses corresponding to the non-sampled units using data from the sampled units. Towards that end, we will formulate various modeling frameworks and will apply those on the above data. These will be illustrated in the next sections.

3. Non-spatial model

Here we describe the general approach to predict the finite population proportion using a non-spatial model.

3.1. Model specification

In order to specify the model framework, we need to define the propensity scores in the context of our setup. The propensity score for a subject/entity is the conditional probability of it being selected in a sample given its covariates. The foundational assumption in this regard is that all pertinent covariates related to the sample units are included in the study. Supposing $\boldsymbol{x}_i$ is the covariate vector corresponding to the i^{th} subject in the population, its propensity score, $\pi(\boldsymbol{x}_i)$ is given by

$$\pi(\boldsymbol{x}_i) = P(R_i = 1|\boldsymbol{x}_i, \boldsymbol{\phi}), \quad i = 1, 2, ..., N, \tag{1}$$

where R_i has been defined in Sec 2 and $\boldsymbol{\phi}$ is a vector of unknown parameters. We use a logistic regression model to model $\pi(\boldsymbol{x}_i)$ *i.e.*.

$$\pi(\boldsymbol{x}_i) = \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}, \tag{2}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ is our target of inference. We assume a non-informative prior on $\boldsymbol{\beta}$ *i.e.*. $\pi(\boldsymbol{\beta}) = 1$. Assuming independence, the conditional distribution of R_i is given by

$$R_i|\boldsymbol{\beta} \sim \text{Bernoulli}\{\pi(\boldsymbol{x}_i)\}.$$

Accordingly, the posterior density of $\boldsymbol{\beta}$ can be expressed as

$$\pi(\boldsymbol{\beta}|\boldsymbol{R}) \propto \prod_{i=1}^{N}\left(\frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}\right)^{R_i}\left(1-\frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}\right)^{1-R_i} = \prod_{i=1}^{N}\left(\frac{e^{R_i\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}\right). \quad (3)$$

It is important to note that our target of inference in this case is the finite population proportion *i.e.*. $\frac{1}{N}\sum_{i=1}^{N} Y_i$, where the sample values $(Y_i, i = 1, 2, ..., n)$ are observed and the non-sample values $(Y_i, i = n+1, ..., N)$ are missing. In the context of non-probability sampling, the missing data mechanism can be assumed to be missing-at-random (MAR), given the covariates (Little and Rubin, 2002). However, this is not a binding condition since inference can be performed on non-probability samples accommodating for both nonignorable nonresponse and selection biases (Nandram and Choi, 2010; Nandram, 2022).

3.2. Bayesian computation

Since the above posterior is not in closed form, we will need to perform the Metropolis-Hastings algorithm (Hastings, 1970) in order to draw samples from it. For that purpose, we need to define a suitable proposal density. We use Laplace approximation for that purpose. The advantage of the Laplace approximation is that for small degrees of freedom, it has increased flexibility to accommodate skewness, thus enhancing its effectiveness as an approximation. It is worth noting that the Laplace approximation is simply used as a proposal density (first approximation) in the Metropolis sampler. Accordingly, we assume that $\boldsymbol{\beta}$ approximately follows a multivariate t distribution parametrized as

$$\boldsymbol{\beta}|\sigma^2 \sim N(\hat{\boldsymbol{\beta}}, \gamma^2\hat{\Sigma}); \ \frac{\nu}{\gamma^2} \sim \chi^2_\nu$$
$$\text{where} \quad \hat{\Sigma} = -(H(\hat{\boldsymbol{\beta}}))^{-1}, \ \hat{\boldsymbol{\beta}} \text{ being the mode of } \boldsymbol{\beta}.$$

Here ν is the degrees of freedom of the multivariate t distribution and acts as a tuning parameter. The values of $\hat{\boldsymbol{\beta}}$ and $-(H(\hat{\boldsymbol{\beta}}))^{-1}$ are obtained through numerical approximation. For posterior simulation, we use Metropolis-Hastings algorithm with the following candidate density for $\boldsymbol{\beta}$

$$p(\boldsymbol{\beta}) \propto \frac{1}{\left[1+\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\hat{\Sigma}^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{\nu}\right]^{\frac{5+\nu}{2}}}.$$

We first draw 10,000 sets of $\boldsymbol{\beta} = (\beta_0, ..., \beta_4)$. Then we drop the first $B = 5000$ iterates and take every 5^{th} of the remaining iterates *i.e.*. we take iterate number $B+1, B+1+k, B+1+2k, ..., B+1+m\times k$ where $k = 5$ and $m = 1000$ being the final sample size. Table 2 depicts the posterior summaries of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ obtained from the simulated samples, which are provided under Section 3.2 in the Annexure.

We use various diagnostics to assess the convergence of the chains, like trace and autocorrelation plots, Geweke test and effective sample sizes. The trace and auto-correlation plots are shown under Section 3.2 in the annexure. The plots and the diagnostics tests indicate satisfactory mixing and convergence of the chains. In the context of simulated data, the above estimates indicate that all the predictors have a significant effect on the response, R_i. Specifically, being younger, being black, being male or having a high school or lower degree significantly increase the odds of being sampled.

Table 2: Posterior summaries of β for Beta-Bernoulli model

Parameter	Mean	Standard Deviation	95% HPD Interval
β_0	9.957	0.448	(9.089, 10.775)
β_1	-0.196	0.006	(-0.219, -0.189)
β_2	-0.785	0.073	(-0.935, -0.652)
β_3	0.264	0.071	(0.122 0.398)
β_4	-0.416	0.074	(-0.551, -0.261)

3.3. Poststratification and prediction

As mentioned before, our principal aim is to predict the responses for the non-sampled subjects using data from the sampled subjects. Towards that end, it is imperative to balance (or adjust) the covariate distributions between the sampled and non-sampled groups. We will achieve that through a combination of propensity scores and poststratification (Baker *et al.*, 2013) as depicted by Rubin (1979) and in Nandram and Choi (2010) who applied it in the analysis of body mass index data in a small area context.

3.3.1. Poststratification

The poststratification procedure will be described in this section.

For the h^{th} set of simulated values of $\boldsymbol{\beta}$, the corresponding propensity score values are given by

$$\pi_i^{(h)} = \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}^{(h)}}}{1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}^{(h)}}}, \quad h = 1, 2, ..., 1000; \;\; i = 1, 2, ..., 10,000.$$

Thus, we will have $m = 1000$ propensity score values for each of $N = 10,000$ simulated population units resulting in a $N \times m = 10,000 \times 1000$ matrix of propensity scores. Part of this matrix is shown under Sec 3.3.1 in the Annexure. Given the simulated values of the propensity scores, we create ten (10) strata by forming ten intervals from their deciles for implementing the poststratification procedure. The ten intervals are shown in Table 3, where I_j denotes the j^{th} interval. Now, for each simulated value of $\boldsymbol{\beta}$, we allocate the 10,000 population units into these strata/intervals based on their respective propensity score value. Table 4 depicts the number of subjects allocated to each of these strata corresponding to the sampled and non-sampled groups for four simulated values of $\boldsymbol{\beta}$. Note that the sample frequencies vary across the sub-strata because the deciles are based on $10^7(10,000 \times 1000)$ propensity score values.

Table 3: Propensity score intervals

I_1	I_2	I_3	I_4	I_5
(0, 0.0143]	(0.0143, 0.0237]	(0.0237, 0.0337]	(0.0337, 0.0459]	(0.0459, 0.0613]
I_6	I_7	I_8	I_9	I_{10}
(0.0613, 0.0815]	(0.0815, 0.1110]	(0.1110, 0.1550]	(0.1550, 0.2366]	(0.2366, 0.9302]

Table 4: Stratum allocation frequencies for sampled and non-sampled groups for different values of $\boldsymbol{\beta}$

		Stratum Frequency									
$\boldsymbol{\beta}$	Group	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}
	Sampled	8	20	28	36	36	70	105	133	225	339
$\boldsymbol{\beta}^{(1)}$	Non-sampled	985	916	955	1050	825	1042	911	874	820	622
	Sampled	8	30	23	38	38	77	92	115	240	339
$\boldsymbol{\beta}^{(2)}$	Non-sampled	1033	1208	852	967	825	1066	712	805	910	622
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Sampled	9	20	32	35	48	59	101	134	205	357
$\boldsymbol{\beta}^{(999)}$	Non-sampled	1012	947	1057	967	960	926	846	823	783	679
	Sampled	13	25	27	37	40	74	83	133	206	362
$\boldsymbol{\beta}^{(1000)}$	Non-sampled	1203	1038	985	924	890	833	781	871	774	701

It can be easily verified that the cumulative frequencies of the sampled and non-sampled units across all the strata/intervals are 1000 and 9000 respectively for all simulated values of $\boldsymbol{\beta}$. Conditional on the above poststratification, the covariate distribution for the sampled and non-sampled groups can be assumed to be similar for each stratum. Hence, for each stratum/interval, we can predict the response for the non-sampled units using data for the sampled units. Prediction will be carried out using the superpopulation approach mentioned in Section 1 by modeling the conditional density of the response vector (say, $\boldsymbol{Y}_s$) given the covariate vector (say, $\boldsymbol{X}_s$) for the sampled group respectively.

3.3.2. Prediction

The prediction procedure is described in this section.

Let y_{ij} denote the response for the j^{th} unit in the i^{th} stratum for the sampled group, where $i = 1, 2, ..., 10$ and $j = 1, 2, ..., n_i$, n_i being the number of sampled subjects in stratum i. For example, for $\boldsymbol{\beta}^{(1)}$, the number of sampled subjects in the 1^{st} stratum is 8 *i.e.*. $n_i = 8$. Let p_i be the probability of success (*i.e.*. $y_{ij} = 1$) for the i^{th} stratum. We have the following model specification

$$\begin{aligned} Y_{ij}|p_i &\sim \text{Bernoulli}(p_i), \;\; i = 1, 2, ..., 10; j = 1, 2, ..., n_i, \\ p_i &\sim \text{Beta}(0, 0), \;\; i = 1, 2, ..., 10. \end{aligned} \tag{4}$$

The above prior is clearly improper and is also known as the Haldane prior. We have chosen this prior for p_i in order to make the inference as data driven as possible. The posterior of p_i is

$$\begin{aligned} \pi(p_i|\boldsymbol{y}_i) &\propto f(\boldsymbol{y}_i|p_i)\pi(p_i) = p_i^{\sum_{j=1}^{n_i} y_{ij}-1}(1-p_i)^{n_i-\sum_{j=1}^{n_i} y_{ij}-1} \\ i.e.. \; p_i|\boldsymbol{y}_i &\sim \text{Beta}\left(\sum_{j=1}^{n_i} y_{ij}, n_i - \sum_{j=1}^{n_i} y_{ij}\right). \end{aligned} \tag{5}$$

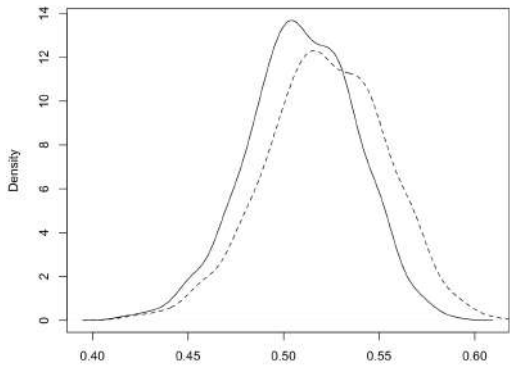
Here $\sum_{j=1}^{n_i} y_{ij}$ and $n_i - \sum_{j=1}^{n_i} y_{ij}$ are respectively the number of 1's and 0's of the response variable for the sampled data in the i^{th} stratum. As an illustration, for $\boldsymbol{\beta}^{(1)}$, there are 8

sampled units and 985 non-sampled units in the 1^{st} stratum. The response values for the 8 sampled units are $(1,0,0,1,1,0,1,1)$ *i.e.*. $\sum_{j=1}^{n_i} y_{ij} = 5$ and $n_i - \sum_{j=1}^{n_i} y_{ij} = 3$. Hence, the posterior distribution of p_1 will be Beta(5, 3). Now, we can draw a random sample of p_1 from a Beta(5, 3) distribution (say $\tilde{p_1}$) and finally draw 985 values of y_{ij} from Bernoulli($\tilde{p_1}$). The simulated values of y so drawn will be the predicted values of Y corresponding to the 1^{st} stratum. Similarly, we can predict all the non-sampled response observations for all the strata.

Once the above prediction is complete, we compute the proportion of successes *i.e.*. $Y = 1$ for the sampled and non-sampled groups separately as well as for all the N subjects in the combined set corresponding to each $\boldsymbol{\beta}^{(h)}, h = 1, 2, ..., 1000$. These are given by

$$P_{all}^{(h)} = \frac{\sum_{k=1}^{N} Y_k^{(h)}}{N} \quad \text{and} \quad P_{ns}^{(h)} = \frac{\sum_{k=n}^{N} Y_k^{(h)}}{N-n}, \quad h = 1, 2, ..., 1000.$$

The true values of the above quantities for the sampled, non-sampled and all individuals taken together are 0.398, 0.509 and 0.498 respectively. The kernel density plots of the above quantities are given in Figure 1. Here the bold (dashed) curve correspond to $P_{all}^{(h)}(P_{ns}^{(h)})$ respectively.



(a) Combined (bold: $P_{all}^{(h)}$; dashed: $P_{ns}^{(h)}$)

Figure 1: Kernel density plots of the proportion of positive responses predicted for all individuals and non-sampled individuals for Beta-Bernoulli model

3.4. Model accuracy

To evaluate the accuracy of our prediction, we compute the 95% highest posterior density (HPD) intervals of $P_{all}^{(h)}$ and $P_{ns}^{(h)}$. If the true proportion values, reported above, lies within and near the centre of the above intervals, it would indicate an accurate fit. However, if the true value lies outside the intervals or towards the edge, that would be indicative of a sub-optimal fit. The HPD interval for the true population proportion of positive responses for all the sampled units taken together ($P_{all}^{(h)}$) is found to be (.450, .559) while that for the proportion of non-sampled subjects is (0.456, 0.577). In both cases, the true values *i.e.*. 0.498 and 0.509 falls within and near the centre of the corresponding intervals. This indicates that our prediction is pretty accurate.

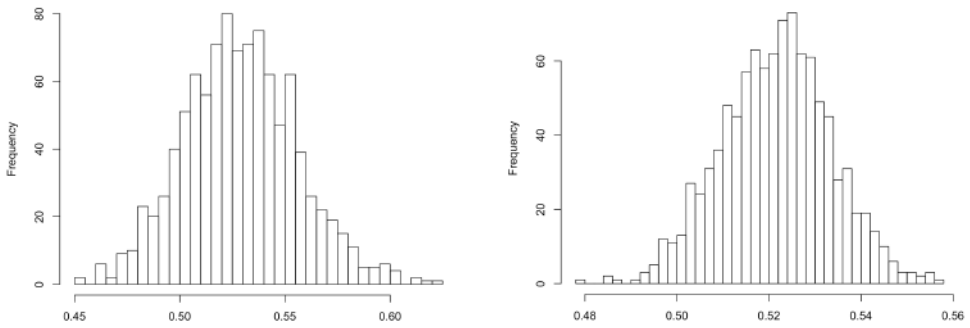
In addition to testing prediction accuracy, we also compare the predictive ability of our modeling framework, based on the superpopulation methodology, with that of quasi-randomization strategy, mentioned in Section 1, using the Horvitz-Thompson (Horvitz and Thompson, 1952) and the Hajek estimators of the population proportion (of positive responses for all the N units). These are given by

$$\hat{P}_{HT}^{(h)} = \frac{\hat{Y}_{HT}^{(h)}}{N} = \frac{1}{N}\sum_{i=1}^{n}\frac{Y_i}{Pr_i^{(h)}} \text{ and } \hat{P}_{H}^{(h)} = \frac{\sum_{i=1}^{n} Y_i/Pr_i^{(h)}}{\sum_{i=1}^{n} 1/Pr_i^{(h)}}, \quad h = 1, 2, ..., 1000,$$

where "HT" and "H" in the suffix denotes "Horvitz-Thompson" and "Hajek" respectively while

$$Pr_i^{(h)} = \frac{n\pi_i^{(h)}}{\sum_{i=1}^{N}\pi_i^{(h)}}, \quad h = 1, 2, ..., 1000; \quad i = 1, 2, ..., N.$$

Here $n = 1000$, $N = 10,000$ while $\pi_i^{(h)}$ is the propensity score for the i^{th} subject corresponding to the h^{th} case. The histograms of the 1000 simulated values of $\hat{P}_{HT}^{(h)}$ and $\hat{P}_{H}^{(h)}$ for the above estimators are shown in Figure 2. The corresponding 95% H.P.D interval of the true population proportion (for all subjects taken together) are (0.478, 0.584) and (0.496, 0.543) respectively. In both cases, the true value, 0.498, falls within the above intervals but more towards one of the edges. This is specially true for the Hajek estimator as the true population proportion is nearly equal to the lower bound of the H.P.D interval, 0.496. Thus, we can conclude that our proposed model has superior predictive properties compared to the Horvitz-Thompson and Hajek estimators. This also indicates that the superpopulation approach fares better than the quasi-optimization approach in predicting the response values for the non-sampled units.



(a) Horvitz-Thompson (b) Hajek

Figure 2: Histograms of the proportion of positive responses for Horvitz-Thompson and Hajek estimators

4. Standard spatial model

The Beta-Bernoulli framework developed in Section 3 enabled us to predict the proportion of positive responses in the non-sampled group for each stratum. In doing so, it was

assumed that the strata are independent of each other (*i.e.*. are uncorrelated). However, since the boundaries of strata are fuzzy, subjects close to the edges of two adjacent strata may have non-negligible correlation. Hence, it may be worthwhile to develop a modeling framework taking into account the spatial relationship between neighbouring strata. In this section, we will develop a Bayesian hierarchical model that incorporates this spatial association. Accordingly, we would like to test whether incorporating this spatial dimension in the modeling framework improves the ability of the model to predict the responses for the non-sampled individuals. This is a novel contribution in non-probability sampling.

4.1. Hierarchical model specification

For the proposed spatial modeling framework, the data and stratum-specific model specification remain the same as for the Beta-Bernoulli model, depicted in (3.4). As mentioned in He and Sun (2000), we specify the following logistic mixed model for p_i,

$$\log\left(\frac{p_i}{1-p_i}\right) = \theta + \nu_i,$$

where p_i is the i^{th} stratum-specific success probability for the sampled group, θ is the fixed effect and ν_i is the i^{th} stratum-specific random effect. Following He and Sun (2000), we use a simultaneous conditional autoregressive model (SCAR) to specify the prior of ν_i. Towards this end, we define the following 10×10 symmetric adjacency matrix (as we have 10 strata)

$$\mathbf{C} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

where $\mathbf{C}_{jk} = 0(1)$ means the j^{th} and k^{th} strata are non-adjacent(adjacent) *i.e..*, does not share (share) a boundary. According to He and Sun (2000), the eigenvalues of the adjacency matrix $\mathbf{C}$, given by $\lambda = (\lambda_1, \lambda_2, ..., \lambda_{10})$, can be defined so that the following inequality holds

$$\tfrac{1}{\lambda_{min}} \leq \rho \leq \tfrac{1}{\lambda_{max}}.$$

where λ_{min} and λ_{max} are the minimum and maximum eigenvalues of C. Then, based on the SCAR properties, discussed by Clayton and Kaldor (1987), the prior distribution of $\boldsymbol{\nu}$ can be shown to be

$$\boldsymbol{\nu} \sim MVN(\mathbf{0}, \delta^2(\mathbf{I} - \rho\mathbf{C})^{-1}), \tag{6}$$

where $\boldsymbol{\nu} = (\nu_1, \nu_2, \nu_3, \nu_4, \nu_5, \nu_6, \nu_7, \nu_8, \nu_9, \nu_{10})$ and $\mathbf{I}$ is a 10×10 identity matrix. In order to determine the prior distributions for θ, we employ the empirical logistic transform (Cox,

2018). Suppose we have g sets of binary observations and in the j^{th} set ($j = 1, 2, ..., g$), the success probability, p_j, is constant for that set. Let there be n_j trials and M_j successes in those trials. Then the empirical logistic transform is defined as

$$Z_j = \log\left(\frac{M_j + \frac{1}{2}}{n_j - M_j + \frac{1}{2}}\right) \text{ with mean } \phi_j = \log\left(\frac{p_j}{1-p_j}\right).$$

As per Gart and Zweifel (1967), an approximate unbiased estimator of the variance of Z_j is given by

$$V_j^2 = \frac{(n_j+1)(n_j+2)}{n_j(M_j+1)(n_j-M_j+1)}.$$

In fact, it can be shown that Z_j approximately follows a normal distribution with mean ϕ_j and variance V_j^2 *i.e.*. $Z_j \sim N(\phi_j, V_j^2)$ (McCullagh, 2019). Using the above transformation, the prior distribution of θ can be expressed as

$$\pi(\theta) = \frac{1}{V\pi\left(1 + \left(\frac{\theta - \hat{\theta}}{V}\right)^2\right)}, \quad -\infty < \theta < \infty, \tag{7}$$

which is a location-scale Cauchy distribution. Now, $y_k \sim \text{Bernoulli}\left(\frac{e^\theta}{1+e^\theta}\right)$ for $k = 1, 2, ..., n$ implying that $\hat{p} = \bar{y} = 0.398$. Thus, $\hat{\theta} = \log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.4138$. On the other hand, V is obtained as

$$V = \sqrt{\frac{(n+1)(n+2)}{n(M+1)(n-M+1)}} = 0.0646,$$

where $M = \sum_{k=1}^{n} y_k = 398$ which is the total number of positive responses in the sampled group. Finally, the prior for (δ^2, ρ) is given by

$$\pi(\delta^2, \rho) = \frac{1}{(1+\delta^2)^2}, \ \ \delta^2 > 0, \ \ \frac{1}{\lambda_{min}} \le \rho \le \frac{1}{\lambda_{max}}. \tag{8}$$

Combining the likelihood and priors specified in (6-9), the joint posterior density of $(\theta, \delta^2, \boldsymbol{\nu}, \rho)$ is given by

$$\pi(\theta, \delta^2, \boldsymbol{\nu}, \rho | \mathbf{Y}) \propto f(\mathbf{Y}|\theta, \nu_i)\pi(\boldsymbol{\nu}|\delta^2, \rho)\pi(\delta^2, \rho)\pi(\theta),$$

where

$$\begin{aligned} f(\mathbf{Y}|\theta, \nu_i) &= \prod_{i=1}^{10}\prod_{j=1}^{n_i}\left\{\frac{e^{\theta+\nu_i}}{1+e^{\theta+\nu_i}}\right\}^{y_{ij}}\left\{1 - \frac{e^{\theta+\nu_i}}{1+e^{\theta+\nu_i}}\right\}^{1-y_{ij}} \\ &= \prod_{i=1}^{10}\prod_{j=1}^{n_i}\left\{\frac{e^{(\theta+\nu_i)y_{ij}}}{1+e^{\theta+\nu_i}}\right\} \end{aligned}$$

and

$$\pi(\boldsymbol{\nu}|\delta^2,\rho) \;=\; \frac{1}{\sqrt{|\delta^2(I-\rho C)^{-1}|}}\exp\left\{-\frac{1}{2}\boldsymbol{\nu}^T(\delta^2(I-\rho C)^{-1})^{-1}\boldsymbol{\nu}\right\}.$$

Combining these forms, the joint posterior density becomes

$$\begin{aligned}\pi(\theta,\delta^2,\boldsymbol{\nu},\rho|\mathbf{Y}) \propto \prod_{i=1}^{10}\prod_{j=1}^{n_i}\left\{\frac{e^{(\theta+\nu_i)y_{ij}}}{1+e^{\theta+\nu_i}}\right\} &\times \frac{1}{\sqrt{|\delta^2(I-\rho C)^{-1}|}}\exp\left\{-\frac{1}{2}\boldsymbol{\nu}^T(\delta^2(I-\rho C)^{-1})^{-1}\boldsymbol{\nu}\right\}\\ &\times \frac{1}{(1+\delta^2)^2}\times\frac{1}{V\pi((1+(\frac{\theta-\hat{\theta}}{V})^2)},\end{aligned} \tag{9}$$

where $\delta^2>0, \hat{\theta}-10\times V<\theta<\hat{\theta}+10\times V$ (the entire unimodal density lies in a narrower interval) and $\frac{1}{\lambda_{min}}<\rho<\frac{1}{\lambda_{max}}$.

4.2. Bayesian computation

Based on the full posterior density specified in (9) above, the full conditional posterior densities are given by

$$\boldsymbol{\nu}|\theta,\delta^2,\rho,\mathbf{Y}\propto\prod_{i=1}^{10}\left\{\frac{e^{(\theta+\nu_i)R_i}}{[1+e^{\theta+\nu_i}]^{n_i}}\right\}\exp\left\{-\frac{1}{2}\boldsymbol{\nu}^T(\delta^2(I-\rho C)^{-1})^{-1}\boldsymbol{\nu}\right\}; \tag{10}$$

$$\theta|\boldsymbol{\nu},\delta^2,\rho,\mathbf{Y}\propto\prod_{i=1}^{10}\left\{\frac{e^{(\theta+\nu_i)R_i}}{[1+e^{\theta+\nu_i}]^{n_i}}\right\}\times\frac{1}{V\pi((1+(\frac{\theta-\hat{\theta}}{V})^2)}; \tag{11}$$

$$\delta^2|\theta,\rho,\boldsymbol{\nu},\mathbf{Y}\propto\frac{1}{\sqrt{\delta^2}}\exp\left\{-\frac{1}{2}\boldsymbol{\nu}^T(\delta^2(I-\rho C)^{-1})^{-1}\boldsymbol{\nu}\right\}\times\frac{1}{(1+\delta^2)^2}; \tag{12}$$

$$\rho|\delta^2,\theta,\boldsymbol{\nu},\mathbf{Y}\propto\frac{1}{\sqrt{|(I-\rho C)^{-1}|}}\exp\left\{-\frac{1}{2}\boldsymbol{\nu}^T(\delta^2(I-\rho C)^{-1})^{-1}\boldsymbol{\nu}\right\}. \tag{13}$$

Since the full conditionals are not in closed form, we need to use a combination of specialized sampling schemes to draw sample from those. Specifically, we use

1. Metropolis-Hastings algorithm to sample from $\pi(\boldsymbol{\nu}|\theta,\delta^2,\rho,\mathbf{Y})$.
2. Grid method to sample from the remaining three full conditionals, namely $\pi(\theta|\boldsymbol{\nu},\delta^2,\rho,\mathbf{Y})$, $\pi(\delta^2|\theta,\rho,\boldsymbol{\nu},\mathbf{Y})$ and $\pi(\rho|\delta^2,\theta,\boldsymbol{\nu},\mathbf{Y})$.

In the first case, we need to determine the candidate generating density to be able to apply Metropolis-Hastings algorithm. We apply the empirical logistic transformation towards this end. As per this procedure,

$$Z_i|\nu_i \sim N(\theta+\nu_i, V_i^2) \text{ where } V_i^2 = \frac{(n_i+1)(n_i+2)}{n_i(M_i+1)(n_i-M_i+1)} \text{ and } Z_i = \log\left(\frac{M_i+0.5}{n_i-M_i+0.5}\right).$$

This implies that

$$\boldsymbol{\nu} \sim \text{MVN}((\mathbf{Z}-\theta), \boldsymbol{\Sigma}),$$

where MVN implies multivariate normal distribution and $\boldsymbol{\Sigma} = \text{diag}(V_1^2, V_2^2, ..., V_{10}^2)$. Assuming $H = \delta^2(I-\rho C)^{-1}$, the proposal density will be

$$\begin{aligned}
\boldsymbol{\nu}|\theta, \delta^2, \rho, \mathbf{Y} &\propto \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\nu}-(\mathbf{Z}-\theta))^T\Sigma^{-1}(\boldsymbol{\nu}-(\mathbf{Z}-\theta))\right]\right\} \times \exp\left\{-\frac{1}{2}[\boldsymbol{\nu}^T H^{-1}\boldsymbol{\nu}]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\nu}^T\Sigma^{-1}\boldsymbol{\nu} - 2\boldsymbol{\nu}^T\Sigma^{-1}(\mathbf{Z}-\theta) + (\mathbf{Z}-\theta)^T\Sigma^{-1}(\mathbf{Z}-\theta) + \boldsymbol{\nu}^T H^{-1}\boldsymbol{\nu}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\nu}^T(\Sigma^{-1}+H^{-1})\boldsymbol{\nu} - 2\boldsymbol{\nu}^T(\Sigma^{-1}+H^{-1})(\Sigma^{-1}+H^{-1})^{-1}\Sigma^{-1}(\mathbf{Z}-\theta)\right]\right\},
\end{aligned}$$

which implies that the proposal density of $\boldsymbol{\nu}|\theta, \delta^2, \rho, \mathbf{Y}$ is

$$\boldsymbol{\nu}|\theta, \delta^2, \rho, \mathbf{Y} \sim MVN\left\{(\Sigma^{-1}+H^{-1})^{-1}\Sigma^{-1}(\mathbf{Z}-\theta), (\Sigma^{-1}+H^{-1})^{-1}\right\}. \tag{14}$$

We use the grid method to draw samples from $\theta|\boldsymbol{\nu}, \delta^2, \rho, \mathbf{Y}$ and $\rho|\delta^2, \theta, \boldsymbol{\nu}, \mathbf{Y}$. This is particularly straightforward in the first case since θ and ρ are bounded, that is

$$\hat{\theta} - 10 \times V < \theta < \hat{\theta} + 10 \times V, \text{ and } \tfrac{1}{\lambda_{min}} < \rho < \tfrac{1}{\lambda_{max}}.$$

For the conditional posterior density $\rho|\delta^2, \theta, \boldsymbol{\nu}, \mathbf{Y}$, we apply the following transformation on δ^2, since δ^2, being positive, does not have an upper bound.

$$\phi = \frac{\delta^2}{1+\delta^2},\ 0 < \phi < 1,$$

which results in the transformed density

$$\phi|\theta, \rho, \boldsymbol{\nu}, \mathbf{Y} \propto \sqrt{\frac{1-\phi}{\phi}}\exp\left\{-\frac{1}{2}\boldsymbol{\nu}^T\left(\frac{\phi}{1-\phi}(I-\rho C)^{-1}\right)^{-1}\boldsymbol{\nu}\right\}.$$

Once we have simulated the values of ϕ, we can back-transform to obtain the corresponding values of δ^2 since $\delta^2 = \dfrac{\phi}{1-\phi}$.

Given the above discussion, it will now be straightforward to simulate observations from the respective full conditionals. In doing so, we randomly select 100 sets of propensity scores among the 1000 and the Gibbs sampler is run for each such set as follows:

1. Initial values for the parameters are selected as: $\rho^{(0)} = 0, \delta^{2(0)} = 1, \theta^{(0)} = 0$.

2. Given the intital values, a sample is drawn from $\boldsymbol{\nu}|\theta^{(0)}, \delta^{2(0)}, \rho^{(0)}, \mathbf{Y}$ using the Metropolis-Hastings sampler through the candidate density derived in (15). Let the sampled value be $\boldsymbol{\nu}^{(1)}$.

3. Given $\boldsymbol{\nu}^{(1)}$, a sample is drawn from $\theta|\boldsymbol{\nu}^{(1)}, \delta^{2(0)}, \rho^{(0)}, \mathbf{Y}$ through the Grid method. Let the sampled value be denoted as $\theta^{(1)}$.

4. Given $\boldsymbol{\nu}^{(1)}$ and $\theta^{(1)}$ obtained above, we sample from $\rho|\boldsymbol{\nu}^{(1)}, \theta^{(1)}, \delta^{2(0)}, \mathbf{Y}$ again using the Grid method. Let the sampled value be denoted as $\rho^{(1)}$.

5. Given the sampled values of $\boldsymbol{\nu}^{(1)}$, $\theta^{(1)}$ and $\rho^{(1)}$, we draw a sample from $\phi|\theta^{(1)}, \rho^{(1)}, \boldsymbol{\nu}^{(1)}, \mathbf{Y}$ by applying Grid method again and perform the transformation $\delta^2 = \dfrac{\phi}{1-\phi}$ to get the corresponding value of δ^2.

6. For implementing the grid samplers in steps (3 - 5), we use the upper and lower bounds of the respective parameters and come up with the grid points.

7. At the completion of the above iteration, we obtain the first set of simulated values of the parameters vis $(\boldsymbol{\nu}^{(1)}, \theta^{(1)}, \rho^{(1)}, \delta^{2(1)})$.

We repeat the above steps 2 to 5 step for 11,000 times and do a burn-in of the first 2000 iterates. Then we do some thinning and keep the following iterates,

$$(\boldsymbol{\nu}^{(2001+9m)}, \theta^{(2001+9m)}, \rho^{(2001+9m)}, \delta^{2(2001+9m)}),$$

where $m = 1, 2, ..., 1000$. In doing so, we are finally left with 1,000 sets of iterates.

As usual, we verify the convergence of the chains using trace and autocorrelation plots along with Gweke test and effective sample sizes. The associated plots and tables are shown under Section 4.2 in the Annexure. The plots are indicative of satisfactory convergence of the chains. Posterior summaries are shown in Table 5.

Some notable observations can be made from the above table. For instance, random effects corresponding to the second, fourth and tenth strata are significant which implies that observations/subjects within these sub-classes have significant dependence. In addition the fixed effect component, θ is also significantly negative. More importantly, the SCAR model of He and Sun (2000) cannot capture the spatial correlation as the 95% credible interval of ρ is (-0.342, 0.488) and hence contains zero. We will address this issue using an improved modeling framework discussed in the next section.

Table 5: Posterior summaries of parameters for first spatial model

Parameter	Mean	Standard deviation	95% Credible interval
ν_1	0.817	0.708	(-0.559, 2.165)
ν_2	0.886	0.458	(0.059, 1.846)
ν_3	0.703	0.396	(-0.061, 1.453)
ν_4	1.204	0.365	(0.420, 1.886)
ν_5	0.539	0.297	(-0.093, 1.131)
ν_6	0.438	0.260	(-0.046, 0.968)
ν_7	0.334	0.195	(-0.051, 0.697)
ν_8	0.007	0.174	(-0.371, 0.313)
ν_9	-0.223	0.141	(-0.505, 0.051)
ν_{10}	-0.431	0.119	(-0.665, -0.192)
θ	-0.413	0.016	(-0.440, -0.378)
ρ	0.0744	0.231	(-0.342, 0.488)
δ^2	3.925	3.554	(0.243, 11.423)

4.3. Prediction

Given the sampled values of the parameters obtained above, it is straightforward to predict the responses for the non-sampled units. For each $\boldsymbol{\beta}$, the number of non-sampled individuals for each stratum is known (see Table 4). Moreover, for the i^{th} stratum,

$$y_{ij}|p_i \sim \text{Bernoulli}(p_i), \;\; i = 1, 2, ..., 10; j = 1, 2, ..., n_i.$$

Hence, we can get a sample of the responses corresponding to the non-sampled group for the i^{th} stratum by drawing the requisite number of y_{ij}'s from Bernoulli(p_i). Based on the sampled values, we can evaluate the proportion of positive responses. This exercise should be repeated for other sets of propensity scores as well. Accordingly, we randomly selected 100 sets of propensity scores and obtained 100 proportion values (of positive responses in the non-sampled group). Based on those values, we form the highest posterior density (HPD) intervals of the true proportion of positive responses as was done for the Beta-Bernoulli model. The resulting interval is (0.449, 0.551) which is clearly narrower than those corresponding to the Beta-Bernoulli, Hajek and Horvitz-Thompson estimators. In addition, the true value of the proportion for all the subjects and for the non-sampled subjects lie near the centre of the interval corresponding to the spatial model. Both of these implies that the predictive ability of the spatial model is superior to the other models *i.e.*. the predicted values of the response in the non-sampled group and the corresponding proportions obtained from the spatial model is more accurate compared to those obtained from the other models, namely Beta-Bernoulli, Horvitz-Thompson and Hajek. Histograms and density plots of the proportions are shown under Section 5.2 in the Annexure.

5. Modified spatial model

In this section, we show how to improve the standard spatial model.

5.1. Model specification

The spatial regression model outlined in Section 4 is motivated by the work of He and Sun (2000). One shortcoming of their formulation is that it fails to account for positive and monotonically weakening spatial correlation. To account for that, we introduce a modified spatial model in this section for which we define the following 10×10 adjacency matrix:

$$\mathbf{A} = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 & \rho^6 & \rho^7 & \rho^8 & \rho^9 \\ \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 & \rho^6 & \rho^7 & \rho^8 \\ \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 & \rho^6 & \rho^7 \\ \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 & \rho^6 \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 \\ \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho^6 & \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^7 & \rho^6 & \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^8 & \rho^7 & \rho^6 & \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^9 & \rho^8 & \rho^7 & \rho^6 & \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}, \quad 0 < \rho < 1.$$

The structure of the adjacency matrix distinguishes it from the spatial model discussed in Section 4. Specifically, the underlying assumption for the above structure is that subjects belonging to strata in close proximity have higher dependence than those belonging to strata which are further apart. The logistic mixed model specification for p_i remains the same as was done for the standard spatial model in Sec 4.1 *i.e.*.

$$\log\left(\frac{p_i}{1-p_i}\right) = \theta + \nu_i \quad i = 1, 2, ..., 10,$$

θ and ν_i having the same connotation as before. The conditional distribution of y_{ij} remains the same as for the standard spatial model *i.e.*.

$$y_{ij}|\theta, \nu_i \sim \text{Ber}\left(\frac{e^{\theta+\nu_i}}{1+e^{\theta+\nu_i}}\right), \quad i = 1, 2, ..., 10, j = 1, 2, ..., n_i.$$

The following priors are specified for the parameters $(\boldsymbol{\nu}, \theta, \delta^2, \rho)$

$$\begin{aligned} \boldsymbol{\nu}|\theta, \delta^2, \rho &\sim MVN(\theta\boldsymbol{j}, \delta^2\mathbf{A}), \\ \pi(\theta, \delta^2, \rho) &\propto \frac{1}{(1+\delta^2)^2}, \end{aligned}$$

where $\boldsymbol{j}$ is a 10×1 dimensional vector of 1's while $0 < \theta < 1$ and $0 < \rho < 1$. Combining the likelihood and priors, the joint posterior density of $(\boldsymbol{\nu}, \theta, \delta^2, \rho)$ is given by

$$\pi(\theta, \delta^2, \boldsymbol{\nu}, \rho|\mathbf{Y}) \propto f(\mathbf{Y}|\nu_i)\pi(\boldsymbol{\nu}|\theta, \delta^2, \rho)\pi(\theta, \delta^2, \rho),$$

where

$$\begin{aligned} f(\boldsymbol{Y}|\theta, \nu_i) &= \prod_{i=1}^{10}\prod_{j=1}^{n_i}\left\{\frac{e^{\theta+\nu_i}}{1+e^{\theta+\nu_i}}\right\}^{y_{ij}}\left\{1-\frac{e^{\theta+\nu_i}}{1+e^{\theta+\nu_i}}\right\}^{1-y_{ij}} \\ &= \prod_{i=1}^{10}\prod_{j=1}^{n_i}\left\{\frac{e^{(\theta+\nu_i)y_{ij}}}{1+e^{\theta+\nu_i}}\right\} \end{aligned}$$

and

$$\pi(\boldsymbol{\nu}|\theta,\delta^2,\rho) = \frac{1}{\sqrt{|\delta^2\mathbf{A}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{\nu}-\theta\boldsymbol{j})^T(\delta^2\mathbf{A})^{-1}(\boldsymbol{\nu}-\theta\boldsymbol{j})\right\}.$$

Thus, the joint posterior density is

$$\begin{aligned}\pi(\theta,\delta^2,\boldsymbol{\nu},\rho|\mathbf{Y}) &\propto \prod_{i=1}^{10}\prod_{j=1}^{n_i}\left\{\frac{e^{(\theta+\nu_i)y_{ij}}}{1+e^{\theta+\nu_i}}\right\}\frac{1}{\sqrt{|\delta^2\mathbf{A}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{\nu}-\theta\boldsymbol{j})^T(\delta^2\mathbf{A})^{-1}(\boldsymbol{\nu}-\theta\boldsymbol{j})\right\}\times\frac{1}{(1+\delta^2)^2}\\ &= \prod_{i=1}^{10}\left\{\frac{e^{(\theta+\nu_i)M_i}}{(1+e^{\theta+\nu_i})^{n_i}}\right\}\frac{1}{\sqrt{|\delta^2\mathbf{A}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{\nu}-\theta\boldsymbol{j})^T(\delta^2\mathbf{A})^{-1}(\boldsymbol{\nu}-\theta\boldsymbol{j})\right\}\times\frac{1}{(1+\delta^2)^2},\end{aligned}$$

where $M_i=\sum_{j=1}^{n_i}y_{ij}$ is the total number of positive responses in the i^{th} subclass of the sampled group.

5.2. Bayesian computation

The following full conditional posterior densities can be derived from the full posterior shown above

$$\boldsymbol{\nu}|\theta,\delta,\rho,\mathbf{Y} \propto \prod_{i=1}^{10}\left\{\frac{e^{(\theta+\nu_i)M_i}}{(1+e^{\theta+\nu_i})^{n_i}}\right\}\frac{1}{\sqrt{|\delta^2\mathbf{A}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{\nu}-\theta\boldsymbol{j})^T(\delta^2\mathbf{A})^{-1}(\boldsymbol{\nu}-\theta\boldsymbol{j})\right\}; \quad (15)$$

$$\theta|\boldsymbol{\nu},\delta,\rho,\mathbf{Y} \sim N\left(\frac{\boldsymbol{j}^T(\delta^2\mathbf{A})^{-1}\boldsymbol{\nu}}{\boldsymbol{j}^T(\delta^2\mathbf{A})^{-1}\boldsymbol{j}},\frac{1}{\boldsymbol{j}^T(\delta^2\mathbf{A})^{-1}\boldsymbol{j}}\right); \quad (16)$$

$$\delta^2|\theta,\rho,\boldsymbol{\nu},\mathbf{Y} \propto \frac{1}{(\delta^2)^5}\exp\left\{-\frac{1}{2}(\boldsymbol{\nu}-\theta\boldsymbol{j})^T(\delta^2\mathbf{A})^{-1}(\boldsymbol{\nu}-\theta\boldsymbol{j})\right\}\times\frac{1}{(1+\delta^2)^2}; \quad (17)$$

$$\rho|\delta^2,\theta,\boldsymbol{\nu},\mathbf{Y} \propto \frac{1}{\sqrt{|\mathbf{A}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{\nu}-\theta\boldsymbol{j})^T(\delta^2\mathbf{A})^{-1}(\boldsymbol{\nu}-\theta\boldsymbol{j})\right\}. \quad (18)$$

Following a similar method that was detailed in Section 4.2, we obtain the following proposal density of $\boldsymbol{\nu}|\theta,\delta,\rho,\mathbf{Y}$

$$\boldsymbol{\nu}|\theta,\delta,\rho,\mathbf{Y} \sim MVN\left\{(\Sigma^{-1}+H^{-1})^{-1}(\Sigma^{-1}\mathbf{Z}+H^{-1}\theta\boldsymbol{j}),(\Sigma^{-1}+H^{-1})^{-1}\right\},$$

where $H=\delta^2\mathbf{A}$ and $\Sigma=\mathrm{diag}(V_1^2,V_2^2,...,V_{10}^2)$ and

$$Z_i=\log\frac{M_i+0.5}{n_i-M_i+0.5},\quad V_i^2=\frac{(n_i+1)(n_i+2)}{n_i(M_i+1)(n_i-M_i+1)},$$

The simulation steps will be similar to those mentioned in Section 4.2. As usual, convergence is verified using trace plots, autocorrelation plots, Geweke test and effective sample size procedures. All these tests indicate adequate convergence. Tables showing the p-values for the Geweke test and effective sample sizes are shown under Section 5.2 in the Annexure along with trace and kernel density plots of all the parameters. As shown in that table, the effective sample sizes of all but one parameter is 1000, the same length as the chain, thus indicating satisfactory convergence. Table 6 depicts the posterior summaries of all the parameters.

Table 6: Posterior summaries for modified spatial model

Parameter	Mean	Standard deviation	95% Credible interval
ν_1	0.429	0.442	(-0.491, 1.229)
ν_2	0.722	0.340	(0.061, 1.382)
ν_3	0.691	0.299	(0.125, 1.264)
ν_4	0.449	0.271	(-0.131, 0.955)
ν_5	0.171	0.240	(-0.316, 0.614)
ν_6	-0.029	0.212	(-0.426, 0.418)
ν_7	-0.210	0.170	(-0.547, 0.137)
ν_8	-0.377	0.151	(-0.684, -0.100)
ν_9	-0.417	0.141	(-0.671, -0.127)
ν_{10}	-0.914	0.113	(-1.116, -0.687)
θ	-0.024	0.463	(-0.963, 0.897)
ρ	0.664	0.203	(0.274, 0.972)
δ^2	0.405	0.254	(0.077, 0.864)

A comparison of Tables 5 and 6 results in some important observations. Firstly, in the modified spatial model, the random effects corresponding to five sub-classes are significant, namely those for second, third, eighth, ninth and tenth subclasses. For the first spatial model, this was true for only three subclasses. This indicates that the modified spatial model has better discriminatory ability in capturing intra-subclass-specific spatial dependence compared to the first spatial model. Secondly, the credible intervals for the modified spatial model are in general narrower than those corresponding to the previous spatial model. This implies that the modified spatial model generates more precise estimates of the parameters relative to the original spatial model. Moreover, the correlation parameter (ρ) is significant for the modified spatial model but was insignificant in the previous model. This is a major finding since it implies that the modified model is more capable of capturing the underlying spatial dependence between the sub-strata compared to the previous model. Thirdly, the estimate for the variance component δ^2 is much smaller for the modified model as compared to the previous model. This indicates that the modified model has superior ability to control for variance inflation of the strata specific random effects which indicate a better predictive ability of the responses for the non-sampled units.

5.3. Prediction

Since the chains have converged, we can use the parameter estimates to predict the responses corresponding to the non-sampled units as was done for the Beta-Bernoulli and standard spatial models. The 95% highest posterior density intervals corresponding to the modified spatial model along with those for the Beta-Bernoulli model, standard spatial model and those of Horvitz-Thompson and Hajek estimators are shown in Table 7. All the intervals relate to the prediction of the proportion of positive responses for all the subjects (sampled + non-sampled). It is clear from Table 7 that the modified spatial model has superior predictive ability compared to all the other models since it results in the narrowest HPD interval among the model-based intervals; the width under the Hajek model is much too small. Moreover, the true value of the proportion of positive responses for all the units *viz.* 0.4976, lies near the centre of the above interval as well. So, we conclude that the modified spatial model is

the optimal model for prediction. It is important to note here that the Hajek estimator is usually more precise in a design-based situation Särndal *et al.* (1992). However, it is difficult to evaluate the standard errors because it involves second-order inclusion probabilities. So, we have used an output analysis from the Metropolis sampler to get repeated values of the Hajek estimator like a bootstrap sample. These may not be the best estimates of the standard errors as they might be small.

Table 7: 95% credible intervals for the different models

Model	95% HPD interval	Width
Horvitz-Thompson	(0.478, 0.584)	0.106
Hajek	(0.497, 0.543)	0.046
Beta-Bernoulli	(0.455, 0.559)	0.104
Spatial	(0.433, 0.517)	0.084
Modified spatial	(0.456, 0.537)	0.081

6. Model comparison

Based on the discussion in the previous section, specifically with regard to the parameter estimates and credible intervals depicted in Tables 5 and 6, it is evident that the modified spatial model is more robust and has better predictive ability than the Beta-Bernoulli and standard spatial models. In this section, we will use two more diagnostic tools, namely conditional predictive ordinate (CPO) and log-pseudo marginal likelihood (LPML) to validate this fact.

The conditional predictive ordinate (CPO), introduced by Geisser (1980), is used to detect observations which are fitted poorly by a given parametric model. The CPO values can be calculated based on the output of the Markov chain Monte Carlo simulation procedure. The Monte Carlo approximation of CPO for the i^{th} stratum is given by

$$\hat{CPO}_i = \left[\frac{1}{M}\sum_{h=1}^{M}\frac{1}{f(y_i|p_i^{(h)})}\right]^{-1}, \quad i = 1, 2, ..., 10; \quad h = 1, 2, ..., 1000,$$

where $\hat{CPO}_i$ is the harmonic mean of $f(y_i|p_i^{(h)})$. For the Beta-Bernoulli model, $M = 1000$, $p_i^{(h)}$ is the h^{th} sample drawn from the posterior density of $p_i|\mathbf{y}_i$ while $\mathbf{y}_i \sim \text{Binomial}(n_i, p_i)$ for i^{th} stratum ($i = 1, 2, ..., 10,\ h = 1, 2, ..., 1000$). For the spatial model, $M = 100$ and p_i is obtained from the following expression

$$p_i = \frac{e^{\theta+\nu_i}}{1 + e^{\theta+\nu_i}},$$

where (θ, ν_i) are drawn from their respective posterior densities through the Monte Carlo simulation. Here also, $\mathbf{y}_i|p_i \sim \text{Binomial}(n_i, p_i)$ for the i^{th} stratum ($i = 1, 2, ..., 10$). For our proposed frameworks, each CPO value will correspond to a particular stratum and will indicate which, if any, stratum is an outlier in terms of model fit. Table 8 depicts the CPO values for each strata corresponding to the Beta-Bernoulli, basic spatial and modified spatial

Table 8: CPO values for the proposed models

Stratum	Beta-Bernoulli	Spatial I	Spatial II
1	0.061	0.119	0.206
2	0.029	0.041	0.071
3	0.046	0.021	0.074
4	0.041	0.077	0.082
5	0.037	0.037	0.063
6	0.039	0.047	0.047
7	0.028	0.039	0.041
8	0.023	0.035	0.034
9	0.021	0.027	0.027
10	0.0004	0.009	0.017

models. The spatial models are denoted as Spatial I (basic spatial) and Spatial II (modified spatial) respectively.

In terms of assessing model fit, observations with CPO values less than 0.025 are deemed as possible outliers while those with values less than .014 are regarded as extreme observations (Ntzoufras, 2011). From the CPO values depicted in Table 8, it can be concluded that for Beta-Bernoulli model, there are three outlying strata, namely strata 8, 9 and 10. Of this, stratum 10 seems to be an influential point since the CPO value is lesser than 0.014. For Spatial model I, there are two outlying strata (strata 3 and 10). Again, stratum 10 seems to be an influential point. Finally, for Spatial model II, only the last stratum is identified as an outlier but not an influential point. Hence, it is apparent that the modified spatial model (Spatial II) performs better than the other models as per this diagnostic measure since it has the lowest number of outlier strata and no influential strata.

In order to have a confirmatory assessment of model fit, we next calculate the log-pseudo marginal likelihood (LMPL), which is a function of CPO, given by

$$\text{LPML} = \sum_{i=1}^{N} \log(\hat{CPO}_i).$$

Larger values of LMPL indicate a better fit. The following table depicts the values of LMPL for all the three models. Since the modified spatial model has the highest value of LMPL, we conclude that it has superior predictive ability compared to the Beta-Bernoulli and the standard spatial models. This validates the findings derived in Section 5.

Model	LMPL
Beta-Bernoulli	-38.11
Spatial	-32.97
Modified spatial	-29.39

7. Discussion

The standard method of obtaining a representative sample from a target population is through a probability sampling scheme which involves the selection of population units

according to a certain specified probability distribution. The most common of these methods is simple random sampling in which each and every population unit is assigned the same probability of selection. Having said that, implementation of an ideal probability sampling scheme in a real life setting is prohibitively difficult due to restrictions on costs, manpower and time among other things. This has led to the popularization of alternate sampling schemes which are easier to implement on the field as well in the online space. Some examples are convenience sample, volunteer sample, online polls etc.

However, one major disadvantage of these schemes is that selection of units are heavily dependent on the choice and preference of the survey designer and is often guided by convenience rather than an underlying probabilistic framework. Hence, these kinds of samples are known as non-probability samples and the generating scheme, a non-probability sampling scheme. Consequently, the sample, so chosen, often comes with various biases which may lead to a unreliable estimate of the parameter of interest. Selection bias is one such bias which results in a sample that may lack representation of certain segments of the target population. This results in a sample that is not a proper representation of the target population.

Regardless of the above shortcomings, non-probability sampling schemes are becoming increasingly popular due to the ease with which they can be implemented, both on the ground and in the virtual space. However, it is equally critical not to sacrifice on the "representativeness" of the final sample and the unbiasedness of the final estimate as it reflects the true population parameter. Hence, it is utmost importance to come up with a general framework that would enable us to predict the responses of sample units which should have been sampled but were left out in a non-probability sampling scheme.

In this article, we have proposed three modeling frameworks that will enable us to predict the non-sampled individuals responses from information obtained from the sampled units. The underlying idea behind each of these frameworks was to first balance the covariate distributions of the sampled and non-sampled groups/units. This was implemented using the propensity scores for those units. The propensity scores quantified the probability that a particular unit is incorporated in a sample given the values of its covariates and were obtained using a Bayesian hierarchical model. Ten strata were constructed based on the quantiles of the propensity scores so obtained. Finally, prediction of the unknown responses of the non-sampled units were carried out using three models - a Beta-Bernoulli model and two spatial models which accounted for possible spatial autocorrelation between the strata. Between the two spatial models we proposed, one incorporated postive and gradually weakening correlation structure while the other did not. We tested our models on a simulated dataset. Estimation was carried out through Markov chain Monte Carlo simulation and Bayesian bootstrap.

A comparison of the predictive abilities of the aforementioned models unambiguously indicated the superiority of the spatial modeling framework over the non-spatial ones, namely the Beta-Bernoulli, Horvitz-Thompson and Hajek estimators. Moreover, the spatial model incorporating the gradually weakening spatial correlation structured performed considerably better than the one which did not incorporate this feature and had the best predictive ability of all the models. This points to the veracity of our assumption about the presence of long range but diminishing spatial autocorrelation between strata, which was an interesting

finding in its own right. Overall, we believe that our proposed methodology will contribute to ongoing research in this important field of research. Our proposed methodology was built on the superpopulation modeling framework. An interesting extension of our work would be the formulation of predictive approaches combining the superpopulation and quasi-randomisation frameworks.

Acknowledgements

It is an honour and privilege to be invited to submit this paper for the upcoming special issue in honour of Prof. C. R. Rao. The authors gratefully acknowledge the assistance received from Hanqi Cao and Zhiqing Xu on an earlier version of this paper. Balgobin Nandram was supported by a grant from the Simons Foundation (#353953, Balgobin Nandram). Last but not the least, we would like to thank the editor, Prof. Vinod Gupta and the reviewers for their careful reading of the manuscript and for their encouraging and insightful comments which led to significant improvement of the article.

References

Beaumont, J. F. (2020). Are probability surveys bound to disappear for the production of official statistics. *Survey Methodology*, **46**, 1–29.

Chen, Y., Li, P., and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, **115**, 2011–2021.

Choi, S., Nandram, B., and Kim, D. (2021). Bayesian predictive inference of small area proportions under selection bias. *Survey Methodology*, **47**, 91–123.

Clayton, D. and Kaldor, J. (1987). Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **1**, 671–681.

Cochran, W. G. and Chambers, S. P. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A*, **128**, 234–266.

Cox, D. R. (2018). *Analysis of Binary Data*. Routledge.

Elliot, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, **2**, 813–845.

Elliott, M. N. and Haviland, A. (2007). Use of a web-based convenience sample to supplement a probability sample. *Survey Methodology*, **33**, 211–215.

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, **32**, 249–264.

Gart, J. J. and Zweifel, J. R. (1967). On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika*, **1**, 181–187.

Geisser, S. (1980). Discussion on sampling and Bayes' inference in scientific modeling and robustness (by gep box). *Journal of the Royal Statistical Society A*, **143**, 416–417.

Hastings, W. K. (1970). *Monte Carlo Sampling Methods using Markov Chains and their Applications*. Oxford University Press.

He, Z. and Sun, D. (2000). Hierarchical Bayes estimation of hunting success rates with spatial correlations. *Biometrics*, **56**, 360–367.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Kohut, A., Keeter, S., Doherty, C., Dimock, M., and Christian, L. (2012). Assessing the representativeness of public opinion surveys. Washington, DC: Pew Research Center.

Little, R. J. (1982). Models for nonresponse in sample surveys. *Journal of the American statistical Association*, **77**, 237–250.

Little, R. J. (1993). Post-stratification: a modeler's perspective. *Journal of the American Statistical Association*, **88**, 1001–1012.

Little, R. J. and Rubin, D. B. (2002). Bayes and multiple imputation. *Statistical Analysis with Missing Data*, , 200–220.

Marella, D. (2023). Adjusting for selection bias in nonprobability samples by empirical likelihood approach. *Journal of Official Statistics*, **39**, 151–172.

McCullagh, P. (2019). *Generalized Linear Models*. Routledge.

Meng, X. L. (2018). Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics*, **12**, 685–726.

Nandram, B. (2022). A Bayesian assessment of non-ignorable selection of a non-probability sample. *Indian Bayesians News Letter*, **14**, 7–20.

Nandram, B. and Choi, J. W. (2005). Hierarchical Bayesian nonignorable nonresponse regression models for small areas: An application to the nhanes data. *Survey Methodology*, **31**, 73–84.

Nandram, B. and Choi, J. W. (2010). A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection. *Journal of the American Statistical Association*, **105**, 120–135.

Nandram, B., Choi, J. W., and Liu, Y. (2021). Integration of nonprobability and probability samples via survey weights. *Journal of Statistics and Probability*, **10**, 1–5.

Nandram, B. and Rao, J. (2021). A Bayesian approach for integrating a small probability sample with a non-probability sample. In *Proceedings of the American Statistical Association (JSM2021-Virtual Conference)*, pages 1568–1603.

Nandram, B. and Rao, J. (2023). Bayesian predictive inference when integrating a non-probability sample and a probability sample. *arXiv preprint arXiv:2305.08997*, .

Nandram, B. and Rao, J. N. K. (2024). Bayesian integration for small areas by supplementing a probability sample with a non-probability sample. *Statistics and Applications*, **22**, 345–376.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, **97**, 558–625.

Ntzoufras, I. (2011). *Bayesian Modeling using WinBUGS*, volume 698. John Wiley & Sons.

Rafei, A., Flannagan, C. A., West, B. T., and Elliott, M. R. (2022). Robust Bayesian inference for big data: Combining sensor-based records with traditional survey data. *The Annals of Applied Statistics*, **16**, 1038–1070.

Rao, J. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, **83**, 242–272.

Robbins, M. W., Ghosh-Dastidar, B., and Ramchand, R. (2021). Blending probability and nonprobability samples with applications to a survey of military caregivers. *Journal of Survey Statistics and Methodology*, **9**, 1114–1145.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, **74**, 318–328.

Sakshaug, J. W., Wiśniowski, A., Ruiz, D. A. P., and Blom, A. G. (2019). Supplementing small probability samples with nonprobability samples: A Bayesian approach. *Journal of Official Statistics*, **35**, 653–681.

Salvatore, C., Biffignandi, S., Sakshaug, J. W., Wiśniowski, A., and Struminskaya, B. (2024). Bayesian integration of probability and nonprobability samples for logistic regression. *Journal of Survey Statistics and Methodology*, **12**, 458–492.

Smith, T. (1983). On the validity of inferences from non-random samples. *Journal of the Royal Statistical Society: Series A (General)*, **146**, 394–403.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, **25**, 1.

Särndal, C. E., Swensson, B., Wretman, J., and et al. (1992). *Model Assisted Survey Sampling*. NewYork: Springer-Verlag.

Wang, L., Valliant, R., and Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, **40**, 5237–5250.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, **31**, 980–991.

Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A., and Blom, A. G. (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, **8**, 120–147.

Xu, Z. and Nandram, B. (2019). Bayesian inference of non-probability samples. In *Proceedings of the American Statistical Association, Bayesian Statistics Section)*, pages 2585–2593.

Xu, Z., Nandram, B., and Manandhar, B. (2020). Bayesian inference of a finite population mean under length-biased sampling. *Statistical Methods and Applications in Forestry and Environmental Sciences*, 79–103.

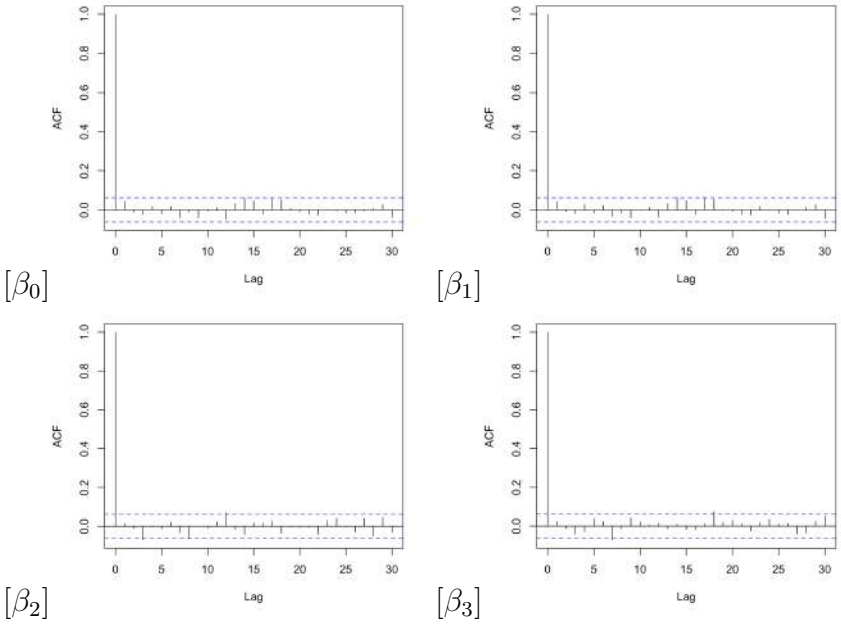
ANNEXURE BY SECTIONS

Section 3.2

The following table depicts part of the 1000 simulated values of $\boldsymbol{\beta}$ obtained using the Metropolis-Hastings sampler on $\pi(\boldsymbol{\beta}|\boldsymbol{R})$.

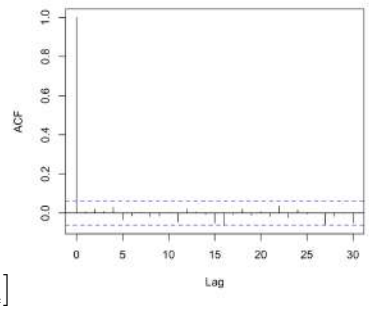
m	β_0	β_1	β_2	β_3	β_4
1	9.7473	-0.2002	-0.8234	0.2845	-0.4375
2	10.3579	-0.2096	-0.8713	0.2202	-0.4657
3	9.5018	-0.1960	-0.6962	0.1437	-0.5264
⋮	⋮	⋮	⋮	⋮	
998	9.1859	-0.1921	-0.6716	-0.2068	-0.3391
999	10.1561	-0.2094	-0.6304	0.3223	-0.4608
1000	10.3528	-0.2090	-0.9777	0.1989	-0.5132

The following figures depict the autocorrelation plots, trace plots and the kernel density plots for the simulated values of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ obtained from the Markov Chain Monte Carlo run of the Bayesian Bootstrap model.



$[\beta_0]$ $[\beta_1]$

$[\beta_2]$ $[\beta_3]$



$[\beta_4]$

Section 3.3.1

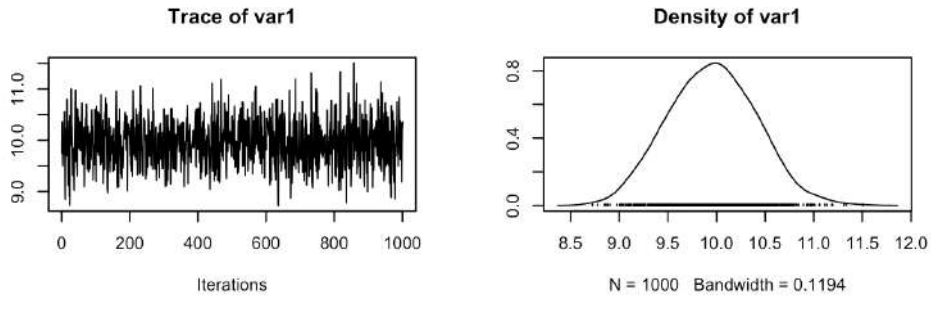
Following is part of the propensity score matrix where the rows correspond to the subjects ($N = 10000$) and columns correspond to 1000 simulated values of $\boldsymbol{\beta}$.

i	$\boldsymbol{\beta}^{(1)}$	$\boldsymbol{\beta}^{(2)}$	$\boldsymbol{\beta}^{(3)}$	⋮	$\boldsymbol{\beta}^{(1000)}$
1	0.5331	0.5614	0.5552		0.5836
2	0.0158	0.0146	0.0178		0.0134
3	0.4336	0.4593	0.4518		0.4790
⋮	⋮	⋮	⋮	⋮	⋮
9998	0.0199	0.0188	0.0246		0.0181
9999	0.0135	0.0125	0.0168		0.0119
10000	0.0526	0.0519	0.0632		0.0497

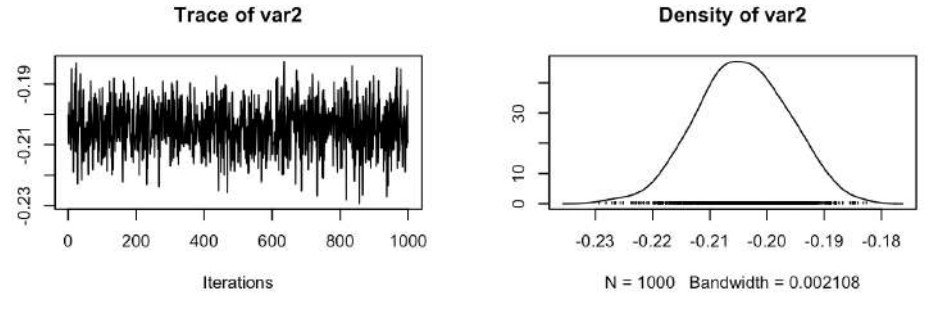
Section 4.2

The following table shows the p-values corresponding to the Gweke test and the effective sample sizes for $(\boldsymbol{\nu}, \theta, \rho, \delta^2)$ of the spatial model. All of the effective sample sizes are close to the size of chain $1,000$, which is desirable.

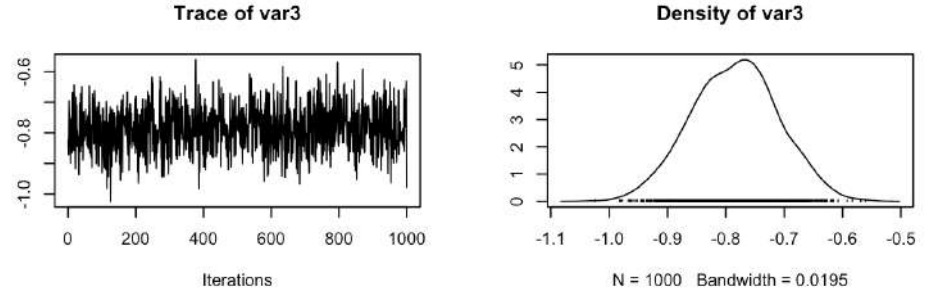
Parameter	P-value	Effective sample size
ν_1	0.098	1000
ν_2	0.186	1000
ν_3	0.459	874
ν_4	0.357	1000
ν_5	0.881	1000
ν_6	0.752	1000
ν_7	0.978	1000
ν_8	0.049	899
ν_9	0.285	1000
ν_{10}	0.768	1000
θ	0.667	1000
ρ	0.796	890
δ^2	0.721	926



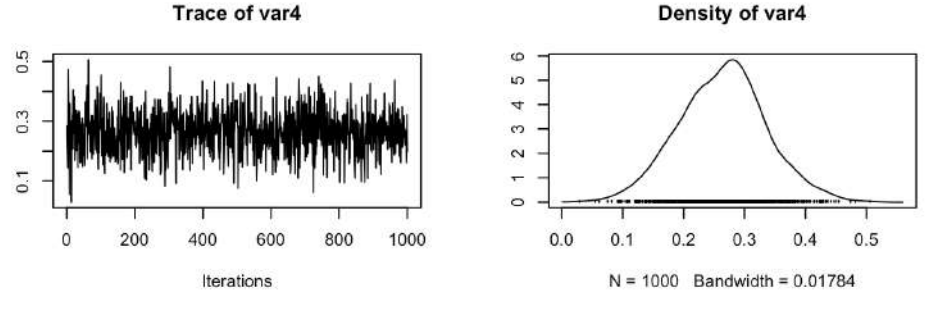
$[\beta_0]$



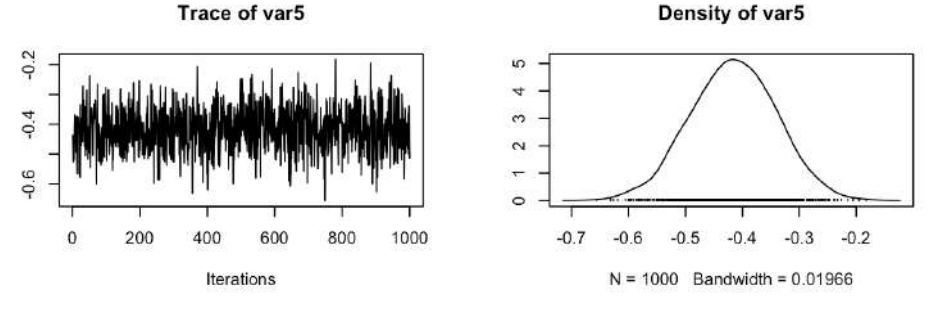
$[\beta_1]$



$[\beta_2]$

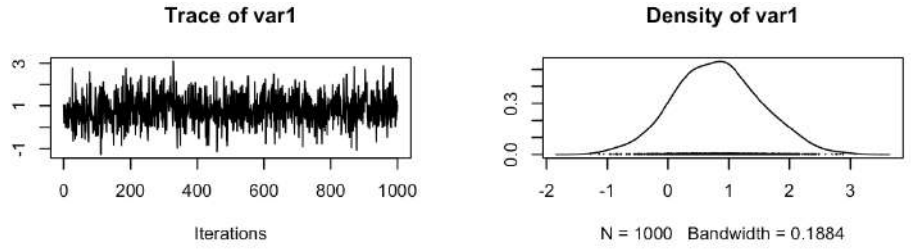


$[\beta_3]$

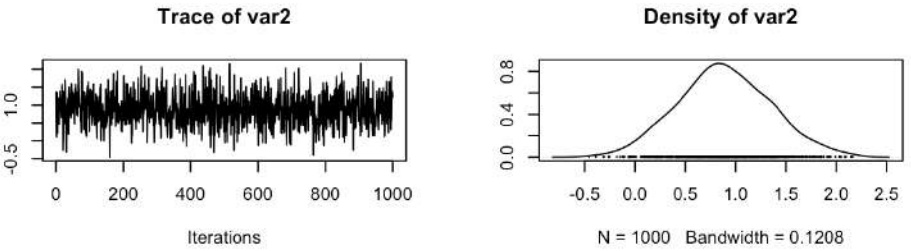


$[\beta_4]$

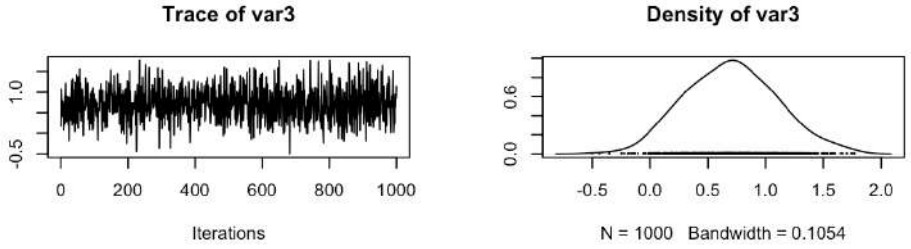
The following figures depict the trace plots, autocorrelation plots and the kernel density plots for the simulated values of $(\boldsymbol{\nu}, \theta, \rho, \delta^2)$ obtained from the Markov Chain Monte Carlo run of the Spatial model.



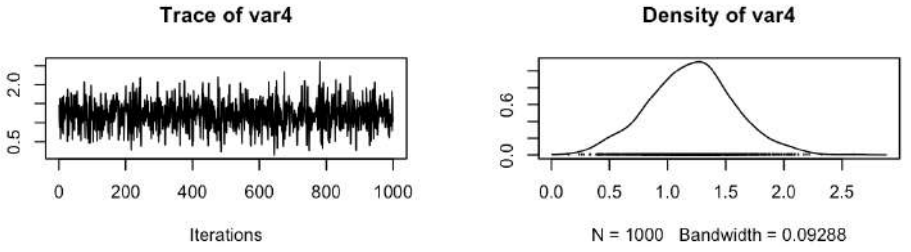
$[\nu_1]$



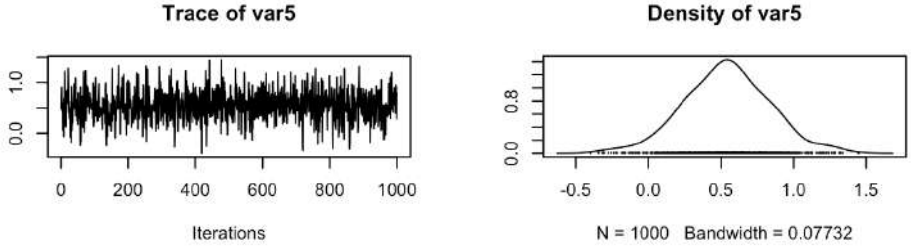
$[\nu_2]$



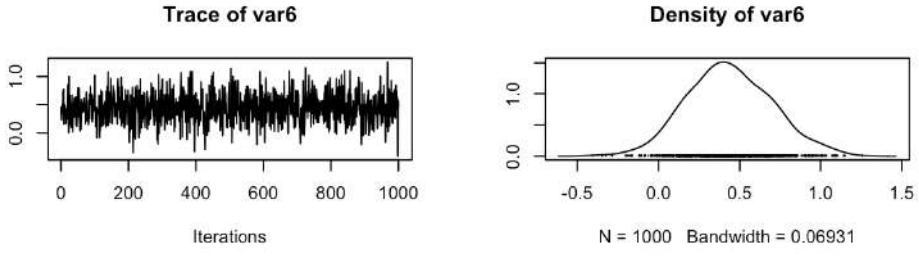
$[\nu_3]$



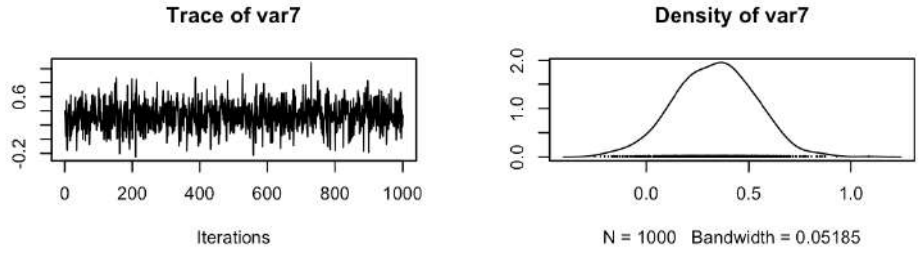
$[\nu_4]$



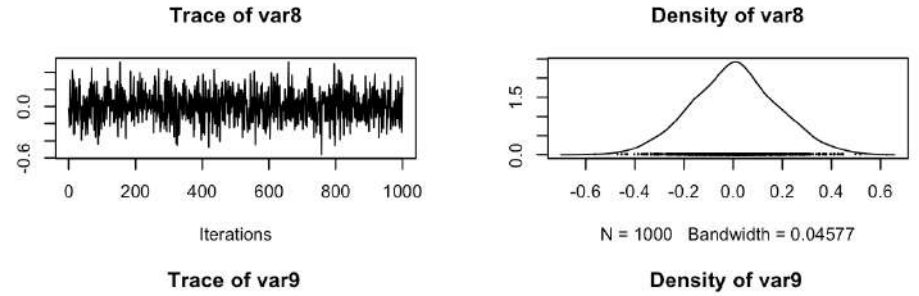
$[\nu_5]$



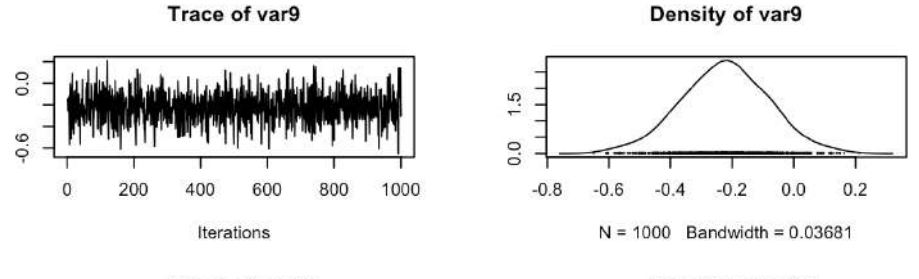
$[\nu_6]$



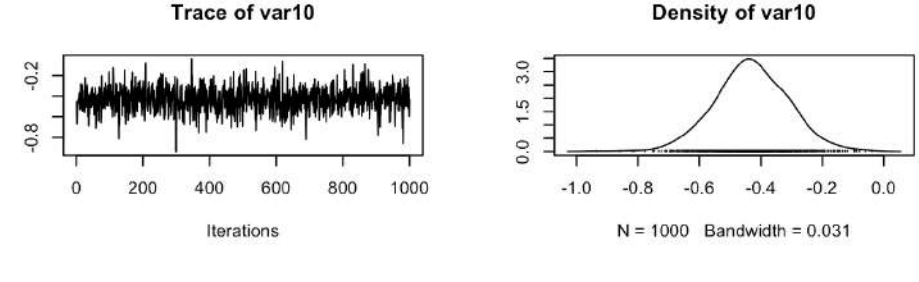
$[\nu_7]$



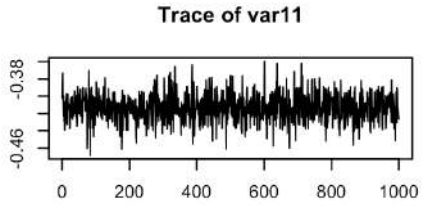
$[\nu_8]$

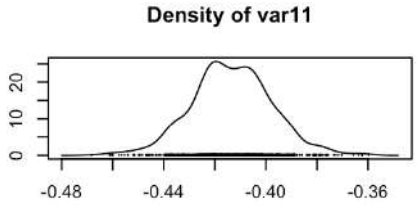


$[\nu_9]$

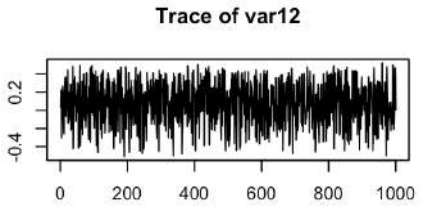


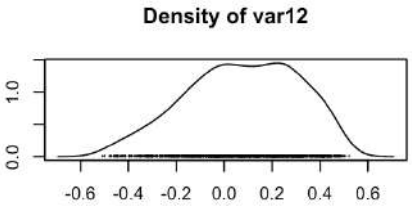
$[\nu_{10}]$



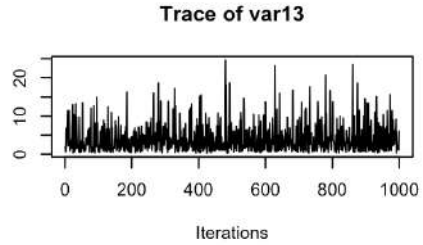


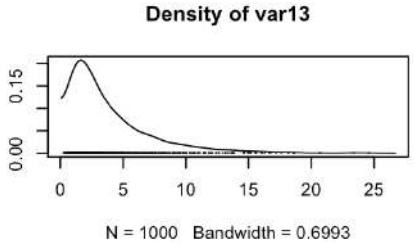
$[\nu_{11}]$



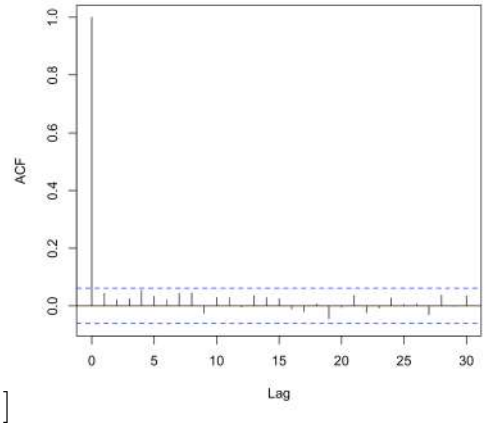


$[\nu_{12}]$

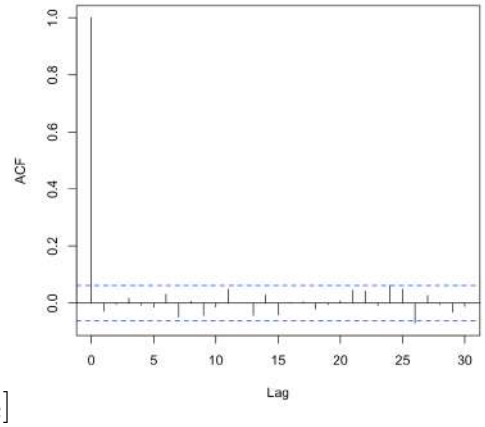




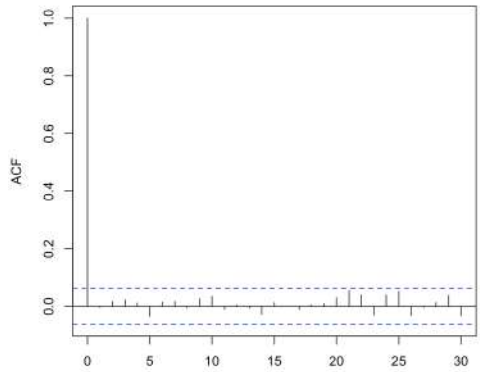
$[\nu_{13}]$



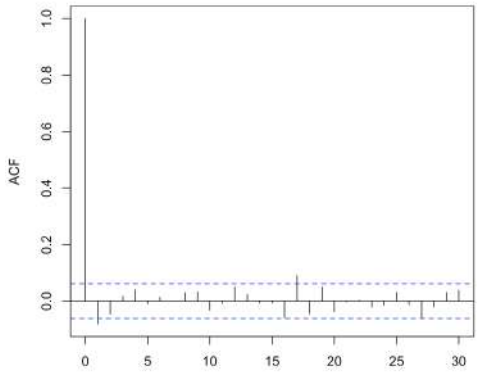
$[\nu_1]$



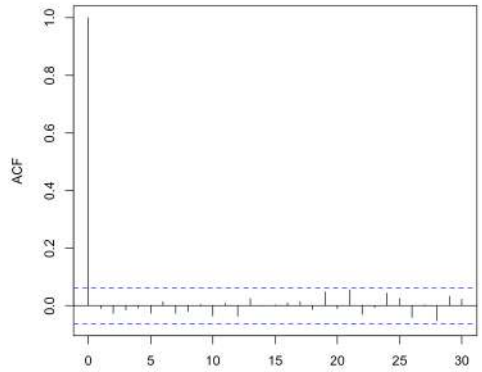
$[\nu_2]$



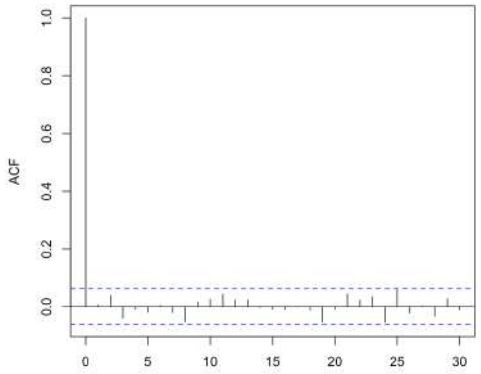
$[\nu_3]$



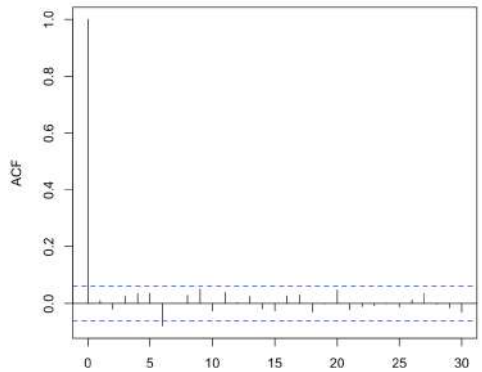
$[\nu_4]$



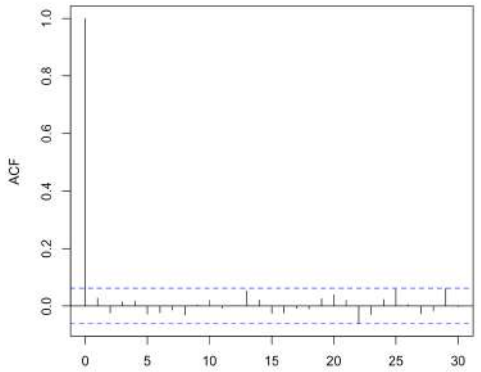
$[\nu_5]$



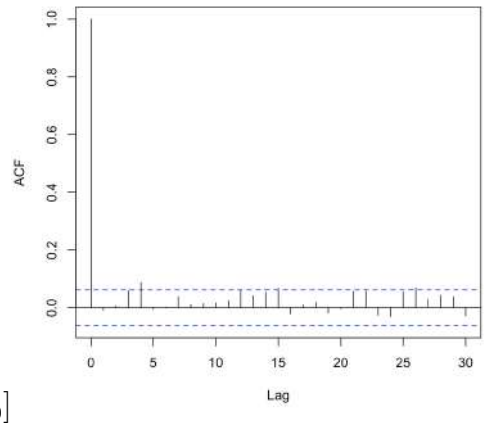
$[\nu_6]$



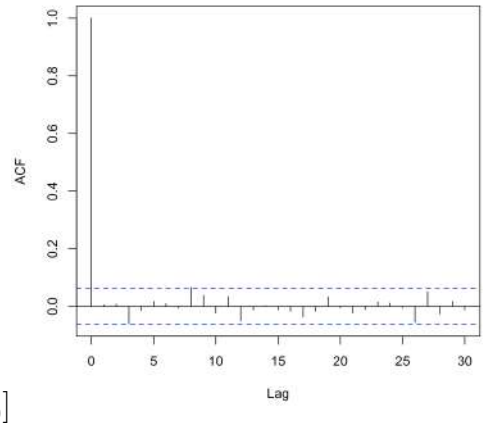
$[\nu_7]$



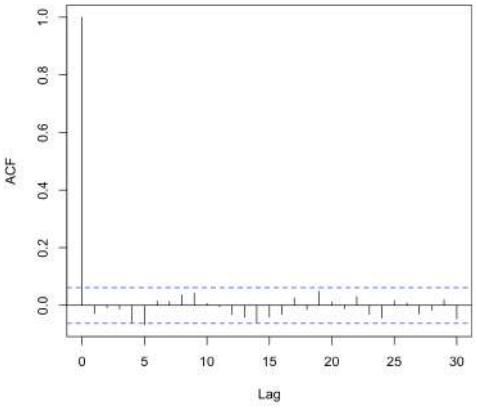
$[\nu_8]$



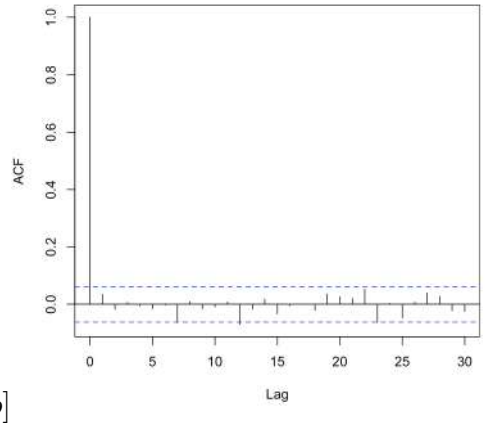
$[\nu_9]$



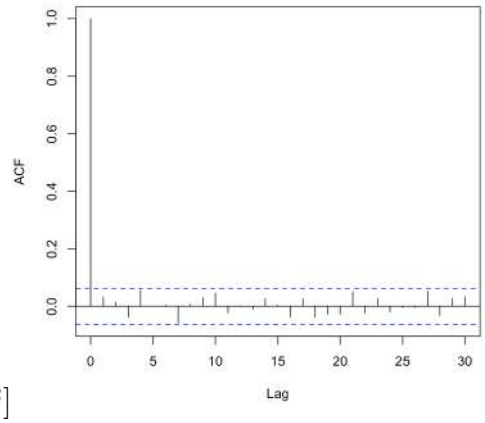
$[\nu_{10}]$



$[\theta]$



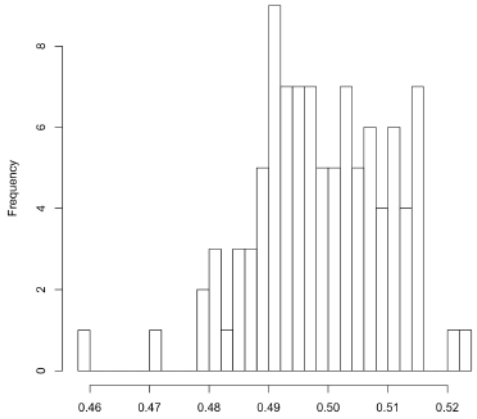
$[\rho]$



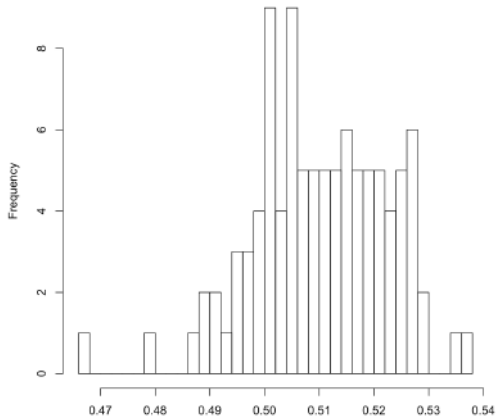
$[\delta^2]$

Section 4.3

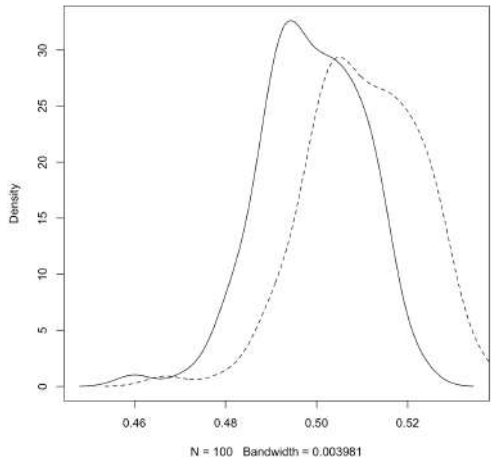
The following figures depict the histogram and kernel density plots of the proportions of positive responses predicted for i) all individuals $P_{all}^{(h)}$ and ii) non-sampled individuals ($P_{ns}^{(h)}$) based on the Spatial model. In the kernel density plot, the bold (dashed) curve corresponds to $P_{all}^{(h)}(P_{ns}^{(h)})$.



(a) All individuals



(b) Non-sampled individuals



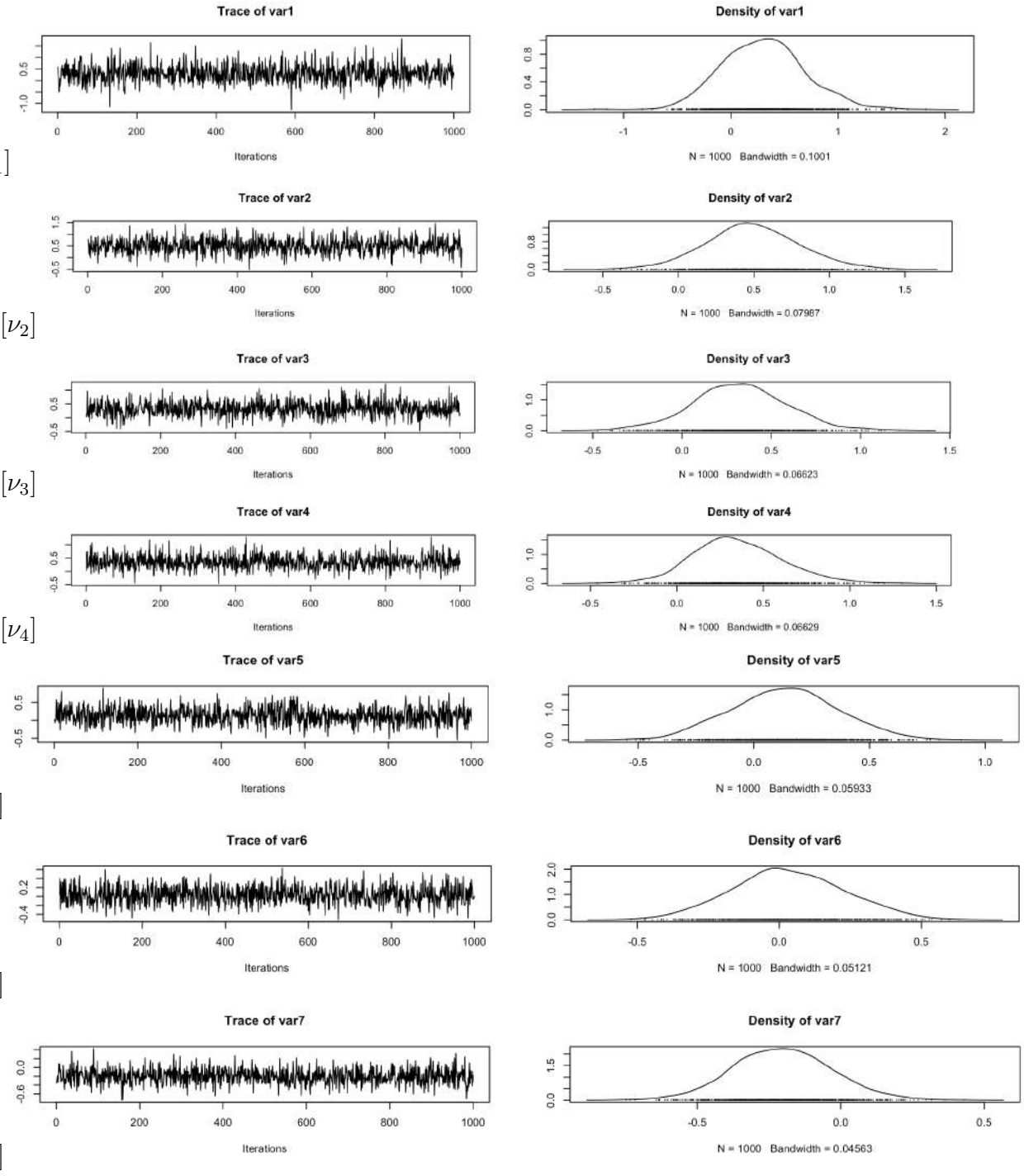
(c) Combined (bold: all subjects; dashed: non-sampled subjects)

Section 5.2

The following table shows the p-values corresponding to the Gweke test and the effective sample sizes for $(\boldsymbol{\nu}, \theta, \rho, \delta^2)$ corresponding to the modified spatial model. All but one of the effective sample sizes are equal to the size of chain *i.e.*. $1,000$, which is desirable.

Parameter	P-value	Effective sample size
ν_1	0.10080200	1000
ν_2	0.95000090	1000
ν_3	0.45674993	1000
ν_4	0.28452094	1000
ν_5	0.91671578	1000
ν_6	0.22139337	1000
ν_7	0.47038949	1000
ν_8	0.06734535	1000
ν_9	0.11521862	1000
ν_{10}	0.34214527	1000
θ	0.38160805	1000
ρ	0.74996683	905
δ^2	0.94700833	1000

The following figures depict the trace and kernel density plots for the simulated values of $(\boldsymbol{\nu}, \theta, \rho, \delta^2)$ obtained from the Markov Chain Monte Carlo run of the modified Spatial model.



$[\nu_1]$

$[\nu_2]$

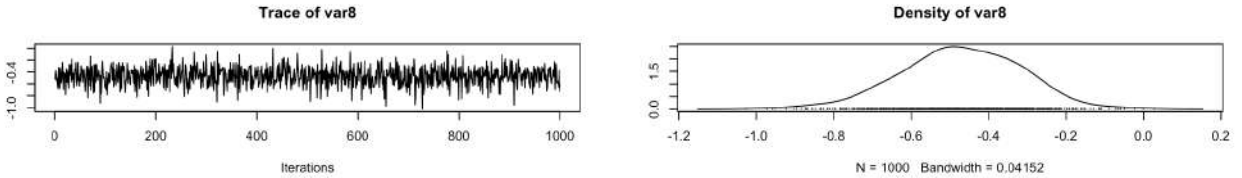
$[\nu_3]$

$[\nu_4]$

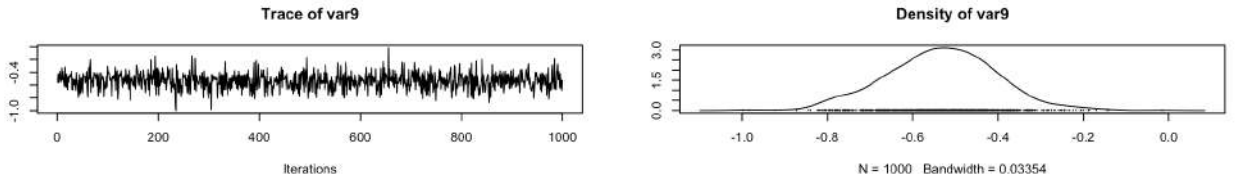
$[\nu_5]$

$[\nu_6]$

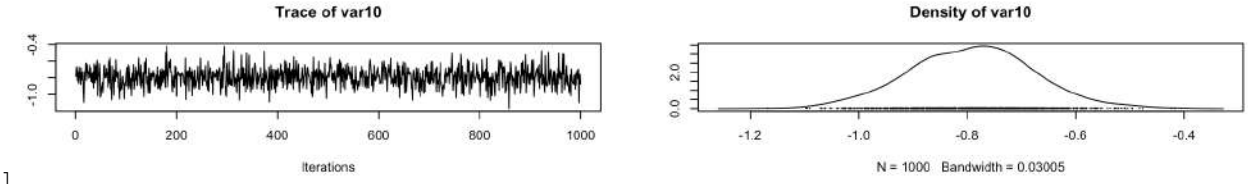
$[\nu_7]$



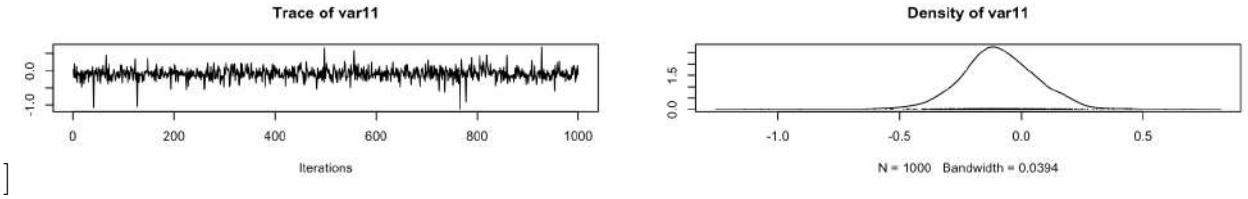
$[\nu_8]$



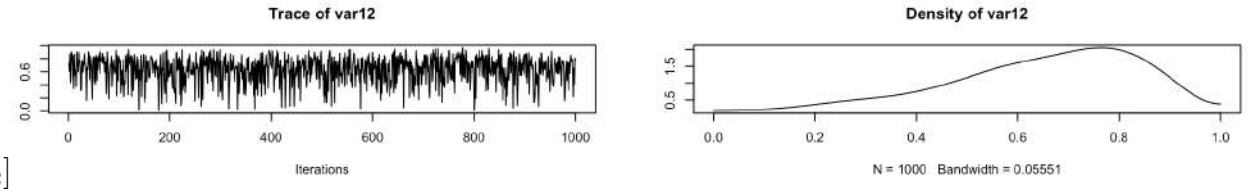
$[\nu_9]$



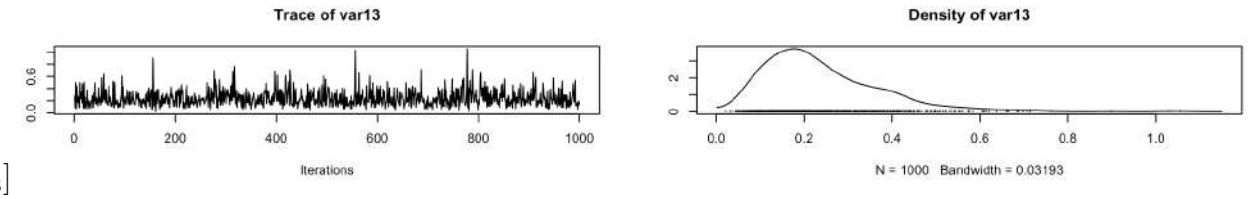
$[\nu_{10}]$



$[\nu_{11}]$



$[\nu_{12}]$



$[\nu_{13}]$

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 401–428
http://www.ssca.org.in/journal

Three Score and 15 Years (1948-2023) of Rao's Score Test: A Brief History

Anil K. Bera[1] **and Yannis Bilias**[2]
[1] *University of Illinois at Urbana-Champaign, U.S.A.*
[2] *Athens University of Economics and Business, Greece*

Received: 15 June 2024; Revised: 03 September 2024; Accepted: 16 September 2024

आचख्युः कवयः केचित्संप्रत्याचक्षते परे ।
आख्यास्यन्ति तथैवान्ये इतिहासमिमं भुवि ॥
ācakhyuḥ kavayaḥ kecitsaṃpratyācakṣate pare |
ākhyāsyanti tathaivānye itihāsamimaṃ bhuvi ||

Mahabharata (∼500BC) by Krishna-Dwaipayana Vyasa, Verse 1.1.24

"Some bards have already published this history, some are now teaching it, and others, in like manner, will hereafter promulgate it upon the earth."

Abstract

Rao (1948) introduced the score test statistic as an alternative to the likelihood ratio and Wald test statistics. In spite of the optimality properties of the score statistic shown in Rao and Poti (1946), the Rao score (RS) test remained unnoticed for almost 20 years. Today, the RS test is part of the "Holy Trinity" of hypothesis testing and has found its place in Statistics and Econometrics textbooks and related software. Reviewing the history of the RS test we note that remarkable test statistics proposed in the literature earlier or around the time of Rao (1948) mostly from intuition, such as Pearson (1900) goodness-fit-test, Moran (1948) **I** test for spatial dependence and Durbin and Watson (1950) test for serial correlation, can be given RS test statistic interpretation. At the same time, recent developments in the robust hypothesis testing under certain forms of misspecification, make the RS test an active area of research in Statistics and Econometrics. From our brief account of the history of the RS test we conclude that its impact in science goes far beyond its calendar starting point with promising future research activities for many years to come.

Key words: Applications to Econometrics and Statistics; Hypothesis testing; Rao's score; Robust tests; Sequential testing.

Corresponding Author: Anil K. Bera
Email: abera@illinois.edu

1. Prologue

C. R. Rao's work was always inspired by some practical problems. In 1946, he was deputed from the Indian Statistical Institute (ISI), Calcutta, to work on an anthropometric project in the Museum of Anthropology and Ethnology at the Cambridge University, U.K. While at Cambridge, Rao took the opportunity to contact R. A. Fisher, then the Belfour Professor of Genetics, and registered for a Ph.D. degree in Statistics under Fisher's guidance. As recollected in Rao (2001), Fisher agreed under the condition that Rao spends time in the Genetics Laboratory where Fisher was breeding mice to map their chromosomes. Rao started by mating mice of different genotypes to collect the necessary data and additionally, he was trying to develop appropriate statistical methodology to analyze the experimental data. The problem was estimation of linkage *parameters* (recombination probabilities in the various segments of the chromosomes) using data sets from different experiments, designed in such a way that each data set had information on the *same parameters*. It was thus necessary to test whether the parameters in different experiments are the same or not.

Rao wrote and published two papers based on this work. The *first* paper, Rao (1948), deals with the *general problem of testing* simple and composite hypotheses concerning a vector parameter. The test was based on the *scores*, derivatives of the log-likelihood function with respect to the individual parameters. The paper was published in the *Proceedings of the Cambridge Philosophical Society*, where he termed the test principle as a *score test*. In this paper, we will refer to it as the Rao score (RS) test. The *other* paper, Rao (1950a), contains the detail steps for *analyzing the data* involving the segregation of several factors in mating of different genotypes. And it used the RS test for the meta-analysis of testing the equality of parameters coming from different experimental data sets. That paper was published in Fisher's new journal *Heredity*. For more, see Rao (2001).

The rest of the paper is organized as follows. In Section 2, we start with the first principle of testing, namely the Neyman-Pearson Lemma and derive the simplest version of RS test and then discuss it in its full generality. There, we also provide RS test interpretation to some of the classic tests in Econometrics and Statistics, such as the quintessential Pearson (1900) goodness-fit-test, which was suggested mostly by pure intuition, but its theoretical foundation can be buttressed by RS test principle. In Sections 3 and 4, we list a (somewhat incomplete) catalogue of RS tests in Econometrics and Statistics. In Section 5, we outline some of the possible ways an assumed probability model can be misspecified, and discuss how the various RS tests can be robustified to make them valid under misspecification. We close the paper in Section 6 (Epilogue) with some concluding remarks. At the outset let us mention that while compiling the 75 years (from 1948 to 2023) history of the RS test, we have included here some of our own past historical accounts and cited accordingly. Our aim is to have a comprehensive review as far as possible at one place, like a one-stop-shopping for the RS test history.

2. Score as an optimal test function: Rao and Poti (1946)

We start by introducing some notation and concepts. Suppose we have n independent observations $y_1, y_2, ..., y_n$ on a random variable Y with density function $f(y;\theta)$, where θ is a $p \times 1$ parameter vector with $\theta \in \Theta \subseteq \mathbb{R}^p$. It is assumed that $f(y;\theta)$ satisfies the regularity conditions stated in Rao (1973, p.364) and Serfling (1980, p.144). The likelihood function is

given by

$$\mathbf{L}(\theta, \mathbf{y}) \equiv \mathbf{L}(\theta) = \prod_{i=1}^{n} f(y_i; \theta), \tag{1}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ denotes the sample. Suppose we want to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ based on the sample $\mathbf{y}$.

The foundation of the theory of hypothesis testing was laid by Neyman and Pearson (1933) fundamental lemma. This lemma provides a way to find the most powerful (MP) and uniformly most powerful (UMP) tests. According to the Neyman-Pearson (N-P) Lemma, the MP critical region for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ having size α, is given by

$$\omega(\mathbf{y}) = \{\mathbf{y} \mid \mathbf{L}(\theta_1) > \kappa \mathbf{L}(\theta_0)\}, \tag{2}$$

where κ is such that $\Pr[\omega(\mathbf{y})|H_0] = \alpha$.

If an MP test maximizes powers uniformly in $\theta_1 \in \Theta_1 \subseteq \Theta$, the test is called UMP test. Unfortunately, an UMP test rarely exists, and when it does not, there is no single critical region best for all alternatives. We, therefore, try to find a critical region that is good for alternatives *close* to the null hypothesis, called *local* alternatives, hoping that the region will also be good for alternatives away from the null. Lehmann (1999, p.529) advocated for such critical region when the sample size n is large, stating, "if the true value is at some distance from θ_0, a large sample will typically reveal this so strikingly that a formal test may be deemed unnecessary."

For a critical region $\omega(\mathbf{y})$, let us define the power function as

$$\gamma(\theta) = \Pr[\omega(\mathbf{y})|\theta] = \int_{\omega(\mathbf{y})} \mathbf{L}(\theta) d\mathbf{y}. \tag{3}$$

Assuming a scalar θ and that $\gamma(\theta)$ admits Taylor series expansion, we have

$$\gamma(\theta) = \gamma(\theta_0) + (\theta - \theta_0)\gamma'(\theta_0) + \frac{(\theta - \theta_0)^2}{2}\gamma''(\theta^*), \tag{4}$$

where θ^* is a value in between θ and θ_0. If we consider local alternatives of the form $\theta = \theta_0 + n^{-\frac{1}{2}}\delta, 0 < \delta < \infty$, the third term will be of the order $\mathbf{O}(n^{-1})$. Therefore, from (4), to obtain the highest power, we need to maximize

$$\gamma'(\theta_0) = \frac{\partial}{\partial\theta}\gamma(\theta)\bigg|_{\theta=\theta_0} = \frac{\partial}{\partial\theta}\int_{\omega(\mathbf{y})} \mathbf{L}(\theta_0) d\mathbf{y} = \int_{\omega(\mathbf{y})} \frac{\partial}{\partial\theta}\mathbf{L}(\theta_0) d\mathbf{y}, \tag{5}$$

for $\theta > \theta_0$, assuming the regularity conditions that allow differentiation under the sign of intergration.

Using the generalized N-P Lemma given in Neyman and Pearson (1936), it is easy to see that the locally most powerful (LMP) critical region for $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$, is given by

$$\omega(\mathbf{y}) = \left\{\mathbf{y} \;\middle|\; \frac{\partial}{\partial\theta}\mathbf{L}(\theta_0) > \kappa \mathbf{L}(\theta_0)\right\}$$

or

$$\omega(\mathbf{y}) = \left\{ \mathbf{y} \;\middle|\; \frac{\partial}{\partial \theta} \ln(\mathbf{L}(\theta_0)) = \frac{\partial}{\partial \theta} l(\theta_0) > \kappa \right\}, \tag{6}$$

where $l(\theta)$ denotes the log-likelihood function and κ is a constant such that the size of the test is α. The quantity $S(\theta) = \partial l(\theta)/\partial \theta$ is called the Fisher-Rao *score function*. The above result in (6) was first discussed in Rao and Poti (1946), who stated that a LMP critical region for $H_0 : \theta = \theta_0$ is given by

$$\omega(\mathbf{y}) = \{\mathbf{y} \mid \kappa_1 S(\theta_0) > \kappa_2\}, \tag{7}$$

where κ_2 is so determined that the size of the test is equal to a preassigned value α with κ_1 as $+1$ or -1, respectively, for alternatives $\theta > \theta_0$ and $\theta < \theta_0$.

Let us define the Fisher information as

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2 l(\theta)}{\partial \theta^2}\right] = Var[S(\theta)]. \tag{8}$$

The result that under H_0, $S(\theta_0)$ is asymptotically distributed as normal with mean zero and variance $\mathcal{I}(\theta_0)$, led Rao and Poti (1946) to suggest a test based on $S(\theta_0)/\sqrt{\mathcal{I}(\theta_0)}$ as standard normal [or $S^2(\theta_0)/\mathcal{I}(\theta_0)$ as χ^2_1], for large n.

2.1. From Rao and Poti (1946) to Rao (1948): Test for the multiparameter case

Rao and Poti (1946) can be viewed as a precursor to Rao (1948). Generalization of the LMP test in (7) to the multiparameter case ($p \geq 2$) is not trivial. There will be scores for each individual paramter, and the problem is to combine them in an "optimal" way. Let $H_0 : \theta = \theta_0$, where now $\theta = (\theta_1, \theta_2, ..., \theta_p)'$ and $\theta_0 = (\theta_{10}, \theta_{20}, ..., \theta_{p0})'$. Consider a *scalar* linear combination

$$\sum_{j=1}^{p} \delta_j \frac{\partial l(\theta)}{\partial \theta_j} = \delta' S(\theta), \tag{9}$$

where $\delta = (\delta_1, \delta_2, ..., \delta_p)'$ is a fixed vector and test the hypothesis $H_{0\delta} : \delta'\theta = \delta'\theta_0$ against $H_{1\delta} : \delta'\theta \neq \delta'\theta_0$, $\delta \in \mathbb{R}^p$.

We rewrite the Fisher information in (8) as

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta'}\right]. \tag{10}$$

Asymptotically, under H_0, $\delta' S(\theta_0)$ is distributed as normal, with mean zero and variance $\delta' \mathcal{I}(\theta_0)\delta$. Thus if δ's were known, a test could be based on

$$\frac{[\delta' S(\theta_0)]^2}{\delta' \mathcal{I}(\theta_0)\delta}, \tag{11}$$

which under H_0 will be distributed as χ^2_1 as in Rao and Poti (1946). Note that our $H_0 : \theta = \theta_0$ for $p \geq 2$ can be expressed as $H_0 \equiv \bigcap_{\delta \in \mathbb{R}^p} H_{0\delta}$, i.e., the multiparameter testing problem can be

decomposed into a series of *single-parameter* problems. To obtain a linear function like (9), Rao (1948) maximized (11) with respect to δ. Using the Cauchy-Schwarz inequality

$$\frac{(u'v)^2}{u'Au} \leq v'A^{-1}v, \tag{12}$$

where u and v are column vectors and A is a non-singular matrix, we have

$$\sup_{\delta \in \mathbb{R}^p} \frac{[\delta' S(\theta_0)]^2}{\delta' \mathcal{I}(\theta_0)\delta} = S(\theta_0)' \mathcal{I}(\theta_0)^{-1} S(\theta_0). \tag{13}$$

In (13), the supremum reaches at $\delta = \mathcal{I}(\theta_0)^{-1} S(\theta_0)$ and this provides an optimal linear combination of scores.

Roy (1953) used Rao's maximization technique (13) to develop his *union-intersection* (UI) method of testing. Let $H_0 \equiv \bigcap_{j \in J} H_{0j}$, where J is an index set. Roy's UI method gives the rejection region for H_0 as the *union* of rejection regions for all H_{0j}, $j \in J$. Consider testing $H_{0\delta} : \delta'\theta = \delta'\theta_0$ against $H_{1\delta} : \delta'\theta \neq \delta'\theta_0$, $\delta \in \mathbb{R}^p$. Let $H_0 = \bigcap_{\delta \in \mathbb{R}^p} H_{0\delta}$ and $H_1 \equiv \bigcap_{\delta \in \mathbb{R}^p} H_{1\delta}$. If T_δ is the likelihood ratio (LR) statistic for testing $H_{0\delta}$ against $H_{1\delta}$, then

$$T = \sup_{\delta \in \mathbb{R}^p} T_\delta \tag{14}$$

is Roy's LR statistic for testing H_0 against H_1. This is the same principle that was used by Rao (1948) to convert a "multivariate" problem into a series of "univariate" ones, as we have seen in equation (13).

When the null hypothesis is composite, like $H_0 : h(\theta) = c$, where $h(\theta)$ is an $r \times 1$ vector function of θ with $r \leq p$, the general form of the RS test statistic is

$$RS = S(\widetilde{\theta})' \mathcal{I}(\widetilde{\theta})^{-1} S(\widetilde{\theta}), \tag{15}$$

where $\widetilde{\theta}$ is the restricted maximum likelihood estimator (MLE) of θ, i.e., $h(\tilde{\theta}) = c$. Asymptotically, under H_0, the RS test statistic is distributed as χ^2_r. Therefore, we observe *two* optimality principles behind the RS test; first, in terms of LMP test as given in (6), and second, in deriving the "optimal" direction for the multiparameter case, as in (13).

Rao (1948) suggested the score test as an alternative to the Wald (1943) statistic, which for testing $H_0 : h(\theta) = c$ is given by

$$W = \left[h(\hat{\theta}) - c\right]' \left[H(\hat{\theta})' \mathcal{I}(\hat{\theta})^{-1} H(\hat{\theta})\right]^{-1} \left[h(\hat{\theta}) - c\right], \tag{16}$$

where $\hat{\theta}$ is the unrestricted MLE of θ, and $H(\theta) = \partial h(\theta)/\partial\theta$ is a $r \times p$ matrix with full column rank r. Rao (1948, p.53) stated that his test (15), "besides being simpler than Wald's has some theoretical advantages." For more on this see Bera (2000) and Bera and Bilias (2001).

Neyman and Pearson (1928) suggested their LR test as

$$LR = 2\left[ln\frac{\mathbf{L}(\hat{\theta})}{\mathbf{L}(\widetilde{\theta})}\right] = 2\left[l(\hat{\theta}) - l(\widetilde{\theta})\right]. \tag{17}$$

Their suggestion did not come from any search procedure satisfying an optimality criterion. It was purely based on intuitive grounds; as Neyman (1980, p.6) stated, "The intuitive background of the likelihood ratio test was simply as follows: if among the contemplated admissible hypotheses there are some that ascribe to the facts observed probabilities much larger than that ascribed by the hypothesis tested, then it appears 'reasonable' to reject the null hypothesis."

The three statistics RS, W, and LR, given respectively in (15), (16), and (17) are referred to as the "Holy Trinity." These tests can be viewed as three different *distance measures* between H_0 and H_1. When H_0 is true, we should expect the restricted and unrestricted MLEs of θ, namely $\widetilde{\theta}$ and $\hat{\theta}$ to be close, and likewise the log-likelihood functions $l(\widetilde{\theta})$ and $l(\hat{\theta})$, respectively. The LR statistic in (17) measures the distance through the log-likelihood function and is based on the difference $l(\hat{\theta}) - l(\widetilde{\theta})$. To see the intuition behind the RS test, note that $S(\hat{\theta}) = 0$ by construction, and thus we should expect $S(\widetilde{\theta})$ to be cloes to zero if H_0 is true. Therefore, the basis of the RS test is $S(\widetilde{\theta}) - S(\hat{\theta}) = S(\widetilde{\theta})$, distance between $\widetilde{\theta}$ and $\hat{\theta}$ measured through the function $S(\theta)$. Finally to test $H_0 : h(\theta) = c$, W considers the distance directly in terms of $h(\theta)$, and is based on $\left[h(\hat{\theta}) - c\right] - \left[h(\widetilde{\theta}) - c\right] = h(\hat{\theta}) - c$, where $h(\widetilde{\theta}) = c$ by construction, as we see in expression (16). It is interesting to note the similarity between the Wald and the RS tests based on $h(\hat{\theta})$ and $S(\widetilde{\theta})$, respectively. Therefore the RS test statistic is closer to W than LR. Therefore, it makes sense that Rao (1948, p.53) mentioned his test as an alternative to W.

The interrelationships among these three tests can be brought home to the students of Statistics through the following amusing story [see Bera and Premaratne (2001, p.58)]: Once around 1946 Ronald Fisher invited Jerzy Neyman, Abraham Wald, and C.R. Rao to his Cambridge University lodge for afternoon tea. During their conversation, Fisher mentioned the problem of deciding whether his dog, who had been going to an "obedience school" for some time, was disciplined enough. Neyman quickly came up with an idea: leave the dog free for some time and then put him on leash. If there is not much difference in his behavior, the dog can be thought of as having completed the course successfully. Wald, who lost his family in the concentration camps, was adverse to any kind of restrictions and simply suggested leaving the dog free and seeing whether it behaved properly. Rao, who had observed the nuisances of stray dogs in Calcutta streets, did not like the idea of letting the dog roam freely, and suggested keeping the dog on a leash at all times and observing how hard it pulls on the leash. If it pulled too much, it needed more training. That night when Rao was back in his Cambridge dormitory after tending Fisher's mice at the genetics laboratory, he suddenly realized the connection of Neyman and Wald's recommendations to the Neyman–Pearson LR and Wald tests, respectively. He got an idea and the rest, as they say, is history.

At this stage, it will be instructive to provide a geometric illustration highlighting the fundamental connections and contrasts among the three tests [see Bera (1983, pp.56-60)]. For simplicity, let us consider the case of scalar θ, i.e., $p = 1$, and that the null hypothesis is $H_0 : \theta = \theta_0$. In Figure 1, we plot the score function $S(\theta) = dl(\theta)/d\theta$ against θ, the solid curved line. The unrestricted MLE $\hat{\theta}$ is obtained by setting $S(\hat{\theta}) = 0$, i.e., at the point D.

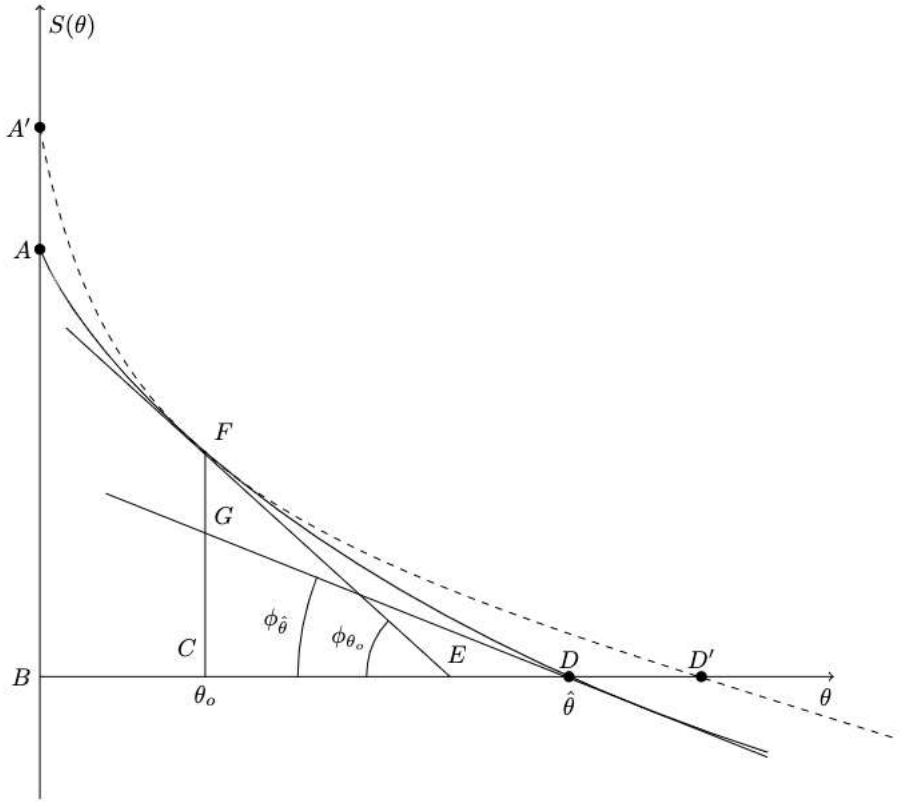


Figure 1: Geometry of LR, W and RS statistics

From the Figure 1, it is easily seen that

$$l(\hat{\theta}) - l(\theta_0) = \int_{\theta_0}^{\hat{\theta}} S(\theta)d(\theta) =$$
$$= \text{Area under the curve } S(\theta) \text{ from } \theta_0(\text{point C}) \text{ to } \hat{\theta}(\text{point D}). \tag{18}$$

Therefore,

$$LR = 2[l(\hat{\theta}) - l(\theta_0)] = 2 \cdot Area(CDF). \tag{19}$$

For our particular case, $h(\theta) = \theta - \theta_0, H(\theta) = 1$ and $c = 0$. Thus, W in (16) can be expressed as

$$W = (\hat{\theta} - \theta_0)^2 \mathcal{I}(\hat{\theta}) = CD^2 \cdot \mathcal{I}(\hat{\theta}). \tag{20}$$

$\mathcal{I}(\hat{\theta})$ can be obtained from $-d^2l(\theta)/d\theta^2 = -dS(\theta)/d\theta$, evaluated at $\theta = \hat{\theta}$, i.e., from $\tan\phi_{\hat{\theta}} = CG/CD$. Therefore,

$$W = CD^2 \cdot \frac{CG}{CD} = CD \cdot CG = 2 \cdot Area(\triangle CDG). \tag{21}$$

On the other hand, the RS test will be based on $S(\theta)$ at θ_0, i.e., on the distance CF. The variance of $S(\theta_0)$ can be estimated by $-dS(\theta_0)/d\theta = \tan\phi_{\theta_0} = CF/CE$. Hence,

$$RS = CF^2. \quad \frac{CE}{CF} = CF \cdot CE = 2 \cdot Area(\triangle CEF). \tag{22}$$

From above, we can note the following features of the three tests. *First*, since the tests are based on three different areas in general, they will yield conflicting inference if the same critical value is used [see Berndt and Savin (1977)]. *Second*, the RS tests depends only on $S(\theta)$ and the slope of $S(\theta)$ at θ_0. We can draw many curved lines through F with the same slope at F, and the dotted line $A'FD'$ is an example. This implies that there may be other $S(\theta)$ functions, i.e., other likelihood function representing different alternative hypothesis, with the same slope at θ_0, giving rise to the same RS test statistic. In the literature, this property is known as invariance property of the RS test principle [see Godfrey (1988, p.70)]. *Finally*, for both the RS and W tests, the variances can be calculated in a number of ways which are asymptotically equivalent. This can lead to different versions of the test statistics. It is not clear which versions will give better results in finite samples.

Example 1: Let us start with a simple example where $y_i \sim IIDN(\theta, 1),\ i = 1, 2, \ldots, n$, and we test $H_0 : \theta = \theta_0 = 0$ against $\theta > 0$. Here the log-likelihood and score functions are respectively

$$\begin{aligned} l(\theta) &= \text{Constant} - \frac{1}{2}\sum_{i=1}^{n}(y_i - \theta)^2, \\ \text{and} \quad S(\theta) &= \sum_{i=1}^{n}(y_i - \theta) = n(\bar{y} - \theta), \end{aligned} \tag{23}$$

where $\bar{y} = \sum_i y_i/n$. Note that here $S(\theta)$ is *linear* in θ, and thus from Figure 1, all the three tests LR, W, and RS will be identical. Given that $S(\theta_0) = n\bar{y}$ with $Var[S(\theta_0)] = n$, we will reject H_0, if $\sqrt{n}\bar{y} > \mathbf{z}_\alpha$, where $\mathbf{z}_\alpha$ is the upper α percent cut-off point of standard normal distribution. For fixed n, the power of this test goes to 1 as $\theta \to \infty$. Hence the score test $\sqrt{n}\bar{y} > \mathbf{z}_\alpha$ is not only LMP, but also UMP for all $\theta > 0$.

Example 2: [Ferguson (1967, p.235)] Consider testing for the median of a Cauchy distribution with density

$$f(y;\theta) = \frac{1}{\pi} \cdot \frac{1}{1 + (y - \theta)^2}, \quad -\infty < y < \infty. \tag{24}$$

Since here $\mathcal{I}(\theta) = n/2$, the RS test will reject $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$, if

$$\frac{S(\theta_0)}{\sqrt{\mathcal{I}(\theta_0)}} = \sqrt{\frac{2}{n}}\sum_{i=1}^{n}\frac{2(y_i - \theta_0)}{1 + (y_i - \theta_0)^2} > \mathbf{z}_\alpha. \tag{25}$$

As $\theta \to \infty$ with n remaining fixed, $\min(y_i - \theta_0) \xrightarrow{p} \infty$, and $S(\theta_0)/\sqrt{\mathcal{I}(\theta_0)} \xrightarrow{p} 0$. Thus the power of the test tends to zero as $\theta \to \infty$. Therefore what works for *local* alternatives may not work for *not-so-local* alternatives. This is in contrast to Example 1 where the LMP test is also the UMP.

In the example below we illustrate one of the most famous tests in the Statistics literature that was suggested long before 1948 and the theoretical foundation of which can be buttressed by the RS test principle.

Example 3: [Pearson (1900) Goodness-of-fit test]. Consider a multinomial distribution with p classes and let the probability of an observation belonging to the j-th class be $\theta_j (\geq 0)$,

$j = 1, 2, \ldots, p$, so that $\sum_{i=1}^{p} \theta_j = 1$. Denote the observed frequency of the j-th class by n_j with $\sum_{j=1}^{p} n_j = n$. We are interested in testing $\theta_j = \theta_{j0}$, $j = 1, 2, \ldots, p$, where θ_{j0} are known constants. Pearson (1900) suggested the statistic

$$P = \sum_{j=1}^{p} \frac{(n_j - n\theta_{j0})^2}{n\theta_{j0}} = \sum \frac{(O-E)^2}{E}, \tag{26}$$

where O and E denote respectively, the observed and expected frequencies. Given the profound importance of P in almost all branches of science, we demonstrate the theoretical underpinnings of P based on the RS test principle. The log-likelihood function, score and information matrix are respectively, given by [see Bera and Bilias (2001, p.17)]

$$l(\theta) = Constant + \sum_{j=1}^{p} n_j \ln(\theta_j) \tag{27}$$

$$\underset{[(p-1)\times 1]}{S(\theta)} = \begin{bmatrix} \frac{n_1}{\theta_1} - \frac{n_p}{\theta_p} \\ \frac{n_2}{\theta_2} - \frac{n_p}{\theta_p} \\ \cdots \\ \frac{n_{p-1}}{\theta_{p-1}} - \frac{n_p}{\theta_p} \end{bmatrix} \tag{28}$$

and

$$\underset{[(p-1)\times(p-1)]}{\mathcal{I}(\theta)} = n\left[diag\left(\frac{1}{\theta_1}, \frac{1}{\theta_2}, \ldots, \frac{1}{\theta_{p-1}}\right) + \frac{1}{\theta_p}\mathbf{1}\mathbf{1}'\right] \tag{29}$$

where $\mathbf{1} = (1, 1, \ldots, 1)'$ is a $(p-1) \times 1$ vector of ones. We end up with effectively $(p-1)$ parameters since $\sum_{j=1}^{p} \theta_j = \sum_{j=1}^{p} \theta_{j0} = 1$. Using the above expressions, it is easy to see that

$$S(\theta_0)'\mathcal{I}(\theta_0)^{-1}S(\theta_0) = P, \tag{30}$$

where $\theta_0 = (\theta_{10}, \theta_{20}, ..., \theta_{p0})'$ [see Rao (1973, p.442) and (Cox and Hinkley, 1974, p.316)]. The coincidence that P is same as the RS test, is an amazing result. Pearson (1900) suggested his test mostly based on intuitive grounds almost 50 years before Rao (1948).

3. Some applications of the RS test in econometrics

RS test was well ahead of its time. It went unnoticed for very many years. It is fair to say that econometricians can claim major credit in recognizing its importance and applying the RS test in several useful contexts and coming up with closed form, neat test statistics. Rao himself acknowledged this fact by writing [see Rao (2005, p.15)] "I am gratified to see the large number of papers contributed by econometricians on the application of the score statistic to problems in econometrics and the extensions and improvements they have made." More recently, statisticians are catching up with innovative applications. To obtain a quantitative perception of the influence of Rao (1948), we plot the yearly citations for the last 75 years in Figure 2. The corresponding cumulative citations are depicted in Figure 3. *First* thing to note is that the total number of citations in the last 75 years is only 980, apparently a very low number for such a seminal paper. Of course, we need to take into consideration of the fact that there are many papers, especially in the Statistics literature, that use the score test without making any reference to Rao (1948). *Second*, there are only a handful citations

during the first thirty years, i.e., until around 1978. That was the time econometricians recognized the usefulness of the Rao test principle, and used it in developing several model specification tests. There was another surge in its use after another 30 years, i.e., around 2008, in both the Statistics and Econometrics literature. *Finally*, from both Figures 2 and 3, it is clear that overall, the number of citations is still going up at an increasing rate, indicating continuing influence of Rao (1948), as far as the citation numbers go.

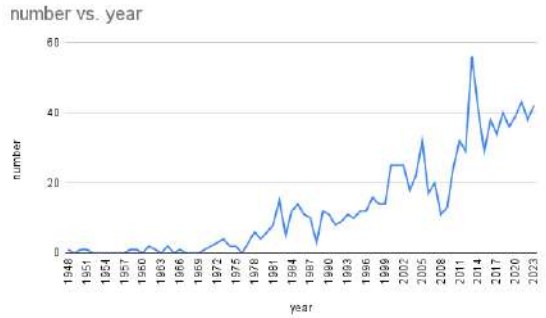


Figure 2: Yearly number of citations of Rao (1948): 1948-2023

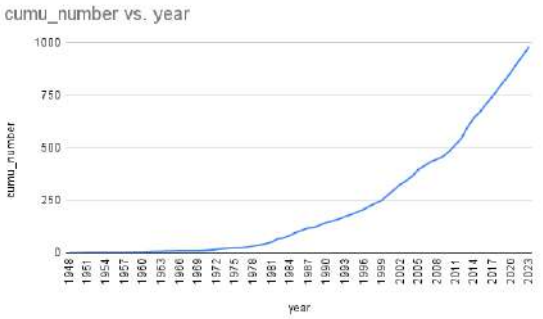


Figure 3: Cumulative number of citations

Byron (1968) was probably the first to apply the RS test in Econometrics. He used Silvey (1959) Lagrange multiplier (LM) version along with the LR statistic for testing homogeneity and symmetry restrictions in the demand system. In the Econometrics literature, the RS test is known as the LM test - the terminology came from Silvey (1959). Note that the restricted MLE $\widetilde{\theta}$ under the restriction $H_0 : h(\theta) = c$ can be obtained from the first order condition of the Lagrangian function

$$\mathcal{L} = l(\theta) - \lambda'[h(\theta) - c], \tag{31}$$

where λ is an $r \times 1$ vector of Lagrange multipliers. The first order conditions are

$$S(\widetilde{\theta}) - H(\widetilde{\theta})\widetilde{\lambda} = 0 \tag{32}$$

$$h(\widetilde{\theta}) = c, \tag{33}$$

where $H(\theta) = dh(\theta)/d\theta$. Therefore, from (32) we have $S(\widetilde{\theta}) = H(\widetilde{\theta})\widetilde{\lambda}$. Given that $H(\theta)$ has full rank, $S(\widetilde{\theta}) = 0$ is equivalent to $\widetilde{\lambda} = 0$. These multipliers can be interpreted as the implicit cost (shadow prices) of imposing the restrictions $h(\theta) = c$. It can be shown that

$$\widetilde{\lambda} = \frac{dl(\widetilde{\theta})}{dc}, \tag{34}$$

i.e., the multipliers give the rate of change of the maximum attainable value of the log-likelihood function with respect to the change in constraints. If $H_0 : h(\theta) = c$ is true and $l(\widetilde{\theta})$ gives the optimal value, $\widetilde{\lambda}$ should be close to zero. Given this "economic" interpretation in terms of Lagrange multipliers, it is not surprising that econometricians prefer the term LM rather than RS. In terms of Lagrange multipliers, (15) can be expressed as

$$RS = LM = \widetilde{\lambda}' H(\widetilde{\theta})' \mathcal{I}(\widetilde{\theta})^{-1} H(\widetilde{\theta})\widetilde{\lambda}. \tag{35}$$

After Byron (1968), it took another decade for econometricians to realize the potential of the RS test. The earlier notable contributions include Savin (1976), Berndt and Savin (1977), Breusch (1978, 1979) and Godfrey (1978a,b,c). Possibly Breusch and Pagan (1980) had been the most influential. They collected relevant research reported in the Statistics literature, presented the RS test in a general framework in the context of evaluating various econometric models, and discussed many applications. In a full length research monograph, Godfrey (1988) provided a comprehensive account of most of the available RS tests in Econometrics. Bera and Ullah (1991) and Bera and Bilias (2001) demonstrated that many of the commonly used specification tests could be given a score-test interpretation. For the last two-score years the RS tests had been the most common items in econometricians' kit for testing tools. It is not hard to understand the popularity of the score test principle in economics. In most cases, the algebraic forms of W and LR tests can hardly be simplified beyond their original formulae (16) and (17). On the other hand, in the majority of the cases the RS test statistics can explicitly be reduced to neat and elegant explicit formulae enabling its easy incorporation into computer software.

We will not make any attempt to provide a comprehensive list of applications of the RS test in Econometrics, for there are far too many. For instance, consider the workhorse of basic econometric modeling, the linear regression model:

$$y_i = x_i'\beta + \epsilon_i, \tag{36}$$

where y_i is the i-th observation on the dependent variable, x_i is the i-th observation on k exogenous variables and $\epsilon_i \sim IIDN(0, \sigma^2),\ i = 1, 2, \ldots, n$. The ordinary least squares (OLS) estimation and the related hypotheses tests are based on the four basic assumptions: correct linear functional form; the assumptions of disturbance normality; homoskedasticity; and serial independence. Just to name some of the uses of the RS test principle, test for normality was derived by Bera and Jarque (1981) and Jarque and Bera (1987); Breusch and Pagan (1979) proposed a test for homoskedasticity; and Godfrey (1978a,b) developed tests for serial independence which are very close to the earlier Durbin and Watson (1950) test.

To see the attractiveness of the RS test, let us briefly consider the popular Jarque and Bera (JB) test for normality. Bera and Jarque (1981) started with the Pearson (1895) family of distributions for the disturbance term ϵ_i in (36). That means if the pdf of ϵ_i is $f(\epsilon_i)$, we can write

$$\frac{d \log f(\epsilon_i)}{d\epsilon_i} = \frac{c_1 - \epsilon_i}{\sigma^2 - c_1\epsilon_i + c_2\epsilon_i^2}, \qquad i = 1, 2, \ldots, n, \tag{37}$$

where c_1 and c_2 are constants. The null hypothesis of normality can be stated as $H_0 : c_1 = c_2 = 0$ in (37). Given the complexity of ML estimation of σ^2, c_1, and c_2 in the Pearson family of distributions, W and LR tests are ruled out from a practical point of view. However, the score functions corresponding to c_1 and c_2 in (37), evaluated under the normality assumption, are given respectively by

$$S(\tilde{c}_1) = \frac{n\sqrt{b_1}}{3} \tag{38}$$

and

$$S(\tilde{c}_2) = \frac{n}{4}(b_2 - 3), \tag{39}$$

where $\sqrt{b_1} = m_3/m_2^{3/2}$ and $b_2 = m_4/m_2^2$ with $m_j = \frac{1}{n}\sum_{i=1}^n \tilde{\epsilon}_i^j$, $\tilde{\epsilon}_i = y_i - x_i'\tilde{\beta}$ as OLS residuals, $j = 2, 3, 4$. For large n, under normality

$$E\left[\sqrt{b_1}\right] = 0, \quad Var\left[\sqrt{nb_1}\right] = 6, \tag{40}$$

$$E\left[b_2\right] = 3, \quad Var\left[\sqrt{n}b_2\right] = 24, \tag{41}$$

and they are asymptotically normally distributed. Thus, a simple test statistic for normality is given by

$$JB = n\left[\frac{(\sqrt{b_1})^2}{6} + \frac{(b_2-3)^2}{24}\right], \tag{42}$$

which is asymptotically distributed as χ_2^2. It turns out that this test was mentioned by Bowman and Shenton (1975) but was hardly used in practice due to its lack of theoretical underpinnings. The RS test principle uncovered the theoretical justification of (42), ensuing the asymptotic optimality of the test. As it is obvious, JB is based on the two moments, third and fourth. One could have started with these two moments directly without going through the full derivations. From that point of view this RS test has a *moment* test interpretation.

It is quite common to express specification tests in Econometrics as *moment* tests. In a way "any" moment test can be obtained as a RS test under a suitably defined density function. To see this, let us write the r moment restrictions as

$$E_f[m(y;\theta)] = 0, \tag{43}$$

where $E_f[\cdot]$ means that (43) is true only when $f(y;\theta)$ is the correct pdf. A test for the hypothesis $H_0 : E_f[m(y;\theta)] = 0$ can be based on the estimate of the sample counterpart of $E_f[m(y;\theta)]$, namely,

$$\frac{1}{n}\sum_{i=1}^n m(y_i;\theta). \tag{44}$$

Now consider an auxiliary density function

$$f^*(y;\theta,\gamma) = f(y;\theta)\exp[\gamma' m(y;\theta) - \phi(\theta,\gamma)], \tag{45}$$

where $\phi(\theta,\gamma) = \ln\int \exp[\gamma' m(y;\theta)]f(y;\theta)dy$, with γ as $(r\times 1)$ parameter vector.

Note that if $f(y;\theta)$ is the correct pdf, then $\gamma = 0$ in (45). The log-likelihood function under the alternative hypothesis is

$$l^*(\theta,\gamma) = \sum_{i=1}^n \ln f^*(y_i;\theta,\gamma). \tag{46}$$

Therefore, the score function for testing $\gamma = 0$ in (45) is given by

$$\left.\frac{\partial l^*(\theta,\gamma)}{\partial\gamma}\right|_{\gamma=0} = \sum_{i=1}^n m(y_i,\theta), \tag{47}$$

and it provides the identical moment test as in (44). This interpretation of the moment test as a score test was first noted by White (1994). It is easy to see that there are many choices of auxiliary pdf $f^*(y;\theta,\gamma)$ and the score test will be invariant with respect to these choices, as depicted in Figuire 1. The LR and W tests, however, will be sensitive to the forms of $f^*(y;\theta,\gamma)$. This ends our coverage of the use of the RS test in Econometrics.

4. Some applications of the RS test in statistics

Rao (1950b) proposed a sequential test of null hypotheses based on the score statistic and his work on locally most powerful (LMP) tests in the case of one-sided alternative hypotheses. His proposal was a reaction to Wald (1945) sequential probability ratio test (SPRT) which was based on the idea of likelihood ratio test for the fixed sample case.

Wald's SPRT statistic was devised to discriminate between different alternative hypotheses for the value of the unknown parameter θ. On the other hand, Rao (1950b) seeks to test a null hypothesis $H_0 : \theta = \theta_0$ against a one-sided alternative hypothesis $H_1 : \theta > \theta_0$ with a test statistic that depends only on the null value.

For the fixed sample case, with sample of size N, the LMP test suggested by Rao and Poti (1946) is defined by [also see equation (6)]

$$P'_N(\theta_0) \geq \mu\, P_N(\theta_0), \tag{48}$$

where $P'_N(\theta_0)$ is the first derivative of $P_N(\theta)$ at $\theta = \theta_0$, with μ chosen so as to maintain Type-I error at a predetermined level. Motivated by this result, Rao (1950b) proposes a sequential test of the form

$$P'_n(\theta_0) \geq A(N)\, P_n(\theta_0), \tag{49}$$

with $n \leq N$, $A(N)$ a properly determined constant depending on the overall level of significance, and N being the upper limit to the number of observations. According to this sequential testing scheme, the sampling stops with rejection of the null hypothesis, at the smallest value of n for which the inequality (49) holds true. If by the Nth sampled unit (49) is not realized, the null is not rejected.

Berk (1953) proved that the sequential score tests against a one-sided alternative, where the stopping rule is the first time a certain random walk exceeds a bounded interval, are LMP tests *asymptotically*.

Sequential testing procedures that perform interim analyses during the evolution of the experiment, with the goal of obtaining the result earlier than the termination time suggested by the fixed sample analysis due to time or monetary cost considerations or ethical reasons, are easier validated with the use of score-based test statistics rather than the analogues of LR statistics. We may refer to chapters 9-11 of Sen (1981) for the role of score processes in sequential nonparametrics, where it is mentioned (p.339) "it is comparatively simpler to verify these regularity conditions [i.e., for the score] than those for the likelihood function."

Lombardi (1951), in a thesis on how to select a panel of judges for taste testing and quality evaluation using scientifically sound methods, appears to be one of the first applications of the sequential testing using Rao (1950b) methodology and its comparison with Wald's approach.

To diagnose the potential ability of candidate judges and to decide on the selection of a taste panel, each candidate judge is required to perform a prespecified number of sample comparisons. The number of sample comparisons that should be performed by each candidate judge before reaching to a decision on who to include in the taste panel is always a concern.

Bradley (1953), in conjunction with Lombardi, adapted Rao's method to binomial distribution, and is an early effort to communicate these statistical procedures for the selection of a taste panel to food technologists.

What makes the Rao procedure relatively more appropriate than that of Wald is that a limit to the testing of any one potential judge may be set. In this application, N denotes the maximum number of tests to be given to any judge. As it is noted by (Bradley, 1953, p.28): "The theory of the procedure needs further investigation since its properties are not well known. However, when it is applied to sequences of triangle tests, apparently satisfactory results are obtained."

In another context, time-series researchers use sequential analysis to determine and test for structural breaks. In a recent application, Bucci (2024) proposes a sequentially computed score statistic to test for the number of regimes in multivariate nonlinear models.

4.1. The role of the score statistic in survival analysis

Another context where the RS test statistic has found fruitful applications for inference is the analysis of survival data. The semiparametric proportional hazards model proposed by Cox (1972) is a standard tool of analysis for time-to-an-event data met in medical, engineering and economic applications. The parameter estimation using the partial likelihoood of Cox (1975) initiated an intense research activity for the validation of inference. The majority of the test statistics are special cases of weighted score statistics for different weighting functions and different type of covariates.

The partial likelihood score statistic has a natural martingale characterization. By rewriting the model within the counting process framework, Andersen and Gill (1982) were able to obtain a general asymptotic theory of the score statistic and the associated estimator. In a research related to sequentially computed score test statistic for repeated significance tests, Tsiatis (1981) established the joint asymptotic normality of efficient scores test for the proportional hazards model calculated over time. In a fundamental breakthrough, Sellke and Siegmund (1983) showed that the score process (over time) of the partial likelihood is approximated by a suitable martingale and thus behaves asymptotically like Brownian motion.

Bilias (2000) offered an application of a repeated significance test in a retrospective analysis of the Pennsylvania 'Reemployment Bonus' controlled experiments conducted by the US Department of Labor. Their main purpose was to determine whether the offer of a bonus amount to the unemployment insurance (UI) claimants, provided that they find a job with some required permanence within a given period of time, can act as an incentive for more intensive job-seeking with subsequent reduction of the unemployment spells. The response of primary interest is the *length of insured unemployment spell* and it is assumed that it follows a proportional hazard regression model. The statistic for measuring the effect of the various bonus packages on the duration of insured unemployment relative to the existing scheme is the partial likelihood score statistic. In carrying out sequential analysis, the score statistic is evaluated repeatedly, at different points in chronological time, each time with the available data. The retrospective sequential analysis concluded that the experiment could be concluded earlier than the fixed sample analysis with gains in time and monetary savings.

5. Robust RS tests under distributional and parametric misspecifications

As we have narrated in the previous sections the success of the RS test had been phenomenal. However the main problem in these specification tests is that they are developed under the assumption that the underlying probability model is *correctly specified*. When the assumed model is misspecified, it is well known that the RS test loses its local optimal properties.

While discussing the problem in statistical hypothesis testing, Haavelmo (1944, pp.65-66) stated, "Whatever be the principles by which we choose a "best" critical region of size α, the essential thing is that a test is always developed with respect to a *given fixed* set of possible *alternatives* Ω^0." Haavelmo called Ω^0, the *a priori admissible hypotheses* and according to him, a test is not robust if we shift our attention to another admissible set Ω' (that may be obtained by extending Ω^0 to include new/different alternatives), for which the proposed test has poor size and power properties.

Very often it is difficult to interpret the results of a test applied to a *misspecified* model. For instance, while testing the significance of some of the regression coefficients in the *linear* regression models, the results are not easily interpretable when a *nonlinear* model is the appropriate one [see, White (1980), Bera and Byron (1983) and Byron and Bera (1983)]. In the Statistics and Econometrics literature, most emphasis has been put on the minimization of type-I and type-II error probabilities. There are, however, only a few works that seriously consider the consequences and suggest remedies of misspecifying the a priori admissible hypothesis – which can be called the type-III error.

Note that the model under our a priori admissible hypothesis could be misspecified in a variety of ways. Here we consider only two kinds: distributional and parametric. In the former case, the assumed probability density function differs from the true data generating process (DGP). Kent (1982) and White (1982) analyzed this case and suggested a modified version of the RS test that involves adjustment of the variance of the score function. In the parametric misspecification case, the dimension of the assumed parameter space does not match with the true one. Bera and Yoon (1993) developed a modified RS test that is valid under the *local* parametric misspecification.

5.1. Robust RS test under distributional misspecification

Let the true DGP be described by the unknown density $g(y)$ and $f(y;\theta)$ be our assumed distribution. The RS test statistic given in (15) is not valid when $g(y)$ and $f(y;\theta)$ differ. This is because some of the standard results breakdown under distributional misspecification. For instance, consider the information matrix (IM) equality:

$$E_f\left[\frac{\partial \ln f(y;\theta)}{\partial \theta}\cdot\frac{\partial \ln f(y;\theta)}{\partial \theta'}\right]=E_f\left[-\frac{\partial^2 \ln f(y;\theta)}{\partial\theta\partial\theta'}\right], \tag{50}$$

where $E_f[\cdot]$ denotes expectation under $f(y;\theta)$. Let us now define

$$J(\theta_g)=nE_g\left[\frac{\partial \ln f(y;\theta)}{\partial \theta}\cdot\frac{\partial \ln f(y;\theta)}{\partial \theta'}\right] \tag{51}$$

$$K(\theta_g)=nE_g\left[-\frac{\partial^2 \ln f(y;\theta)}{\partial\theta\partial\theta'}\right], \tag{52}$$

where θ_g minimizes the Kullback-Leibler information criterion [see White (1982)]

$$\mathcal{I}_{KL} = E_g\left[\ln \frac{g(y)}{f(y;\theta)}\right]. \tag{53}$$

One can easily see that $J(\theta_g) \neq K(\theta_g)$, in general.

Example 4: Suppose we take $f(y;\theta) \equiv N(\mu, \sigma^2)$, and let the DGP $g(y)$ satisfy $E_g[y] = \mu$, $E_g[y-\mu]^2 = \sigma^2$, $E_g[y-\mu]^3 = \mu_3$ and $E_g[y-\mu]^4 = \mu_4$. Then it is easy to show that

$$J(\theta_g) = \begin{bmatrix} \frac{1}{\sigma^2} & \frac{\mu_3}{2\sigma^6} \\ \frac{\mu_3}{2\sigma^6} & \frac{\mu_4}{4\sigma^2} - \frac{1}{4\sigma^4} \end{bmatrix} \tag{54}$$

and

$$K(\theta_g) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^2} \end{bmatrix}. \tag{55}$$

Hence, $J(\theta_g) = K(\theta_g)$ if and only if $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$. We can clearly see the connection of these conditions and the JB test for normality given in (42).

Due to this divergence between J and K, and noting that we defined the information matrix $\mathcal{I}(\theta)$ in (10) by taking expectation under $f(y;\theta)$ instead of under the DGP $g(y)$, the standard RS test in (15) is not valid. Let us define an estimator of θ by maximizing a likelihood function based on the misspecified density $f(y;\theta)$ in place of the unknown DGP $g(y)$. Such an estimator is called quasi-MLE (QMLE). An early reference to QMLE can be found in Koopmans *et al.* (1950, p.135) [for more on this see Bera *et al.* (2020)]. We will denote QMLE of θ (under H_0) by $\tilde{\theta}$. Kent (1982) and White (1982) suggested the following robust form of the RS test statistic for testing the $H_0 : h(\theta) = c$:

$$RS^*(D) = S(\tilde{\theta})' K(\tilde{\theta})^{-1} H(\tilde{\theta}) [H(\tilde{\theta})' B(\tilde{\theta}) H(\tilde{\theta})] H(\tilde{\theta})' K(\tilde{\theta})^{-1} S(\tilde{\theta}), \tag{56}$$

where $H(\theta) = \partial h(\theta)/\partial\theta$, $B(\theta) = K(\theta)^{-1} J(\theta) K(\theta)^{-1}$ and the notation $RS^*(D)$ is used to signify *robust* RS test statistic under *distributional* misspecification.

Under $H_0 : h(\theta) = c$, $RS^*(D)$ is asymptotcally distributed as χ^2_r even under distributional misspecification, that is, when the assumed density $f(y;\theta)$ does not coincide with the true DGP $g(y)$. This approach of finding the asymptotically correct formula for variance has its origin in Koopmans *et al.* (1950, pp.148-150); for more on this see Bera *et al.* (2021). Expression (56) can be simplified if the parameter vector θ $(p \times 1)$ can be partitioned as $\theta = (\gamma', \psi')'$ where γ and ψ have dimensions m and r, respectively, $m + r = p$, and we test $H_0 : \psi = \psi_*$ (say). Let us also partition the score function $S(\theta)$ and $J(\theta)$ [similarly $K(\theta)$] as

$$S(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial l(\theta)}{\partial \gamma} \\ \frac{\partial l(\theta)}{\partial \psi} \end{bmatrix} = \begin{bmatrix} S_\gamma(\theta) \\ S_\psi(\theta) \end{bmatrix} \qquad \text{(say)} \tag{57}$$

and

$$J(\theta) = \begin{bmatrix} J_\gamma & J_{\gamma\psi} \\ J_{\psi\gamma} & J_\psi \end{bmatrix}. \tag{58}$$

While testing $H_0 : \psi = \psi_*$, under this setup $h(\theta) = \psi - \psi_*$ and $H(\theta) = [0_{r\times(p-r)}, I_{(r\times r)}]$, and we can express $RS^*(D)$ in (56) as [see also Bera *et al.* (2020)]

$$\begin{aligned} RS^*_\psi(D) &= S'_\psi(\widetilde{\theta})\Big[K_\psi(\widetilde{\theta}) + J_{\psi\gamma}(\widetilde{\theta})J^{-1}_\gamma(\widetilde{\theta})K_\gamma(\widetilde{\theta})J^{-1}_\gamma(\widetilde{\theta})J_{\gamma\psi}(\widetilde{\theta}) \\ &\quad -J_{\psi\gamma}(\widetilde{\theta})J^{-1}_\gamma(\widetilde{\theta})K_{\gamma\psi}(\widetilde{\theta}) - K_{\psi\gamma}(\widetilde{\theta})J^{-1}_\gamma(\widetilde{\theta})J_{\gamma\psi}(\widetilde{\theta})\Big]^{-1} S_\psi(\widetilde{\theta}), \end{aligned} \tag{59}$$

where $\tilde{\theta} = (\tilde{\gamma}', \psi'_*)'$, the restricted MLE under $H_0 : \psi = \psi_*$.

Example 5: In (37) the parameters c_1 and c_2 of the Pearson family of distributions can be treated, respectively, as the "skewness" and "kurtosis" parameters. Suppose we test the symmetry ignoring the (excess) kurtosis. Then we can start with the system (37) with $c_2 = 0$, that is,

$$\frac{d\log f(\epsilon_i)}{d\epsilon_i} = \frac{c_1 - \epsilon_i}{\sigma^2 - c_1\epsilon_i}. \tag{60}$$

After some derivation, it can be shown that the standard RS test for $c_1 = 0$ is given by

$$RS_{c_1} = n\frac{(\sqrt{b_1})^2}{6}, \tag{61}$$

which is essentially the first part of JB in (42). If $f(\epsilon_i)$ in (60) is not the true DGP, RS_{c_1} will not be valid; in particular, the asymptotic variance formula used in (61), $Var(\sqrt{n}b_1) = 6$ is incorrect [see also equation (40)]. For instance in the presence of excess kurtosis, there will be proportionately more outliers, resulting in higher variance, and thus "6" will underestimate the true variance of $\sqrt{n}b_1$. After incorporating the variance correction as in $RS^*_\psi(D)$ in (59) the robust RS test statistic can be written as [for further details see Premaratne and Bera (2017)]:

$$RS^*_{c_1}(D) = n\frac{(\sqrt{b_1})^2}{[9 + m_6 m_2^{-3} - 6m_4 m_2^{-2}]}, \tag{62}$$

where $m_j = \frac{1}{n}\sum_{i=1}^n \tilde{\epsilon}_i^j$, $j = 2, 4, 6$. From (62) we can write the population counterpart of the $Var(\sqrt{n}b_1)$ as

$$Var(\sqrt{n}b_1) = 9 + \mu_6\mu_2^{-3} - 6\mu_4\mu_2^{-2}, \tag{63}$$

where μ_j denotes the j-th population moment of ϵ. Therefore, the construction of the robust RS test statistic $RS^*_{c_1}(D)$ indicates that the true variance of $\sqrt{n}b_1$ that is valid under excess kurtosis is given by (63). If we impose normality, $\mu_6 = 15\sigma^6$ and $\mu_4 = 3\sigma^4$, then with $\mu_2 = \sigma^2$, (63) reduces to $Var(\sqrt{n}b_1) = 9 + 15 - 6 \times 3 = 6$, as in (61).

Example 6: Consider the test for homoskedasticity under the regression framework of (36), where we now explicitly specify the heteroskedastic structure as $Var(\epsilon_i) = \sigma_i^2 = \sigma^2 + \delta' z_i$, where δ is a $r \times 1$ vector and z_i's are fixed exogenous variables, $i = 1, 2, \ldots, n$. Assuming *normality* of ϵ_i the RS statistic for testing homoskedasticity hypothesis $H_0 : \delta = 0$ is given by [see Breusch and Pagan (1979)]

$$RS_\delta = \frac{\nu' Z(Z'Z)^{-1}Z'\nu}{2\tilde{\sigma}^4}, \tag{64}$$

where $\nu_i = \tilde{\epsilon}_i^2 - \tilde{\sigma}^2$, $\nu = (\nu_1, \nu_2, \ldots, \nu_n)'$ and $Z = (z_1, z_2, \ldots, z_n)'$. The factor "$2\tilde{\sigma}^4$" is the consequence of the normality assumption, and therefore, the test in (64) will not be valid

even asymptotically if ϵ_i's are not distributed as normal. Using (61), the robust form of RS test statistics can be derived as

$$RS^*_\delta(D) = \frac{\nu' Z(Z'Z)^{-1}Z'\nu}{\frac{\nu'\nu}{n}}. \tag{65}$$

This is the same modification suggested by Koenker (1981). Note that the modification amounts to replacing $Var(\epsilon_i^2) = \mu_4 - \mu_2^2 = 3\sigma^4 - \sigma^4 = 2\sigma^4$ (derived under normality) by a robust estimate, namely, by, $\frac{1}{n}\sum_{i=1}^n(\tilde{\epsilon}_i^2 - \tilde{\sigma}^2)^2 = (\nu'\nu)/n$. For other applications of $RS^*(D)$ see for instance, Lucas (1998) and Premaratne and Bera (2017).

In a similar fashion the Wald statistic in (16) can be robustified as [see Kent (1982), White (1982), and Pace and Salvan (1997)]:

$$W^* = [h(\hat{\theta}) - c]'[H(\hat{\theta})B(\hat{\theta})H(\hat{\theta})]^{-1}[h(\hat{\theta}) - c], \tag{66}$$

and asymptotically it has χ^2_r distribution under the null hypothesis $H_0 : h(\theta) = c$. Thus, robust RS^* and W^* are obtained by robustifying the variance expressions, respectively, of $S(\tilde{\theta})$ and $h(\hat{\theta})$. However, similar robustification of LR statistic in (17) is not possible. Kent (1982) showed that under distributional misspecification LR statistic is asymptotically distributed as a weighted sum of r independent χ^2_1 variables, and thus no obvious "variance" adjustment is possible.

5.2. Robust RS tests under parametric misspecification

Consider a general statistical model represented by the log-likelihood function $l(\gamma, \psi, \phi)$ where $\gamma, \psi,$ and ϕ are parameter vectors with dimensions $(m \times 1)$, $(r \times 1)$ and $(q \times 1)$, respectively. Thus our $(p \times 1)$ parameter vector is $\theta = (\gamma', \psi', \phi')'$ and $p = m + r + q$. Suppose an investigator sets $\phi = 0$ and tests $H_0 : \psi = 0$ using the log-likelihood function $l_1(\gamma, \psi) = l(\gamma, \psi, 0)$. We will denote the RS statistic for testing H_0 in $l_1(\gamma, \psi)$ by RS_ψ. Let us also denote $\widetilde{\theta} = (\tilde{\gamma}', 0, 0)'$, where $\tilde{\gamma}$ is MLE of γ when $\psi = 0$ and $\phi = 0$. The score vector and the information matrix are defined, respectively, as

$$S(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial l(\theta)}{\partial \gamma} \\ \frac{\partial l(\theta)}{\partial \psi} \\ \frac{\partial l(\theta)}{\partial \phi} \end{bmatrix} = \begin{bmatrix} S_\gamma(\theta) \\ S_\psi(\theta) \\ S_\phi(\theta) \end{bmatrix} \qquad \text{(say)} \tag{67}$$

$$\mathcal{I}(\theta) = E_\theta\left[-\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}\right] = \begin{bmatrix} \mathcal{I}_\gamma & \mathcal{I}_{\gamma\psi} & \mathcal{I}_{\gamma\phi} \\ \mathcal{I}_{\psi\gamma} & \mathcal{I}_\psi & \mathcal{I}_{\psi\phi} \\ \mathcal{I}_{\phi\gamma} & \mathcal{I}_{\phi\psi} & \mathcal{I}_\phi \end{bmatrix}. \tag{68}$$

If $l_1(\gamma, \psi)$ were correctly specified, then the RS test statistic of (15), in the current context can be written as

$$RS_\psi = S_\psi(\widetilde{\theta})'\mathcal{I}^{-1}_{\psi\cdot\gamma}(\widetilde{\theta})S_\psi(\widetilde{\theta}), \tag{69}$$

where $\mathcal{I}_{\psi\cdot\gamma} = \mathcal{I}_\psi - \mathcal{I}_{\psi\gamma}\mathcal{I}_\gamma^{-1}\mathcal{I}_{\gamma\psi}$ and it will be asymptotically distributed as *central* χ^2_r. Under this set-up, asymptotically RS_ψ will have the correct size and will be locally optimal.

Let us now consider the case of *parametric misspecification*. Suppose the true log-likelihood function is $l_2 = (\gamma, \phi) = l(\gamma, 0, \phi)$, so that the alternative $l_1(\gamma, \psi)$ becomes misspecified. Using the sequence of *local* DGP $\phi = \delta/\sqrt{n}$, Davidson and MacKinnon (1987) and Saikkonen (1989) showed that under $l_2(\gamma, \phi)$ with $\phi = \delta/\sqrt{n}$, RS_ψ in (69), under $H_0 : \psi = 0$ will be distributed as *non-central* χ^2_r with noncentrality parameter,

$$\lambda(\delta) = \delta' \mathcal{I}_{\phi\psi\cdot\gamma} \mathcal{I}^{-1}_{\psi\cdot\gamma} \mathcal{I}_{\psi\phi\cdot\gamma} \delta, \tag{70}$$

with $\mathcal{I}'_{\phi\psi\cdot\gamma} = \mathcal{I}_{\psi\phi\cdot\gamma} = \mathcal{I}_{\psi\phi} - \mathcal{I}_{\psi\gamma}\mathcal{I}^{-1}_{\gamma}\mathcal{I}_{\gamma\phi}$. Owing to the presence of this non-centrality parameter, RS_ψ will reject the null hypothesis $H_0 : \psi = 0$ more often than allowed by the preassigned size of the test, even when $\psi = 0$. Therefore, under parametric misspecification, RS_ψ will have an excessive size. For the expression of $\lambda(\delta)$ in (70), we note that the crucial quantity is $\mathcal{I}_{\psi\phi\cdot\gamma}$, which can be interpreted as the conditional covariance between the scores S_ψ and S_ϕ given S_γ. If $\mathcal{I}_{\psi\phi\cdot\gamma} = 0$, then the local presence of the misspecified parameter $\phi = \delta/\sqrt{n}$ will have no effect on the performance of RS_ψ.

Using the expression in (70), Bera and Yoon (1993) suggested a modification to RS_ψ so that the resulting test is robust to the presence of ϕ. The modified statistic is given by

$$\begin{aligned} RS^*_\psi(P) =& [S_\psi(\widetilde{\theta}) - \mathcal{I}_{\psi\phi\cdot\gamma}(\widetilde{\theta})\mathcal{I}^{-1}_{\phi\cdot\gamma}(\widetilde{\theta})S_\phi(\widetilde{\theta})]' \\ & [\mathcal{I}_{\psi\cdot\gamma}(\widetilde{\theta}) - \mathcal{I}_{\psi\phi\cdot\gamma}(\widetilde{\theta})\mathcal{I}^{-1}_{\phi\cdot\gamma}(\widetilde{\theta})\mathcal{I}_{\phi\psi\cdot\gamma}(\widetilde{\theta})]^{-1} \\ & [S_\psi(\widetilde{\theta}) - \mathcal{I}_{\psi\phi\cdot\gamma}(\widetilde{\theta})\mathcal{I}^{-1}_{\phi\cdot\gamma}(\widetilde{\theta})S_\phi(\widetilde{\theta})]. \end{aligned} \tag{71}$$

Here the notation $RS^*(P)$ is used to signify *robust* RS test statistic under *parametetric* misspecification. Under $H_0 : \psi = 0$, $RS^*_\psi(P)$ is asymptotically distributed as *central* χ^2_r, i.e., $RS^*_\psi(P)$ has the *same* asymptotic distribution as of RS_ψ in (69) based on the correct specification. Thus, $RS^*_\psi(P)$ provides an asymptotically correct-size test under the locally misspecified alternative $l_2(\gamma, \phi)$.

$RS^*_\psi(P)$ essentially adjusts the asymptotic mean and variance of standard (unadjusted) RS_ψ. Another way to look at $RS^*_\psi(P)$ is to view the quantity, $\mathcal{I}_{\psi\phi\cdot\gamma}(\widetilde{\theta})\mathcal{I}_{\phi\cdot\gamma}(\widetilde{\theta})^{-1}S_\phi(\widetilde{\theta})$ as the prediction of $S_\psi(\widetilde{\theta})$ by $S_\phi(\widetilde{\theta})$. Here $S_\phi(\widetilde{\theta})$ is the score function of the parameter vector ϕ whose effect we want to take into account in constructing the robust version of the test. Therefore, the net score $S^*_\psi(\widetilde{\theta}) = S_\psi(\widetilde{\theta}) - \mathcal{I}_{\psi\phi\cdot\gamma}(\widetilde{\theta})\mathcal{I}^{-1}_{\phi\cdot\gamma}(\widetilde{\theta})S_\phi(\widetilde{\theta})$ is the part of $S_\psi(\widetilde{\theta})$ that remains after eliminating the effect of $S_\phi(\widetilde{\theta})$. In summary, $S^*_\psi(\widetilde{\theta} \perp S_\phi(\widetilde{\theta})$, though $S_\phi(\widetilde{\theta})$ has "peer" effect on $S_\psi(\widetilde{\theta})$. Three more things regarding $RS^*_\psi(P)$ are worth noting. *First*, $RS^*_\psi(P)$ requires estimation only under the joint null, namely for the constrained model in which both $\psi = 0$ and $\phi = 0$. Given the full specification of the model $l(\gamma, \psi, \phi)$, it is of course possible to derive a RS test for $H_0 : \psi = 0$ in the presence of ϕ. However, that requires the MLE of ϕ, which could be difficult to obtain in some cases. *Second*, when $\mathcal{I}_{\psi\phi\cdot\gamma} = 0$, $RS^*_\psi(P) = RS_\psi$. This is a simple condition to check in practice. As mentioned earlier, if this condition is true, RS_ψ is an asymptotically valid test in the local presence of ϕ. *Finally*, Bera and Yoon (1993) showed that for local misspecification $RS^*_\psi(P)$ is asymptotically equivalent to Neyman (1959) $C(\alpha)$ test, and therefore, shares its optimality properties.

Example 7: To illustrate the usefulness of the robust score statistic $RS^*_\psi(P)$, we now consider the tests developed in Anselin *et al.* (1996) for the mixed regressive - spatial au-

toregressive (SAR) model with a SAR disturbance

$$\begin{aligned} y &= \phi W y + X\gamma + u \\ u &= \psi W u + \epsilon \\ \epsilon &\sim N(0, I\sigma^2). \end{aligned} \tag{72}$$

In this model, y is an $(n \times 1)$ vector of observations on a dependent variable recorded at each of n locations, X is an $(n \times m)$ matrix of exogenous variables, and γ is a $(m \times 1)$ vector of parameters, ϕ and ψ are scalar spatial parameters and W is an observable spatial weight matrix with positive elements, associated with the spatially lagged dependent variable and SAR disturbance u. This spatial weight matrix represents "degree of potential interactions" among neighboring locations and are scaled so that the sum of the each row elements of W is equal to one.

The conventional RS statistic for testing $H_0 : \psi = 0$ is given by

$$RS_\psi = \frac{[\widetilde{u}'W\widetilde{u}/\widetilde{\sigma}^2]^2}{T}, \tag{73}$$

where $\widetilde{u} = y - X\widetilde{\gamma}$ are the OLS residuals, $\widetilde{\sigma}^2 = \widetilde{u}'\widetilde{u}/n$ and $T = tr[(W' + W)W]$. One very interesting observation here is that RS_ψ is essentially same as the widely used Moran (1948) **I** test. Let us now consider testing H_0 under the local presence of ϕ. First, the crucial quantity to consider is $\mathcal{I}_{\psi\phi\cdot\gamma}$ which is equal to T and that can never be zero. Therefore robustification of RS_ψ is needed. Anselin *et al.* (1996) derived the robust test as

$$RS^*_\psi(P) = \frac{[(\widetilde{u}'W\widetilde{u})/\widetilde{\sigma}^2 - T(\mathcal{I}_{\phi\cdot\gamma})^{-1}(\widetilde{u}'Wy)/\widetilde{\sigma}^2]^2}{T[1 - T(\mathcal{I}_{\phi\cdot\gamma})^{-1}]}, \tag{74}$$

where

$$\mathcal{I}_{\phi\cdot\gamma} = \frac{[(WX\widetilde{\gamma})'M(WX\widetilde{\gamma}) + T\widetilde{\sigma}^2]}{\widetilde{\sigma}^2}, \tag{75}$$

with $M = I - X(X'X)^{-1}X'$. A comparison of (73) and (74) clearly reveals that $RS^*_\psi(P)$ modifies the standard RS_ψ by correcting the asymptotic mean and variance of the score function S_ψ.

In a similar way we can find RS_ϕ and $RS^*_\phi(P)$ which are given, respectively, by

$$RS_\phi = \frac{[(\widetilde{u}'Wy)/\widetilde{\sigma}^2]^2}{\mathcal{I}_{\phi\cdot\gamma}} \tag{76}$$

and

$$RS^*_\phi(P) = \frac{[(\widetilde{u}'Wy)/\widetilde{\sigma}^2 - (\widetilde{u}'W\widetilde{u})/\widetilde{\sigma}^2]^2}{\mathcal{I}_{\phi\cdot\gamma} - T}, \tag{77}$$

where $\mathcal{I}_{\phi\cdot\gamma} = \mathcal{I}_\phi - \mathcal{I}_{\phi\gamma}\mathcal{I}_\gamma^{-1}\mathcal{I}_{\gamma\phi}$ using the submatrices of the partinioned form of $\mathcal{I}(\theta)$ given in (68). Anselin (1988) derived a joint RS test for $H_0 : \psi = \phi = 0$ under the framework of (72) and that takes the following form

$$RS_{\psi\phi} = \frac{[(\widetilde{u}'W\widetilde{u})/\widetilde{\sigma}^2]^2}{T} + \frac{[(\widetilde{u}'Wy)/\widetilde{\sigma}^2 - (\widetilde{u}'W\widetilde{u})/\widetilde{\sigma}^2]^2}{\mathcal{I}_{\phi\cdot\gamma} - T}. \tag{78}$$

This statistic is asymptotically distributed χ^2_2. It is easy to verify that [see Bera *et al.* (2020, Corollary 1)]

$$RS_{\psi\phi} = RS_{\psi} + RS^*_{\phi}(P) = RS_{\phi} + RS^*_{\psi}(P). \tag{79}$$

In other words, the directional RS test for ψ and ϕ can be decomposed into sum of the unadjusted one-directional test for one type of alternative and the adjusted form for the other alternative. Equalities in (79) can facilitate computations of the adjusted (robust) RS tests after having the unadjusted versions which are easy to obtain and are reported in most of the spatial software.

Anselin and Florax (1995) and Anselin *et al.* (1996) provided simulation results on the finite sample performance of the unadjusted and adjusted RS tests and some related tests. The adjusted tests $RS^*_{\psi}(P)$ and $RS^*_{\phi}(P)$ performed remarkably well. Those had very reasonable empirical sizes, remaining within the confidence intervals in all cases. In terms of power they performed exactly the way they were supposed to.

5.3. Robust RS tests under *both* the distributional and parametric misspecifications

Now we combine the results of Sections 5.1 and 5.2 and develop robust tests $RS^*_{\psi}(DP)$ which provides a two-way protection against both types of misspecifications, distributional (D) and parametric (P). As we have noted in (59), $RS^*_{\psi}(D)$ involves both the $J(\theta)$ and $K(\theta)$ matrices in the variance expression of $S_{\psi}(\tilde{\theta})$. While to account of the parametric misspecification, as we did in (71), $\mathcal{I}_{\psi\phi\cdot\gamma}(\tilde{\theta})\mathcal{I}^{-1}_{\phi\cdot\gamma}(\tilde{\theta})S_{\phi}(\tilde{\theta})$ must be subtracted from $S_{\psi}(\tilde{\theta})$ to center its mean to zero. The expression for $RS^*_{\psi}(DP)$ is given by [for details see Bera *et al.* (2020)]:

$$\begin{aligned} RS^*_{\psi}(DP) \;=\; & \Big[S_{\psi}(\tilde{\theta}) - J_{\psi\phi\cdot\gamma}(\tilde{\theta})J^{-1}_{\phi\cdot\gamma}(\tilde{\theta})S_{\phi}(\tilde{\theta})\Big]' \\ & \Big[B_{\psi\cdot\gamma}(\tilde{\theta}) + J_{\psi\phi\cdot\gamma}(\tilde{\theta})J^{-1}_{\phi\cdot\gamma}(\tilde{\theta})B_{\phi\cdot\gamma}(\tilde{\theta})J^{-1}_{\phi\cdot\gamma}(\tilde{\theta})J_{\phi\psi\cdot\gamma}(\tilde{\theta}) \\ & - J_{\psi\phi\cdot\gamma}(\tilde{\theta})J^{-1}_{\phi\cdot\gamma}(\tilde{\theta})B_{\phi\psi\cdot\gamma}(\tilde{\theta}) - B_{\psi\phi\cdot\gamma}(\tilde{\theta})J^{-1}_{\phi\cdot\gamma}(\tilde{\theta})J_{\phi\psi\cdot\gamma}(\tilde{\theta})\Big]^{-1} \\ & \Big[S_{\psi}(\tilde{\theta}) - J_{\psi\phi\cdot\gamma}(\tilde{\theta})J^{-1}_{\phi\cdot\gamma}(\tilde{\theta})S_{\phi}(\tilde{\theta})\Big], \end{aligned} \tag{80}$$

where

$$B_{\psi\cdot\gamma} = K_{\psi} + J_{\psi\gamma}J^{-1}_{\gamma}K_{\gamma}J^{-1}_{\gamma}J_{\gamma\psi} - J_{\psi\gamma}J^{-1}_{\gamma}K_{\gamma\psi} - K_{\psi\gamma}J^{-1}_{\gamma}J_{\gamma\psi}, \tag{81}$$

similarly $B_{\phi\cdot\gamma}$ and

$$B_{\psi\phi\cdot\gamma} = K_{\psi\phi} - J_{\psi\gamma}J^{-1}_{\gamma}K_{\gamma\phi} - K_{\psi\gamma}J^{-1}_{\gamma}J_{\gamma\phi} + J_{\psi\gamma}J^{-1}_{\gamma}K_{\gamma}J^{-1}_{\gamma}J_{\gamma\phi}, \tag{82}$$

and similarly $B_{\phi\psi\cdot\gamma}$. Expressions for the general forms of $J(\theta)$ and $K(\theta)$ are given in (51) and (52) and here we are using their partitioned forms for $\theta = (\gamma', \psi', \phi')'$. Under $H_0 : \psi = 0$, the $RS^*_{\psi}(DP)$ test statistic will be asymptotically distributed as χ^2_r in the presence of both distributional and parametric misspecifications. Although $RS^*_{\psi}(DP)$ has rather a lengthy expression as in (80), it is actually easy to compute requiring only $\tilde{\theta} = (\tilde{\gamma}', 0', 0')'$. It can be

easily seen that under no distributional misspecification, i.e., when $f(y;\theta) \equiv g(y)$, resulting in $K(\widetilde{\theta}) = J(\widetilde{\theta})$,

$$RS^*_\psi(DP) = RS^*_\psi(P), \tag{83}$$

and similarly under no parametric misspecification, i.e., when $\delta = 0$ in $\phi = \delta/\sqrt{n}$,

$$RS^*_\psi(DP) = RS^*_\psi(D). \tag{84}$$

Finally, trivially when $K = J$ and $\delta = 0$,

$$RS^*_\psi(DP) = RS_\psi \tag{85}$$

as given in (69).

Example 8: Let us briefly go back to Example 7 and now introduce distributional misspecification along with the presence of parametric misspecification. This case has been rigorously considered by Fang *et al.* (2014) and they demostrated both analytically and through extensive simulations that $RS^*_\psi(P)$ and $RS^*_\phi(P)$ as given, respectively in (74) and (77) are valid under non-normality. Therefore, $RS^*_\psi(DP) = RS^*_\psi(P)$ and $RS^*_\phi(DP) = RS^*_\phi(P)$. This is a somewhat unsual situation. For this model as given in (72), information matrix equality does not hold, i.e., $J(\theta_g) \neq K(\theta_g)$ [see equations (51)-(53)]. However, still $J^{-1}KJ^{-1} = J^{-1}$. This is a serendipitous situation, since no additional adjustment is needed for the distributional misspecification. The intuition behind this serendipity is that the hypotheses $\psi = 0$ and $\phi = 0$ relate to the conditional *mean* (first moment) of y in (72) (conditional on the neighborhood as captured by the W matrix). However, in general, only tests for variance (second moment) and higher moments get affected by non-normality. A similar case appeared in Bera *et al.* (2020) where they considered testing for random effects and serial correlation within an error component model. Extensive simulation results are also given in Koley and Bera (2022, 2024) demonstrating the robustness of the RS tests under non-normality in finite sample in *spatial regression* model set up.

6. Epilogue

We started this survey paper by stating that C.R. Rao's work was always inspired by some practical problems. In his 2003 *Econometric Theory (ET)* Interview [see Bera (2003, p.349)], on the RS test, Rao had the following to say, "The test evolved in a natural way while I was analyzing some genetic data. As I recall, the problem was the estimation of a linkage parameter using data sets from different experiments designed in such a way that each data set had information on the same linkage parameter. It was, however, necessary to test whether such an assumption could be made because of unforeseen factors affecting the experiments. This required a test for consistency of estimates derived from different experimental data sets." Thus we had a new statistical test principle, after LR and W, motivated by a practical problem in *genetics*. However, as we have narrated here, the resulting RS test principle has a far reaching influence even beyond the Statistics, and in particular, it has become one of the most useful model misspecification testing tools in the Econometrics literature. In the history of any scientific field, once in a while there comes a moment for a major breakthrough. Appearance of Rao (1948) was such a historical moment. In fact, after that we have not witnessed any new test principle, beyond the trinity, LR, W and RS.

To keep our exposition simple and to be close to the spirit of Rao (1948), we have sticked to the *likelihood framework*. However, it is easy to extend the RS test and its various ramifications to the generalized method of moments (GMM) and estimating functions (EF) frameworks [for more on these, see for instance Basawa (1991) and Bera *et al.* (2010)]. We have also largely confined ourselves to the asymptotic properties and distributions of the tests. However there is a huge literature on the investigation of the finite sample performance of the RS, particularly, in relation to that of LR and W and finding bootstrap critical values; for example see, Mukerjee (1990, 1993) and Horowitz (1997).

To conclude, we can only speculate what is stored in the future. Given the current vastness of the field we have lost "sharp moments of birth", like that of Rao (1948). However, considering that 75 years have already been passed, it might be a time for a brand new equally good test principle.

Acknowledgements

We are profoundly thankful to an anonymous referee for her/his careful reading and offering many pertinent comments which led to improvement of the paper. An earlier version of the paper was presented at the Invited Memorial Session for Professor C. R. Rao, Joint Statistical Meetings (JSM), Portland, Oregon, August 3-8, 2024. We are thankful to the participants of that conference for their comments, especially to the organizer of the Session, Professor Ronald L. Wasserstein, Executive Director of the American Statistical Association (ASA), for giving us the opportunity to present our paper. This paper was also presented at the Department of Statistics, University of Illinois at Urbana-Champaign (UIUC). We would like to thank the attendees of the UIUC seminar for their constructive feedback that further helped in producing an improved version. We are grateful to our research assistant (RA) Anirudh Adhikary. Without his assistance this paper wouldn't have taken off. We are also grateful to our other RAs, Rong Yuwen, Tiancheng Guo, Scarlett He, Yice Zhang and Wenqi Zeng for their diligent work on compiling the citation numbers and a very careful analysis. We thank Professor Osman Doğan for reading an earlier version of the paper with great care and thoroughness, and providing many pertinent comments with detail suggestions for improvements. We wish we could incorporate all his suggestions. Thanks are also due to Dr. Malabika Koley for comments that improved the exposition of the paper. At the publication stage, the Chair Editor, Professor V. K. Gupta paid painstaking attention to the finest details. Indeed, his suggestions, we believe, greatly improved the final presentation. We, however, retain the responsibility of any remaining errors. Part of this work was completed when the second author visited UIUC.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, **10**, 1100–1120.

Anselin, L. (1988). Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical Analysis*, **20**, 1–17.

Anselin, L., Bera, A. K., Florax, R., and Yoon, M. J. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, **26**, 77–104.

Anselin, L. and Florax, R. (1995). Small sample properties of tests for spatial dependence in regression models: Some further results. In Anselin, L. and Florax, R., editors, *New Directions in Spatial Econometrics*, pages 21–74. Springer-Verlag, Berlin.

Basawa, I. V. (1991). Generalized score tests for composite hypotheses. In Godambe, V. P., editor, *Estimating Functions*, pages 121–131. Clarendon Press, Oxford.

Bera, A. K. (1983). Aspects of econometric modeling. Unpublished Ph.D. Dissertation, The Australian National University, Canberra.

Bera, A. K. (2000). Hypothesis testing in the 20th century with a special reference to testing with misspecified models. In Rao, C. and Szekely, G., editors, *Statistics for the 21st Century*, pages 33–92. Marcel Dekker, New York.

Bera, A. K. (2003). The ET interview: Professor C.R. Rao: Interviewed by Anil K. Bera. *Econometric Theory*, **19**, 331–400.

Bera, A. K. and Bilias, Y. (2001). Rao's score, Neyman's C(α) and Silvey's LM tests: An essay on historical developments and some new results. *Journal of Statistical Planning and Inference*, **97**, 9–44.

Bera, A. K., Bilias, Y., Yoon, M. J., Doğan, O., and Taşpınar, S. (2020). Adjustments of Rao's score test for distributional and local parametric misspecifications. *Journal of Econometric Methods*, **9**. doi:10.1515/jem-2017-0022.

Bera, A. K. and Byron, R. P. (1983). A note on the effects of linear approximation on hypothesis testing. *Economics Letters*, **12**, 251–254.

Bera, A. K., Doğan, O., and Taşpınar, S. (2021). Asymptotic variance of test statistics in the ML and QML frameworks. *Journal of Statistical Theory and Practice*, **15**. doi:10.1007/s42519-020-00137-0.

Bera, A. K. and Jarque, C. M. (1981). An efficient large-sample test for normality of observations and regression residuals. Working Paper in Economics and Econometrics, No. 40, The Australian National University, Canberra.

Bera, A. K., Montes-Rojas, G., and Sosa-Escudero, W. (2010). General specification testing with locally misspecified models. *Econometric Theory*, **26**, 1838–1845.

Bera, A. K. and Premaratne, G. (2001). General hypothesis testing. In Baltagi, B. H., editor, *A Companion to Theoretical Econometrics*, pages 38–61. Blackwell Publishing Ltd.

Bera, A. K. and Ullah, A. (1991). Rao's score test in econometrics. *Journal of Quantitative Economics*, **7**, 189–220.

Bera, A. K. and Yoon, M. J. (1993). Specification testing with locally misspecified alternatives. *Econometric Theory*, **9**, 649–658.

Berk, R. H. (1953). Locally most powerful sequential tests. *The Annals of Statistics*, **3**, 373–381.

Berndt, E. R. and Savin, N. E. (1977). Conflict among criteria for testing hypotheses in the multivariate linear regression model. *Econometrica*, **45**, 1263–1277.

Bilias, Y. (2000). Sequential testing of duration data: The case of the pennsylvania 'reemployment bonus' experiment. *Journal of Applied Econometrics*, **15**, 575–594.

Bowman, K. O. and Shenton, L. R. (1975). Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2. *Biometrika*, **62**, 243–250.

Bradley, R. A. (1953). Some statistical methods in taste testing and quality evaluation. *Biometrics*, **9**, 22–38.

Breusch, T. S. (1978). Testing for autocorrelation in dynamic linear models. *Australian Economic Papers*, **17**, 334–355.

Breusch, T. S. (1979). Conflict among criteria for testing hypotheses: Extensions and comments. *Econometrica*, **47**, 203–207.

Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, **47**, 1287–1294.

Breusch, T. S. and Pagan, A. R. (1980). The Lagrange Multiplier test and its applications to model specification in econometrics. *The Review of Economic Studies*, **47**.

Bucci, A. (2024). A sequential test procedure for the choice of the number of regimes in multivariate nonlinear models. arXiv:2406.02152v1.

Byron, R. P. (1968). Methods for estimating demand equations using prior information: a series of experiments with Australian data. *Australian Economic Papers*, **7**, 227–248.

Byron, R. P. and Bera, A. K. (1983). Least squares approximation to unknown regression functions: A comment. *International Economic Review*, **24**, 255–260.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269–276.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall/CRC.

Davidson, R. and MacKinnon, J. G. (1987). Implicit alternatives and the local power of test statistics. *Econometrica*, **55**, 1305–1329.

Durbin, J. and Watson, G. S. (1950). Testing for serial correlation in least squares regression: I. *Biometrika*, **37**, 409–428.

Fang, Y., Park, S. Y., and Zhang, J. (2014). A simple spatial dependence test robust to local and distributional misspecifications. *Economics Letters*, **124**, 203–206.

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. New York and London: Academic Press, Inc.

Godfrey, L. G. (1978a). Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica*, **46**, 1293–1301.

Godfrey, L. G. (1978b). Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica*, **46**, 1303–1310.

Godfrey, L. G. (1978c). Testing for multiplicative heteroskedasticity. *Journal of Econometrics*, **8**, 227–236.

Godfrey, L. G. (1988). *Misspecification Tests in Econometrics*. Cambridge University Press, Cambridge.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica*, **12**, iii–115.

Horowitz, J. L. (1997). Bootstrap methods in econometrics: theory and numerical performance. In Kreps, D. M. and Wallis, K. F., editors, *Advances in Economics and Econometrics: Theory and Applications*, volume 3, pages 188–222. Cambridge University Press.

Jarque, C. M. and Bera, A. K. (1987). Test for normality of observations and regression residuals. *International Statistical Review*, **55**, 163–172.

Kent, J. T. (1982). Robust properties of likelihood ratio test. *Biometrika*, **69**, 19–27.

Koenker, R. (1981). A note on studentizing a test for heteroscedasticity. *Journal of Econometrics*, **17**, 107–112.

Koley, M. and Bera, A. K. (2022). Testing for spatial dependence in a spatial autoregressive (SAR) model in the presence of endogenous regressors. *Journal of Spatial Econometrics*, **3**. doi:10.1007/s43071-022-00026-7.

Koley, M. and Bera, A. K. (2024). To use, or not to use the spatial Durbin model?–that is the question. *Spatial Economic Analysis*, **19**, 30–56.

Koopmans, T. C., Rubin, H., and Leipnik, R. B. (1950). Measuring the equation systems of dynamic economics. In Koopmans, T. C., editor, *Statistical Inference in Dynamic Economic Models*, Cowles Commission for Research in Economics, Monograph No. 10, pages 53–237. John Wiley and Sons, Inc.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. Springer-Verlag, New York, Inc.

Lombardi, G. J. (1951). The sequential selection of judges for organoleptic testing. Unpublished thesis, Virginia Polytechnic Institute.

Lucas, A. (1998). Inference on cointegrating ranks using LR and LM tests based on pseudo-likelihoods. *Econometric Reviews*, **17**, 185–214.

Moran, P. A. P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B (Methodological)*, **10**, 243–251.

Mukerjee, R. (1990). Comparison of tests in the multiparameter case i: Second order power. *Journal of Multivariate Analysis*, **33**, 17–30.

Mukerjee, R. (1993). Rao's score test: Recent asymptotic results. In G. S. Maddala, C. R. R. and Vinod, H., editors, *Handbook of Statistics 11*, pages 363–379. North- Holland Science Publishers, Amsterdam.

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypothesis. In Grenander, U., editor, *Probability and Statistics*, pages 213–234. John Wiley, New York.

Neyman, J. (1980). Some memorable incidents in probabilistic/statistical studies. In Chakravarti, I. M., editor, *Asymptotic Theory of Statistical Tests and Estimation*, pages 1–32. Academic Press, New York.

Neyman, J. and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, **20**, 175–240.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transanctions of the Royal Society A*, **231**, 694–706.

Neyman, J. and Pearson, E. S. (1936). Contributions to the theory of testing statistical hypotheses. *Statistical Research Memoirs*, **1**, 1–37.

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective*. Wolrd Scientific, River Edge, N.J.

Pearson, K. (1895). Contribution to the mathematical theory of evolution– ii. skewed variation in homogeneous material. *Philosophical Transanctions of the Royal Society A*, **186**, 343–414.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **50**, 157–175.

Premaratne, G. and Bera, A. K. (2017). Adjusting the tests for skewness and kurtosis for distributional misspecifications. *Communications in Statistics - Simulation and Computation*, **46**, 3599–3613.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society*, **44**, 50–57.

Rao, C. R. (1950a). Methods of scoring linkage data giving the simultaneous segregation of three factors. *Heredity*, **4**, 37–59.

Rao, C. R. (1950b). Sequential tests of null hypotheses. *Sankhyā*, **10**, 361–370.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. John Wiley & Sons, Inc.

Rao, C. R. (2001). Two score and 10 years of score tests. *Journal of Statistical Planning and Inference*, **97**, 3–7.

Rao, C. R. (2005). Score test: Historical review and recent developments. In Balakrishnan, N., Nagaraja, H. N., and Kannan, N., editors, *Advances in Ranking and Selection, Multiple Comparisons, and Reliability: Methodology and Applications*, pages 3–20. Birkhäuser Boston, Boston, MA.

Rao, C. R. and Poti, S. J. (1946). On locally most powerful tests when alternative are one sided. *Sankhyā*, **7**, 439.

Roy, S. N. (1953). On a Heuristic Method of Test Construction and its use in Multivariate Analysis. *The Annals of Mathematical Statistics*, **24**, 220 – 238.

Saikkonen, P. (1989). Asymptotic relative efficiency of the classical test statistics under misspecification. *Journal of Econometrics*, **42**, 351–369.

Savin, N. E. (1976). Conflict among testing procedures in a linear regression model with autoregressive disturbances. *Econometrica*, **44**, 1303–1315.

Sellke, T. and Siegmund, D. (1983). Sequential analysis of the proportional hazards model. *Biometrika*, **70**, 315–326.

Sen, P. K. (1981). *Sequential Nonparametrics: Invariance Principles and Statistical Inference*. Wiley, New York.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Silvey, S. D. (1959). The Lagrangian Multiplier test. *The Annals of Mathematical Statistics*, **30**, 389–407.

Tsiatis, A. A. (1981). The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika*, **68**, 311–315.

Wald, A. (1943). Tests of statistical hypothesis concerning several parameters when the number of observation is large. *Transanctions of the American Mathematical Society*, **54**, 426–482.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, **16**, 117–186.

White, H. (1980). Using least squares to approximate unknown regression functions. *International Economic Review*, **21**, 149–170.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–25.

White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 429–448
http://www.ssca.org.in/journal.html

r-Power for Multiple Hypotheses Testing under Dependence

Swarnita Chakraborty, Adebowale Sijuwade and Nairanjana Dasgupta
Department of Mathematics and Statistics, Washington State University, United States

Received: 28 March 2024; Revised: 13 September 2024; Accepted: 17 September 2024

Abstract

In an era of "big data" the challenge of managing large-scale multiplicity in statistical analysis has become increasingly crucial. The concept of r-power, introduced by Dasgupta *et al.* (2016), presents an innovative approach to addressing multiplicity with a focus on the reliability of selecting a relevant list of hypotheses. This manuscript advances the r-power conversation by relaxing the original assumption of independence among hypotheses to accommodate a block diagonal correlation structure. Through analytical exploration and validation via simulations, we unveil how the underlying dependence structure influences r-power. Our findings illuminate the nuanced role that dependence plays in the reliability of hypothesis selection, offering a deeper understanding and novel perspectives on managing multiplicity in large datasets. Furthermore, we highlight the practicality and applicability of our results in the context of a Genome-Wide Association Study (GWAS).

Key words: r-power; Multiplicity; Multiple hypotheses testing; Dependence; False positives; Genome-wide association study.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Multiple hypotheses testing has always been a concern in scientific research due to the challenge of increasing false discoveries with growing multiplicity. However, the increasing prevalence of large-scale testing has brought this topic front and center. Despite the progress made, researchers continue to seek the "holy grail" that balances statistical power and control of false discoveries. The review of the literature, in this work aims to contribute to the understanding and advancement of multiple hypotheses testing, fostering the development of practical, feasible, and sensible methods in this field.

Multiple hypotheses testing has gained significant importance in various scientific disciplines, including chemistry (metabolomics), biology (genomics, proteomics), medicine (fMRI), and social sciences. As the scale of testing has expanded, controlling for multiplicity has become a critical concern due to the inflation of Type I error rates resulting from simultaneous testing. Traditional approaches like the Bonferroni Family-wise Error Rate (FWER), which has been used for a long time, are conservative and

Corresponding Author: Swarnita Chakraborty
Email: s.chakraborty4@wsu.edu

hence, impractical when dealing with millions of hypotheses being tested. This research area starting with Holm (1979), Simes (1986), Hochberg (1988), Hommel (1988), Sarkar (1998) has persevered for the "gold standard" and has see-sawed between techniques that are very conservative or too liberal. The introduction of the FDR (False Discovery Rate) by Benjamini and Hochberg (1995) has marked a significant advancement, providing a flexible and powerful framework for controlling false discoveries. The False Discovery Rate (FDR) is a measure that estimates the expected proportion of false discoveries among all the hypotheses that are rejected. Since the introduction of the FDR, numerous studies have enhanced and refined its methodology. Benjamini and Yekutieli (2001) proposed a modified procedure to accommodate dependence structures, ensuring valid control of the FDR. Efron *et al.* (2001) introduced empirical Bayes methods that borrow strength across hypotheses to improve FDR estimation. Building on these ideas, Efron (2004) developed the "local FDR" approach, allowing for more precise estimation of the FDR. Storey (2002) introduced the concept of q-values, which provide an intuitive interpretation of the FDR, enabling researchers to control the proportion of false discoveries at various thresholds. Furthermore, Storey (2003), Storey (2007), Heller *et al.* (2006), Dudoit *et al.* (2002), Dudoit *et al.* (2003), Pan (2002), Nichols and Holmes (2002), Nichols and Hayasaka (2003), Worsley (2003), Ge *et al.* (2003), Storey (2011) provides a general discussion of further developments related to FDR. These studies have significantly enhanced our understanding of the False Discovery Rate (FDR) framework, shedding light on its practical applications and uncovering its potential limitations. Storey (2011) also provided a comprehensive review of these advancements, offering a valuable resource for researchers in the field. While the FDR has been widely adopted and has greatly influenced the field of multiple hypotheses testing, researchers continue to search for an optimal method that balances statistical power and control of false discoveries.

Looking from a different perspective, practitioners often rely on available software and commonly used packages in R, which incorporate a ranked "top-table" approach following multiplicity corrections. Smyth *et al.* (2003) and Smyth and Speed (2003) highlight this practice and emphasize the importance of revisiting the top-table approach through the lens of multiplicity control. The fundamental question posed by practitioners is how to design studies that allow the identification of features of interest without being overwhelmed by multiplicity corrections and rigid notions of statistical significance. In response to this question, researchers have explored selection-based-on-ranking approaches within the multiplicity framework. Notable contributions in this line of research include works by Smyth *et al.* (2003), Smyth (2005), Kuo and Zaykin (2011), Kuo and Zaykin (2013), Knecht *et al.* (2003), Abbott *et al.* (2010).

Continuing in the same vein of research, Dasgupta *et al.* (2016) introduced the notion of "r-power" to provide a mathematical framework for the top-table approach. r-power is defined as the probability that no false positives exist among the test candidates included in the top-table. However, their analysis assumes independence among the hypothesis testing units, which is often an unreal assumption to implement in practice. Our study aims to relax the assumption of independence and re-formulate r-power under dependence making it applicable to real-life scenarios. It begins by considering the simplest case of equicorrelation and subsequently extends the analysis to more realistic scenarios involving block diagonal correlation structures.

While numerous approaches have been proposed to assess dependence among test candidates in multiple hypotheses problems, our method based on r-power offers a fresh perspective on this issue. Recent methods, such as the one proposed by Leek and Storey (2008), construct a dependence kernel to ensure independence of test statistics. Kim

and van de Wiel (2008) propose a method that assesses dependence using a constrained random correlation matrix. Sun and Tony Cai (2009) introduce a data-driven approach to minimize the false non-discovery rate, assuming a two-state hidden Markov model for the observed data. Additionally, Friguet *et al.* (2009) propose a conditional false discovery rate (FDR) based on a factor model. Furthermore, Liu *et al.* (2016) develop a method to assess dependence in multiple hypotheses testing using graphical models, where latent binary Markov random fields represent the underlying true states of hypotheses, and the observed test statistics appear as coupled mixture variables.

In contrast, our method takes a different perspective. We focus specifically on estimating the probability of false positives within the selected list of hypotheses, rather than considering the entire dataset and we incorporate block diagonal correlation structure to assess dependence among the test candidates. By adopting this approach, our computational framework becomes efficient and easily understandable from a practitioner's point of view.

2. Introducing r-power

In the following, we reintroduce r-power, dropping the assumption of independence. We only present the one-sided case: one-sided hypotheses, for one-sample problems, as it is the foundation of our main results in the sections to follow. Further details on the formulation of r-power in the two-sided case are available in Dasgupta *et al.* (2016). A practical approach to large scale testing, r-power focuses on selection-based ranking and answer the question: can one merely rank a test-statistic and identify the top r candidates from a set of hypotheses generated? By determining r-power, we measure the reliability of this "top table", with a focus on prioritizing features of interest over multiplicity corrections. We now present the underlying multiple testing problem in its canonical form.

2.1. Testing for normal means

Let $\vec{X}$ be a random vector following a multivariate normal distribution such that $\vec{X} \sim N(\boldsymbol{\mu}, \vec{\Sigma})$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$ is the mean vector and $\vec{\Sigma}$ is the covariance matrix, which can be one of the following: (i) an identity matrix (ii) an equicorrelated matrix (iii) a block diagonal matrix with each block being equi-correlated.

Focusing on t-tests as the statistic of interest, we assume that the number of observations is large, and that the t statistics can be approximated by normal z statistics. Without loss of generality, we define our alternative hypothesis to be in the greater than direction. Comparing to a known mean μ_0, our hypotheses of interest are given by $H_{0i} : \mu_i \leq \mu_0$ and $H_{Ai} : \mu_i > \mu_0$. We assume that for K of them are from $\mu_i = \mu_0$ and $N - K$ of them are from $\mu_i = \mu_1$, where $\mu_1 > \mu_0$. Letting $\bar{y}_i, s_i$ denote the sample mean and sample standard deviation for the i^{th} hypothesis of interest, in the one-sided case, we define our test statistic $t_i = \sqrt{n}(\bar{y}_i - \mu_0)/s_i$. We assume that the number of observations is large, and thus, t_i are approximately $N(0, 1)$ for K of the hypotheses and $N(\delta, 1)$ for $N - K$ of the hypotheses, with corresponding effect size $\delta = \sqrt{n}(\mu_1 - \mu_0)/\sigma$.

2.2. Determining r-power

To determine r-power in the testing problem above, we consider the top r hypotheses among the test statistics $t_1, \ldots, t_N$. Let G_0, G_A denote the groups of hypotheses

supporting the null and alternative hypotheses respectively, with respective test statistics t_i^0 and t_i^A. In the case of independence, it is assumed that the N hypotheses are independent, with equal variances, reducing the covariance structure to the identity matrix, and t_i^0 and t_i^A are i.i.d. $N(0,1)$ and $N(\delta,1)$, respectively. We denote the respective null and alternative order statistics by $Z_{(i)}, U_{(j)}, i = 1,\ldots,K, j = 1,\ldots,N-K$. Misclassification occurs if the largest member of G_0 is greater than or equal to the $(N-K-r)^{th}$ order statistic from G_A, or equivalently, $Z_{(K)} \geq U_{(N-K-r)}$. We define r-power as the probability of correct classification, that is,

$$r_P = P(Z_{(K)} < U_{(N-K-r)}). \tag{1}$$

Assuming independence, we can write $\vec{\Sigma} = \sigma^2 \vec{I}_N$, resulting in an r-power of

$$r_P{}^{(1)}(N,K,r,\delta) = \int_0^1 \Phi(\Phi^{-1}(t)+\delta)^K \beta(N-K-r, r+1, t)\, dt, \tag{2}$$

where Φ, ϕ denote the respective standard normal distribution and density. In practice, $r \leq N-K$ is chosen by the researcher. Ranking selection based methods such as r-power require some knowledge on the true number of null hypotheses. There have been various methods proposed for estimating the null proportion K/N, such as Jin (2008), Chen (2018), Sijuwade *et al.* (2023).

3. Incorporating dependence

With growing dimension and complexity comes an increased risk of Type I error inflation and thus the assumption of independence becomes less realistic. We consider a more general but practical option: a block diagonal correlation structure. This approach is inspired by the success or similar methods from omics studies, in which genes, lipids and metabolites tend to be related based on common chemical or biological properties. Some compelling examples include the following. Perrot-Dockès *et al.* (2019) estimated block diagonal covariance matrices to study seed quality based on omics information. Pacini *et al.* (2017) established a method to reduce false discoveries in gene expression studies using block diagonal correlation structures. To reduce computation complexity in a sensitivity analysis problem, Broto *et al.* (2020) developed a method to estimate high dimensional block-diagonal covariance matrices for Gaussian data. In practice, unstructured dependence is most realistic to consider for multiple testing, however, we show that our proposed method is general enough to approximate it, but simple enough to obtain reasonable estimates of r-power for implementation.

4. Motivating example: A GWAS study

Genome-Wide Association Studies (GWAS) aim to identify associations between genetic variants, specifically Single-nucleotide polymorphisms (SNPs), and observed traits or phenotypes. This study focuses on the association between SNPs and human cholesterol levels. The dataset used in this study is based on 323 individuals from India, China, and Malaysia, with 2,527,458 SNPs and cholesterol level measurements based on the Singapore Integrative Omics Study Saw *et al.* (2017). The purpose of this study was to use this data as an example and understand the performance of r-power when the test candidates (here, SNPs) are dependent. We focused our analysis on a subset consisting of 316 individuals and 32,010 SNPs from Chromosome 1.

Data description:

We downloaded the dataset from a public github repository basing their GWAS analysis and tutorial on data from the Singapore Integrative Omics Study `https://github.com/monogenea/GWAStutorial/tree/master/public`. Along with their methods in this tutorial, we also followed the GWAS methods for data pre-processing from Reed *et al.* (2015). The dataset includes three sub-parts:

- Genotype: A SNP matrix with columns representing SNPs and rows representing sample IDs. Genotype values range from 0 to 2, indicating different allele combinations.
- Mapping File: Contains sample IDs, SNP IDs, chromosome numbers, SNP positions, and allele types.
- Phenotype: Includes sample IDs and continuous-scale cholesterol level measurements.

Data pre-processing

SNPs with high missingness, low variability and genotyping errors were filtered out. We conducted our entire analysis in R and utilized the libraries **SNPRelate** and **snpStats** from the **BiocManager** package in R alongside commonly used packages for data handling, visualization and parallel processing - **tidyverse, doParallel, foreach** and wrote our own function for conducting the GWAS, based on the following

- **Call Rate:** The percentage of individuals in the study with available SNP information. SNPs with a call rate below 1 were discarded, removing missing information.
- **Minor Allele Frequency (MAF):** MAF denotes the proportion of least common alleles for each SNP. SNPs with MAF below 0.1 were discarded, focusing on those with a higher frequency of less common alleles.
- **Heterozygosity & Hardy Weinberg Equilibrium (HWE):** Heterozygosity occurs when each of the two alleles are present at a given SNP within an individual. HWE is a condition where the population does not evolve over generations. More specifically, this means that the alleles and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences.

  A measure of HWE is given by the Inbreeding Coefficient: $|F| = |1 - H/H_{\text{exp}}|$, where H is the observed heterozygosity, $H_{\text{exp}} = 2pq$ is the expected heterozygosity and p, q are the frequencies of the respective dominant and recessive alleles 'A' and 'a'. We retain samples that are not too heterozygous (affecting sample quality) or too homozygous (indicating inbreeding), discarding those with $|F| > 0.1$.
- **Linkage Disequilibrium (LD):** The presence of a statistical association between allelic variants within a population due to the history of recombination, mutation, and selection in a genomic region.
- **Kinship Coefficient:** A measure of relatedness among the individuals. It denotes the probability that a pair of randomly sampled homologous alleles is identical by descent. SNPs with a kinship coefficient above 0.2 were discarded, reducing relatedness bias.

After filtering based on call rate and MAF, 795668 SNPs remained. Following preprocessing, 316 individuals and 32010 SNPs were retained for analysis.

Analysis and results

We fitted a generalized linear model for each of the 32010 SNPs using the top 20 principal components and the Origin variable (dichotomized) as the covariates with our model structure. Our approach was motivated by Reed *et al.* (2015), Lipka *et al.* (2012); Price *et al.* (2006), and Wang and Zhang (2021). We conducted principal component analysis on a LD pruned dataset with an LD cut-off of 0.2. To understand the population structure, we have conducted a principal component analysis on the SNPs. We have pruned the SNPs with a linkage disequilibrium value higher than 0.2. We did so to understand the underlying population substructure, if any, through principal components. In our analysis, the top 21 principal components explained approximately 70% of the variability and we included these PCs as covariates in our model. The first two principal

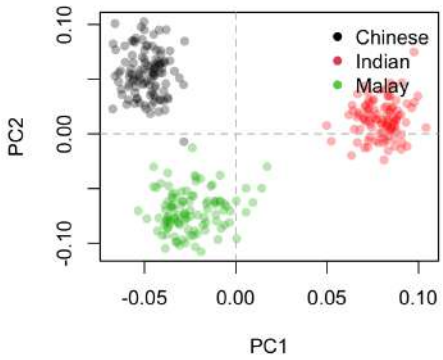


Figure 1: PCA Plot and Difference by Origin

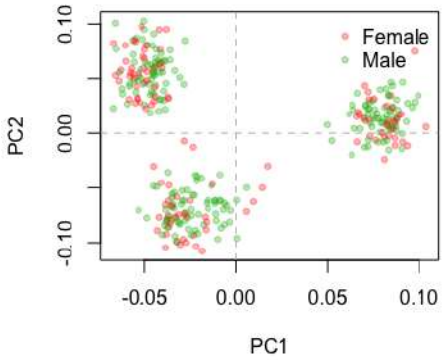


Figure 2: PCA Plot and Difference by Gender

components account for 2.42% of the variability in the data. When we plot PC1 versus PC2, we observe that SNPs originating from the same population tend to cluster together (see Figure 1). Interestingly, when we examine the distribution of gender, we find it to be fairly uniform across the different origins (see Figure 2). To ensure the reliability of our findings and minimize the occurrence of false positives, we employed a robust set of statistical procedures, including the Benjamini-Hochberg, Holmes, Sidak, and Benjamini-

Yukuteli methods. SNPs identified by these methods are shown in Table 1.

Table 1: Comparison of Methods to Control False Discovery

Method	SNPs Identified	SNP ID
No Correction	1502	rs7527051, rs12406924, rs4844688,...
Bonferroni	1	rs7527051
Benjamini Hochberg	2	rs7527051, rs12140539
Benjamini & Yekutieli	0	-
Holm	1	rs7527051
Sidak Single-Step	1	rs7527051

The effect of the multiplicity correction methods on the number of identified SNPs can been seen in 1, with 1502 SNP selections in the absence of any correction. Bonferroni's method, known for its stringent control of family-wise error rate (FWER), only selected 1 SNP. On the other hand, Benjamini Hochberg's method, aimed at controlling the false discovery rate (FDR), chose 2 SNPs. Interestingly, the Benjamini-Yekutieli method, which considers positive regression dependency among SNPs instead of assuming independence, did not select any SNPs. Additionally, Holm's and Sidak's procedures each reported 1 SNP. The discrepancy between the number of significant SNPs obtained from the various correction methods raises an important question: How many SNPs should we follow up on? Whilst no multiplicity correction resulted in a large number of significant SNPs, FDR & FWER corrections, yielded one or two significant SNPs. Striking the right balance is essential to ensure the accuracy and reliability of our findings. Hence, in addition to these established methods, we introduce our formulation of r-power allowing dependence in our formulation. In the following section, we generalize the idea of r-power under dependence.

5. Equicorrelation

Returning to the testing problem 2.1, we consider the equicorrelated case corresponding to a joint distribution $N_N(\boldsymbol{\mu}, \vec{\Sigma})$, where $\vec{\Sigma} = \sigma^2[\rho \vec{1}_N \vec{1}_N^T + (1-\rho)\vec{I}_N], 0 \leq \rho < 1$. We present the following result, from which the probability of correct classification can readily be determined.

Theorem 1: r-power under Compound Symmetry

Under the equicorrelated testing scenario, we have

1. $\vec{\Sigma}^{-1/2} = a\vec{1}_N\vec{1}_N^T + (b-a)\vec{I}_n$, where

$$a = \frac{1}{\sigma N}\left(\frac{1}{\sqrt{1+(N-1)\rho}} - \frac{1}{\sqrt{1-\rho}}\right),\ b = a + \frac{1}{\sigma\sqrt{1-\rho}}.$$

2. $\vec{\Sigma}^{-1/2}\vec{Y} \sim N_N(\vec{\Sigma}^{-1/2}\boldsymbol{\mu}, \vec{I}_N)$.

3. For $1 - \sigma^{-2} < \rho < 1$, the probability of misclassification is always less than the equivalent probability under the independent testing scenario 2.1 with equivalent dimensions, with corresponding classification probability $r_P{}^{(1)}(N, K, r, (b-a)\delta)$, following (2).

Proof:

1. Let $\vec{e}_k$ denote the k^{th} standard basis vector of $\mathbb{R}^N$, $\vec{S}_k = \sum_{i=1}^{k} \vec{e}_i$. Since $\rho \neq (N-1)^{-1}$, the Sherman-Morrison formula $(\vec{A} + \vec{u}\vec{v}^T)^{-1} = \vec{A}^{-1} - (\vec{A}^{-1}\vec{u}\vec{v}^T\vec{A}^{-1})/(1 + \vec{v}^T\vec{A}^{-1}\vec{u})$ with $\vec{u} = \vec{v} = \vec{1}_N, \vec{A} = (1-\rho)\vec{I}_N$, implies that $\vec{\Sigma}$ has symmetric positive definite inverse $\vec{\Sigma}^{-1} = C_1(\vec{I}_N - C_2\rho\vec{1}_N\vec{1}_N^T)$, $C_1 = \sigma^{-2}(1-\rho)^{-1}, C_2 = (1+(N-1)\rho)^{-1}$, since $\vec{1}_N\vec{1}_N^T$ has spectrum $\lambda_1 = N, \lambda_2 = \cdots = \lambda_N = 0$, and eigenvectors $\vec{u}_1 = \vec{1}_N, \vec{u}_j = \vec{e}_1 - \vec{e}_j, j = 2, \ldots, N$.

   Let $\vec{D} = \text{diag}(l_1, \ldots, l_N)$ where $l_j = 1/\sqrt{C_1(1 - C_2\rho\lambda_j)}$, and let $\vec{U}$ denote the matrix with columns $\vec{u}_j$. Letting $\vec{H} = (\vec{1}_N\vec{1}_N^T + \text{diag}(0, -N, \ldots, -N))/N$, $\vec{H}\vec{U} = [\vec{1}_N, \vec{0}_N, \ldots, \vec{0}_N]^T + [(\vec{e}_1 - \vec{S}_N), \vec{e}_2, \ldots, \vec{e}_N]^T = \vec{I}_N$. Since $\vec{H}, \vec{U}$ and their product are symmetric, they commute, and we write $\vec{H} = \vec{U}^{-1}$.

   Let $d_1 = \vec{D}_{11} = 1/(\sigma\sqrt{1+(N-1)\rho}), d_2 = \vec{D}_{22} = 1/(\sigma\sqrt{1-\rho})$ and determine the inverse square root $\vec{\Sigma}^{-1/2} = \vec{U}\vec{D}\vec{U}^{-1}$, as indeed by the Spectral Theorem, $(\vec{\Sigma}^{-1/2})^2 = \vec{U}\vec{D}^2\vec{U}^{-1} = \vec{\Sigma}^{-1}$. Finally, we have the inverse square root $\vec{\Sigma}^{-1/2} = \vec{U}[d_1\vec{1}_N, d_2(\vec{S}_N - N\vec{e}_2), \ldots, d_2(\vec{S}_N - N\vec{e}_N)]^T/N$, so set $a = (d_1 - d_2)/N, b = (d_1 + (N-1)d_2)/N = a + d_2$.

2. Using characteristic functions, let $i = \sqrt{-1}$ denote the imaginary unit and $\vec{r} = (r_1, \ldots, r_N)$ denote an arbitrary deterministic vector. Setting $\vec{s} = (\vec{\Sigma}^{-1/2})^T\vec{r}$, we have $\mathbb{E}(e^{i\vec{r}^T\vec{Y}^*}) = \mathbb{E}(e^{i\vec{r}^T\vec{\Sigma}^{-1/2}\vec{Y}}) = \mathbb{E}(e^{i\vec{s}^T\vec{Y}})$ since $\vec{Y} \sim N_N(\boldsymbol{\mu}, \vec{\Sigma})$. Observing that $\vec{\Sigma}^{-1/2}\vec{\Sigma}(\vec{\Sigma}^{-1/2})^T = \vec{I}_N$ and $\vec{s}^T\vec{s} = \vec{r}^T\vec{r}$ due to the symmetry of $\vec{\Sigma}^{-1/2}$, we have $e^{i\vec{s}^T\boldsymbol{\mu} - \vec{s}^T\vec{\Sigma}\vec{s}/2} = e^{i\vec{r}^T(\vec{\Sigma}^{-1/2}\boldsymbol{\mu}) - \vec{r}^T\vec{\Sigma}\vec{r}/2}$, as desired.

3. Without loss of generality, the mean vector can be written as $\boldsymbol{\mu} = \mu_0(\vec{S}_N - \vec{S}_k) + \mu_1(\vec{S}_N - \vec{S}_{N-k})$. Then, $\vec{\Sigma}^{-1/2}\boldsymbol{\mu} = (\mu_0{}^*)(\vec{S}_N - \vec{S}_k) + (\mu_1{}^*)(\vec{S}_N - \vec{S}_{N-k})$, where $\mu_0{}^* = \mu_0(a + (k-1)b) + \mu_1 b(N-k)$, $\mu_1{}^* = \mu_0 bk + \mu_1(a + b(N-k-1))$. The result follows from the monotonicity of Φ and the observation that in the two-sided case, with effect size $\delta^* = \sqrt{b}(\mu_1{}^* - \mu_0{}^*)/\sigma$, we have $\delta^*/|\mu_1 - \mu_0| = (b-a) = (\sigma\sqrt{1-\rho})^{-1} > 1$ when $1 - \sigma^{-2} < \rho < 1$. We then compare with equation (2).

□

6. Block diagonal approach

We now move to a more general scenario based on the equicorrelated testing problem in our theorem, as equicorrelation is still too restrictive for practical applications. As we mentioned on in Section 3, we extend the equicorrelated case to obtain an analytic form for the probability of correct classification. Based on this recent research, and its adjacency to r-power in application, we consider a block-diagonal correlation structure based on the equicorrelated case which tends toward unstructured as the block size increases. The probability of misclassification can be determined based on the distribution of the within-block order statistics. We assume independent blocks in which each block corresponds to test candidates belonging to either the null or alternative hypotheses or a mix of both.

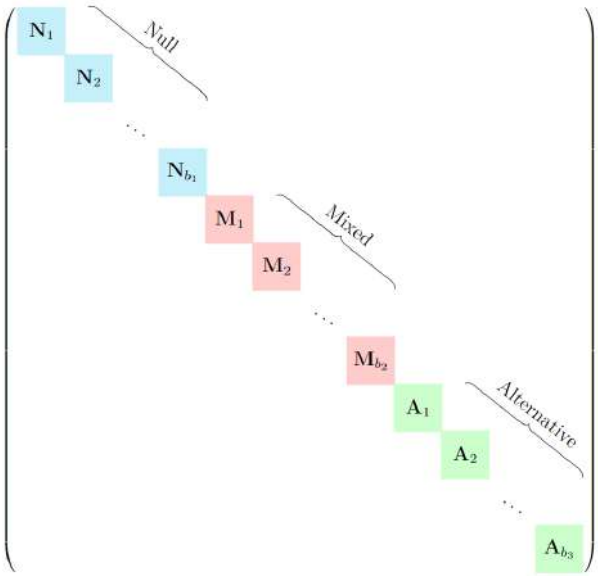
We now define the $N \times N$ block diagonal correlation matrix $\vec{B}$. For $j = 1, 2, 3$, we assume respective null, mixed and alternative test candidate counts $b_j : \sum_j b_j = N$

and define the respective null, mixed and alternative major blocks and indices by $\vec{B}^{(j)} = diag(\vec{B}_1^{(j)}, \ldots, \vec{B}_{b_j}^{(j)})$, defined as follows: each minor block $\vec{B}_i^{(j)}$ is compound symmetric, corresponding to test statistic vectors $\vec{t}_i^{(j)}$ as in the normal means problem, with assumed joint distribution $N_{N_i^{(j)}}(\boldsymbol{\mu}_i^{(j)}, \vec{B}_i^{(j)}), i = 1, \ldots, b_j$ for $j = 1, 2, 3$, based on a common effect size $\delta > 0$:

$$\boldsymbol{\mu}_i^{(j)} = \delta \begin{cases} \vec{0}_{K_i^{(j)}} & j = 1 \\ \vec{S}_{N_i^{(j)}} - \vec{S}_{K_i^{(j)}} & j = 2 \\ \vec{1}_{N_i^{(j)}} & j = 3 \end{cases} \tag{3}$$

$$\vec{B}_i^{(j)} = \sigma_{ij}^2[\rho_{ij}\vec{1}_{N_i^{(j)}}\vec{1}_{N_i^{(j)}}^T + (1 - \rho_{ij})\vec{I}_{N_i^{(j)}}].$$

Here, $\vec{S}_k$ is defined as in our theorem, $0 < \rho_{ij} < 1, \sigma_{ij}^2 > 0$ denote the respective correlation and variance of $\vec{B}_i^{(j)}$. Without loss of generality, we order the minor blocks such that $\vec{B} = diag(\vec{B}^{(1)}, \vec{B}^{(2)}, \vec{B}^{(3)})$. We assume that each block corresponds to $N_i^{(j)}$ test candidates, null candidates $K_i^{(j)} : \sum_{i,j} K_i^{(j)} = K$ and research hypotheses $r_i^{(j)} : \sum_{i,j} r_i^{(j)} = r$, noting that $N_i^{(1)} = K_i^{(1)}, K_i^{(3)} = 0$, since $j = 1, 3$ correspond to only null or alternative candidates respectively. We denote the null and alternative order statistics corresponding to each sub-block $\vec{B}_i^{(j)}$ by $(Z_i^{(j)})_{(k)}, (U_i^{(j)})_{(l)}, k = 1, \ldots, K_i^{(j)}, l = 1, \ldots, N_i^{(j)} - K_i^{(j)} - r_i^{(j)}$ and for convenience, denote the null, mixed and alternative major blocks by $\vec{N}, \vec{M}, \vec{A}$ corresponding to the null, mixed and alternative major blocks $\vec{B}^{(j)}$: $\vec{N} = \vec{B}^{(1)}, \vec{M} = \vec{B}^{(2)}, \vec{A} = \vec{B}^{(3)}$. Write $\vec{B} = diag(\vec{N}, \vec{M}, \vec{A})$, where $\vec{N} = diag(\vec{N}_1, \ldots \vec{N}_{b_1})$, and likewise for $\vec{M}, \vec{A}$. We visualize the block diagonal $\vec{B}$ below,



adopting the notation $\prod_{\vec{N}} = \prod_{i,j:\vec{B}_i^{(j)} \in \vec{N}}$ and likewise for $\vec{M}, \vec{A}$. Under the above assumptions, misclassification occurs when any members of the top r table come from the null, which occurs when any members the $r_i^{(j)}$-th hypotheses come from the corresponding null group of any sub-block $\vec{B}_i^{(j)}$: the largest of the $K_i^{(j)}$-th null order statistics is at least

the smallest of the $N_i^{(j)} - K_i^{(j)} - r_i^{(j)}$-th alternative test statistics. We compute r-power corresponding to the block diagonal $\vec{B}$ as the classification probability

$$r_P = P\left(\max_{\vec{N},\vec{M}}(Z_i^{(j)})_{(K_i^{(j)})} < \min_{\vec{A}}(U_i^{(j)})_{(N_i^{(j)}-K_i^{(j)}-r_i^{(j)})}\right). \tag{4}$$

We note that due to the structure of the means $\boldsymbol{\mu}_i^{(j)}$, the null and alternative test statistics corresponding to the *mixed* block $\vec{M}$ are not exchangeable. However, the test statistics corresponding to the null block $\vec{N}$ and alternative block $\vec{A}$ are (respectively), and their distributions can be readily determined based on the normality in (3), via Theorem 5.3.1 of Tong (1990b) to obtain the r-power analytically. We show that if $b_2 \leq 1$, r_P is completely determined by the distribution

$$F_{i,j,k,\mu}(x) = \int_{\mathbb{R}} H_{ijk}\left(\frac{(x-\mu)/\sigma_{ij} + z\sqrt{\rho_{ij}}}{\sqrt{1-\rho_{ij}}}\right)\phi(z)\,dz, \tag{5}$$

where $H_{ijk}(z) = \sum_{m=k}^{N_i^{(j)}} \binom{N_i^{(j)}}{m}\Phi(z)^m\Phi(-z)^{N_i^{(j)}-m}$.

6.1. One-sided case, no mixed candidates

We use the shorthand $Z_{ij} = (Z_i^{(j)})_{(K_i^{(j)})}, U_{ij} = (U_i^{(j)})_{(N_i^{(j)}-K_i^{(j)}-r_i^{(j)})}, K_{ij} = (K_i^{(j)})$ and likewise for $N_i^{(j)}, r_i^{(j)}$. Define $M_{ij} = N_{ij} - K_{ij} - r_i^{(j)} > 0$ and let G_{ij} denote the distribution function of U_{ij} with corresponding density g_{ij}. Starting with the probability of misclassification, due to our assumptions (3), we use the exchangeability of the $t_i^{(j)}$ and proceed as in Tong (1990a). The density corresponding to the distribution (5) is given by

$$f_{i,j,k,\mu} = \sigma_{ij}^{-1}(1-\rho_{ij})^{-1/2}\int_{\mathbb{R}} h_{ijk}\left(\frac{(x-\mu)/\sigma_{ij} + z\sqrt{\rho_{ij}}}{\sqrt{1-\rho_{ij}}}\right)\phi(z)\,dz, \tag{6}$$

where $h_{ijk}(z) = k\binom{N_{ij}}{k}\Phi^{k-1}(z)\Phi^{N_{ij}-k}(-z)\phi(z)$. Since $b_2 = 0$, integrating by parts, we obtain

$$\begin{aligned}
r_P &= P\left(\max_{\vec{N}} Z_{ij} \leq \min_{\vec{A}} U_{ij}\right) = \prod_{\vec{N}} P\left(Z_{i1} \leq \min_{\vec{A}}(U_{k3})\right) = \prod_{\vec{N}}\int_{\mathbb{R}} P\left(Z_{i1} \leq u|u\right) g_{k3}(u)du \\
&= \prod_{\vec{N}}\int_{\mathbb{R}}\left(\int_{-\infty}^{u} P\left(\sigma_{i1}(Z_{i1}\sqrt{1-\rho_{i1}} + Z\sqrt{\rho_{i1}}) \leq x\right)dx\right) g_{k3}(u)du \\
&= \prod_{\vec{B}_i\in\vec{N}}\int_{\mathbb{R}}\left(\int_{-\infty}^{u} f_{i,1,K_{i1},0}(x)dx\right)\cdot\frac{\partial}{\partial u}\left(1 - \prod_{\vec{B}_k\in\vec{A}}(1 - F_{k,3,M_{k3},\delta}(u))\right)du \\
&= \prod_{\vec{B}_i\in\vec{N}}\int_{\mathbb{R}} -\left(\int_{-\infty}^{u} f_{i,1,K_{i1},0}(x)dx\right)\left(\frac{\partial}{\partial u}\prod_{\vec{B}_k\in\vec{A}}(1 - F_{k,3,M_{k3},\delta}(u))\right)du \\
&= \prod_{\vec{N}}\int_{\mathbb{R}} f_{i,1,N_{i1},0}(u)\prod_{\vec{A}}(1 - F_{k,3,N_{k3}-r_{k3},\delta}(u))\,du.
\end{aligned}$$

6.2. One-sided case, one mixed candidate

Let p denote the probability that the minimum of the top r table corresponds to the mixed block $\vec{M}$. Since the test candidates corresponding to the mixed block $\vec{M}$ are not exchangeable, to streamline our analytical formulation, we assume the existence of at most one mixed block. Since $b_2 = 1$, using the density (6) and performing the change of variables $t = \Phi\left(\frac{(x-\mu)/\sigma_{ij} + z\sqrt{\rho_{ij}}}{\sqrt{1-\rho_{ij}}}\right)$, we obtain

$$
\begin{aligned}
p &= P(\min_{\vec{A}} U_{ij} = U_{12}) = \prod_{\vec{A}} P(U_{i3} \geq U_{12}) \\
&= \prod_{\vec{A}} \int_{\mathbb{R}} P(U_{i3} \geq u|u) g_{12}(u)\, du \\
&= \frac{1}{\sigma_{12}\sqrt{\rho_{12}}} \prod_{\vec{A}} \int_{\mathbb{R}} P(U_{i3} \geq u|u) \int_0^1 \phi\left(\frac{\Phi^{-1}(t) + (\delta_{12} - u)\sigma_{12}^{-1}}{\rho_{12}^{1/2}(1-\rho_{12})^{-1/2}}\right) \\
&\beta(M_{12}, K_{12} + r_{12} + 1, t)\, dt du \\
&= \prod_{\vec{A}} \int_{\mathbb{R}} f_{i,3,N_{i3}-r_{i3},\delta}(u) F_{1,2,M_{12},\delta_{12}}(u)\, du,
\end{aligned}
$$

where $\delta_{ij} = \delta 1_{\{i,j:\vec{B}_i^{(j)} \in \vec{M}, i > K_{ij}\}}$. We apply our theorem to obtain

$$
\begin{aligned}
P_1 &= P(\max_{\vec{N},\vec{M}} Z_{ij} \geq U_{12}) \\
&= P\left(Z_{i2} \geq U_{12}\right) \prod_{\vec{N}} \int_{\mathbb{R}} \left(\int_u^\infty f_{i,1,K_{i1},0}(x) dx\right) f_{1,2,M_{12},\delta_{12}}(u)\, du \\
&= \left[1 - r_P^{(1)}\left(N_{12}, K_{12}, r_{12}, \delta/\left(\sigma_{12}\sqrt{1-\rho_{12}}\right)\right)\right] \prod_{\vec{N}} \int_{\mathbb{R}} \left(1 - F_{i,1,K_{i1},0}(u)\right) f_{1,2,M_{12},\delta_{12}}(u)\, du, \\
P_2 &= P(\max_{\vec{N}} Z_{ij} \geq \min_{\vec{A}} U_{ij}) \\
&= \prod_{i,\vec{N}} \int_{\mathbb{R}} - \left(\int_s^\infty f_{i,1,K_{i1},0}(x) dx\right) \left(\frac{\partial}{\partial s} \prod_{k,\vec{A}} (1 - F_{k,3,M_{k3},\delta}(s))\right) ds, \\
&= \prod_{\vec{N}} \left(1 - \int_{\mathbb{R}} f_{i,1,N_{i1},0}(s) \prod_{\vec{A}} (1 - F_{k,3,N_{k3}-r_{k3},\delta}(s))\, ds\right)
\end{aligned}
$$

and finally, we have $r_P = p(1 - P_1) + (1 - p)(1 - P_2)$.

6.3. Limiting behavior and the two-sided case

We examine 6.1 to determine the limiting behavior of r_P due to the structural similarity in each case. As the null proportion K/N tends to 1, since $r \leq N - K$, across $\vec{A}, F_{i,j,M_{ij},\mu}$ tends to $F_{i,j,0,\mu} = \int_{\mathbb{R}} \phi(z) dz = 1$ (using the binomial theorem), resulting in vanishing products over $\vec{A}$ and an r-power of zero. The situation in which $r \to N - K$ is similar. Likewise, as $\delta \to \infty$, $F_{i,j,k,\delta} \to 0$, following the limiting behavior of the terms $\Phi^m(z)$ as $z \to -\infty$, r_P tends to $\prod_{\vec{N}} \int_{\mathbb{R}} f_{i,1,N_{i1},0}(u)\, du = 1$. This aligns with our intuition

from the independent case that as the effect size increases, it is easier to distinguish the alternative from the null, and vice versa with increasing null proportion.

In the two-sided case, the test statistics $t_i^{(j)} = \sqrt{N_{ij}}|X_{ij} - \bar{X}|/\sigma$ are assumed to be jointly distributed according to the folded normal, with mean vector entries 0 or δ depending on whether or not they correspond to the null or alternative groups for their respective blocks. We assume a Gaussian copula $C(U_1, \ldots, U_N) = \Phi_{\vec{B}}(\Phi^{-1}(F_1(X_1)), \ldots, \Phi^{-1}(F_N(X_N)))$ with covariance matrix $\vec{B}$ as in (3) and $\sigma_{ij} \equiv 1$. For $1 \leq l \leq N$, the distributions F_l are given by $2\Phi(z) - 1$ and $\Phi(z+\delta) + \Phi(z-\delta) - 1$ respectively. We then determine the r-power as $r_P = P(\max_{N,\vec{M}} \Phi^{-1}(F_k(Z_{ij})) < \max_{\vec{A}} \Phi^{-1}(F_l(U_{ij})), 1 \leq k \leq b_1 + b_2 \leq l \leq N$ and proceed as in 6.1 and 6.2, replacing Φ, ϕ with F_l and its derivative in $H_{ijk}(z)$ from (5). Since r_P has no closed form expression in the block diagonal scenarios, we approximate it numerically. One approach is to reexamine $F_{i,j,k,\mu}(x)$ in (5):

$$\begin{aligned}
&\int_{\mathbb{R}} \sum_{m=k}^{N} \binom{N}{k} \Phi(Az+B)^m (1 - \Phi(Az+B))^{N-m} \phi(z) dz \\
&= \sum_{m=k}^{N} \sum_{j=0}^{N-k} \binom{N}{k} \binom{N-k}{j} (-1)^j \int_{\mathbb{R}} \Phi^{m+j}(Az+B)\phi(z) dz.
\end{aligned}$$

As in Owen (1980), Hartmann (2017), an application of the Fubini-Tonelli theorem and a change of variables $z_k = y_k + x - B/A, k = 1, \ldots, m, \vec{s} = (x, y_1, \ldots, y_m)$ yields

$$\begin{aligned}
&\int_{\mathbb{R}} \Phi^m(Az+B)\phi(z)dz \\
&= \int_{\mathbb{R}} \prod_{k=0}^{m} \Phi(Az+B)\phi(z)dz \\
&= \int_{\mathbb{R}} \int_{-\infty}^{z} \cdots \int_{-\infty}^{z} \prod_{k=0}^{m} \phi(Az_k+B)\phi(x)\, dz_1 \ldots dz_k dx. \\
&\frac{1}{\sqrt{(2\pi)^{m+1}|\vec{V}|}} \int_{\mathbb{R}} \int_{-\infty}^{z} \cdots \int_{-\infty}^{z} \exp\left(\frac{-1}{2} \vec{s}^T \vec{V}^{-1/2} \vec{s}\right) d\vec{y} dx, \\
&= \frac{1}{\sqrt{(2\pi)^m |\vec{\Sigma}_A|}} \int_{-\infty}^{-B/A} \cdots \int_{-\infty}^{-B/A} \exp\left(\frac{-1}{2} \vec{s}^T \vec{\Sigma}_A^{-1} \vec{s}\right) d\vec{y}, \\
&\vec{V} = \vec{e}_1^T \vec{e}_1 - 2(\vec{e}_1^T \vec{1}_m + \vec{1}_m^T \vec{e}_1) + 4(\vec{1}_m^T \vec{1}_m) + \text{diag}(0, 1, \ldots, 1)/A^2, \vec{\Sigma}_A = \vec{1}_m \vec{1}_m^T + A^{-2} \vec{I}_m.
\end{aligned}$$

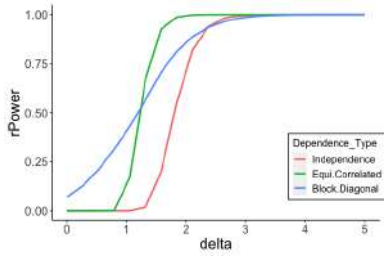
We obtain the multivariate normal distribution function $\vec{F}_m(\vec{0}_m, -(B/A)\vec{1}_m, \vec{\Sigma}_A)$, which can be accurately approximated, as in Genz (1992). The two-sided case can be handled similarly using $F_l(z) = 2\Phi(z) - 1$ or $\Phi(z+\delta) + \Phi(z-\delta) - 1$ respectively, depending on correspondence with the null or alternative.

7. Simulation study

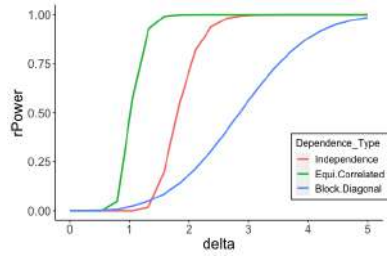
In Section 6, we provided analytical formulae and examined the structure of r-power under our block diagonal assumptions. To support our results, we conducted

an empirical simulation study. We simulated r_P under different scenarios for the one and two-sided cases, letting $b_1 = b_3 = 2, b_2 = 1, N_{ij} \equiv 10^4$. We also varied the number of nulls ($K_{ij} = (1, 3, 5, 7, 9) \cdot 10^3$), null proportions $10^{-4}K_{ij}$, top-table size ($r = 50, 100, 250, 500, 750$) and effect size ($\delta = 0.1, 0.5, 1, 2, 3$). We compare the results for the block diagonal against the independence and equicorrelated scenarios.

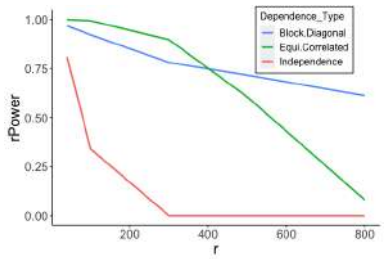
Our results are plotted in Figure 3 for the one-sided case and Figure 4 for the two-sided case. In each figure, we look at r-power as a function of effect size, δ and top-table/list size r. For illustration, we have provided the case with $N = 5$ blocks, assuming $b_1 = b_3 = 2$ and writing $\rho_k, k = 1, \ldots, N$ as the correlation corresponding to block $\vec{B}_k$, *i.e.*, $\rho_1 = \rho_{11}, \rho_2 = \rho_{21}, \rho_3 = \rho_{12}, \rho_4 = \rho_{13}, \rho_5 = \rho_{23}$. To highlight the impact of changing correlation, we vary ρ across the null and alternative blocks respectively, starting with ($\rho_1 = \rho_2 = .7$ and $\rho_4 = \rho_5 = .6$ with $\rho_3 = .5$). Our findings support our expectations from part 3 of our theorem and the tendency of the misclassification probability to increase with r, mentioned in section (6.3). Since $1 - \sigma_{ij}^{-2} = 0 < \rho < 1$, equicorrelation overtakes independence, given equal top table size and effect size with the block diagonal case generally falling somewhere in between the two. The situation changes depending on how the null and alternative correlations compare to each other. Additional results are available upon request from the authors.



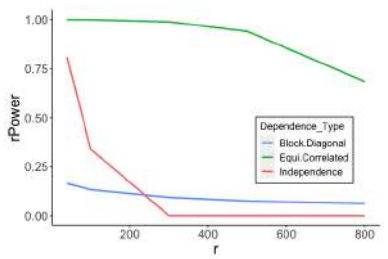
$\rho_1 = 0.7, \rho_2 = 0.7, \rho_3 = 0.4, \rho_4 = 0.5, \rho_5 = 0.5$



$\rho_1 = 0.4, \rho_2 = 0.4, \rho_3 = 0.8, \rho_4 = 0.9, \rho_5 = 0.9$



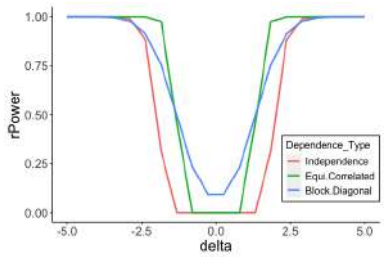
$\rho_1 = 0.9, \rho_2 = 0.9, \rho_3 = 0.4, \rho_4 = 0.4, \rho_5 = 0.4$



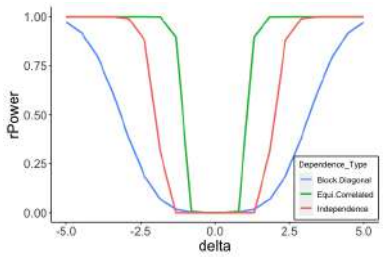
$\rho_1 = 0.4, \rho_2 = 0.4, \rho_3 = 0.9, \rho_4 = 0.9, \rho_5 = 0.9$

Figure 3: One-sided case: r-power vs effect size δ (top left, top right) and hypothesis selection size r for (bottom left, bottom right)

Since r-power depends upon the size of the top-table r and the number of nulls k, if $r > N - k$, the top-table becomes unreliable as its length exceeds that of the number of alternatives, thus containing members from the null. This supports our findings in Figure 3 (bottom left): as we increase r, we run the risk of this scenario occurring regardless of the dependence structure. However, as the dependence among the alternatives increases, dimensionality is impacted, reducing $N-K$, and when the true $N-K$ tends to be smaller than a given estimate, we are more likely to undershoot for a given choice of top-table size r, reducing the r-power. On the other hand, if the dependence among the alternatives is much smaller than that of the nulls ($\max_{\vec{A}} \rho_{ij} \ll \min_{\vec{N}} \rho_{kl}$), we see a reversal and expect



$\rho_1 = 0.7, \rho_2 = 0.7, \rho_3 = 0.4, \rho_4 = 0.5, \rho_5 = 0.5$



$\rho_1 = 0.4, \rho_2 = 0.4, \rho_3 = 0.9, \rho_4 = 0.9, \rho_5 = 0.9$

Figure 4: r-power vs effect size

larger r-power than the compound symmetry case. This is supported by Figure 4, also where the null correlations are lower than that of the alternatives (*e.g.* .4 vs .9, as shown), r-power for the block diagonal performs worse than the independent case.

8. Revisiting GWAS results

We have seen in 4 that there was a clear discrepancy between the number of significant SNPs obtained from the various correction methods which raised an important question: How many SNPs should we follow up on? Whilst no multiplicity correction resulted in a large number of significant SNPs, FDR & FWER corrections yielded one or two significant SNPs. Hence, in addition to these established methods, we introduced r-power for the block diagonal testing setup as in (3), comparing the reliability of selecting top-5 SNPs. The Manhattan Plots denote the position of top-1 SNP with Bonferroni's cut-off and with Top-5 SNPs based on the ranked test statistic value, respectively.

The Manhattan Plots in Figure 5 show the SNP that was selected from the existing methods (Bonferroni,Holm,Sidak,Benjamini-Hochberg) (above) and the position of the top-5 SNPs (below). From these, we determine the confidence of these selected lists based on r-power. First, we need to estimate the proportion of null hypotheses before using r-power. For this study, we employed the Laplace-transform-based estimator from Sijuwade *et al.* (2023) for its low mean square error in comparison to other estimators. The resulting estimate yielded a null proportion of $\pi_0 = 0.9017$, indicating that there are 28,864 null hypotheses and 3,146 alternative hypotheses. Before evaluating r-power, we also need to determine the block diagonal correlation approximation from the SNP correlation matrix. We construct this by dividing the null and alternative groups and performing variable clustering. The steps involved in determining the parameters of r-power are as follows:

1. Conduct a cluster analysis on the SNPs in both the alternative and null groups based on their mean values, assuming that the null and the alternative groups are well separated.
2. Perform a clustering analysis using CLARA, an extension of the k-medoids algorithm, which is suitable for handling large-scale data Kaufman and Rousseeuw (2008).
3. Assess cluster quality using the widely adopted silhouette method to determine the optimal number of clusters.

We analyzed the alternative group, identifying the top $r = 5$ SNPs, as depicted in Figure 6 and observed two clusters displaying a wide range of correlation values.

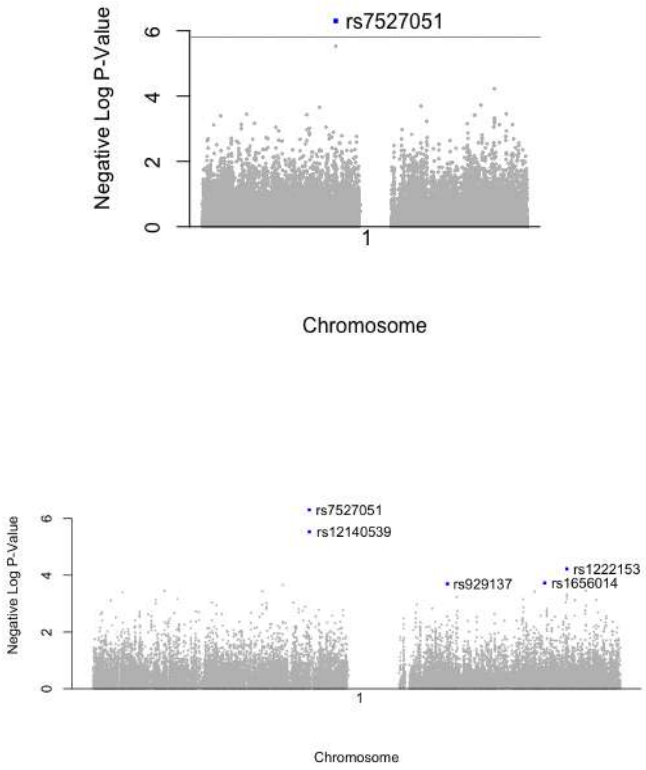


Figure 5: Bonferroni SNPs (top), Top 5 SNPs(bottom)

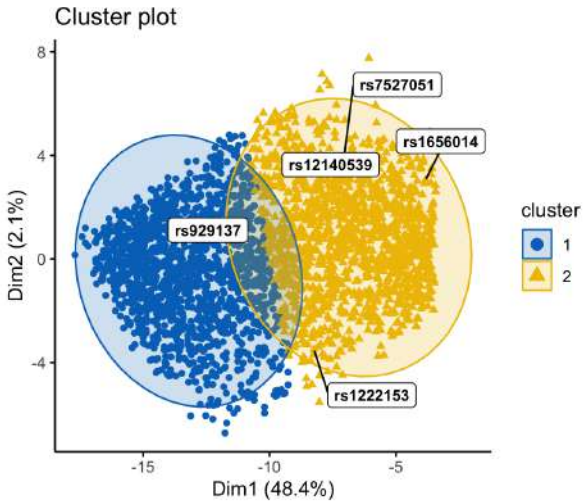


Figure 6: Cluster Plot of the Test Candidates

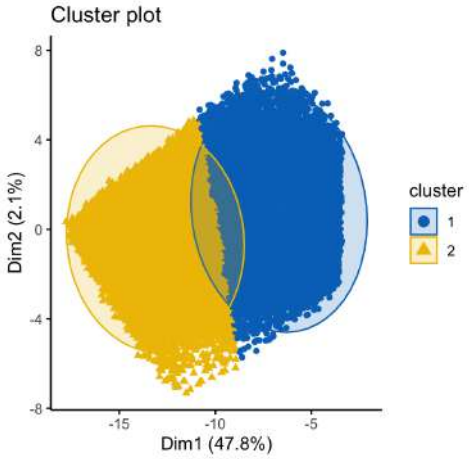


Figure 7: Cluster Plot of the Null Candidates

To enhance the analysis, we subdivided each of these clusters into different blocks of approximately equicorrelated variables using LD-pruning, resulting in an approximated block diagonal correlation matrix. LD is calculated based on R^2 values, and we considered the absolute correlation values of the SNPs since our formulation on r-power is based on positive correlation. Thus, for evaluating r-power, we consider 8 blocks, as illustrated in Table 2.

In the analysis of the null group, we also identified two clusters. However, due to the large number of candidates in the null hypothesis, calculating the correlation matrix for all SNPs was not feasible. Instead, we focused on the number of null candidates with a linkage disequilibrium (measured as R-squared) less than or equal to 0.4 from each cluster. Since our goal in calculating r-power is to ensure that none of the selected $top-r$ candidates are from the null group, we want to avoid over-penalizing the probability by considering all SNPs in the null group, regardless of their correlation. To do so, we choose SNPs with low correlation within the null group to calculate r-power. Cluster 1 in the null group which originally consisted of 14639 SNPS, has 5675 markers, with a linkage disequilibrium threshold value of 0.4. Cluster 2 from the null group, which originally consisted of 14225 SNPs and also has 5682 markers with linkage disequilibrium threshold value 0.4.

We calculated r_P for a block diagonal correlation with 10 blocks - 8 blocks from alternative and 2 from the null as illustrated in Table 2. Although we have 10 blocks, we needed to find the block allocation of the top-5 SNPs to calculate r-power. In our study, the top-5 selected SNPs are "rs7527051" , "rs12140539" ,"rs1222153" , "rs1656014" and "rs929137". The cluster allocation is described in Table 2.

Under the assumption of block diagonal correlation, the r-power for selecting the top 5 significant SNPs with an effect size of 4 was reported to be 91%, indicating a high probability of correctly identifying the relevant SNPs. At an effect size of 3, the r-power was reported to be 60%. Thus, the r-power method not only provides a powerful tool for confidently selecting relevant SNPs but also offers valuable insights into the relationship between effect size and r-power. By visualizing the r-power as a type of power curve, researchers can gain a better understanding of how to choose the optimal value for r in their r-power analysis.

Table 2: Cluster Information and Correlation Among SNPs

Cluster	Hypothesis Group	Block Size	Cluster Correlation	No of SNPs
Cluster 1	Null	5675	0.4	0
Cluster 2	Null	5682	0.4	0
Cluster 1	Alternative	56	0.1	0
Cluster 1	Alternative	921	0.5	0
Cluster 1	Alternative	232	0.7	1
Cluster 1	Alternative	437	1	0
Cluster 2	Alternative	45	0.1	0
Cluster 2	Alternative	855	0.5	2
Cluster 2	Alternative	142	0.7	1
Cluster 2	Alternative	466	1	1

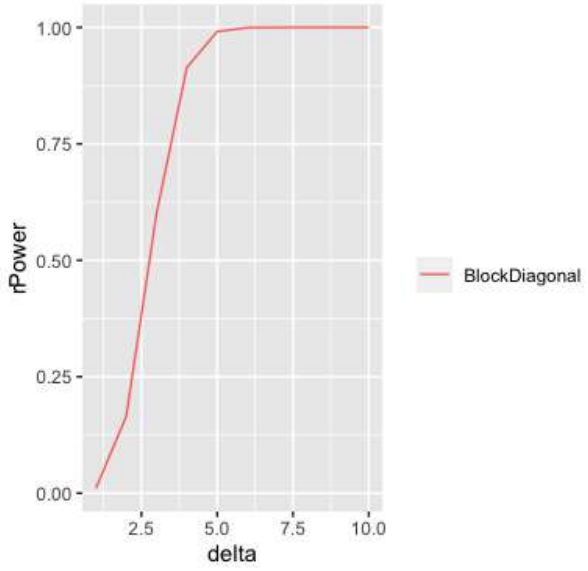


Figure 8: r-power of the top 5 selected SNPs

9. Conclusion

In this article, we addressed a fundamental issue concerning dependence with respect to the normal means problem, making positive steps towards addressing the complexity of the unstructured scenario by investigating dependence patterns, deriving analytical formulae and offering practical solutions to multiplicity issues in large-scale multiple-hypothesis testing problems. Our comprehensive simulation experiments serve to support our findings and demonstrate robustness. Our simulation results consistently show that a positive equicorrelated structure yields higher r-power compared to independence among hypotheses and that the correlation structure within blocks significantly effects the classification probability calculation.

Focusing on top tables, r-power offers insights into the robustness of the systematic selection of candidates based on combinatorial methods. We find that high within-group correlation reduces the effective dimensionality of the top-r table, in which case testing becomes more conservative and in this way, r-power provides insight into test reliability. From our findings, the correlation within the null group surpasses that of the research group, r-power under the block diagonal setup tends to outperform the equicorrelated scenario. Our formulation is built to address scenarios in which sources of variation are

difficult to identify and various features are clustered. Examples of relevant domains for future consideration and applications include but are not limited to large-scale testing within genomics, metabolomics, proteomics and fMRI studies. Our GWAS results in particular, highlight the advantage of our approach in determining test reliability compared to traditional methods, especially in in SNP detection and we are developing an R library for its implementation.

References

Abbott, D. F., Waites, A. B., Lillywhite, L. M., and Jackson, G. D. (2010). fMRI assessment of language lateralization: An objective approach. *NeuroImage*, **50**, 1446–1455.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165–1188.

Broto, B., Bachoc, F., Clouvel, L., and Martinez, J. M. (2020). Block-diagonal covariance estimation and application to the Shapley effects in sensitivity analysis. *arXiv:1907.12780 [math, stat]*, . arXiv: 1907.12780.

Chen, X. (2018). Estimators of the proportion of false null hypotheses: I "universal construction via Lebesgue-Stieltjes integral equations and uniform consistency under independence". *arXiv preprint arXiv:1807.03889*, .

Dasgupta, N., Lazar, N. A., and Genz, A. (2016). A look at multiplicity through misclassification. *Sankhya B*, **78**, 96–118. Publisher: Springer.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**, 71–103.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–139.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, **99**, 96–104.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**, 1151–1160.

Friguet, C., Kloareg, M., and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, **104**, 1406–1415.

Ge, Y., Dudoit, S., and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test*, **12**, 1–77.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, **1**, 141–149.

Hartmann, M. (2017). Extending Owen's integral table and a new multivariate bernoulli distribution. *arXiv preprint arXiv:1704.04736*, .

Heller, R., Stanley, D., Yekutieli, D., Rubin, N., and Benjamini, Y. (2006). Cluster-based analysis of FMRI data. *NeuroImage*, **33**, 599–608.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**, 800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, **75**, 383–386.

Jin, J. (2008). Proportion of non-zero normal means: Universal oracle equivalences and uniformly consistent estimators. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **70**, 461–493.

Kaufman, L. and Rousseeuw, P. J. (2008). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Inc, New Jersey.

Kim, K. I. and van de Wiel, M. A. (2008). Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics*, **9**, 1–12.

Knecht, S., Jansen, A., Frank, A., Van Randenborgh, J., Sommer, J., Kanowski, M., and Heinze, H. (2003). How atypical is atypical language dominance? *NeuroImage*, **18**, 917–927.

Kuo, C.-L. and Zaykin, D. (2013). The ranking probability approach and its usage in design and analysis of large-scale studies. *Plos One*, **8**, e83079.

Kuo, C.-L. and Zaykin, D. V. (2011). Novel rank-based approaches for discovery and replication in genome-wide association studies. *Genetics*, **189**, 329–340.

Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, **105**, 18718–18723.

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., Gore, M. A., Buckler, E. S., and Zhang, Z. (2012). GAPIT: Genome association and prediction integrated tool. *Bioinformatics*, **28**, 2397–2399.

Liu, J., Zhang, C., and Page, D. (2016). Multiple testing under dependence via graphical models. *Annals of Applied Statistics*, **1**, 1699–1724.

Nichols, T. and Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: A comparative review. *Statistical Methods in Medical Research*, **12**, 419–446.

Nichols, T. E. and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, **15**, 1–25.

Owen, D. B. (1980). A table of normal integrals: A table. *Communications in Statistics-Simulation and Computation*, **9**, 389–419.

Pacini, C., Ajioka, J. W., and Micklem, G. (2017). Empirical Bayes method for reducing false discovery rates of correlation matrices with block diagonal structure. *BMC Bioinformatics*, **18**, 213.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **18**, 546–554.

Perrot-Dockès, M., Lévy-Leduc, C., and Rajjou, L. (2019). Estimation of large block structured covariance matrices: Application to "multi-omic" approaches to study seed quality. *arXiv:1806.10093 [stat]*, . arXiv: 1806.10093.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904–909.

Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M. P., and Foulkes, A. S. (2015). A guide to genome-wide association analysis and post-analytic interrogation. *Statistics in Medicine*, **34**, 3769–3792.

Sarkar, S. K. (1998). Some probability inequalities for ordered mtp2 random variables: a proof of the simes conjecture. *Annals of Statistics*, , 494–504.

Saw, W.-Y., Tantoso, E., Begum, H., Zhou, L., Zou, R., He, C., Chan, S. L., Tan, L. W.-L., Wong, L.-P., Xu, W., et al. (2017). Establishing multiple omics baselines for three southeast asian populations in the singapore integrative omics study. *Nature Communications*, **8**, 653.

Sijuwade, A. J., Chakraborty, S., and Dasgupta, N. (2023). An inverse Laplace transform oracle estimator for the normal means problem. *Metrika*, **1**, 1–18.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.

Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer.

Smyth, G. K. and Speed, T. (2003). Normalization of cDNA microarray data. *Methods*, **31**, 265–273.

Smyth, G. K., Yang, Y. H., and Speed, T. (2003). Statistical issues in cDNA microarray data analysis. In *Functional Genomics*, pages 111–136. Springer.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 479–498.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics*, **31**, 2013–2035.

Storey, J. D. (2007). The optimal discovery procedure: A new approach to simultaneous significance testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**, 347–368.

Storey, J. D. (2011). FDR. In *International Encyclopedia of Statistical Science*, pages 504–508. Springer.

Sun, W. and Tony Cai, T. (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **71**, 393–424.

Tong, Y. L. (1990a). Order statistics of normal variables. In Tong, Y. L., editor, *The Multivariate Normal Distribution*, Springer Series in Statistics, pages 123–149. Springer, New York, NY.

Tong, Y. L. (1990b). Positively dependent and exchangeable normal variables. In Tong, Y. L., editor, *The Multivariate Normal Distribution*, Springer Series in Statistics, pages 91–122. Springer, New York, NY.

Wang, J. and Zhang, Z. (2021). GAPIT version 3: Boosting power and accuracy for genomic association and prediction. *Genomics, Proteomics & Bioinformatics*, **19**, 629–640.

Worsley, K. (2003). Detecting activation in fMRI data. *Statistical Methods in Medical Research*, **12**, 401–418.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 449–469
http://www.ssca.org.in/journal

Hierarchical Bayes Small Area Estimation from Aggregated Data using Various Spatial Models

Jiacheng Li[1], Hee Cheol Chung[2], David Okech[3] and Gauri S. Datta[4,5]
[1] *Wells Fargo Bank, Charlotte, NC*
[2] *Department of Mathematics and Statistics, University of North Carolina, Charlotte, NC*
[3] *School of Social Work, University of Georgia, Athens, GA*
[4] *Department of Statistics, University of Georgia, Athens, GA*
[5] *Center for Statistical Research and Methodology, U.S. Census Bureau, Suitland, MD*

Received: 04 August 2024; Revised: 17 September 2024; Accepted: 20 September 2024

Abstract

Small area estimation methods are important tools for applied statisticians to help policymakers in need of reliable statistics for lower level disaggregated populations. While aggregated statistics at the higher level may be available from surveys, they are not useful to estimate characteristics for lower level subpopulations. Often useful covariates for these subpopulations are available, which can be integrated through innovative small area estimation methodology to leverage aggregated data to produce better estimates and measures of uncertainty for the disaggregated subpopulation means.

To serve our need we generalize the celebrated Fay-Herriot model, which has been extensively used for several decades by many National Statistical Offices around the world, to produce reliable small area statistics. We consider the traditional independence for the Fay-Herriot linking model errors as well as various important spatially dependent models for these errors. We conduct a hierarchical Bayesian analysis for all these models based on a popular class of noninformative improper prior densities for the linking model parameters. We illustrate the usefulness of our proposal by producing estimates of statewide four-person family median incomes for the U.S. states for the year 1990. We create for our illustration the aggregated statistics from the 1990 Current Population Survey. We evaluate the accuracy of our state predictions against the corresponding incomes, deemed to be reliable, produced by the 1990 Census. For all models and for all improper prior densities for the model parameters considered here we prove the propriety of the resulting posterior distributions. The result in Corollary 1 of Chung and Datta (2022, *Survey Methodology*, vol. 48, No. 2, pp. 463-489) follows as a special case. Our empirical assessments amply demonstrate the usefulness of our novel approach.

Key words: Aggregated statistics; Conditional autoregression; Current Population Survey; Fay-Herriot model; LCAR; Simple CAR; Simultaneous AR.

AMS Subject Classifications: 62K05, 05B05

Corresponding Author: Gauri S. Datta
Email: gauri@uga.edu

1. Introduction

Preparation and implementation of effective social welfare and human development policy proposals require reliable statistics measuring important population characteristics, for example, income, employment, education, healthcare, agricultural productions and environmental safety. National statistical offices (NSO's) around the world collect relevant data and produce these statistics. Many nations and international organizations recognize the need for these statistics at the national level as well as at su-bnational/sub-population levels. These sub-populations may be geographic (states, counties or districts), demographic (gender, race, age) or cross-classification of geographic and demographic factors (state level poverty rates for the school-age children).

The NSOs and international organizations, for example, the World Bank, rely on appropriate data to produce relevant statistics. Since national censuses are carried out every five or ten years, these data will fail to capture the current state of the population when the last census gets outdated. Decennial or quinquennial censuses are expensive. To gather timely and less expensive data the NSOs conduct carefully planned sample surveys to collect data from only a fraction of the population. It is well-documented in the statistics literature that carefully planned surveys with reasonably large samples can be as accurate as a census.

Even if a nation may be doing well overall, often various segments of the nation may not be doing as well. While any functioning government that cares to serve its people requires accurate data for the entire nation, it also needs reliable disaggregated data for various segments of the nation. For example, the U.S. government has mandated it by law to produce timely and accurate disaggregated statistics measuring income, employment and health service for various demographic groups at the county or state level. The European Union and the United Nations have many programs that require accurate poverty and income information for many geographic/demographic sub-populations. Production of reliable, disaggregated statistics is known as small area estimation in survey sampling.

Sample surveys are generally designed to provide useful data in estimating various characteristics of a population of interest. Sample sizes are so chosen to ensure that traditional design-based estimators are adequately accurate. Sample size is usually the key thing, and when it comes to estimating a sub-population characteristic, based solely on the part of the original sample which is in the sub-population, the sub-sample may be small or empty. The version of the national level design-based direct estimate from the sub-sample for a sub-population, if it has enough sample to be computed, may be highly variable, or may be non-existent due to lack of sample. Sub-populations with low or no sample size to produce reliable direct estimates are known as small areas. Due to limited resources, a survey, by design, may not allocate any sample to many sub-populations. For example, the American Community Survey (ACS) is conducted to produce reliable statistics for nearly three thousand U.S. counties. However, the ACS usually samples about one-third of the counties, resulting in many non-sampled small areas. Post-surveys some sub-populations may also be defined for the current need, and there may not be any units selected from these sub-populations. Again, resource constraints do not permit selection of new sample to transform unreliable or unavailable small area estimates to reliable ones. To increase the accuracy of inadequate direct estimates of small areas (or to produce estimates for non-sampled areas), statistical methods advocate model-based approach to enable borrowing information

from direct estimates of other domains and other data sources. In many applications, other related surveys and administrative data provide useful covariates. A model-based estimate of an area is produced by suitably shrinking a direct estimate (if available) to a synthetic estimate of a regression function based on auxiliary variables.

In small area estimation if unit-level data are available, a unit-level small area model by Battese *et al.* (1988) is often recommended for modeling. However, in many applications to protect confidentiality of the respondents the organization conducting the survey releases only summary data at the area-level for the areas sampled. In this setup, Fay and Herriot (1979) introduced an area-level model. This popular model is known in small area estimation as the Fay-Herriot model. In this model estimating the small area mean θ_i for a small area i, if its direct estimator Y_i is available, it is called a *sampled area*. We assume that Y_i is unbiased for θ_i. No direct estimator is available for an *unsampled area*.

Fay and Herriot (1979) proposed a linking model for all m small area means θ_i based on a multiple linear regression of the θ_i's on some available suitable covariates $\mathbf{x}_i$. For a sampled area the model-based estimator of θ_i is obtained by shrinking its direct estimator Y_i to the synthetic regression estimator $\mathbf{x}_i^T\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is an estimator of the regression coefficient $\boldsymbol{\beta}$ in the regression mean function $\mathbf{x}_i^T\boldsymbol{\beta}$. If an area is unsampled, synthetic estimator $\mathbf{x}_i^T\hat{\boldsymbol{\beta}}$ is the small area estimator of θ_i.

In small area estimation a population is partitioned into m sub-populations, and a survey design samples $m - m_1$ sub-populations and does not sample the other m_1 sub-populations (sometimes $m_1 = 0$ but for the ACS it is positive). The Fay-Herriot model described above uses the $m - m_1$ direct estimators and covariate $\mathbf{x}_i$ from all m areas to estimate θ_i, ith sub-population mean, $i = 1, \ldots, m$.

From cost and administrative considerations a survey, by design, may merge t_1 sub-populations and select a sample from this combined bigger sub-population. Suppose a direct estimator S_1 from this sample estimates η_1, where, for example, η_1 may represent the total employment or total healthcare expenditure, then it is equal to the sum of θ_i's for these t_1 sub-populations. In general, we assume that η_1 is a known linear combination of the t_1 θ_i's. Similarly, t_2 other sub-populations may be merged for sampling, and a direct estimator S_2 from a sample from this merged sub-populations may be formed which estimates the corresponding population characteristic η_2. Again, we assume that η_2 is a known linear combination of t_2 θ_i's. In this way, an estimator S_r is obtained which is an unbiased estimator of η_r, where η_r is a known linear combination of θ_i's. This setup is the motivation of the problem that we will consider here. We assume that we have an $r \times 1$ vector of estimators $\mathbf{S}$ with its associated variance-covariance matrix $\mathbf{D}_S$. We assume that $\mathbf{S}$ is an unbiased estimator of $\mathbf{C}\boldsymbol{\theta}$ for an $r \times m$ known matrix $\mathbf{C}$. We assume that rank of $\mathbf{C}$ is r and that $\mathbf{D}_S$ is a known, positive definite (p.d.) matrix. If $r = m - m_1$ and each of the rows of $\mathbf{C}$ has all elements 0 and one element 1 (first element in the first row, the second in the second row, etc.), then $\eta_1 = \theta_1,\ \eta_2 = \theta_2$, etc. and we get the traditional Fay-Herriot setup (cf. Fay and Herriot (1979)).

Alternatively, in the Fay-Herriot setup, suppose an area i is an union of n_i sub-areas and we are interested in estimating the sub-area mean θ_{ij} based on available covariates $\mathbf{x}_{ij}$ from that area. The ith area mean η_i is a known linear combination of the sub-area means θ_{ij}'s. A direct estimator Y_i is available for η_i but there are no direct estimators

of θ_{ij} for the sub-areas. Our goal is to estimate the θ_{ij}'s based on the survey estimates Y_i's and the sub-area covariates $\mathbf{x}_{ij}$'s. To address this problem we are expanding the scope of the traditional Fay-Herriot model. Note that there is no direct estimate of θ_{ij}. We use $\boldsymbol{\theta}_i$ to denote the vector $(\theta_{i1}, \ldots, \theta_{in_i})^T$ and use the traditional independent Fay-Herriot linking model where $\theta_{ij} \overset{ind}{\sim} N(\mu_{ij}, \sigma^2), j = 1, \ldots, n_i, i = 1, \ldots, m$, where for simplicity of presentation we assume that μ_{ij}'s and σ^2 known. Suppose the survey estimator Y_i is normally distributed variance D_i and mean $\sum_{j=1}^{n_i} c_{ij}\theta_{ij}$, the coefficients c_{ij}'s are known. Under this setup, simple algebra shows that (if we invoke a Bayesian setup), the posterior mean of θ_{ij} is $\tilde{\theta}_{ij} = \mu_{ij} + \{\sigma^2 c_{ij}/(D_i + \sigma^2 \sum_{j=1}^{n_i} c_{ij}^2)\}(Y_i - \sum_{j=1}^{n_i} c_{ij}\mu_{ij})$, and the posterior variance

$$\tilde{\sigma}_{ij}^2 = \frac{\sigma^2\{D_i + \sigma^2 \sum_{k\neq j} c_{ik}^2\}}{D_i + \sigma^2 \sum_{j=1}^{n_i} c_{ij}^2}. \tag{1}$$

This result makes sense. Since a θ_{ij} appears only in the distribution of Y_i and since all the θ_{ij}'s are independent, it follows that $Y_i|\theta_{ij} \sim N(c_{ij}(\theta_{ij} - \mu_{ij}) + \sum_{k=1}^{n_i} c_{ik}\mu_{ik}, D_i + \sigma^2 \sum_{k\neq j} c_{ik}^2)$ and $\theta_{ij} \sim N(\mu_{ij}, \sigma^2)$. These two distributions imply that $\theta_{ij}|y_i \sim N(\tilde{\theta}_{ij}, \tilde{\sigma}_{ij}^2)$. If $r = [m/2]$, and $n_i = 2$ for $i = 1, \ldots, r$, and $c_{i1} = c_{i2} = 1$, then $\tilde{\theta}_{i1} = \tilde{\theta}_{i2} + \mu_{i1} - \mu_{i2}$, and $\tilde{\sigma}_{i1}^2 = \tilde{\sigma}_{i2}^2$. If $n_i = 1$ and $c_{i1} = 1$, the above expressions for the posterior mean and the variance for θ_{i1} will reduce to the results from the regular independent Fay-Herriot model.

For a comprehensive literature on small area estimation we refer to Rao and Molina (2015) who documented the need for reliable small area statistics in many applications in agriculture, education, healthcare, economy and industry. Here is an outline of the article. In Section 2 we presented a generalized Fay-Herriot model for aggregated small area statistics. We introduced the hierarchical Bayes (HB) model as well as the distribution of Fay-Herriot linking model error under various spatial models. In Subsection 2.1, we introduced the neighborhood matrix, an important element in spatial modeling. We outlined some useful properties of the eigenvalues of this matrix and those of a couple of other matrices defined from this matrix. In Section 3, we presented a set of sufficient conditions to ensure the propriety of all the posterior distributions that result from the class of HB models and a class of noninformative improper prior pdf's introduced the last section. We illustrated our novel ideas in Section 4 to the estimation of four-person households median incomes of the forty-nine contiguous states of the US. Section 5 reviews the importance of the proposed methodology. Finally, Section 6 presents detailed arguments to prove the propriety of the posterior pdf's for a couple of spatial models, and how these arguments can be modified for the remaining models.

2. A generalized Fay-Herriot model for aggregated statistics

As it was described in Section 1, the aggregated statistics $\mathbf{S}$ is assumed to be an unbiased estimator of $\mathbf{C}\boldsymbol{\theta}$. We present below an extended version of the popular Fay-Herriot model to draw inference for $\boldsymbol{\theta}$ based on the aggregated data $\mathbf{S}$. The $r \times m$ matrix $\mathbf{C}$ is an appropriate known matrix, described further in Remark 1.

The HB model:

(a) $\mathbf{S}|\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2, \rho \sim N(\mathbf{C}\boldsymbol{\theta}, \mathbf{D}_S)$,

(b) $\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2, \rho \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{\Omega}^{-1}(\rho))$,

(c) The prior pdf for $\boldsymbol{\beta}, \sigma^2$ and ρ is

$$\pi(\boldsymbol{\beta}, \sigma^2, \rho) = \pi(\boldsymbol{\beta}) \times g(\sigma^2) \times h(\rho), \tag{2}$$

where $\pi(\boldsymbol{\beta})$ is a bounded positive function corresponding to a prior pdf (may be improper), $g(\sigma^2)$ is an appropriate (may also be improper) prior, and $h(\rho)$ is a proper pdf for ρ defined on an appropriate finite interval.

For the model above, $\mathbf{D}_S$ is a known p.d. matrix. Also, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^T$ is an $m \times p$ matrix of covariates, with rank p. The regression coefficient $\boldsymbol{\beta}$ is a $p \times 1$ vector. For a particular model in (b), namely, the independent Fay-Herriot model, the matrix $\boldsymbol{\Omega} = \mathbf{I}_m$ is free from ρ. In this case, the model being free from ρ, a prior for ρ is not required. However, we can use any proper prior for ρ and the posterior pdf of ρ will be the same as the prior pdf. An improper uniform prior $\pi(\boldsymbol{\beta}) = 1$ is extensively used in the Bayesian literature (see, for example, Berger (1985) and Ghosh (1992)).

Remark 1: The part (b) of the above hierarchical model is known as the *linking model* (see Rao and Molina (2015)). In order for the sampling and the linking models in the above hierarchical model to be capable of producing inference for $\boldsymbol{\beta}$ under the frequentist setup (without part (c) for prior specification), the matrix $\mathbf{C}$ needs to have certain structure. In particular, the row space of $\mathbf{CX}$ must be the same as that of $\mathbf{X}$. It is equivalent to $rank(\mathbf{CX}) = rank(\mathbf{X})$, the estimability requirement of $\boldsymbol{\beta}$ based on the design matrix for $\mathbf{S}$. It implies that $r \geq rank(\mathbf{C}) \geq rank(\mathbf{CX}) = rank(\mathbf{X}) = p$ is a necessary condition on r.

The part (b) of the above hierarchical model implies a representation for the ith component of $\boldsymbol{\theta}$, which is given by

$$\theta_i = \mathbf{x}_i^T\boldsymbol{\beta} + v_i, \ i = 1, \ldots, m, \tag{3}$$

where the v_i's are also called random effects in mixed linear model. This decomposition implies that the random effects vector $\mathbf{v} = (v_1, \ldots, v_m)^T$ is normally distributed with mean vector $\mathbf{0}$ and variance-covariance matrix $\sigma^2\boldsymbol{\Omega}^{-1}(\rho)$. We appropriately choose various forms of $m \times m$ the p.d. matrix $\boldsymbol{\Omega}$ to specify a class of models for $\boldsymbol{\theta}$. For the independent Fay-Herriot model, $\boldsymbol{\Omega} = \mathbf{I}_m$ which means that the θ_i's are independently distributed. It is not unreasonable to anticipate that if effective covariates are available, they can capture most of the variability of the θ_i's. Any unexplained variation among the θ_i's will be modeled by the random effects, and across small areas these random effects will not have any particular pattern. This variability may be modeled through $\boldsymbol{\Omega} = \mathbf{I}_m$.

While the independent Fay-Herriot model is the default model, in a recent paper Chung and Datta (2022) showed that in the absence of good covariates some spatially-dependent models for the random effects vector improve the prediction of θ_i's. In our case when a majority of small areas have no direct estimators, and only a few (no less than p) aggregated statistics are available that estimate some linear combinations of the small area mean vector $\boldsymbol{\theta}$, importance of both effective covariates and good linking models explaining the dependence of the components of $\boldsymbol{\theta}$ cannot be overemphasized.

2.1. A neighborhood matrix for spatial models with some useful results

For the Fay-Herriot model, Chung and Datta (2022) considered four different spatial models for the random effects and showed that many of these models yielded better predictions of small area means for the non-sampled areas than the independent Fay-Herriot model. They considered four spatial models which are determined by suitable structure of the matrix $\mathbf{\Omega}$. These $\mathbf{\Omega}$ matrices depend on a special matrix $\mathbf{W}$, known as neighborhood or *incidence* matrix.

The incidence matrix $\mathbf{W}$ is determined by the neighborhood structure of the small areas. This matrix is an $m \times m$ non-null, symmetric, square matrix. All the diagonal elements of this matrix are zero. In the popular version if two areas i and j are neighbors, then $W_{ij} = 1$, and it is zero otherwise. We now introduce additional matrices derived from $\mathbf{W}$ and describe some properties of these matrices that would be useful in exploration of our spatial models. For $i = 1, \ldots, m$, we define the ith row sum of $\mathbf{W}$ by $W_{i\cdot}$. We assume that $W_{i\cdot} \geq 1$ for all i, and we define the diagonal matrix $\mathbf{L} = \text{diag}(W_{1\cdot}, \ldots, W_{m\cdot})$. Using $\mathbf{L}$ and $\mathbf{W}$, we define two more matrices: $\widetilde{\mathbf{W}} = \mathbf{L}^{-1}\mathbf{W}$ and $\mathbf{R} = \mathbf{L} - \mathbf{W}$. The matrices $\mathbf{W}, \widetilde{\mathbf{W}}$ are non-null. Each matrix must have at least one nonzero eigenvalue. Since $\text{tr}(\mathbf{W}) = 0 = \text{tr}(\widetilde{\mathbf{W}})$, all the eigenvalues of each matrix sum to zero. Since $\mathbf{W}$ is symmetric, all its eigenvalues are real. Let $\mathbf{L}^{-1/2}$ be a diagonal matrix such that the ith diagonal is $W_{i\cdot}^{-1/2}$. Then all the eigenvalues of the matrix $\mathbf{L}^{-1/2}\mathbf{W}\mathbf{L}^{-1/2}$ will be real. Further, these eigenvalues are the same as the eigenvalues of $\widetilde{\mathbf{W}}$. Hence, for both $\mathbf{W}$, $\widetilde{\mathbf{W}}$, the smallest eigenvalue must be negative and the largest must be positive.

Suppose $\tilde{\lambda}_i, i = 1, \ldots, m$, are the eigenvalues of $\widetilde{\mathbf{W}}$, which are real. We can order them as $\tilde{\lambda}_m \leq \cdots \leq \cdots \leq \tilde{\lambda}_1$. Note that the elements of the matrix $\widetilde{\mathbf{W}}$ are nonnegative, diagonals are zero, and each row of the matrix sums to 1. It is a stochastic matrix. That ensures that at least one of its eigenvalues is 1, and all other eigenvalues must be between -1 and 1. Thus, $-1 \leq \tilde{\lambda}_m < 0 < \tilde{\lambda}_1 = 1$.

Similarly, if λ_i's are the eigenvalues of $\mathbf{W}$, then these are finite and real. With the smallest, λ_m, and the largest λ_1, we get $-\infty < \lambda_m < 0 < \lambda_1 < \infty$.

We consider four spatially dependent random effects models with variance-covariance matrix $\sigma^2\mathbf{\Omega}(\rho)^{-1}$, defined through their associated p.d. "precision" matrices, depending on a spatial parameter ρ. These models are simultaneous autoregressive (SAR), conditional autoregressive (CAR), simple CAR (SCAR) and Leroux CAR (LCAR). For these models we have

$$\text{SAR: } \mathbf{\Omega}_2(\rho) = (\mathbf{I}_m - \rho\widetilde{\mathbf{W}})^{\mathrm{T}}(\mathbf{I}_m - \rho\widetilde{\mathbf{W}}), \quad \rho \in (-1, 1), \tag{4}$$

$$\text{SCAR: } \mathbf{\Omega}_3(\rho) = \mathbf{I}_m - \rho\mathbf{W}, \quad \rho \in (\lambda_m^{-1}, \lambda_1^{-1}), \tag{5}$$

$$\text{CAR: } \mathbf{\Omega}_4(\rho) = \mathbf{L} - \rho\mathbf{W}, \quad \rho \in (-1, 1), \tag{6}$$

$$\text{LCAR: } \mathbf{\Omega}_5(\rho) = \rho\mathbf{R} + (1 - \rho)\mathbf{I}_m, \quad \rho \in [0, 1). \tag{7}$$

For all the models the ranges of the parameter ρ are defined above so that the $\mathbf{\Omega}$ matrices are p.d. Even though we have used the same notation σ^2, ρ for the scale and the spatial parameters in all four models (see stage (b) of the HB model), neither they admit the same interpretations nor a combination of their values signifies equal variability and spatial strength of dependence across the models. Finally, the SAR, SCAR and LCAR models

include the traditional independent Fay-Herriot linking model as a special case.

3. The posterior distribution of the small area mean vector

We carry out inference for $\boldsymbol{\theta}$ by conditioning on $\mathbf{S} = \mathbf{s}$ from the HB model given in Section 2. Our approach is computing-based, we will use the Monte Carlo method to generate multiple copies of sample of $\boldsymbol{\theta}$ from its posterior pdf. We use the Hamiltonian Monte Carlo (HMC) algorithm to sample the posterior distribution, and we implement this algorithm using the `RStan` software package (see Stan Development Team (2018)). The samples for $\boldsymbol{\theta}$ from its posterior distribution will be meaningful provided the posterior distribution, $\pi(\boldsymbol{\theta}|\mathbf{s})$, is proper. In the Theorem below we provide a set of sufficient conditions for the propriety of $\pi(\boldsymbol{\theta}|\mathbf{s})$.

We now describe conditions for propriety of the posterior distributions under various spatial small area models given in (4)–(7). Let $I(\cdot)$ be the indicator function taking the value 1 when its argument is true and 0 otherwise. We first provide general conditions for the posterior propriety of the proposed models.

Theorem 1: For all the HB spatial models given above, and equations (2), and (4)–(7), the posterior probability density functions are proper if the following conditions hold for some positive constant $N > 0$:

(a) $\int_0^\infty g(\sigma^2) I(\sigma^2 \leq N) d\sigma^2 < \infty$.

(b) $\int_0^\infty (\sigma^2)^{-(r-p)/2} g(\sigma^2) I(\sigma^2 > N) d\sigma^2 < \infty$.

If $g(\cdot)$ is a proper pdf, then (a) holds true automatically, and (b) is satisfied if $r \geq p$. We explained earlier the obvious necessity of the condition $r \geq p$ since at least p summary statistics are needed to estimate p components of $\boldsymbol{\beta}$ when no substantive information about $\boldsymbol{\beta}$ is available. Note that for all the spatial models we have the conditions (a) and (b) for propriety of the respective posterior distribution. Under the popular family of noninformative priors

$$\pi(\boldsymbol{\beta}, \sigma^2, \rho) \propto (\sigma^2)^{-\alpha} I(l < \rho < u), \quad \boldsymbol{\beta} \in \mathbb{R}^p,\ \sigma^2 > 0, \tag{8}$$

the posterior pdfs are proper under the following conditions.

Corollary 1.1: For any of the HB spatial models given in (4)–(7) and with the prior in (8), the posterior pdf is proper as long as $\alpha < 1$ and $r > p + 2 - 2\alpha$.

For the uniform prior with $\alpha = 0$ (which is used in this paper), the propriety of the posterior distributions for models (4)–(7) are guaranteed as long as $r > p + 2$. We prove the Theorem in Section 6. The Corollary follows easily from the Theorem.

4. An illustration to a data from the current population survey

We illustrate our method to estimation of 1989 four-person family median incomes for the U.S. forty-eight mainland states and the Washington, DC. We consider this application

for two reasons. First, Chung and Datta (2022) used this application and applied the independent Fay-Herriot model and four spatial models to estimate the true median incomes, θ_i's, based on forty-nine direct estimates for these states coming from an annual supplement of the Current Population Survey (CPS). Second, a reliable set of values of these incomes are available from a large sample from the 1990 Census. Many SAE experts, for example, Ghosh *et al.* (1996) treat these values as "true values" or "gold standards" and assess accuracy of various sets of estimates against these values. The Census Bureau annually supplied accurate estimates of median incomes for states to the U.S. Department of Health and Human Service (HHS) agency that needed these estimates to implement a federal welfare program. The annual state-level estimates of these parameters from the CPS data are less reliable due to their large sampling standard deviations. To produce more reliable estimates the U.S. Census Bureau considered model-based small area estimation by using effective auxiliary data from other sources.

In our illustration for the four spatial and the independent Fay-Herriot model, we consider two types of mean functions, specified by the regression function $\mathbf{x}_i^T\boldsymbol{\beta}$. The most effective regression function involves both the covariates x_1 and x_2 that are introduced above. It has been found that x_2 has more predictive power in predicting θ_i,s than x_1. Here, x_1 is a weaker covariate and x_2 is a stronger covariate. We consider two regression functions: one with both the covariates (all covariates, $k = 1$), and the other with x_1 (the weaker covariate, $k = 2$).

Based on use of data types, we have full data case (Y_i's available for all areas, F) and aggregated data case (based on S_j's, A). Within each mean function and data type, we have fitted five versions of the Fay-Herriot model, resulting in a combination of 20 models and 20 sets of predictions of the θ_i's.

Our goal is to estimate θ_i, the true 1989 four-person family median income of the ith state, $i = 1, \ldots, 49$, excluding Alaska and Hawaii. From the 1990 CPS we get Y_i, the direct estimate of θ_i. The Census Bureau statistician Bob Fay found out that the corresponding 1980 Census median income figure (x_{i1}), and an adjusted 1980 Census median income x_{i2}, adjusted by per capita income data from 1979 and 1989, are two powerful covariates for prediction of θ_i. The CPS data also provided D_i, the sampling variance of Y_i. In our illustration we create a set of aggregated statistics $\mathbf{S}$ by grouping 49 states into 25 "super-areas", 24 groups of two states, and one lone state. In our illustration, we create required aggregated statistics by calculating $S_i = Y_{2i-1} + Y_{2i}, D_{Si} = D_{2i-1} + D_{2i}$, $i = 1, \ldots, 24$, and $S_{25} = Y_{49}$, $D_{S25} = D_{49}$. We apply five versions of the Fay-Herriot model mentioned above to this data and compare results from each of these models with the similar results presented by Chung and Datta (2022). We will also compare the five proposed models among themselves in terms of their prediction accuracy when we have only aggregated data but no data for the individual states.

4.1. Four-person family median income estimation with all covariates

We have twenty different settings formed by combination of five types of model variance matrices in the Fay-Herriot model, two linear regressions and two data types for the response. In our Bayesian analysis for these twenty settings we used uniform prior for the regression and variance parameters that appear in the corresponding model. For each setting

we used `Rstan` to generate 24000 representative, nearly independent, Monte Carlo samples of all the parameters from the respective posterior distribution. Based on the posterior samples for the jth model error variance type, kth mean function type, and the Tth data type, we computed Bayes estimate of θ_i, denoted by $\hat{\theta}_{T,j,k,i}$. We also compute the posterior standard deviation $\sigma_{T,j,k,i}$ associated with $\hat{\theta}_{T,j,k,i}$. We also computed summary and relative frequency histograms of the spatial parameters ρ for the models that have this parameter ($j = 2, 3, 4, 5$).

We use g_i as the gold standard for θ_i from the 1990 Census to empirically evaluate performance of $\hat{\theta}_{T,j,k,i}, i = 1, \ldots, m$, we compute for each set of predictions based on data type T, the empirical mean squared error $eMSPE_{T,j,k} = \sum_{i=1}^{49}(\hat{\theta}_{T,j,k,i} - g_i)^2/49$ for $j = 1, \ldots, 5$, $k = 1$ ($k = 2$ is considered in Subsection 4.2). These values for $k = 1$ are presented in the second column of Table 1 (for aggregated data), and Table 2 (for full data). We also computed average posterior standard deviations $\bar{\sigma}_{T,j,k} = \sum_{i=1}^{49} \sigma_{T,j,k,i}/49$, $T = A, F$. These values are given in the sixth column of the tables we created. Additionally, within each model, using appropriate posterior quantiles, we constructed 95% central credible interval for each θ_i. Using these intervals and the gold standard values we calculated empirical coverage rates of these intervals by computing the fraction of the 49 intervals that included the g_i values (presented in the fifth column). We also presented in the fourth column average length of these intervals.

In the absence of a direct estimate for a small area, a synthetic estimate based on the estimated regression function and covariates from that area is a reasonable alternative. In our case where we only have access to aggregated statistics based on data from multiple areas, we do not typically have direct estimates for any areas. In this scenario, synthetic estimates for all the areas may appear to be appealing. A synthetic estimate of θ_i for a typical model is $\hat{\theta}_{syn,T,j,k,i} = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{T,j,k}$, where $T = A, F$, $j = 1, \ldots, 5$, and $k = 1, 2$. Here, $\hat{\boldsymbol{\beta}}_{T,j,k}$ is a Bayes estimator of $\boldsymbol{\beta}$ under the T, j, kth setting. We note that all these results corresponding to $T = F$ for full data were obtained by Chung and Datta (2022).

At the early stage of small area estimation— pre-dating use of random effects or hierarchical models— practitioners used synthetic estimates. For the synthetic estimates we computed empirical MSPE by averaging the squares differences of the estimates from the gold standard values, g_i. We present these measures, represented as "syn MSPE" in the eighth column of the tables. Under any Bayesian model, the accuracy of corresponding synthetic estimates are evaluated by the posterior root mean squared error of each estimate. Averages of these values are reported as synthetic average root posterior mean squared error (syn ARPME) in the last column. Synthetic estimators usually tend to be biased, particularly if the regression function is an inadequate fit for the θ_i's, but they have smaller variances. In the case of poor model fit, the bias term of the synthetic estimator usually gets elevated, and the variance may fail to compensate for the larger bias, resulting in a large posterior MSE of synthetic estimator.

Both Table 1 and Table 2 show that the synthetic estimates for each model have smaller empirical MSPE than for their Bayesian counterparts. This is rather unusual unless the covariates are very effective, which appears to be the case here. However, the Bayes estimates have average posterior standard deviations in column 6 which are smaller than the average root posterior means squared error of their synthetic estimate counterparts,

Table 1: Aggregated data with all covariates

	eMSPE	eMSPE-PI	AL	CP	APSD	APSD-PI	syn MSPE	syn ARPME
FH	4.02	-	14.30	0.9509	3.15	-	2.10	3.47
SAR	3.87	3.70 %	14.21	0.9507	3.15	-0.07 %	2.12	3.70
SCAR	4.10	-2.03 %	14.26	0.9510	3.12	0.88 %	2.10	3.45
CAR	4.41	-9.72 %	14.62	0.9471	3.19	-1.37 %	2.13	3.56
LCAR	3.37	16.09 %	13.74	0.9489	3.09	1.82 %	2.42	3.98

empirical Mean squared prediction error (eMSPE), average posterior standard deviation (APSD), and respective percentage improvements (PI) of spatial models over the independent FH model for Bayes predictor of $\boldsymbol{\theta}$ and synthetic estimator $\mathbf{X}^T\hat{\boldsymbol{\beta}}$, and also average length (AL), coverage probability (CP).

Table 2: Full data from forty-nine states with all covariates

	eMSPE	eMSPE-PI	AL	CP	APSD	APSD-PI	syn MSPE	syn ARPME
FH	2.88	-	7.63	0.9592	1.93	-	1.86	2.57
SAR	2.61	9.55 %	7.58	0.9592	1.94	0.34%	2.00	2.74
SCAR	3.03	-5.14 %	7.66	0.9592	1.95	-0.91%	1.86	2.57
CAR	2.64	8.47 %	7.48	0.9592	1.91	1.24%	1.98	2.63
LCAR	2.47	14.50 %	7.31	0.9592	1.85	4.19%	2.37	3.02

Table 3: Posterior mean/mode (standard deviation) of ρ for various models ans data types.

Data type	Covariate included	SAR	SCAR	CAR	LCAR
Aggregated	x_1,x_2	-0.10 /-0.22(0.44)	-0.09 /-0.09(0.14)	-0.20 /-0.25(0.59)	0.47 / 0.22(0.28)
data	x_1	0.43 / 0.68(0.39)	0.02 / 0.17(0.13)	0.41 / 0.95(0.54)	0.71 / 0.98(0.24)
Full	x_1, x_2	0.10 / 0.38(0.48)	-0.06 / 0.11(0.14)	0.21 / 0.98(0.55)	0.57 / 0.78(0.27)
data	x_1	0.76 / 0.83(0.14)	0.14 / 0.18(0.04)	0.93 / 0.98(0.09)	0.85 / 0.98(0.13)

reported in column 9. Actually, by being the Bayes estimates, they will have smaller posterior means squared error. Since few applications provide any set of gold standards to compare estimates against, it is important to compare various estimates in terms of their variability or concentration.

Even in the presence of powerful predictors Table 1 showed that only the LCAR model emerged to be the best among the five sets (including the independent Fay-Herriot) in terms of eMSPE, AL and APSD. Other spatial models turned out to be less competitive or inferior to the independent FH model. If we turn to Table 2, even when we have direct estimates from all 49 states, the LCAR model still turned out to be the best of the five models in terms of the same measures. Among the other models, the SAR and CAR models also improved over the independent FH model.

To assess efficiency loss due to data compression through aggregation, we compare the results of Table 1 with those of Table 2. Across models the percentages increase in eMSPE's for the aggregated data over their counterparts for the full data, respectively, are 40, 48, 35, 67 and 36; the two smaller of the increases are for the SCAR and the LCAR models. We note that in both the tables that all the CP's are practically at the target 95%. When we compare the average length of the credible intervals, the percentage increases for the aggregated data across models over their counterparts for the full data, respectively, are 87, 87, 86, 95 and 88; this time, the smaller of the increases are for the models other than the

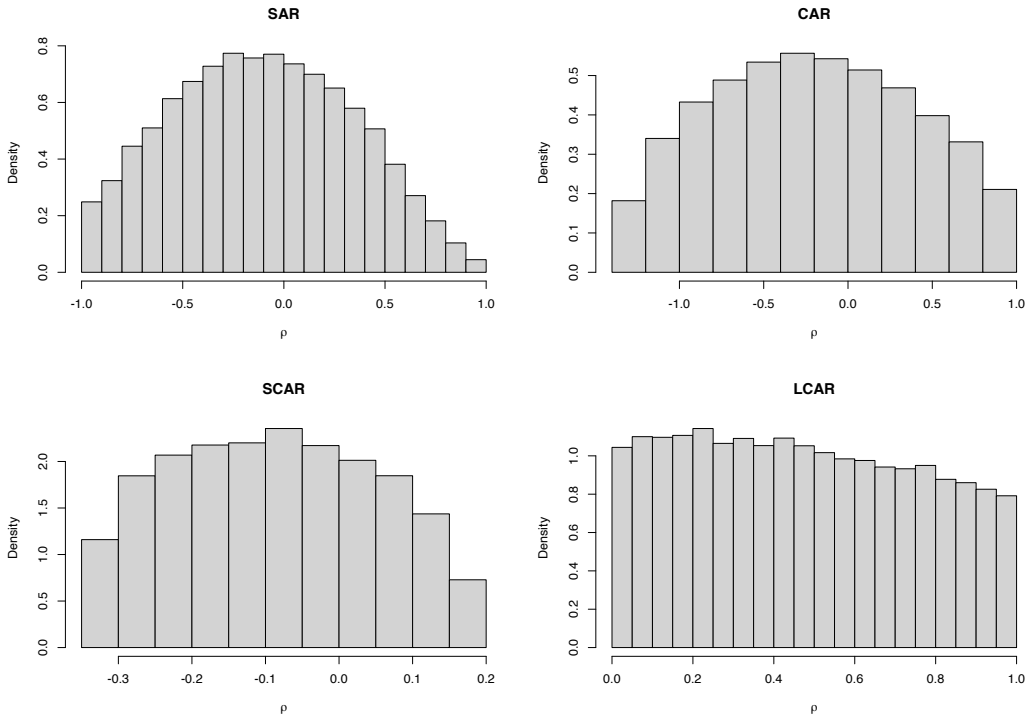


Figure 1: Posterior relative frequency histogram of ρ for aggregated data with all covariates

CAR. Among these spatial models, the LCAR model produced the smallest eMSPE and the AL values. Finally, the rows corresponding to x_1, x_2 in Table 3 show that for both data types the spatial parameter only for the LCAR model appears to be the one that is more likely to be non-zero. The same conclusion emerges about the spatial models from the posterior relative frequency histograms of ρ presented in Figures 1 and 2.

4.2. Four-person family median income estimation with the weaker covariate

Research shows that spatial random effects models tend to have better predictive power than a corresponding independent Fay-Herriot model when no effective covariates are available, see, for example, Chung and Datta (2022) and Vogt *et al.* (2023). For the median income estimation problem based on direct estimates from all 49 mainland states Chung and Datta (2022) showed that in the absence of any covariates some of the spatial models do better than the independent Fay-Herriot model. Usually, the SAR or the LCAR model provides the best prediction. In this section, we plan to investigate based on modeling of only aggregated statistics if any of the spatial models would be better than the independent Fay-Herriot model.

We note from Table 4 and Table 5 that for both aggregated data and full data cases in the absence of powerful predictors of θ_i's, all the spatial models provide better predictions than the independent FH model when compared in terms of eMSPE, AL and APSD. In this setting with low quality predictor, all synthetic estimators of θ_i's have bigger average MSPE's than their Bayesian counterparts (see columns 2 and 8). In this case, the LCAR is the best

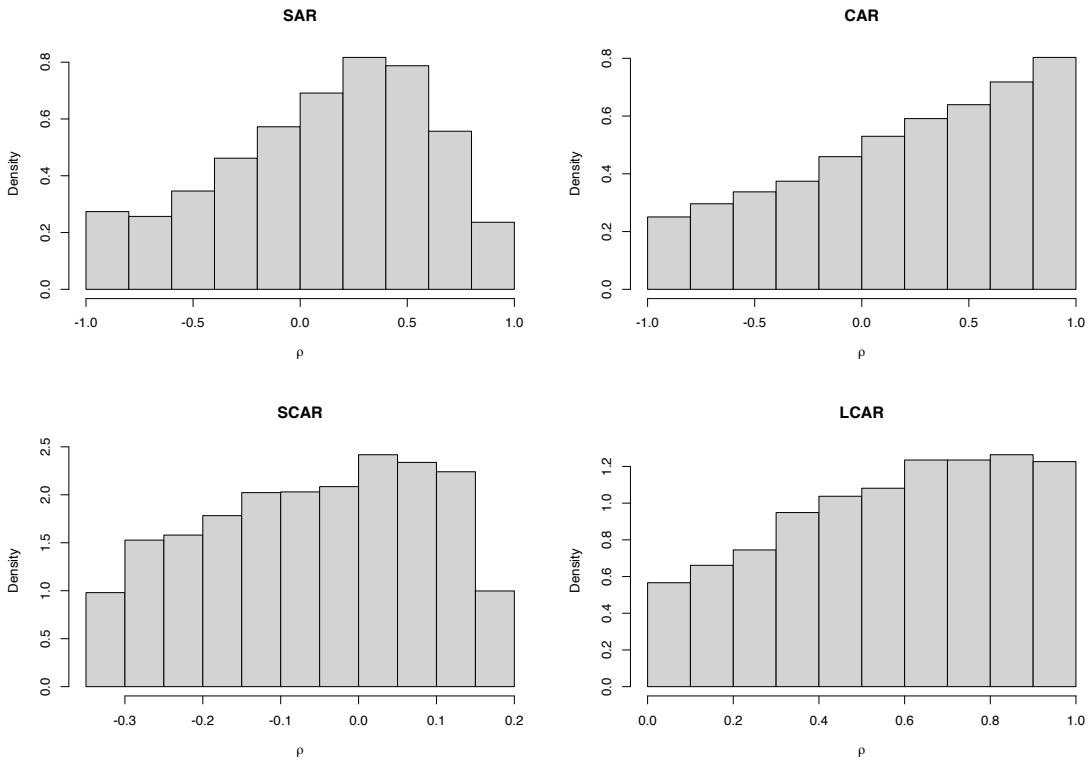


Figure 2: Posterior relative frequency histogram of ρ with all covariates and 49 states data

spatial model across both data settings. Results from Table 4 and those corresponding to the x_1 rows in Table 3 show that the LCAR is the best of the spatial models and the spatial parameter of this model appeared most likely to be different from zero. Moreover, from the two relative frequency histograms of ρ in Figure 3 and Figure 4 it is obvious that for the LCAR model 95% highest posterior density credible intervals of ρ will not include the zero value. For the full data case we also note from the last row of Table 3 and Figure 4 that the respective spatial parameter in all the spatial models appears very likely to be different from zero.

Before concluding Subsection 4.1 and Subsection 4.2, from a quick look at the APSD's for the independent FH model reported in the first row of Table 1, Table 2, Table 4 and Table 5, we found out that the average of the posterior variances of the θ_i's under the aggregated data setting is nearly three times that quantity under the full data case. This substantial increase in the posterior variances of θ_i under the aggregated data setting compared to the full data setting can reasonably be explained by the expression of the posterior variance of θ_i in equation (1) under the assumption of known model parameters.

Table 4: Aggregated data with a weaker covariate

	eMSPE	eMSPE-PI	AL	CP	APSD	APSD-PI	syn MSPE	syn ARPME
FH	11.78	-	20.47	0.9513	4.02	-	14.44	4.77
SAR	6.76	42.64%	17.53	0.9507	3.77	6.09%	14.29	6.63
SCAR	10.60	10.06%	19.88	0.9513	3.99	0.75%	14.73	4.84
CAR	8.52	27.67%	18.58	0.9476	3.85	4.08%	14.17	5.06
LCAR	6.03	48.80%	16.39	0.9480	3.48	13.31%	14.20	6.55

Table 5: Full data from forty-nine states with a weaker covariate

	eMSPE	eMSPE-PI	AL	CP	APSD	APSD-PI	syn MSPE	syn ARPME
FH	7.27	-	9.09	0.9388	2.31	-	14.45	4.04
SAR	4.34	40.22%	7.73	0.9796	1.98	14.25%	14.61	7.65
SCAR	5.62	22.62%	8.75	0.9592	2.22	3.52%	15.36	4.27
CAR	4.62	36.35%	7.84	0.9388	2.01	12.97%	14.70	5.06
LCAR	4.54	37.51%	7.77	0.9592	1.97	14.36%	14.67	6.17

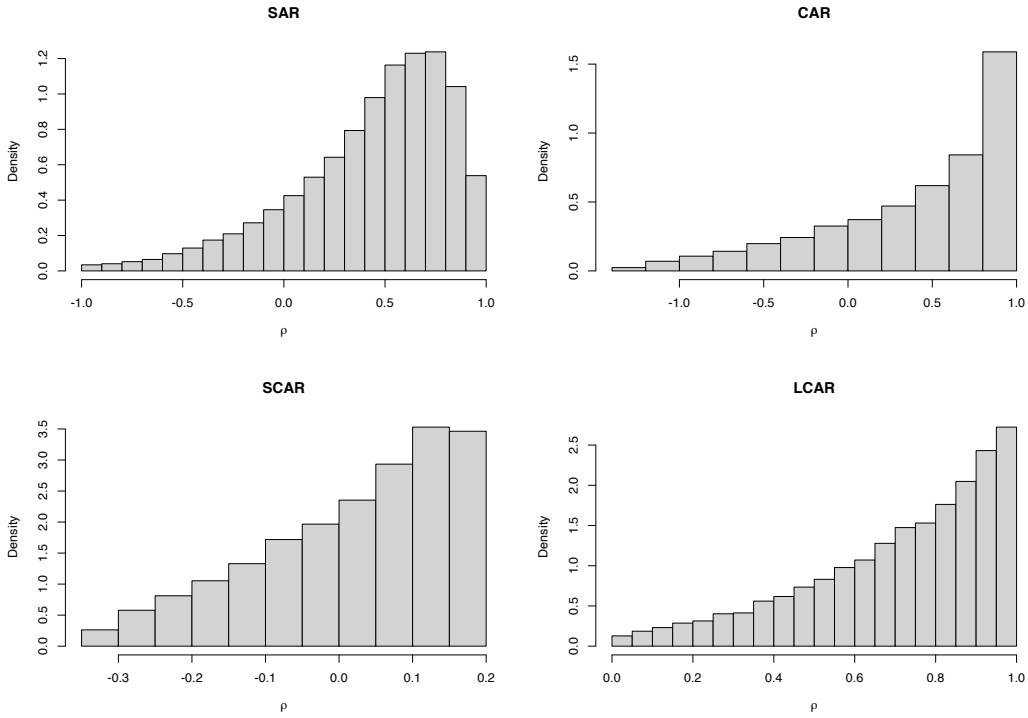


Figure 3: Posterior relative frequency histogram of ρ for aggregated data and a weaker covariate

5. Importance of the study

This study addresses an important problem in area-level small area estimation when most or all of the small areas do not have a direct estimates for θ_i's. Such data can not be had due to not having a survey that collects data from the individual areas. Due to administrative or budgetary considerations, a survey may do stratified sampling where each stratum is formed by merging multiple targeted small areas. If the goal is to estimating

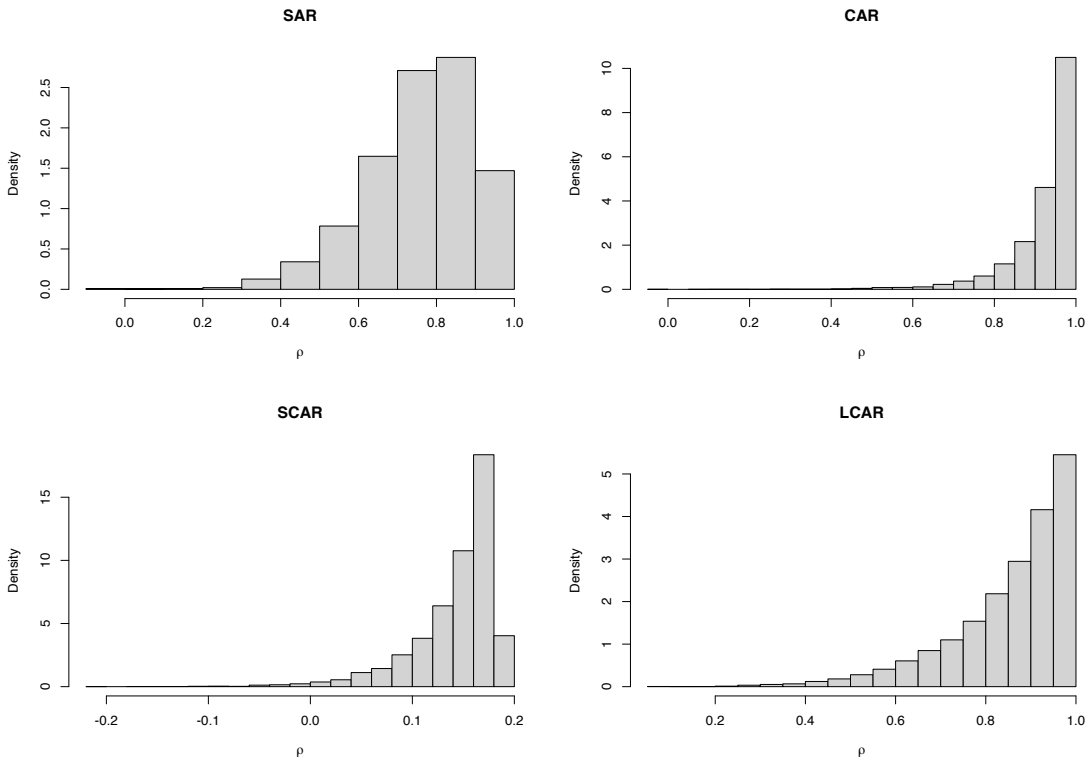


Figure 4: Posterior relative frequency histogram of ρ with a weaker covariate and 49 states data

total agricultural productions or total employments for the strata, our study shows that the stratified means can be leveraged to reliably estimate the means for the original small areas by integrating the strata level means of a response variable with area-level data from covariates that have good predictive power to predict the small area means for the response.

For the setup we are considering here, the success of a generalization of the Fay-Herriot model depends on the availability of effective predictor variables for the response variable. In the absence of effective covariates, from the studies by Chung and Datta (2022) and Vogt *et al.* (2023), it is known that various spatial alternatives to the independent Fay-Herriot model produce significantly better predictions by accounting for the spatial variation of the small area means. Even when no substantial spatial variation exists among the means, the spatial models make marginally better predictions than the independent FH model without sacrificing model fit. We demonstrated the usefulness and the strength of our proposed method by applying this to an application that has been important to both the HHS Department and the Census Bureau of the United States.

6. Proof of propriety of the posterior pdfs

We know that our vector of aggregated statistics $\mathbf{S}$ is $r \times 1$ with $r \geq p$. We assume that

$$\mathbf{S}|\boldsymbol{\theta} \sim N(\mathbf{C}\boldsymbol{\theta}, \mathbf{D}_S), \tag{9}$$

where $\mathbf{C}$ is a known $r \times m$ matrix of rank r, $\boldsymbol{\theta}$ is an $m \times 1$ vector, and $\mathbf{D}_S$ is a known positive definite (p.d.) matrix of rank r.

Suppose the largest eigenvalue of $\mathbf{D}_S$ is δ, which is finite and positive. Let $N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the multivariate normal pdf with mean $\boldsymbol{\mu}$ and p.d. variance-covariance matrix $\boldsymbol{\Sigma}$ at $\mathbf{x}$. Since $\delta^{-1} > 0$ is the smallest eigenvalue of $\mathbf{D}_S^{-1}$, from the property of the minimum eigenvalue we get that

$$\begin{aligned} &(\mathbf{s} - \mathbf{C}\boldsymbol{\theta})^T \mathbf{D}_S^{-1} (\mathbf{s} - \mathbf{C}\boldsymbol{\theta}) \geq \delta^{-1} (\mathbf{s} - \mathbf{C}\boldsymbol{\theta})^T (\mathbf{s} - \mathbf{C}\boldsymbol{\theta}) \\ &\Rightarrow N(\mathbf{s}|\mathbf{C}\boldsymbol{\theta}, \mathbf{D}_S) \leq K N(\mathbf{s}|\mathbf{C}\boldsymbol{\theta}, \delta \mathbf{I}_r), \end{aligned} \tag{10}$$

where $K > 0$ is a generic known suitable constant, dependent on $\mathbf{D}_S$ but free from $\mathbf{s}$ or $\boldsymbol{\theta}$.

We can select a matrix $\mathbf{F}((m-r) \times m)$, dependent on $\mathbf{C}$ but known so that the $m \times m$ matrix $\mathbf{M} = (\mathbf{C}^T, \mathbf{F}^T)^T$ is non-singular. This implies that the rank of $\mathbf{F}$ is $m - r$. For an $(m - r) \times 1$ vector $\mathbf{h}_2$ note that

$$\int_{R^{m-r}} N(\mathbf{h}_2|\mathbf{F}\boldsymbol{\theta}, \delta \mathbf{I}_{m-r}) d\mathbf{h}_2 = K < \infty, \tag{11}$$

where K is a generic and positive constant. By (10)-(11) we get that

$$\begin{aligned} N(\mathbf{s}|\mathbf{C}\boldsymbol{\theta}, \mathbf{D}_S) &\leq K N(\mathbf{s}|\mathbf{C}\boldsymbol{\theta}, \delta \mathbf{I}_r) \int_{R^{m-r}} N(\mathbf{h}_2|\mathbf{F}\boldsymbol{\theta}, \delta \mathbf{I}_{m-r}) d\mathbf{h}_2 \\ &= K \int_{R^{m-r}} N(\mathbf{h}|\mathbf{M}\boldsymbol{\theta}, \delta \mathbf{I}_m) d\mathbf{h}_2, \end{aligned} \tag{12}$$

where $\mathbf{h} = (\mathbf{s}^T, \mathbf{h}_2^T)^T$ is an $m \times 1$ vector.

Let $\mathbf{M}^{-1} = \mathbf{B}$. Let k be the smallest eigenvalue of the p.d. matrix $\mathbf{M}^T\mathbf{M}$. Using $\mathbf{h} - \mathbf{M}\boldsymbol{\theta} = \mathbf{M}(\mathbf{B}\mathbf{h} - \boldsymbol{\theta})$ we get

$$\begin{aligned} (\mathbf{h} - \mathbf{M}\boldsymbol{\theta})^T (\mathbf{h} - \mathbf{M}\boldsymbol{\theta}) &= (\mathbf{B}\mathbf{h} - \boldsymbol{\theta})^T \mathbf{M}^T \mathbf{M} (\mathbf{B}\mathbf{h} - \boldsymbol{\theta}) \\ &\geq k (\mathbf{B}\mathbf{h} - \boldsymbol{\theta})^T (\mathbf{B}\mathbf{h} - \boldsymbol{\theta}). \end{aligned}$$

From the above, using $k > 0$, we get that

$$N(\mathbf{h}|\mathbf{M}\boldsymbol{\theta}, \delta \mathbf{I}_m) \leq K N(\mathbf{B}\mathbf{h}|\boldsymbol{\theta}, \delta k^{-1} \mathbf{I}_m). \tag{13}$$

By (12)-(13), writing $\delta k^{-1} = \delta^*$, we get

$$\begin{aligned} N(\mathbf{s}|\mathbf{C}\boldsymbol{\theta}, \mathbf{D}_S) &\leq K \int_{R^{m-r}} N(\mathbf{h}|\mathbf{M}\boldsymbol{\theta}, \delta \mathbf{I}_m) d\mathbf{h}_2 \\ &\leq K \int_{R^{m-r}} N(\mathbf{B}\mathbf{h}|\boldsymbol{\theta}, \delta^* \mathbf{I}_m) d\mathbf{h}_2. \end{aligned} \tag{14}$$

Recall that for the class of spatial models, the *linking model* is given by

$$\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2, \rho \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{\Omega}^{-1}), \tag{15}$$

where $\mathbf{X}$ is a known $m \times p$ matrix of covariates of rank p, and $\mathbf{\Omega}$ is an $m \times m$ p.d. matrix that depends on a parameter ρ which varies on a known finite interval.

Let $f_{\mathbf{S}}(\mathbf{s}|\boldsymbol{\beta}, \sigma^2, \mathbf{\Omega}) = \int_{R^m} N(\mathbf{s}|\mathbf{C}\boldsymbol{\theta}, \mathbf{D}_S) N(\boldsymbol{\theta}|\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{\Omega}^{-1}) d\boldsymbol{\theta}$ be the pdf of $\mathbf{S}$ given $\boldsymbol{\beta}, \sigma^2, \rho$. Then from (14) we get

$$\begin{aligned} f_{\mathbf{S}}(\mathbf{s}|\boldsymbol{\beta}, \sigma^2, \mathbf{\Omega}) &\leq K \int_{R^m} \int_{R^{m-r}} N(\mathbf{Bh}|\boldsymbol{\theta}, \delta^* \mathbf{I}_m) N(\boldsymbol{\theta}|\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{\Omega}^{-1}) d\mathbf{h}_2 d\boldsymbol{\theta} \\ &= K \int_{R^{m-r}} N(\mathbf{Bh}|\mathbf{X}\boldsymbol{\beta}, \delta^* \mathbf{I}_m + \sigma^2\mathbf{\Omega}^{-1}) d\mathbf{h}_2 \\ &= K \int_{R^{m-r}} N(\mathbf{Bh} - \mathbf{X}\boldsymbol{\beta}|\mathbf{0}, \delta^* \mathbf{I}_m + \sigma^2\mathbf{\Omega}^{-1}) d\mathbf{h}_2. \end{aligned} \tag{16}$$

Now, $\mathbf{Bh} - \mathbf{X}\boldsymbol{\beta} = \mathbf{B}(\mathbf{h} - \mathbf{MX}\boldsymbol{\beta})$. Let $\mathbf{d} = (\mathbf{h}_1^T, \mathbf{0}^T)^T$ and

$$\mathbf{G} = \begin{bmatrix} \mathbf{CX} & \mathbf{0} \\ \mathbf{FX} & -\mathbf{I}_{m-r} \end{bmatrix}.$$

Then, we have

$$\mathbf{h} - \mathbf{MX}\boldsymbol{\beta} = \mathbf{d} - \mathbf{G}\boldsymbol{\phi}, \tag{17}$$

where $\boldsymbol{\phi} = (\boldsymbol{\beta}^T, \mathbf{h}_2^T)^T$ is a $(p + m - r) \times 1$ vector. Now, define submatrices $\mathbf{G}_1$ and $\mathbf{G}_2$ to introduce a column partition of the matrix $\mathbf{G}$, where $\mathbf{G}_1$ is given by the first p columns of $\mathbf{G}$, and $\mathbf{G}_2$ is given by the last $m - r$ columns of $\mathbf{G}$. Columns of $\mathbf{G}_2$ are linearly independent. So $rank(\mathbf{G}_2) = m - r$. Also, since we require $\mathbf{CX} \neq \mathbf{0}$, the columns of $\mathbf{G}_1$ cannot be linearly expressed by the columns of $\mathbf{G}_2$. However, $\mathbf{G}_1 = \mathbf{MX}$ implies $rank(\mathbf{G}_1) = rank(\mathbf{X}) = p$. Hence, $rank(\mathbf{G}) = rank(\mathbf{G}_1) + rank(\mathbf{G}_2) = p + m - r$.

Let $\mathbf{Bd} = \mathbf{d}_*$, $\mathbf{BG} = \mathbf{G}_*$. Then, by (17)

$$\mathbf{Bh} - \mathbf{X}\boldsymbol{\beta} = \mathbf{B}(\mathbf{h} - \mathbf{MX}\boldsymbol{\beta}) = \mathbf{B}(\mathbf{d} - \mathbf{G}\boldsymbol{\phi}) = \mathbf{d}_* - \mathbf{G}_*\boldsymbol{\phi}. \tag{18}$$

Using (16) and (18) we get,

$$f_{\mathbf{S}}(\mathbf{s}|\boldsymbol{\beta}, \sigma^2, \mathbf{\Omega}) \leq K \int_{R^{m-r}} N(\mathbf{d}_* - \mathbf{G}_*\boldsymbol{\phi}|\mathbf{0}, \delta^* \mathbf{I}_m + \sigma^2\mathbf{\Omega}^{-1}) d\mathbf{h}_2. \tag{19}$$

Further,

$$\begin{aligned} & N(\mathbf{d}_* - \mathbf{G}_*\boldsymbol{\phi}|\mathbf{0}, \delta^* \mathbf{I}_m + \sigma^2\mathbf{\Omega}^{-1}) = K|\delta^* \mathbf{I}_m + \sigma^2\mathbf{\Omega}^{-1}|^{-1/2} \\ \times \quad & \exp[-\frac{(\mathbf{d}_* - \mathbf{G}_*\boldsymbol{\phi})^T (\delta^* \mathbf{I}_m + \sigma^2\mathbf{\Omega}^{-1})^{-1} (\mathbf{d}_* - \mathbf{G}_*\boldsymbol{\phi})}{2}]. \end{aligned} \tag{20}$$

We consider four spatially dependent random effects models with variance-covariance matrix $\sigma^2\mathbf{\Omega}(\rho)^{-1}$, defined through their associated p.d. "precision" matrices, depending on a spatial parameter ρ: for all the models the parameter ρ is defined on an appropriate *finite* interval so that the $\mathbf{\Omega}$ matrices are p.d.

To continue our propriety proof, for convenience of notation, we denote $\mathbf{\Omega}_k(\rho)$ by $\mathbf{\Omega}_k$, for $k = 1, \ldots, 5$. Here, $\mathbf{\Omega}_1(\rho) = \mathbf{I}_m$ is for the independent Fay-Herriot model. In the next

two subsections we present detail arguments establishing the propriety of the posterior pdfs for the SCAR and the SAR models. Under the same conditions, similar arguments can be made for proving the propriety of the posterior pdfs for the CAR and the LCAR models; see also Appendices A.3 and A.4 of Chung and Datta (2022). Result for the independent model follows from the SAR or the SCAR model with $\rho = 0$.

Finally, suppose $\mathbf{C} = \begin{bmatrix} \mathbf{0} & \mathbf{I}_r \end{bmatrix}$, an $r \times m$ matrix. This is a special case of the general setup considered in this paper. This special case was considered in Chung and Datta (2022).

6.1. The propriety for the SCAR model

For the eigenvalues λ_i's of $\mathbf{W}$, let $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_m)$ and $\mathbf{P}_W$ be an orthogonal matrix such that $\mathbf{W} = \mathbf{P}_W \mathbf{\Lambda} \mathbf{P}_W^T$. For the SCAR model, $\mathbf{\Omega} = \mathbf{\Omega}_3$ and

$$\mathbf{\Omega}_3^{-1} = \mathbf{P}_W [\mathbf{I}_m - \rho \mathbf{\Lambda}]^{-1} \mathbf{P}_W^T. \tag{21}$$

From this we get

$$(\delta^* \mathbf{I}_m + \sigma^2 \mathbf{\Omega}_3^{-1})^{-1} = \mathbf{P}_W [\delta^* \mathbf{I}_m + \sigma^2 \{\mathbf{I}_m - \rho \mathbf{\Lambda}\}^{-1}]^{-1} \mathbf{P}_W^T,$$

which implies that

$$\begin{aligned} & (\mathbf{d}_* - \mathbf{G}_* \boldsymbol{\phi})^T (\delta^* \mathbf{I}_m + \sigma^2 \mathbf{\Omega}_3^{-1})^{-1} (\mathbf{d}_* - \mathbf{G}_* \boldsymbol{\phi}) \\ = \; & (\mathbf{d}_{**} - \mathbf{G}_{**} \boldsymbol{\phi})^T [\delta^* \mathbf{I}_m + \sigma^2 \{\mathbf{I}_m - \rho \mathbf{\Lambda}\}^{-1}]^{-1} (\mathbf{d}_{**} - \mathbf{G}_{**} \boldsymbol{\phi}) \\ = \; & \sum_{i=1}^{m} \frac{\{d_{**i} - \mathbf{g}_{**i}^T \boldsymbol{\phi}\}^2}{\delta^* + \sigma^2 (1 - \rho \lambda_i)^{-1}} \end{aligned} \tag{22}$$

where $\mathbf{d}_{**} = \mathbf{P}_W^T \mathbf{d}_*$, $\mathbf{G}_{**} = \mathbf{P}_W^T \mathbf{G}_*$, d_{**i} is the ith element of $\mathbf{d}_{**}$ and $\mathbf{g}_{**i}^T$ is the ith row of $\mathbf{G}_{**}$.

Clearly,

$$rank(\mathbf{G}_{**}) = rank(\mathbf{G}_*) = rank(\mathbf{G}) = p + m - r = q \text{ (say)}.$$

We can select q linearly independent rows of $\mathbf{G}_{**}$. By rearrangement of those rows we can assume that the first q rows of $\mathbf{G}_{**}$ could be taken as linearly independent. Then from (22) we get

$$(\mathbf{d}_* - \mathbf{G}_* \boldsymbol{\phi})^T (\delta^* \mathbf{I}_m + \sigma^2 \mathbf{\Omega}_3^{-1})^{-1} (\mathbf{d}_* - \mathbf{G}_* \boldsymbol{\phi}) \geq \sum_{i=1}^{q} \frac{\{d_{**i} - \mathbf{g}_{**i}^T \boldsymbol{\phi}\}^2}{\delta^* + \sigma^2 (1 - \rho \lambda_i)^{-1}}.$$

Using this inequality in equations (19)-(20) we get

$$\begin{aligned}
&\int_{R^p} f_{\mathbf{S}}(\mathbf{s}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\Omega})\pi(\boldsymbol{\beta})d\boldsymbol{\beta} \le K|\delta^*\mathbf{I}_m+\sigma^2\boldsymbol{\Omega}^{-1}|^{-1/2}\\
\times\ &\int\int \pi(\boldsymbol{\beta})\exp[-\frac{(\mathbf{d}_*-\mathbf{G}_*\boldsymbol{\phi})^T(\delta^*\mathbf{I}_m+\sigma^2\boldsymbol{\Omega}_3^{-1})^{-1}(\mathbf{d}_*-\mathbf{G}_*\boldsymbol{\phi})}{2}]d\mathbf{h}_2 d\boldsymbol{\beta}\\
\le\ & K\prod_{i=1}^{m}\{\delta^*+\sigma^2(1-\rho\lambda_i)^{-1}\}^{-1/2}\\
&\times\int\int \pi(\boldsymbol{\beta})\exp[-\frac{1}{2}\sum_{i=1}^{q}\frac{\{d_{**i}-\mathbf{g}_{**i}^T\boldsymbol{\phi}\}^2}{\delta^*+\sigma^2(1-\rho\lambda_i)^{-1}}]d\boldsymbol{\beta}d\mathbf{h}_2\\
\le\ & K\prod_{i=1}^{m}\{\delta^*+\sigma^2(1-\rho\lambda_i)^{-1}\}^{-1/2}\\
\times\ &\int_{R^q}\exp[-\frac{1}{2}\sum_{i=1}^{q}\frac{\{d_{**i}-\mathbf{g}_{**i}^T\boldsymbol{\phi}\}^2}{\delta^*+\sigma^2(1-\rho\lambda_i)^{-1}}]d\boldsymbol{\phi}\\
=\ & K\prod_{i=q+1}^{m}\{\delta^*+\sigma^2(1-\rho\lambda_i)^{-1}\}^{-1/2}, \qquad (23)
\end{aligned}$$

where we assumed that $\pi(\boldsymbol{\beta})$ is bounded above, which is satisfied by a uniform prior on R^p.

Now, we notice that for any positive constant N

$$\begin{aligned}
\delta^*+\sigma^2(1-\rho\lambda_i)^{-1} &\ge \delta^*I(\sigma^2\le N)+\sigma^2(1-\rho\lambda_i)^{-1}I(\sigma^2>N)\\
\Rightarrow \{\delta^*+\sigma^2(1-\rho\lambda_i)^{-1}\}^{-1/2} &\le KI(\sigma^2\le N)+(\sigma^2)^{-1/2}(1-\rho\lambda_i)^{1/2}I(\sigma^2>N). \qquad (24)
\end{aligned}$$

Since λ_i's are finite and ρ is integrated over a finite interval, it follows that $1-\rho\lambda_i$ is a finite positive quantity. Then, using $q=m-r+p$,

$$\begin{aligned}
\prod_{i=q+1}^{m}\{\delta^*+\sigma^2(1-\rho\lambda_i)^{-1}\}^{-1/2} &\le K[I(\sigma^2\le N)+(\sigma^2)^{-(m-q)/2}I(\sigma^2>N)]\\
&\le K[I(\sigma^2\le N)+(\sigma^2)^{-(r-p)/2}I(\sigma^2>N)]. \qquad (25)
\end{aligned}$$

Now using $\pi(\sigma^2,\rho)=g(\sigma^2)h(\rho)$ and that $h(\rho)$ is a pdf, we get

$$\begin{aligned}
&\int_0^\infty\int_l^u\prod_{i=q+1}^{m}\{\delta^*+\sigma^2(1-\rho\lambda_i)^{-1}\}^{-1/2}g(\sigma^2)h(\rho)d\rho d\sigma^2\\
\le\ & K\int_0^N g(\sigma^2)d\sigma^2+K\int_N^\infty g(\sigma^2)(\sigma^2)^{-(r-p)/2}d\sigma^2<\infty, \qquad (26)
\end{aligned}$$

by sufficient conditions (a) and (b) in Theorem 1.

In particular, if $g(\sigma^2)=(\sigma^2)^{-\alpha}$, $1-\alpha>0$ ensures (a), and $(r-p)/2+\alpha>1$ ensures (b). Equivalently, we need $\alpha<1$ and $r>p+2-2\alpha$.

6.2. The propriety of the posterior pdf under the SAR model

We now consider the SAR model. For this model

$$\mathbf{\Omega}_2(\rho) = (\mathbf{I}_m - \rho\widetilde{\mathbf{W}})^{\mathrm{T}}(\mathbf{I}_m - \rho\widetilde{\mathbf{W}}),\ -1 < \rho < 1.$$

Note that $\mathrm{tr}[\mathbf{\Omega}_2(\rho)] = m + \rho^2 \sum\sum \tilde{w}_{ij}^2 \leq 2m$. Let $\mathbf{W}_* = \mathbf{L}^{-1/2}\mathbf{W}\mathbf{L}^{-1/2}$. Again,

$$\mathbf{\Omega}_2(\rho) = \mathbf{L}^{1/2}(\mathbf{I} - \rho\mathbf{W}_*)\mathbf{L}^{-1}(\mathbf{I} - \rho\mathbf{W}_*)\mathbf{L}^{1/2}. \tag{27}$$

Let $\nu_1 \geq \cdots \geq \nu_m$ be the eigenvalues of $\mathbf{W}_*$. From our discussions in Subsection 2.1 all ν_i's are real. Moreover, $\nu_1 = 1$ and $|\nu_i| \leq 1$. Since $1 - \rho\nu_i$ are the eigenvalues of $\mathbf{I} - \rho\mathbf{W}_*$, for $-1 < \rho < 1$, these eigenvalues are all positive. Hence the matrix is p.d. Actually, for all i, $0 < 1 - \rho\nu_i < 2$.

Let $l_{(1)} = \min W_{i\cdot}$ and $l_{(m)} = \max W_{i\cdot}$. Note that $1 \leq l_{(1)} \leq l_{(m)} < m$. Define the matrix

$$\mathbf{H} = \delta^*\mathbf{L} + \sigma^2(\mathbf{I} - \rho\mathbf{W}_*)^{-1}\mathbf{L}(\mathbf{I} - \rho\mathbf{W}_*)^{-1}.$$

Since the matrix $\mathbf{L} - \mathbf{I}$ is n.n.d., the matrix $\mathbf{H} - \{\delta^*\mathbf{I} + \sigma^2(\mathbf{I} - \rho\mathbf{W}_*)^{-2}\}$ is n.n.d. It easily follows that

$$|\mathbf{H}| \geq |\delta^*\mathbf{I} + \sigma^2(\mathbf{I} - \rho\mathbf{W}_*)^{-2}| = \prod_{i=1}^{m}\{\delta^* + \sigma^2(1 - \rho\nu_i)^{-2}\}.$$

Let $\mathbf{\Sigma}_2 = \delta^*\mathbf{I} + \sigma^2\mathbf{\Omega}_2^{-1}$. Note that $\mathbf{\Sigma}_2 = \mathbf{L}^{-1/2}\mathbf{H}\mathbf{L}^{-1/2}$, and $|\mathbf{L}| < m^m$. Using these, and if we use K to denote a suitable finite, positive and generic constant, not depending on any parameters, we get that

$$|\mathbf{\Sigma}_2|^{-1/2} \leq K\prod_{i=1}^{m}\{\delta^* + \sigma^2(1 - \rho\nu_i)^{-2}\}^{-1/2}. \tag{28}$$

Let $\mathbf{P}_{\mathbf{W}_*}$ be the matrix of eigenvectors of $\mathbf{W}_*$ such that $\mathbf{P}_{\mathbf{W}_*}^T\mathbf{W}_*\mathbf{P}_{\mathbf{W}_*} = \mathrm{diag}(\nu_1, \ldots, \nu_m) = \mathbf{N}_*$. Then,

$$\begin{aligned}
& (\mathbf{d}_* - \mathbf{G}_*\phi)^T(\delta^*\mathbf{I}_m + \sigma^2\mathbf{\Omega}_2^{-1})^{-1}(\mathbf{d}_* - \mathbf{G}_*\phi) \\
= \ & (\mathbf{L}^{1/2}\mathbf{d}_* - \mathbf{L}^{1/2}\mathbf{G}_*\phi)^T\mathbf{H}^{-1}(\mathbf{L}^{1/2}\mathbf{d}_* - \mathbf{L}^{1/2}\mathbf{G}_*\phi) \\
\geq \ & l_{(m)}^{-1}(\mathbf{r} - \mathbf{F}\boldsymbol{\phi})^T\{\delta^*\mathbf{I} + \sigma^2(\mathbf{I} - \rho\mathbf{W}_*)^{-2}\}^{-1}(\mathbf{r} - \mathbf{F}\boldsymbol{\phi}) \\
= \ & (\tilde{\mathbf{r}} - \tilde{\mathbf{S}}\boldsymbol{\phi})^T\{\delta^*\mathbf{I} + \sigma^2(\mathbf{I} - \rho\mathbf{N}_*)^{-2}\}^{-1}(\tilde{\mathbf{r}} - \tilde{\mathbf{S}}\boldsymbol{\phi}),
\end{aligned} \tag{29}$$

where $\mathbf{r} = \mathbf{L}^{1/2}\mathbf{d}_*$, $\mathbf{F} = \mathbf{L}^{1/2}\mathbf{G}_*$, $\tilde{\mathbf{r}} = l_{(m)}^{-1/2}\mathbf{P}_{\mathbf{W}_*}\mathbf{r}$, and $\tilde{\mathbf{S}} = l_{(m)}^{-1/2}\mathbf{P}_{\mathbf{W}_*}\mathbf{F}$.

Suppose $\{i_1, \ldots, i_q\}$ is a subset of $\{1, \ldots, m\}$ so that the matrix $\tilde{\mathbf{S}}_1$ formed by plucking the rows of $\tilde{\mathbf{S}}$ corresponding to the indices $\{i_1, \ldots, i_q\}$ is non-singular. Note that this matrix is determined by $\mathbf{W}$.

From (29) we get

$$(\mathbf{d}_* - \mathbf{G}_*\phi)^T(\delta^*\mathbf{I}_m + \sigma^2\mathbf{\Omega}_2^{-1})^{-1}(\mathbf{d}_* - \mathbf{G}_*\phi) \geq \sum_{j=1}^{q}\frac{(\tilde{r}_{i_j} - \tilde{\mathbf{s}}_{i_j}^T\boldsymbol{\phi})^2}{\delta^* + \sigma^2(1 - \rho\nu_{i_j})^{-2}}. \tag{30}$$

Using equations (29)-(30) we get

$$\begin{aligned}
&\int \pi(\beta)\exp[-\frac{(\mathbf{d}_* - \mathbf{G}_*\phi)^T(\delta^*\mathbf{I}_m + \sigma^2\mathbf{\Omega}_2^{-1})^{-1}(\mathbf{d}_* - \mathbf{G}_*\phi)}{2}]d\boldsymbol{\phi} \\
\leq\ & K\int \exp[-\frac{1}{2}\sum_{j=1}^{q}\frac{(\tilde{r}_{i_j} - \tilde{\mathbf{s}}_{i_j}^T\boldsymbol{\phi})^2}{\delta^* + \sigma^2(1-\rho\nu_{i_j})^{-2}}]d\boldsymbol{\phi} \\
=\ & K\prod_{i=1}^{q}\{\delta^* + \sigma^2(1-\rho\nu_{i_j})^{-2}\}^{1/2}. \qquad (31)
\end{aligned}$$

Hence we get

$$\begin{aligned}
&|\mathbf{\Sigma}_2|^{-1/2}\int \pi(\beta)\exp[-\frac{(\mathbf{d}_* - \mathbf{G}_*\phi)^T(\delta^*\mathbf{I}_m + \sigma^2\mathbf{\Omega}_2^{-1})^{-1}(\mathbf{d}_* - \mathbf{G}_*\phi)}{2}]d\boldsymbol{\phi} \\
\leq\ & K\prod_{i\notin\{i_1,\ldots,i_q\}}\{\delta_* + \sigma^2(1-\rho\nu_i)^{-2}\}^{1/2} \\
\leq\ & K[I(\sigma^2 \leq N) + I(\sigma^2 > N)(\sigma^2)^{-(m-q)/2}\prod_{i\notin\{i_1,\ldots,i_q\}}(1-\rho\nu_i)] \\
\leq\ & K[I(\sigma^2 \leq N) + I(\sigma^2 > N)(\sigma^2)^{-(r-p)/2}], \qquad (32)
\end{aligned}$$

where we use the facts that $-1 < \rho < 1$ and $-1 \leq \nu_i \leq 1$ to claim that $0 < 1 - \rho\nu_i < 2$ for all i. From equation (32) if we continue our proof along the lines of the proof for the SCAR model, we will get the propriety of the posterior pdf for the SAR model under the same conditions.

Disclaimer and acknowledgments

This research was funded by a grant from the United States Department of State (SSJTIP18CA0015). This report is released to inform interested parties of ongoing research and to encourage discussion. The views, opinions, findings and conclusions expressed on statistical, methodological, technical, or operational issues herein are those of the authors and not those of the United States Department of State, the U.S. Census Bureau, Wells Fargo Bank, or the University of Georgia.

References

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28–36.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.

Chung, H. C. and Datta, G. S. (2022). Bayesian spatial models for estimating means of sampled and non-sampled small areas. *Survey Methodology*, **48**, 463–489.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269–277.

Ghosh, M. (1992). Hierarchical and empirical Bayes multivariate estimation. In Ghosh, M. and Pathak, P. K., editors, *Current Issues in Statistical Inference: Essays in honor of D. Basu*, pages 151–177. Institute of Mathematical Statistics.

Ghosh, M., Nangia, N., and Kim, D. H. (1996). Estimation of median income of four-person families: a Bayesian time series approach. *Journal of the American Statistical Association*, **91**, 1423–1431.

Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation*. John Wiley & Sons.

Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.17.3.

Vogt, M., Lahiri, P., and Munnich, R. (2023). Spatial prediction in small area estimation. *Statistics in Transition new series*, **24**.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 471–483
http://www.ssca.org.in/journal

On Retrieving Multivariate Data Sets from Their Moments

Serge B. Provost[1], **S. Ejaz Ahmed**[2] **and Zhaoqi Yang**[1]

[1]*Department of Statistical and Actuarial Sciences, The University of Western Ontario, London, Canada*
[2]*Department of Mathematics and Statistics, Brock University, Saint Catharines, Canada*

Received: 12 June 2024; Revised: 15 September 2024; Accepted: 23 September 2024

Abstract

This paper introduces several methodologies that solve the inverse problem of recovering a multivariate sample from subsets of its associated marginal and joint integer moments. These results rely in part on their univariate counterpart, which is examined in some detail. It is also explained that some of them also apply to complex-valued data sets. Several illustrative examples are presented.

Key words: Inverse problem; Multivariate samples; Joint moments; Complex-valued observations.

AMS Subject Classifications: 62H05; 11P70; 47A57.

1. Introduction

Evidently, one can readily evaluate sample moments from a given data set. The problem being considered herein, which consists of retrieving a sample of multivariate observations from certain of its marginal and joint sample moments, can be regarded as an inverse problem.

Inverse problems generally involve determining certain causes from some effects. They currently constitute a rich field of research. For instance, they appear in the Mathematics Subject Classification index in connection with quantum theory, optics, harmonic analysis, trigonometry, linear operators, and electromagnetic theory. Inverse problems of various nature have, for example, also found applications in geophysics (Zhdanov, 2015), acoustics (Klyuchinskiy *et al.* 2020), image processing (Zou *et al.* 2021), astronomy (Escárate *et al.* 2023), system identification (Blanken and Oomen, 2020), language processing (Nakanishi, 2024), machine learning (Koffer *et al.* 2023), signal processing (Giovannelli and Idier, 2015) and tomography (Mohamad-Djafari, 2013).

The results introduced in this paper imply that a certain number of marginal and joint moments actually hold all the information that is contained in a given data set since the latter

Corresponding Author: Serge B. Provost
Email: provost@stats.uwo.ca

can be entirely retrieved from the former. Accordingly, such moments constitute sufficient statistics. To some extent, this remark provides a justification for making use of moment-based statistical methodologies such as the density function estimation techniques advocated in Provost and Zheng (2015), Provost and Ha (2016), Jin *et al.* (2016), Zareamoghaddam *et al.* (2017), Kang *et al.* (2019), Provost *et al.* (2020) and Provost and Zang (2024).

The problem of recovering a univariate sample of size n from its first n moments is considered in Section 2 where the applicability of the result is discussed. The case of bivariate observations and their sample moments is addressed in Section 3 where generalizations to complex-valued and multivariate data sets are explored. All the results and their extensions are illustrated by means of numerical examples. Lastly, some concluding remarks are offered in Section 4.

2. A theorem relating a univariate data set to its moments

In this section, we state a result that was established in Provost *et al.* (2020), explain that it holds in the complex domain, and discuss related considerations. Two numerical examples are provided as well.

Theorem 1: A data set of size n can be recovered from the first n moments of the sample.

The proof of this result is given in the Appendix for the sake of completeness. The following example illustrates the steps to follow when applying Theorem 1.

Example 1: Let $n = 5$ and the sample be $\{1.2, 3.4, 6.7, 8.1, 11.9\}$. The moments of orders zero to five are $1, 6.26, 53.022, 511.6790, 5301.7767, 57492.260726$ and, for $j = 0, 1, 2, 3, 4, 5$, the e_j's as defined in the Appendix, are $1, 31.3, 357.29, 1814.543, 3910.731, 2634.91704$. According to equation (1), the resulting polynomial is then $-2634.91704+3910.731\,x-1814.543\,x^2 +357.29\,x^3 - 31.3\,x^4 + x^5$, its five roots being $\{1.2, 3.4, 6.7, 8.1, 11.9\}$.

We note that the proof of Theorem 1 remains valid in the complex domain. It should also be observed that any loss of precision can be avoided by making use of fractions.

Example 2: Let $n = 3$ and the sample be $\{2.4 + 5.1\,i,\ 6.7 - 9.5\,i, 11.8 + 1.4\,i\}$, that is, $\{\frac{12}{5} + \frac{51\,i}{10}, \frac{67}{10} - \frac{19\,i}{2}, \frac{59}{5} + \frac{7\,i}{5}\}$ in fractional form. The moments of orders zero, one, two and three are $1, \frac{209}{30} - i, \frac{2389}{100} - \frac{1163\,i}{50}$, and $-\frac{56531}{1500} + \frac{38517\,i}{1000}$, and for $j = 0, 1, 2, 3$, the e_j's as defined in the Appendix are $1, \frac{209}{10} - 3\,i, \frac{17807}{100} - \frac{2781\,i}{100}, \frac{93192}{125} + \frac{56127\,i}{250}$. The polynomial, $x^3 - \left(\frac{209}{10} - 3\,i\right)x^2 + \left(\frac{17807}{100} - \frac{2781\,i}{100}\right)x - \left(\frac{93192}{125} + \frac{56127\,i}{250}\right)$, is then obtained from equation (1) and, as expected, its three roots are $\{\frac{12}{5} + \frac{51\,i}{10}, \frac{67}{10} - \frac{19\,i}{2}, \frac{59}{5} + \frac{7\,i}{5}\}$.

Since there exists a one-to-one correspondence between the observations and their associated empirical distribution function, the following corollary to Theorem 1 holds.

Corollary 1: Given a simple random sample of size n from a continuous distribution, its empirical distribution function F_n is uniquely specified by the first n sample moments.

In light of the strong law of large numbers, for every fixed x, the empirical distribution function $F_n(x)$ will converge almost surely to the underlying distribution function $F(x)$. Moreover, given a simple random sample of size n, the Glivenko-Cantelli theorem states

that

$$\sup_{x\in\Re}|F_n(x) - F(x)|$$

tends to zero almost surely, and that the convergence of $F_n(x)$ to $F(x)$ is uniform. However, as was aptly pointed out by Ričardas Zitikis, a colleague of the first author, a contradiction would ensue if one were to let n tend to infinity in Corollary 1 as this result would then imply that, given the integer moments of a random variable, its distribution could be specified uniquely. This is clearly not the case since there exists distinct distributions whose integer moments are all identical.

Consider for example the following density functions:

$$f_1(x) = \frac{1}{4}e^{-\sqrt{|x|}}, \quad x \in \Re,$$

and

$$f_2(x) = \frac{1}{4}e^{-\sqrt{|x|}}\Big(\cos(\sqrt{|x|}) + 1\Big), \quad x \in \Re,$$

which are plotted in Figure 1.

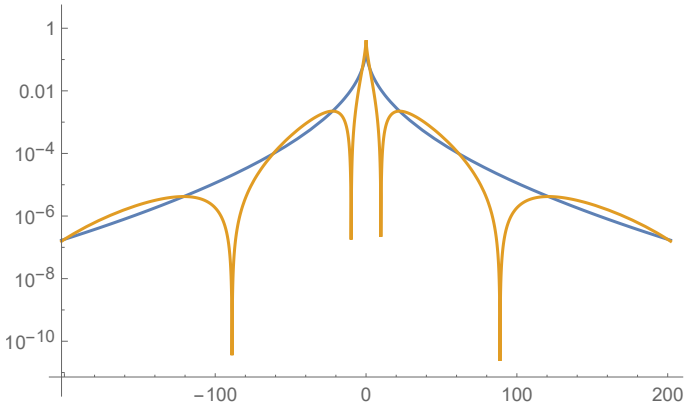


Figure 1: Plots of $f_1(x)$ and $f_2(x)$ on a logarithmic scale for $-200 < x < 200$

Although these two distributions are clearly distinct, their k^{th} moment,

$$m_1(k) = \frac{1}{2}\Big((-1)^k + 1\Big)\,\Gamma(2k+2)$$

and

$$m_2(k) = \frac{1}{2}\Big((-1)^k + 1\Big)\,\Gamma(2k+2)\Big(1 - \frac{\sin(k\,\pi/2)}{2^{k+2}}\Big),$$

happen to coincide for $k = 0, 1, 2, \ldots$.

To summarize, in the limit, F_n can specify the underlying population distribution function. However, as previously illustrated, a population distribution function F may not be uniquely specified by an infinite sequence of its integer moments. Thus, Corollary 1 cannot be extended beyond finite values of n.

It should also be pointed out that moment-based methodologies lend themselves to the modeling of massive data sets since only a moderate number of moments are needed to

apply such techniques, as opposed to other approaches such as those based on likelihoods for which all the observations are required. Actually, ample information can generally be secured from a fairly limited number of moments, whereas each data point contains an equal amount of information that is inversely proportional to the sample size. Moreover, once a new set of observations, $\{x_{n_1+1}, \ldots, x_n\}$, becomes available in addition to an initial dataset, $\{x_1, \ldots, x_{n_1}\}$, there is no need to make use of each of the n_1 original data points to compute the moments since the h^{th} updated moment will then be $(n_1 m_h + \sum_{i=n_1+1}^{n} x_i^h)/n$ where m_h denotes the h^{th} sample moment as evaluated from the initial data set.

3. On recovering multivariate samples from their moments

The four propositions introduced in this section enable one to retrieve bivariate sets of observations from some of their marginal and joint moments—or those of their component-wise ranks, the observations on each variable being assumed to be distinct. It is explained that each of the proposed methodologies also apply to multivariate data sets and that two of them hold in the complex domain. Several numerical examples are provided.

Proposition 1: A bivariate sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ can be retrieved from the first n marginal moments of the first variable, that is,

$$m_{1,0}, \ldots, m_{n-1,0}, m_{n,0},$$

in conjunction with the following bivariate sample moments:

$$m_{0,1}, m_{1,1}, \ldots, m_{n-1,1},$$

where $m_{j,k}$ denotes the moment of orders j and k, which is equal to $\sum_{i=1}^{n} x_i^j \, y_i^k/n$.

Proof: In light of Theorem 1, the observations on the first variable, namely, $x_1, \ldots, x_n$ can be retrieved from the given marginal moments. The remainder of the proof relies on a representation of the joint moments that involves a Vandermonde matrix.

It is assumed that the following joint moments are known:

$$m_{j,1} = \frac{1}{n} \sum_{1=1}^{n} x_i^j \, y_i, \;\; j = 0, \ldots, n-1.$$

This system of equations can be equivalently expressed as follows:

$$\frac{1}{n} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{n-1} & x_2^{n-1} & \cdots & x_n^{n-1} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} m_{0,1} \\ m_{1,1} \\ m_{2,1} \\ \vdots \\ m_{n-1,1} \end{pmatrix}$$

where the above matrix is a Vandermonde matrix, which is nonsingular since the x_i's are assumed to be nonidentical. Note that the vector of y_j's which is the unique solution of this linear system, enables one to pair each of them appropriately with the corresponding x_i. □

Remark 1: Given the definition of $m_{j,k}$, it is apparent that the order of the n bivariate sample points is immaterial. Thus, in applications, it suffices to set a certain order for the x_i's, and the y_j's to be associated with these x_i's will be properly ordered in the solution vector of the linear system.

Additionally, we note that the $2n$ moments that are specified in Proposition 1 are jointly sufficient statistics, since they provide enough information to recover the entire bivariate sample of ordered observations—which, incidentally, requires $3n-1$ pieces of information, namely, the observations on each variable and the ranks of $n-1$ observations on the second component relative to those on the first.

Example 3: Let the sample be $\{(1,7),(2,2),(5,3)\}$. Given the marginal moments on the first variable, one can determine that the observations on the first variable are 1, 2 and 5. Additionally, let the joint moments of orders (0,1), (1,1) and (2,1), that is, $m_{0,1}=4$ $m_{1,1}=26/3$ and $m_{2,1}=30$, be available. The solution of the following system, which is (7,2,3), yields the values of the y_j's to be associated with the x_i's:

$$\frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 5 \\ 1^2 & 2^2 & 5^2 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}=\begin{pmatrix} 4 \\ 26/3 \\ 30 \end{pmatrix}.$$

As is the case for univariate observations, complex-valued bivariate or multivariate observations can also be recovered. This can be readily achieved by initially implementing Theorem 1 and then, solving a linear system of equations involving complex values.

Example 4: Let $\{(2.4+5.1\,i,7.3-1.8\,i),(6.7-9.5\,i,2.2),(11.8+1.4\,i,9.8\,i)\}$ be the sample to recover. Note that if the first three marginal moments of the first variables are given, one can retrieve the three observations on the first component, which happens to be the univariate data set utilized in Example 2. Now, assume that the joint moments of orders (0,1), (1,1) and (2,1), namely, $m_{0,1}=19/6+(8\,i)/3$, $m_{1,1}=231/25+(851\,i)/20$ and $m_{2,1}=-(105469/600)+(640219\,i)/1500$ are available. As expected, the solution of the linear system,

$$\frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 12/5+51\,i/10 & 67/10-19\,i/2 & 59/5+7\,i/5 \\ -81/4+612\,i/25 & -1134/25-1273\,i/10 & 3432/25+826\,i/25 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

$$=\begin{pmatrix} 19/6+8\,i/3 \\ 231/25+851\,i/20 \\ -105469/600+640219\,i/1500 \end{pmatrix}$$

is $\{y_1,y_3,y_4\}=\{73/10-(9\,i)/5,11/5,(49\,i)/5\}$.

A trivariate observation vector $(x_i,y_i,z_i),\ i=1,\ldots,n$, can be similarly recovered if, in addition to the the first n marginal moments of the first variable from which the x_i's can be specified, one knows $m_{0,1,0},m_{1,1,0},\ldots,m_{n-1,1,0}$ which will yield the y_j's associated with the x_i's, as well as $m_{0,0,1},\ldots,m_{0,n-1,1}$ which will then yield the z_k's associated with the y_j's. By proceeding in like fashion, Proposition 1 can extended to sets of multivariate observations.

Example 5: Let the sample be $\{(2,4,6),(7,3,1),(5,6,3)\}$. Given the marginal moments on the first variable, we can determine that the observations on that variable are 2, 5 and 7 and, in light of Remark 1, we may let $\{x_1,x_2,x_3\}=\{2,7,5\}$ (or any other permutation thereof). The joint moments of orders (0,1,0), (1,1,0) and (2,1,0) are $m_{0,1,0}=13/3$ $m_{1,1,0}=59/3$ and $m_{2,1,0}=313/3$, and the joint moments of orders (0,0,1), (0,1,1) and (0,2,1) are $m_{0,0,1}=10/3$ $m_{0,1,1}=15$ and $m_{0,2,1}=71$. The solutions of the systems of equations,

$$\frac{1}{3}\begin{pmatrix}1&1&1\\2&7&5\\2^2&7^2&5^2\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}=\begin{pmatrix}13/3\\59/3\\313/3\end{pmatrix}$$

and

$$\frac{1}{3}\begin{pmatrix}1&1&1\\4&3&6\\4^2&3^2&6^2\end{pmatrix}\begin{pmatrix}z_1\\z_2\\z_3\end{pmatrix}=\begin{pmatrix}10/3\\15\\71\end{pmatrix},$$

yield the values of the y_j's to be paired with the x_i's, that is, $\{y_1,y_2,y_3\}=\{4,3,6\}$, and then those of the z_k's to be paired with the y_j's namely, $\{z_1,z_2,z_3\}=\{6,1,3\}$).

Proposition 1 can also be extended as follows: Given the marginal moments of the first variable up to order n, any additional set of n joint moments that does not include any of the first variable marginal moments can be utilized to recover the sample. The resulting system of equations can be solved by making use of an array of computing packages. This flexibility in the selection of joint moments also applies in the case of multivariate observations.

Example 6: Let the sample be $\{(2,4),(5,6),(7,3)\}$. Given the first three marginal moments on the first variable which are $\{14/3,26,476/3\}$, it can be determined from Theorem 1 that the observations on that variable are 2, 5 and 7. Now, assume that the joint moments of orders (0,1), (1,2) and (2,3), namely, $m_{0,1}=13/3$, $m_{1,2}=275/3$ and $m_{2,3}=6979/3$ are available. It then suffices to solve of system, $\{y_1+y_2+y_3=13,\ 2y_1^2+5y_2^2+7y_3^2=275,\ 4y_1^3+25y_2^3+49y_3^3=6979\}$ to obtain the corresponding values for the second variable, that is, (4,6,3).

Proposition 2: A bivariate sample of size n can be retrieved from the first n marginal sample moments of each variable, that is, $m_{i,0},\ i=1,\ldots,n,$ and $m_{0,j},\ j=1,\ldots,n$, where $m_{i,j}$ denotes the sample moment of orders i and j, *in conjunction with* the ranks of the observations within each variable—or equivalently those of the corresponding pseudo-observations.

Pseudo-observations are the component-wise ranks of the data points divided by n. Note that all the pseudo-observations originating from a given sample can be secured from the associated empirical copula, as originally defined by Deheuvels (1979).

Proof: As previously explained, the data on each variable can be retrieved from the marginal moments by appealing to Theorem 1. Then, given the ranks of the observations on each variable, the observations can be appropriately paired. □

Example 7: Let the original sample be {(1,7), (2,2), (5,3)}. First, it can be determined from the first three marginal moments of each variable that the observations on the first and second variables are respectively {1, 2 ,5}, and {2, 3, 7}. If in addition, it is known that the ranks of the observations on each component are $[r_1,s_1]=[1,3],[r_2,s_2]=[2,1]$ and

$[r_3, s_3] = [3, 2]$, then, it can readily be determined that the sample points are (1,7), (2,2) and (5,3).

This approach can be directly extended to sets of multivariate observations.

Example 8: Consider the following sample of trivariate observations: {(2, 5, 7), (3, 4, 8),(1, 3, 6)}. Given the first three marginal moments of each of the three variables, it can be determined from Theorem 1 that the observations on the first, second and third components are {1, 2, 3}, {3, 4, 5} and {6, 7, 8}, respectively. If it is also known that the ranks of these component-wise observations are $[r_1, s_1, t_1] = [2, 3, 2]$, $[r_2, s_2, t_2] = [3, 2, 3]$, and $[r_3, s_3, t_3] = [1, 1, 1]$, it can then be readily determined that the sample points are (2, 5, 7), (3, 4, 8) and (1, 3, 6).

Proposition 3: A random sample of size n arising from a continuous bivariate distribution can be retrieved from the first n marginal moments of each variable, that is, $m_{i,0},\ i = 1, \ldots, n$ and $m_{0,j},\ i, j = 1, \ldots, n,$ *in conjunction with* the joint moments, $m^*_{0,1}, \ldots, m^*_{n-1,1}$, of the ranks of the observations.

Proof: In light of Theorem 1, the observations on each variable, namely, $x_1, \ldots, x_n$, and $y_1, \ldots, y_n$, can be recovered from the marginal moments. The remainder of the proof relies on a representation of the joint moments of the ranks that involves a Vandermonde matrix. Let again r_i and s_i denote the ranks of the observations with respect to the first and second variables. By assumption, the joint moments, $m^*_{0,1}, \ldots, m^*_{n-1,1}$, of the ranks are known with, in general,

$$m^*_{j,k} = \frac{1}{n} \sum_{i=1}^{n} s_i^j \, r_i^k, \ \ j = 0, \ldots, n-1.$$

Note that $m^*_{0,1} = (n+1)/2$. This system of equations can be equivalently expressed as follows:

$$\frac{1}{n} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ r_1 & r_2 & \cdots & r_n \\ r_1^2 & r_2^2 & \cdots & r_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ r_1^{n-1} & r_2^{n-1} & \cdots & r_n^{n-1} \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ \vdots \\ s_n \end{pmatrix} = \begin{pmatrix} m^*_{0,1} \\ m^*_{1,1} \\ m^*_{2,1} \\ \vdots \\ m^*_{n-1,1} \end{pmatrix}$$

where the above matrix is a Vandermonde matrix, which is nonsingular since the r_i's are distinct. Note that the unique solution of this linear system will yield $s_1, \ldots, s_n$, and associate each of them appropriately with the corresponding r_i, which will enable one to correctly pair the known x_i's and y_j's. □

Remark 2: Given the definition of $m^*_{j,k}$, it is apparent that the order of the n bivariate sample points does not matter, since the pair of ranks corresponding to a given bivariate observation will remain unchanged. Thus, in applications, it suffices to set a certain order for the r_i's, and the s_j's to be associated with these r_i's will be properly ordered in the solution vector of the linear system.

Example 9: Let the sample be $\{(1, 7), (2, 2), (5, 3)\}$. Given the marginal moments on each variable, one can retrieve the observations on the first variables, namely, 1, 2 and 5, as well

as the observations on the second variables which are 2, 3 and 7. It now remains to pair them using Proposition 3. We have to determine the second component of the following paired ranks: $[r_1, s_1] = [1, 3], [r_2, s_2] = [2, 1]$ and $[r_3, s_3] = [3, 2]$, that is, $[s_1, s_2, s_3] = [3, 1, 2]$. The joint moments of the ranks of orders (0,1), (1,1) and (2,1) are $m^*_{0,1} = 2$, $m^*_{1,1} = 11/3$ and $m^*_{2,1} = 25/3$, respectively. Solving the following system will yield the ranks of the second component, that is, [3,1,2], and enable one to correctly pair the data points:

$$\frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1^2 & 2^2 & 3^2 \end{pmatrix}\begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 11/3 \\ 25/3 \end{pmatrix}.$$

This result can be generalized to the multivariate case by proceeding as in the generalization of the Proposition 1, except that in this case, the joint moments of the ranks are utilized in addition to the marginal sample moments of each variable. As well, joint moments of the ranks other than those specified in Proposition 3 can be utilized as was done in Example 6 in conjunction with certain joint moments of the observations.

Example 10: Let the sample be $\{(2, 4, 6), (7, 3, 1), (5, 6, 3)\}$. Given the marginal moments on each variable, one can retrieve the observations on the first, second and third variables, that is, {2, 5, 7}, {3, 4, 6}, and {1, 3, 6}, respectively. We then have to determine the ranks of the entries in second and third components, namely, $[s_1, s_2, s_3] = [2, 1, 3]$ and $[t_1, t_2, t_2] = [3, 1, 2]$ and end up with the following set of ranks: $[r_1, s_1, t_1] = [1, 2, 3]$, $[r_2, s_2, t_2] = [3, 1, 1]$, and $[r_3, s_3, t_3] = [2, 3, 2]$, which enables us to retrieve the original data set.

The joint moments of the ranks of orders (0,1,0), (1,1,0) and (2,1,0) are $m^*_{0,1,0} = 2$, $m^*_{1,1,0} = 11/3$ and $m^*_{2,1,0} = 23/3$, respectively. Let the given joint moments of the ranks of orders (0,0,1), (0,1,1) and (0,2,1) be $m^*_{0,0,1} = 2$, $m^*_{0,1,1} = 13/3$ and $m^*_{0,2,1} = 31/3$, respectively.

We started off with $r_1 = 1$, $r_2 = 3$ and $r_3 = 2$; however, as per Remark 2, any permutation thereof will lead to the data set with its trivariate observations appearing in a different order. Thus, we first solve the following linear system, which will yield the ranks of the second component entries, that is, [2,1,3]:

$$\frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1^2 & 3^2 & 2^2 \end{pmatrix}\begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 11/3 \\ 23/3 \end{pmatrix}.$$

The solution of the linear system that follows will then yield the ranks of the third component entries, which are [3,1,2]:

$$\frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 3 \\ 2^2 & 1^2 & 3^2 \end{pmatrix}\begin{pmatrix} t_1 \\ t_2 \\ t_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 13/3 \\ 31/3 \end{pmatrix}.$$

Proposition 4: A bivariate sample of size n can be retrieved on the basis of the first n marginal sample moments of each variable in conjunction with any *single* additional joint sample moment that does not involve moments of order zero.

Proof: On applying Theorem 1, the set of observations on each variable can be determined from the marginal moments. Then, given the ordered observations on the first variable, there will be a unique permutation of the observations on the second variable that will yield the given joint moment. □

This assumes that the observations have been recorded with sufficient precision.

Example 11: Consider the sample $\mathcal{S} = \{(1,7),(2,2),(5,3)\}$. Given the marginal moments of each variables, it can be determined that the first and second component values will respectively be {1,2,5} and {2,3,7}. Assuming for instance that, it is known that the joint moment $m_{1,1} = 26/3$, and, for instance, setting the the observations on the first component in increasing order, that is, 1,2,5, we are seeking the permutation of {2,3,7} among the 6 possible ones that will yield the same joint moment of orders 1 and 1. This process will lead to the identification of the correct bivariate data points that constitute the sample $\mathcal{S}$. The 6 possible pairs of observations and their joint moment of order (1,1) are:

$\{(1,7),(2,3),(5,2)\} \Rightarrow\ m_{1,1} = 23/3$
$\{(\mathbf{1},\mathbf{7}),(\mathbf{2},\mathbf{2}),(\mathbf{5},\mathbf{3})\}\ \Rightarrow\ m_{1,1} = \mathbf{26/3}$
$\{(1,2),(2,7),(5,3)\}\ \Rightarrow\ m_{1,1} = 31/3$
$\{(1,2),(2,3),(5,7)\}\ \Rightarrow\ m_{1,1} = 43/3$
$\{(1,3),(2,2),(5,7)\}\ \Rightarrow\ m_{1,1} = 42/3$
$\{(1,3),(2,7),(5,2)\}\ \Rightarrow\ m_{1,1} = 27/3.$

Accordingly, we select the bold-faced set as the original sample since its joint moment of order (1,1) coincides with that of $\mathcal{S}$.

Proposition 4 which, incidentally, is implementable in the case of moderately sized samples, can readily be extended to sets of multivariate observations.

Example 12: Consider the sample $\mathcal{S} = \{(2,4,6),(7,3,1),(5,6,3)\}$. Given the first three marginal moments of each variable, it can be determined that the observations on the first, second and third components are {2,5,7}, {3,4,6}, and {1,3,6}, respectively. Assuming for instance that, it is known that the joint moment $m_{1,1,1} = 53$, and setting the observations on the first component in increasing order, that is, $\{2,5,7\}$, we are seeking the permutation of $\{3,4,6\}$ and that of $\{1,3,6\}$ that will yield the same joint moment. This will enable us to identify the correct triplet of trivariate observations comprising $\mathcal{S}$. The 36 possible sets of observations and their joint moments of order (1,1,1) are:

$\{(2,3,1),\ (5,4,3),\ (7,6,6)\} \Rightarrow m_{1,1,1} = 106,$
$\{(2,3,1),\ (5,4,6),\ (7,6,3)\} \Rightarrow m_{1,1,1} = 84,$
$\{(2,3,3),\ (5,4,1),\ (7,6,6)\} \Rightarrow m_{1,1,1} = \frac{290}{3},$
$\{(2,3,3),\ (5,4,6),\ (7,6,1)\} \Rightarrow m_{1,1,1} = 60,$
$\{(2,3,6),\ (5,4,1),\ (7,6,3)\} \Rightarrow m_{1,1,1} = \frac{182}{3},$
$\{(2,3,6),\ (5,4,3),\ (7,6,1)\} \Rightarrow m_{1,1,1} = 46,$
$\{(2,3,1),\ (5,6,3),\ (7,4,6)\} \Rightarrow m_{1,1,1} = 88,$
$\{(2,3,1),\ (5,6,6),\ (7,4,3)\} \Rightarrow m_{1,1,1} = 90,$
$\{(2,3,3),\ (5,6,1),\ (7,4,6)\} \Rightarrow m_{1,1,1} = 72,$

$\{(2,3,3),\,(5,6,6),\,(7,4,1)\} \Rightarrow m_{1,1,1} = \frac{226}{3}$,
$\{(2,3,6),\,(5,6,1),\,(7,4,3)\} \Rightarrow m_{1,1,1} = 50$,
$\{(2,3,6),\,(5,6,3),\,(7,4,1)\} \Rightarrow m_{1,1,1} = \frac{154}{3}$,
$\{(2,4,1),\,(5,3,3),\,(7,6,6)\} \Rightarrow m_{1,1,1} = \frac{305}{3}$,
$\{(2,4,1),\,(5,3,6),\,(7,6,3)\} \Rightarrow m_{1,1,1} = \frac{224}{3}$,
$\{(2,4,3),\,(5,3,1),\,(7,6,6)\} \Rightarrow m_{1,1,1} = 97$,
$\{(2,4,3),\,(5,3,6),\,(7,6,1)\} \Rightarrow m_{1,1,1} = 52$,
$\{(2,4,6),\,(5,3,1),\,(7,6,3)\} \Rightarrow m_{1,1,1} = 63$,
$\{(2,4,6),\,(5,3,3),\,(7,6,1)\} \Rightarrow m_{1,1,1} = 45$,
$\{(2,4,1),\,(5,6,3),\,(7,3,6)\} \Rightarrow m_{1,1,1} = \frac{224}{3}$,
$\{(2,4,1),\,(5,6,6),\,(7,3,3)\} \Rightarrow m_{1,1,1} = \frac{251}{3}$,
$\{(2,4,3),\,(5,6,1),\,(7,3,6)\} \Rightarrow m_{1,1,1} = 60$,
$\{(2,4,3),\,(5,6,6),\,(7,3,1)\} \Rightarrow m_{1,1,1} = 75$,
$\{(2,4,6),\,(5,6,1),\,(7,3,3)\} \Rightarrow m_{1,1,1} = 47$,
$\{\mathbf{(2,4,6)},\,\mathbf{(5,6,3)},\,\mathbf{(7,3,1)}\} \Rightarrow \mathbf{m_{1,1,1} = 53}$,
$\{(2,6,1),\,(5,3,3),\,(7,4,6)\} \Rightarrow m_{1,1,1} = 75$,
$\{(2,6,1),\,(5,3,6),\,(7,4,3)\} \Rightarrow m_{1,1,1} = 62$,
$\{(2,6,3),\,(5,3,1),\,(7,4,6)\} \Rightarrow m_{1,1,1} = 73$,
$\{(2,6,3),\,(5,3,6),\,(7,4,1)\} \Rightarrow m_{1,1,1} = \frac{154}{3}$,
$\{(2,6,6),\,(5,3,1),\,(7,4,3)\} \Rightarrow m_{1,1,1} = 57$,
$\{(2,6,6),\,(5,3,3),\,(7,4,1)\} \Rightarrow m_{1,1,1} = \frac{145}{3}$,
$\{(2,6,1),\,(5,4,3),\,(7,3,6)\} \Rightarrow m_{1,1,1} = 66$,
$\{(2,6,1),\,(5,4,6),\,(7,3,3)\} \Rightarrow m_{1,1,1} = 65$,
$\{(2,6,3),\,(5,4,1),\,(7,3,6)\} \Rightarrow m_{1,1,1} = \frac{182}{3}$,
$\{(2,6,3),\,(5,4,6),\,(7,3,1)\} \Rightarrow m_{1,1,1} = 59$,
$\{(2,6,6),\,(5,4,1),\,(7,3,3)\} \Rightarrow m_{1,1,1} = \frac{155}{3}$,
$\{(2,6,6),\,(5,4,3),\,(7,3,1)\} \Rightarrow m_{1,1,1} = 51$

Accordingly, we select the bold-faced set as the original sample since its joint moment of order (1,1,1) coincides with that of $\mathcal{S}$.

Proposition 4 can as well be extended to complex-valued samples.

Example 13: Consider the sample $\mathcal{S}$=$\{(5.4 + 6.1\,i, 9 + 3.4\,i),\,(6.7, 3.3\,i), (8\,i, 1.9)\}$. Given the first three marginal moments of the first and second components, which are respectively $\{121/30+(47\,i)/10,\,-679/75+(549\,i)/25,\,-5783/120-(68451\,i)/1000\}$ and $\{109/30+(67\,i)/30, 518/25 + (102\,i)/5,\,423739/3000+\,(750959\,i)/3000\}$, one can determine the three entries in each of the two components as was done in Example 2 for the univariate case. Now, assume that, additionally, $m_{1,1}$=$1393/150 + (11057\,i)/300$, is provided. On keeping the observations on first component in a given order and permuting those of the second component, only one of the six joint moments of orders 1 and 1 so obtained will equal $m_{1,1}$, the corresponding set of paired observations being those included in $\mathcal{S}$.

4. Concluding remarks

Four methodologies were introduced for the purpose of recovering a multivariate data set from certain of its associated marginal and joint moments as evaluated from the ob-

servations or their component-wise ranks. In fact, two of them also hold in the complex domain. For a given multivariate sample, the evaluation of the marginal and joint moments is straightforward and constitutes a direct problem. As explained in the Introduction, the results introduced in this paper actually solve the inverse problem consisting of recovering the original observations on the basis of certain marginal and joint moments.

Interestingly, a parallel can be established between Proposition 2 which makes use of a number of marginal moments and all the component-wise ranks of the observations—or, equivalently, the pseudo-observations—to recover the entire sample, and Sklar's theorem as introduced by Sklar (1959), which states that a joint distribution can be expressed in terms of the marginal distributions and a function that depends only on the pseudo-observations, which is referred to as a copula. In fact, copulas completely account for the dependence between the variables. Several nonparametric copula density estimation techniques were recently proposed in Provost and Zang (2024). For an introduction to copulas and related results, the reader is referred to Nelsen (2006). All the calculations were carried out with the symbolic computing package *Mathematica*, the code being available from the first author upon request.

Acknowledgements

The financial support of the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged.

References

Blanken, L. and Oomen, T. (2020). Kernel-based identification of non-causal systems with application to inverse model control. *Automatica*, **114**, 1–7.

Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés: un test non paramétrique d'indépendance. *Académie Royale de Belgique–Bulletin de la Classe des Sciences*, **65**, 274–292.

Escárate, P., Curé, M., Araya, I., Coronel, M., Cedeño, A. L., Celedon, L., Cavieres, J., Aguero, J. C., Arcos, C., Cidale, L. S., and Levenhagen, R. S. (2023). A method to deconvolve stellar profiles – The non-rotating line utilizing Gaussian sum approximation. *Astronomy & Astrophysics*, **676**, A44. doi.org/10.1051/0004-6361/202346587

Giovannelli, J. F. and Idier, J. (2015). *Regularization and Bayesian Methods for Inverse Problems in Signal and Image Processing*. New York: John Wiley & Sons, Inc.

Jin T., Provost S. B, and Ren J. (2016). A moment-based methodology for approximating the distribution of aggregate losses. *Scandinavian Actuarial Journal*, **16**, 216–245.

Kang, J. S. J., Provost, S. B., and Ren, J. (2019). Moment-based density approximation techniques as applied to heavy-tailed distributions. *International Journal of Statistics and Probability*, **8**, 1–23.

Klyuchinskiy, D., Novikov, N., and Shishlenin, M. (2020). A modification of gradient descent method for solving coefficient inverse problem for acoustics equations. *Computation*, **8**, 1–14. doi:10.3390/computation8030073

Kofler, A., Altekruger, F., Antarou Ba, F., Kolbitsch, C., Papoutsellis, E., Schote, D., and Papafitsoros, K. (2023). Learning regularization parameter-maps for variational image reconstruction using deep neural networks and algorithm unrolling. *SIAM Journal on Imaging Sciences*, **16**, 2202–2246.

Mohamad-Djafari, A. (2013). *Inverse Problems in Vision and 3D Tomography*. New York: John Wiley & Sons. ISBN 978-1-118-60046-7.

Nakanishi, T. (2024). *Detection of Latent Gender Biases in Data and Models Using the Approximate Generalized Inverse Method*. In 2024 IEEE 18th International Conference on Semantic Computing (ICSC), 191–196.

Nelsen, R .B. (2006) *An Introduction to Copulas*, 2nd ed. New York: Springer.

Provost, S. B. and Ha, H.-T. (2016). Distribution approximation and modelling via orthogonal polynomial sequences. *Statistics*, **50**, 454–470.

Provost, S. B. and Zang, Y. (2024). Nonparametric Copula Density Estimation Methodologies. *Mathematics*, **12**, 398. https://doi.org/10.3390/math12030398

Provost, S. B., Zareamoghaddam, H., Ahmed, S. E., and Ha, H.-T. (2020). The generalized Pearson family of distributions and explicit representation of the associated density functions. *Communications in Statistics - Theory and Methods*, **51**, 5590–5606.

Provost, S. B. and Zheng, S. Z. (2015). Polynomially adjusted saddlepoint density approximations. *International Journal of Statistics and Probability*, **4**, 1–11.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publication de l'Institut Statistique de l'Université de Paris*, **8**, 22–231.

Zareamoghaddam, H., Provost, S. B., and Ahmed, E. (2017). A moment-based bivariate density estimation methodology applicable to big data modeling. *Journal of Probability and Statistical Science*, **30**, 155–162.

Zhdanov, M. S. (2015). *Inverse Theory and Applications in Geophysics*. Amsterdam: Elsevier.

Zou, Z., Shi, T., Shi, Z., and Ye, J. (2021). Adversarial training for solving inverse problems in image processing. *IEEE Transactions on Image Processing*, **30**, 2513–2525.

APPENDIX

Proof of Theorem 1

Let $\mathcal{S} = \{x_1, x_2, \ldots, x_n\}$ be a sample of size n and $\mathcal{M} = \{m_1, m_2, \ldots, m_n\}$ where $m_h = \sum_{i=1}^{n} x_i^h/n$. According to the fundamental theorem of algebra, $p(z) = a_0 + a_1 z + \cdots + a_{n-1}z^{n-1} + z^n$ is uniquely defined by its coefficients a_i's and it is also uniquely specified by its n roots x_i's for $i = 1, \ldots, n$. Moreover, given $\mathcal{S}$, the coefficients of $p(x)$ can be expressed in terms of the sequence of moments $\mathcal{M}$ via the Newton-Girard identity. Accordingly, a given polynomial of degree n, say $p(x)$, can be represented as follows:

$$\prod_{i=1}^{n}(x - x_i) = \sum_{k=0}^{n}(-1)^{n-k}\, e_{n-k}\, x^k, \tag{1}$$

where $e_0 = 1$ and

$$e_\ell = \frac{n}{\ell}\sum_{h=1}^{\ell}(-1)^{h-1}\, e_{\ell-h}\; m_h, \quad \ell = 1, \ldots, n. \tag{2}$$

Thus, given the first n sample moments associated with $\mathcal{S}$, a sample of size n, one can express the right-hand side of (1) as a polynomial whose roots are precisely $\{x_1, x_2, \ldots, x_n\}$. This establishes that $\mathcal{S}$ is uniquely specified by $\mathcal{M}$.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 485–508
http://www.ssca.org.in/journal

Bayesian Variable Selection for Ultrahigh-dimensional Sparse Linear Models

Minerva Mukhopadhyay[1] **and Subhajit Dutta**[2]
[1]*Interdisciplinary Statistical Research Unit,*
Indian Statistical Institute Kolkata, 203 B. T. Road, Kolkata – 700108, WB, India.
[2]*Applied Statistics Unit,*
Indian Statistical Institute Kolkata, 203 B. T. Road, Kolkata – 700108, WB, India.
[1,2]*Department of Mathematics and Statistics,*
Indian Institute of Technology Kanpur, Kanpur - 208016, UP, India.

Received: 12 August 2024; Revised: 25 September 2024; Accepted: 30 September 2024

Abstract

We consider the problem of variable selection for the ultrahigh-dimensional linear regression model, allowing the number of covariates p_n to grow exponentially with n. Assuming the true model to be sparse, we propose a set of priors suitable for this regime. In the ultrahigh-dimensional setting, the selection of the unique true model among all the 2^{p_n} possible ones involves prohibitive computation. To cope with this, a two-stage model selection algorithm is proposed. In the first stage, an efficient screening algorithm is employed to find a *good* d_n-dimensional model, where $d_n \ll n$. In the next stage, an explicit model search algorithm is employed on the space of all submodels of the first-stage-selected model. Theoretical investigations justify the two-stage procedure. It is demonstrated that the first-stage screening is expected to select a supermodel of the true model, consequently, the second-stage algorithm identifies the true model with probability tending to one. This procedure is computationally efficient, simple and intuitive. We validate the competitive performance of the proposed algorithm with a variety of simulated and real data sets, and compare with several frequentist as well as Bayesian methods.

Key words: Model selection consistency; Reversible jump MCMC; Screening consistency.

AMS Subject Classifications: 62H05; 11P70; 47A57.

1. Introduction

Variable selection in ultrahigh-dimensional regression setup has become a flourishing area in the contemporary research, due to increasing availability of data in various fields like genetics, finance, machine learning. Consider, for example, in genome-wide association studies (GWAS), where a phenotype is measured for a panel of individuals and a large number of single nucleotide polymorphisms (SNPs) are genotyped for each individual. The goal is to identify SNPs that are statistically associated with the phenotype. Sparsity has

Corresponding Author: Minerva Mukhopadhyay
Email: minervam@iitk.ac.in

frequently been identified as an underlying feature for such data sets, where among a large number of covariates (SNPs) only a small subset are actually important.

Several variable selection methods have been proposed for high-dimensional data in both the frequentist and the Bayesian paradigms. Two predominant classes of methods in frequentist paradigm are penalized likelihood methods and screening based methods. Penalized likelihood methods includes Least Absolute Shrinkage and Selection Operator (LASSO) and its variants like the elastic net of Zou and Hastie (2005), the group LASSO of Yuan and Lin (2006) and the adaptive LASSO of Zou (2006), *etc.*, while the screening based methods include sure independence screening (SIS) of Fan and Lv (2008), iterative SIS (ISIS) of Fan and Song (2010), forward selection-based screening of Wang (2009), nonparametric independence screening (NIS) of Fan *et al.* (2011), iterative varying-coefficient screening (IVIS) of Song *et al.* (2014), *etc.* For a comprehensive review of frequestist variable selection method, see Bühlmann and van de Geer (2011).

In situations with extreme sparsity LASSO-type estimates are outperformed by testing-based subset selection methods (see, for example (Tibshirani, 1996, Section 11)), and tend to overfit. On the other hand, screening based methods focus on marginal association of covariates with the response, and therefore fail to capture the joint structure of the covariates. As a result these methods suffer under presence of multicollinearity, which is almost inenviable in high-dimensional scenario.

In the Bayesian literature, popular methods include the empirical Bayes variable selection (see George and Foster (2000)), where a mixture of testing and optimization is employed to identify the optimal model, fully testing-based methods like spike and slab variable selection (see Ishwaran and Rao (2005)), and optimization and thresholding-based shrinkage prior methods for variable selection like Bayesian LASSO (see Park and Casella (2008)). Among recent developments, the methods of Bondell and Reich (2012), Liang *et al.* (2013), Song and Liang (2015) and Castillo *et al.* (2015) use the idea of penalized credible regions to accomplish variable selection in the ultrahigh-dimensional setting.

Among notable theoretical developments, Castillo *et al.* (2015) proved results related to the posterior consistency for regression parameters, while Liang *et al.* (2013) have shown the equivalence of posterior consistency and model selection consistency under appropriate sparsity assumptions. Narisetty and He (2014) claim to prove the *'strongest selection consistency result'* using the spike and slab prior under under the $\log p_n = o(n)$ setting.

Although the optimization based methods are fast and easily implementable to high-dimensional framework, strong selection consistency property is usually not investigated for these methods. Strong selection consistency, requiring posterior probability of the true model stochastically converging to one, has been shown in Narisetty and He (2014), however, for implementation they rely on the stochastic search variable selection (SSVS) algorithm which is not scalable in high-dimensional situations.

Neighborhood search based SSVS algorithms for the optimal model search are routine for small values of p_n and n, but the resulting computations are quite intensive for higher dimensions due to a large number of possible models. Several authors have developed methods to cope with the high-dimensionality, *e.g.*, Shin *et al.* (2018) proposed a simplified shotgun stochastic search and screening algorithm that employs a variable screening to

reduce neighborhood size in the SSVS algorithm, Li *et al.* (2023) have proposed a highly scalable model-based screening method to explore model space efficiently.

In this paper, we propose a Bayesian method for variable selection and examine its properties both theoretically and numerically, under sparsity assumption. Considering the popular Zellner's g_n-prior (Zellner, 1986) framework, we propose a prior setup suitable for the ultrahigh-dimensional situation. The proposed set of priors has the advantage of generating closed form expressions of the marginals, which makes the resultant method as tractable as the simple information criterion based methods like AIC or BIC.

In a $p_n \gg n$ setting, the size of the model space becomes gigantic and a simple SSVS algorithm can not identify the true model in a finite time. To cope with this situation, we present a two-stage model selection procedure based on an initial screening. The first stage algorithm is intended to select a *good* d_n-dimensional model, where $d_n \ll n$. Under the sparsity assumption, the posterior probability of the class of d_n-dimensional supermodels of the true model uniformly dominates that of of all d_n-dimensional models. Motivated by this result, we first employ a model search algorithm on the space of all d_n-dimensional models. Given an initial model, the algorithm transits to the neighboring d_n-dimensional model with the highest posterior probability. Due to the uniform dominance of the class of supermodels and the less challenging goal of selecting any model in this class, the first-stage algorithm selects a d_n-dimensional supermodel quite efficiently while taking care of joint structure of the covariates, unlike the other screening methods which rely on marginal information.

In the second stage, an SSVS algorithm is employed to search the space of submodels of the first-stage-selected model. Given that a supermodel of the true model is selected at the first stage, the second stage algorithm identifies the true model quite efficiently as $d_n \ll n$. The proposed two-stage algorithm is fast and intuitive. Its good performance is supported by theoretical results under the $\log p_n = O(n)$ settings. To the best of our knowledge, this is the first work on exponential growth of covariates with sample size. The performance of the algorithm is validated extensively with ample simulated and real data sets.

In Section 2, the prior setup and the maximum-a-posteriori (MAP) approach are described. In Section 3, the two-stage algorithm is introduced. Section 4 contains the theoretical results justifying the proposed two-stage algorithm. In Sections 5 and 6, the performance of the proposed algorithm is validated using simulated and real data sets. Section 7 contains concluding remarks. Proofs of all the theoretical results are provided in Section A.

2. The proposed prior setup and the MAP approach

Consider n data points, each consisting of p_n centered regressors $\{x_{1,i}, x_{2,i}, \ldots, x_{p_n,i}\}$ and a centered response y_i with $i = 1, 2, \ldots, n$. The vector of response $\mathbf{y}_n$ is modeled as

$$\mathbf{y}_n = X_n\boldsymbol{\beta} + \mathbf{e}_n, \tag{1}$$

where X_n is the $n \times p_n$ design matrix, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{p_n})'$ is the vector of regression parameters and $\mathbf{e}_n$ is the vector of random errors. For simplicity, we assume that the design matrix X_n is non-stochastic and $\mathbf{e}_n \sim N(\mathbf{0}, \sigma^2 I_n)$.

The space of all models that can be formed by taking at least one covariate is denoted by $\mathcal{G}$, and indexed by γ. Here, $\gamma \in \mathcal{G}$ is a subset of $\{1, \ldots, p_n\}$ of size $p_n(\gamma)$ $(1 \le p_n(\gamma) \le$

p_n), indicating the index set of the covariates corresponding to the model M_γ. Under M_γ, we assume $\mathbf{y}_n = X_\gamma \boldsymbol{\beta}_\gamma + \mathbf{e}_n$, where X_γ is a sub-matrix of X_n consisting of the $p_n(\gamma)$ columns specified by γ and $\boldsymbol{\beta}_\gamma$ is the corresponding vector of regression coefficients. We consider the problem of selecting the sparsest model M_γ with $\gamma \in \mathcal{G}$ that best explains the data.

In a Bayesian approach, each model M_γ is assigned a prior probability and the corresponding set of parameters $\boldsymbol{\theta}_\gamma = (\beta_0, \boldsymbol{\beta}_\gamma, \sigma^2)'$ involved in M_γ, is also assigned a prior distribution. Given prior probability $P(M_\gamma)$ on M_γ and conditional prior density $p(\boldsymbol{\theta}_\gamma|M_\gamma)$ on $\boldsymbol{\theta}_\gamma$ under M_γ, one computes the posterior probability of each model as follows

$$P(M_\gamma|\mathbf{y}_n) = \frac{P(M_\gamma)m_\gamma(\mathbf{y}_n)}{\sum_{\gamma\in\mathcal{G}} P(M_\gamma)m_\gamma(\mathbf{y}_n)}, \qquad \text{where} \qquad m_\gamma(\mathbf{y}_n) = \int p(\mathbf{y}_n|\boldsymbol{\theta}_\gamma, M_\gamma)p(\boldsymbol{\theta}_\gamma|M_\gamma)d\boldsymbol{\theta}_\gamma$$

is the marginal likelihood and $p(\mathbf{y}_n|\boldsymbol{\theta}_\gamma, M_\gamma)$ is the density of $\mathbf{y}_n$ under M_γ. We consider the maximum a-posteriori (MAP) approach which selects the model $\gamma^\star$ in $\mathcal{G}$ with the highest posterior probability as the optimal model.

Throughout this paper, we have considered the following notations and conventions. For two numbers a and b, the notations $a \vee b$ and $a \wedge b$ denote $\max\{a, b\}$ and $\min\{a, b\}$, respectively. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, $a_n \lesssim b_n$ indicates either $a_n/b_n \to 0$ or $a_n \le cb_n$ for all sufficiently large n, and some constant $0 < c < \infty$. Further, if $a_n \gtrsim b_n$ and $b_n \lesssim a_n$, then we write $a_n \sim b_n$. For any square matrix A, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ are the highest and the lowest non-zero eigenvalues of A. For two square matrices A and B of the same order, $A \le B$ means that $B - A$ is positive semidefinite. A model M_γ with dimension $p_n(\gamma) < n$ is said to be of full-rank if $\text{rank}(X'_\gamma X_\gamma) = p_n(\gamma)$.

2.1. Prior specification and posterior probability

Each model M_γ with $\gamma \in \mathcal{G}$ is assigned *Bernoulli* prior $P(M_\gamma) = q_n^{p_n(\gamma)}\,(1 - q_n)^{p_n - p_n(\gamma)}$ with $q_n = 1/p_n$. Given a model M_γ, we consider a conjugate prior on $\boldsymbol{\beta}_\gamma$ as

$$\boldsymbol{\beta}_\gamma|\sigma^2, M_\gamma \ \sim N(\mathbf{0}, g_n\sigma^2 I_{p_n(\gamma)}),$$

where g_n is a hyperparameter. We impose the popular Jeffreys prior $\pi(\sigma^2) \propto 1/\sigma^2$ on σ^2.

The Bernoulli prior is widely used as a model prior probability because of its property of penalizing the models of large dimensions. The choice $q_n = 1/p_n$ has previously been considered by Narisetty and He (2014). This prior is particularly useful for sparse regression models, as it assigns $1/p_n$ weight to each covariate. Thus, the prior probability of a model increases p_n times if one covariate is dropped.

Use of the inverse-gamma prior for error variance is fairly conventional in the literature (see, *e.g.*, George and Mcculloch (1993)). The Jeffreys prior is the limit of inverse-gamma, as both the hyperparameters in the inverse-gamma prior approach zero. The property of invariance under reparametrization makes it suitable as a prior on scale parameter.

For the proposed set of priors, the posterior probability of the model M_γ is

$$P\left(M_\gamma|\mathbf{y}_n\right) \propto \left(\frac{1}{p_n-1}\right)^{p_n(\gamma)} \left|I+g_n X_\gamma' X_\gamma\right|^{-1/2} \left(R_\gamma^{2\star}\right)^{-n/2}, \tag{2}$$

where $R_\gamma^{2\star} = \mathbf{y}_n' \left\{I_n - X_\gamma \left(I_{p_n(\gamma)}/g_n + X_\gamma' X_\gamma\right)^{-1} X_\gamma'\right\} \mathbf{y}_n$.

Our prior choices are simple. Except the choice of g_n, the set of priors is completely specified. Rather than providing a specific choice of g_n, we indicate the optimal order of g_n through theoretical consistency results. The availability of the analytic form of the posterior probability generated by the proposed prior setup (in (2)) makes it easily implementable.

3. Implementation in ultrahigh-dimensional settings

Our model selection procedure is simple as it chooses the model with the highest posterior probability in the model space $\mathcal{G}$, *i.e.*, the MAP model. However, identifying the MAP model is a challenging task in an ultrahigh-dimensional settings. As $p_n = \exp\{O(n)\}$, it is impossible to evaluate all the $2^{p_n}-1$ models in $\mathcal{G}$, even for small values of n. For instance, if $n=5$, the cardinality of $\mathcal{G}$ can be as large as $\exp(45)$. Thus, we need to develop a screening algorithm to discard a large set of *unimportant* covariates initially. Following the implementation of the screening algorithm, ideally, we will be left with a smaller set of covariates which includes all the covariates involved in the MAP model. Then, an exhaustive model search algorithm can be employed in the second stage to find the MAP model. We describe the proposed two-stage algorithm in detail below.

Proposed two-stage algorithm. The proposed two-stage algorithm is based on the *sparsity* assumption, which states that among the large number of available predictors an insignificant fraction of predictors is actually useful. Consequently, the dimension of the MAP model is small. Now, let d_n be a moderately large number, for instance $d_n \sim \log n$. The first step of the two-stage algorithm is devoted towards finding a *good* model of dimension d_n. As the number of useful predictors is small, it is expected that the d_n-dimensional optimal model chosen in first stage includes all the predictors of the MAP model. Towards finding a d_n-dimensional *good* model, a neighborhood-based search algorithm is employed on the space of all d_n-dimensional models. Below, we describe the algorithm.

Stage 1: Screening: The objective of the screening algorithm is to choose a d_n-dimensional model with high posterior probability. Given the choice of d_n, we employ the following steps to achieve this.

1. *Initialization.* Choose a model, say M_{γ_0}, of dimension d_n, where $\boldsymbol{\gamma}_0 \subseteq \{1,\ldots,p\}$ is the index set of the predictors in M_{γ_0}.
2. *Evaluation.* Fix $r \in \boldsymbol{\gamma}_0$. Define

   $$k^\star = \text{argmax}_{l\in\{1,\ldots,p_n\}\setminus\boldsymbol{\gamma}_0} m_{\gamma_0\cup\{l\}\setminus\{r\}}(\mathbf{y}_n), \quad \text{and} \quad \mathrm{u} = \mathbb{I}\left(m_{\gamma_0\cup\{k^\star\}\setminus\{j\}}(\mathbf{y}_n) > m_{\gamma_0}(\mathbf{y}_n)\right)$$

   where $\mathbb{I}(A)$ is the indicator of the event A. If $\mathrm{u}=1$, then replace x_r by $x_{k^\star}$ in $\boldsymbol{\gamma}_0$. If $\mathrm{u}=0$, then keep $\boldsymbol{\gamma}_0$ unaltered.

   Repeat step 2 unless all the components in $\boldsymbol{\gamma}_0$ are evaluated.
3. *Replication.* Repeat Step 2 $N(\geq 1)$ times.

In Step 2, we replace the covariates of M_{γ_0} with the best possible inactive covariates of M_{γ_0}, provided the posterior probabilities increase by the replacements. To obtain the best result, instead of starting with any d_n dimensional model, one may choose the covariates of the initial model M_{γ_0} by a forward regression method.

Finally, we argue that with a good initial model M_{γ_0} the choice of N in Step 3 is expected to be small. We provide the following three intuitive reasons for that: (i) In the screening stage the objective is to arrive at any d_n-dimensional model which contains the MAP-covariates. This is a much easier task than searching for the MAP model. (ii) Under reasonable assumptions, the posterior probability of the class of d_n-dimensional models containing the useful covariates, say $\mathcal{G}_{1,d}$, uniformly dominates the space of all d_n-dimensional models (see Section 4). As the screening algorithm transits to a higher posterior probability model at each move, the complementary class of $\mathcal{G}_{1,d}$, having combined posterior probability close to zero, is stepped aside by the algorithm soon. (iii) Unlike other forward or marginal screening algorithms, the proposed algorithm compares the d_n-dimensional models only. Thus, in one hand the variable dimensional search problem is reduced to a fixed dimensional one, on the other hand the joint structures of the covariates are taken care of.

Stage 2: Model selection: Suppose that the first-stage *screening algorithm* selects the model $M_{\gamma^\star}$. In the next stage, we aim to find the highest-posterior probability model among the $2^{d_n}-1$ models formed by the d_n covariates present in $M_{\gamma^\star}$. Towards that, we employ the reversible jump MCMC (RJMCMC) algorithm described in Chipman *et al.* (2001, Section 3.5), which induces a Markov chain $\mathcal{C}$ with the class of all submodels of $M_{\gamma^\star}$ as the state space, say $\mathcal{G}^\star$. The stationary distribution of $\mathcal{C}$ is the posterior probability distribution of the models restricted to $\mathcal{G}^\star$. Thus, if the covariates of the MAP model of $\mathcal{G}$ is present in $\gamma^\star$, then the MAP model lies in $\mathcal{G}^\star$, and the second stage algorithm reaches the MAP model quite easily, as the cardinality of $\mathcal{G}^\star$ is fairly small.

Remark 1: In practice, the choice of d_n can be as small as possible provided it is larger than the cardinality of the MAP model. A smaller choice of d_n results in faster execution of both the algorithms. The complexity of the first stage screening algorithm is at most of order $O(Nd_np_n)$. Even if one considers all the $2^{d_n}-1$ competing models in $\mathcal{G}^\star$ for comparison in the second stage, the complexity of the second stage algorithm would be at most $O(nd_n^3)$, if $d_n \sim \log n$. Thus the total complexity of the two-stage algorithm is $o(p_n^r)$ for any $r > 1$.

Remark 2: As in the second stage, one could also employ an MCMC algorithm in the first stage. In each iteration, the algorithm would choose a proposal model from the *swap*-neighborhood of the current model and transit to the same according to a Metropolis-Hastings transition function based on the posterior probabilities of the proposal and current models. The algorithm would induce a Markov chain $\mathcal{C}_1$ in the state space $\mathcal{G}_d = \{M_\gamma : p_n(\gamma) = d_n\}$ that would have the posterior probability distribution restricted to $\mathcal{G}_d$ as the stationary distribution. After convergence, it would select model from the high-probability posterior region, *i.e.*, the region of supermodels. However, we avoid taking that path as the proposed screening algorithm is much faster as we will see in the numerical section.

4. Model selection consistency

We consider a frequentist validation approach to theoretically justify the performance of the proposed two-stage algorithm. Towards that, we assume existence of a unique data

generating model, termed as the *true model* (M_{γ_c}), in the model space $\mathcal{G}$. Under M_{γ_c}, $\mathbf{y}_n = \boldsymbol{\mu}_n + \mathbf{e}_n = X_{\gamma_c}\boldsymbol{\beta}_{\gamma_c} + \mathbf{e}_n$, where $\boldsymbol{\mu}_n$ is the expectation of $\mathbf{y}_n$ given X_n. The dimension of M_{γ_c}, denoted by $p(\gamma_c)$, is assumed to be small and free of n. The objective of this section is to show that the two-stage algorithm selects the true model with probability tending to one.

Recall that, the first stage screening algorithm explores the class of all d_n-dimensional models $\mathcal{G}_d$, and at each move it transits to a higher posterior probability model. Thus, it is expected that after sufficient number of moves the algorithm selects a high posterior probability model in $\mathcal{G}_d$. The following subsection (Section 4.1) shows that the posterior probability of the class of all d_n-dimensional supermodels of the true model M_{γ_c}, namely, $\mathcal{G}_{1,d}$, uniformly dominates $\mathcal{G}_d$, with probability tending to one. Thus, with probability tending to one, the high posterior probability model chosen in the first stage will be a supermodel of M_{γ_c}.

In the next stage, we search within the class of all sub-models of the selected model in first stage. As d_n ($\sim \log n$) is small, the second stage RJMCMC algorithm converges to the stationary distribution in finite time. In this case, the stationary distribution is the distribution of posterior probabilities restricted to the sub-models of first stage selected model. Section 4.2 shows that, provided a supermodel of M_{γ_c} is selected at first stage, the restricted posterior distribution converges to a degenerate distribution having non-zero probability mass at M_{γ_c} only, with probability tending to one. Thus, selection of true model is guaranteed with probability tending to one.

Assumptions: Below, we list the assumptions under which our theoretical results hold.

(A1) The number of regressors $p_n = \exp\{b_0 n^r\}$ with $0 < r \leq 1$ and $b_0 > 0$ is free of n.

(A2) The true model M_{γ_c} is unique and its dimension, $p(\gamma_c)$, is free of n. Let $\boldsymbol{\mu}_n = X_{\gamma_c}\boldsymbol{\beta}_{\gamma_c}$ be the true mean of $\mathbf{y}_n$, then $\boldsymbol{\mu}_n'\boldsymbol{\mu}_n = O(n)$.

(A3) Let $\tau_{\max}$ and $\tau_{\min}$ be two positive constants, S be any subset of $\{1, \ldots, p_n\}$ of cardinality $|S| \lesssim \log n$ and X_S be the submatrix of X_n with the columns corresponding to S. Then,

$$n^{-1}\tau_{\min} \leq \inf_S \lambda_{\min}\left(n^{-1}X_S'X_S\right) \leq \sup_S \lambda_{\max}\left(n^{-1}X_S'X_S\right) \leq n\tau_{\max}.$$

(A4) Let $\Delta_0 = \{\delta n^{1-s}\} \vee \{4\sigma^2 p(\gamma_c)\log p_n\}$ for some $\delta > 0$ and $0 < s < 1/2 - \xi$ with $0 < \xi < 1/2$, $\mathcal{G}_0 = \{\gamma \in \mathcal{G} : M_{\gamma_c} \nsubseteq M_\gamma,\ p_n(\gamma) \lesssim \log n\}$ and $P_n(\gamma)$ be the projection matrix onto the span of X_γ. Then, for all sufficiently large n, we have

$$\inf_{\gamma\in\mathcal{G}_0} \boldsymbol{\mu}_n'(I - P_n(\gamma))\boldsymbol{\mu}_n > \Delta_0.$$

Assumption (A1) provides the rate of growth of p_n as a function of n, allowing exponential growth of p_n with respect to n. Assumption (A2) provides the sparsity structure of the true model. Assumption (A3) provides a restriction of the eigenstructure of small dimensional models. By (A3), all models of dimension $O(\log n)$ are of full-rank, although the bounds on the eigenvalues are quite permissive. Assumption (A4) is commonly termed as an identifiability condition for model selection. The quantity $\boldsymbol{\mu}_n'(I - P_n(\gamma))\boldsymbol{\mu}_n$ may be interpreted as the Kullback-Leibler (KL) divergence of the distribution of $\mathbf{y}_n$ under the model M_γ and M_{γ_c}. By Moreno *et al.* (2015, Lemma 3), $\lim_{n\to\infty}\{\boldsymbol{\mu}_n'(I - P_n(\gamma))\boldsymbol{\mu}_n\}/n$ is strictly positive for any non-supermodel of M_{γ_c}. (A3) additionally assumes a uniform lower bound for $\boldsymbol{\mu}_n'(I - P_n(\gamma))\boldsymbol{\mu}_n$ over non-supermodels of small dimension, and fixed a threshold value for the case with $\log p_n \sim b_0 n$. When $\log p_n = b_0 n^{1-r}$ with $r > 0$ the condition is satisfied trivially.

4.1. Consistency of the first-stage screening

Let $\mathcal{G}_d$, $\mathcal{G}_{1,d}$ and $\mathcal{G}_{2,d}$ denote the classes of d_n-dimensional models, supermodels and non-supermodels of M_{γ_c}, respectively. Define $P(M_\gamma \mid \mathcal{G}_d, \mathbf{y}_n)$ as the posterior probability distribution of the models restricted to $\mathcal{G}_d$. The following theorem shows that for any model $\gamma \in \mathcal{G}_d$, the posterior probability of $\gamma \in \mathcal{G}_{1,d}$ uniformly dominates that of $\gamma \in \mathcal{G}_{2,d}$, *i.e.*,

$$P(\gamma \in \mathcal{G}_{1,d} \mid \mathcal{G}_d, \mathbf{y}_n) \to 1, \tag{3}$$

with probability tending to one, as $p_n \to \infty$. This implies that the posterior probability distribution $P(M_\gamma \mid \mathcal{G}_d, \mathbf{y}_n)$ restricted to $\mathcal{G}_d$, assigns nearly 0 probability to $\mathcal{G}_{2,d}$.

Theorem 1: Consider the model stated in (1) with p_n satisfying (A1) and the prior setup discussed in Section 2.1. Suppose there exists a true model M_{γ_c} satisfying (A2) which generates $\mathbf{y}_n$, and let $\mathcal{G}_{1,d}$ and $\mathcal{G}_{2,d}$ be the classes of d_n-dimensional supermodels and non-supermodels of M_{γ_c}. Then, under the assumptions (A3) and (A4) and provided $g_n \gtrsim n$, the following statements hold with a probability at least $1 - \exp\{-c_1 n^\xi\}$, where ξ is as in assumption (A4) and $c_1 > 0$ is some constant free on n.

A. For some constant $c_2 > 0$ and any $\epsilon > 0$,

$$\sup_{\gamma_1 \in \mathcal{G}_{1,d}, \gamma_2 \in \mathcal{G}_{2,d}} \frac{P(M_{\gamma_2}|\mathcal{G}_d, \mathbf{y}_n)}{P(M_{\gamma_1}|\mathcal{G}_d, \mathbf{y}_n)} \leq c_2 n^{d_n} \exp\{-\Delta_0(1-\epsilon)/(2\sigma^2)\}.$$

B. For some constant $c_3 > 0$ and any $\epsilon > 0$,

$$\frac{\sum_{\gamma_2 \in \mathcal{G}_{2,d}} P(M_{\gamma_2}|\mathcal{G}_d, \mathbf{y}_n)}{\sum_{\gamma_1 \in \mathcal{G}_{1,d}} P(M_{\gamma_1}|\mathcal{G}_d, \mathbf{y}_n)} \leq c_3 n^{d_n} p_n^{-(1-2\epsilon)p(\gamma_c)}.$$

C. For any $\gamma \in \mathcal{G}_d$, $P(\gamma \in \mathcal{G}_{1,d} \mid \mathcal{G}_d, \mathbf{y}_n) \to 1$ with probability tending to one, as $n \to \infty$.

In stage 1, the screening algorithm searches for a high-posterior probability model in the restricted model space $\mathcal{G}_d$. By part A of Theorem 1, the posterior probability of the class of models in $\mathcal{G}_{1,d}$ uniformly dominates that of $\mathcal{G}_{2,d}$. Thus, the proposed sequence of $O(Nd_np_n)$ moves in the first-stage algorithm, wherein each move selects a higher posterior probability model, is expected to reach a model in $\mathcal{G}_{1,d}$.

4.2. Consistency of the second-stage selection

As argued in the previous sub-section, the model $M_{\gamma^\star}$ selected in the first stage screening is expected to be a d_n-dimensional supermodel of M_{γ_c}. In the second stage, the RJMCMC algorithm employed explores the class of the all submodels of $M_{\gamma^\star}$, say $\mathcal{G}^\star$. After a sufficient number of iterations, the algorithm selects models as per the posterior distribution restricted to $\mathcal{G}^\star$. The next theorem shows that, if $M_{\gamma^\star}$ is any supermodel of M_{γ_c}, then the posterior distribution restricted to $\mathcal{G}^\star$ limits to a degenerate distribution having non zero probability mass at M_{γ_c}, with probability tending to one. Therefore, provided $M_{\gamma^\star}$ is any supermodel of M_{γ_c}, the second stage algorithm selects M_{γ_c} with probability tending to one.

Theorem 2: Consider the model stated in (1) with p_n satisfying (A1), and the prior setup discussed in Section 2.1 with $g_n \sim p_n^\delta$ with some $0 < \delta < 2$. Suppose there exists a true model M_{γ_c} satisfying (A2), which generates $\mathbf{y}_n$. Let $M_{\gamma^\star}$ be a d_n-dimensional supermodel of M_{γ_c}, $\mathcal{G}^\star = \{M_\gamma : \gamma \subseteq \gamma^\star\}$ be the class of all sub-models of $M_{\gamma^\star}$, and $P\left(M_\gamma \mid \mathcal{G}^\star, \mathbf{y}_n\right)$ be the posterior probability of models restricted to $\mathcal{G}^\star$. Then, under assumptions (A3)-(A4), with a probability at least $1 - cp_n^{-c_0} - \exp\{-cn^\xi\}$, where $c_0 < \delta/2$ and $c > 0$ are two constants, $\xi > 0$ is as in (A4), and $\delta > 0$ is as stated in the choice of g_n, we have

$$\inf_{\gamma^\star \in \mathcal{G}_{1,d}} P\left(M_{\gamma_c} \mid \mathcal{G}^\star, \mathbf{y}_n\right) \geq \left[1 + cn^{p(\gamma_c)+1}\left(\frac{p_n^{2\epsilon}}{g_n}\right)^{1/2} + c\left(\frac{n\sqrt{g_n}}{p_n^{1-2\epsilon}}\right)^{p(\gamma_c)}\right]^{-1}$$

for any $\epsilon < \delta/2$. Consequently, $\inf_{\gamma^\star \in \mathcal{G}_{1,d}} P\left(M_{\gamma_c} \mid \mathbf{y}_n, \mathcal{G}^\star\right) \to 1$, with probability tending to 1.

Theorem 2 states that, provided the first stage algorithm selects any supermodel of M_{γ_c}, the second stage algorithm selects the true model with probability tending to one.

4.3. Consistency of the two-stage procedure

Finally, we argue that the two stage procedure selects the true model with probability tending to one. Towards that, define $P_2(\cdot|\mathbf{y}_n)$ as the probability distribution of the models after the second stage. Let $M_{\gamma^\star}$ be the d_n-dimensional model selected in the first stage, and $\mathcal{G}^\star = \{M_\gamma : \gamma \subseteq \gamma^\star\}$ be the class of all sub-models of $M_{\gamma^\star}$. Then,

$$P_2(M_\gamma|\mathbf{y}_n) = \textstyle\sum_{\gamma^\star \in \mathcal{G}_d} P(M_\gamma|\mathcal{G}^\star, \mathbf{y}_n) P(M_{\gamma^\star}|\mathcal{G}_d, \mathbf{y}_n),$$

if in the first stage a model is selected randomly as per the posterior distribution restricted to $\mathcal{G}_d$, and in the second stage a model is selected randomly as per the posterior distribution restricted to $\mathcal{G}^\star$. The next theorem shows, with probability tending to one, $P_2(M_{\gamma_c}|\mathbf{y}_n) \to 1$.

Theorem 3: Consider the model stated in (1) with p_n satisfying (A1), and the prior setup discussed in Section 2.1 with $g_n \sim p_n^\delta$ with some $0 < \delta < 2$. Suppose there exists a true model M_{γ_c} satisfying (A2), which generates $\mathbf{y}_n$. Further, suppose that a two stage procedure is employed to identify the true model, wherein the first stage selects a d_n-dimensional model $M_{\gamma^\star}$ randomly as per the posterior distribution (2) restricted to $\mathcal{G}_d$ (class of d-dimensional models), and the second stage selects a model randomly from the posterior distribution (2) restricted to the sub-models of $M_{\gamma^\star}$ (*i.e.*, $\mathcal{G}^\star$). Let $P_2(\cdot \mid \mathbf{y}_n)$ be the probability distribution of the models selected at the end of the two-stage procedure then under assumptions (A3)-(A4), $P_2(M_{\gamma_c}|\mathbf{y}_n) \to 1$ as $n \to \infty$, with probability tending to one.

Remark 3: The choice of the only hyperparameter g_n in the prior setup is not specified. However, from the above theoretical developments, we obtain an optimal range of g_n value required for consistency of the two-stage procedure. Theorem 1 holds for any g_n satisfying $g_n \gtrsim n$, while Theorems 2 and 3 requires $g_n \sim p_n^\delta$ with $0 < \delta < 2$. These provide a vast range of plausible choices of g_n. For practical purposes some sensitivity analysis would be useful.

5. Simulation study

We now study the performance of the proposed two-stage variable selection procedure using a wide variety of simulated data sets. Under different simulation schemes, we present

the proportion of times a variable selection algorithm selects the true model.

Our method: Our model selection algorithm is completely described in Section 3, except for the choices of g_n, d_n and N. The choice of d_n is taken to be $[n/4]$ in each case. In the first stage, we choose $g_n = np_n$ and in the second stage, we choose $g_n = d_n^2$. Note that, the theoretical condition on g_n in Theorems 2 and 3 come from the consideration of the two-stages together. However, practically, the task of the second stage is find the MAP model among the $2^{d_n} - 1$ models formed by d_n covariates. Therefore, informed by Fernández *et al.* (2001), the benchmark prior $g_n = \max\{n, d_n^2\}$ is considered in the second stage. Finally, in the first-stage $N = 10$ iterations are considered, and in the second-stage, the RJMCMC algorithm is iterated 6000 times, with a burning period of 3000 iterations. The post-burning most visited model is considered as the optimal model.

Other methods: Among the frequentist variable selection methods, we consider three approaches based on iterative sure independence screening (ISIS). An initial set of variables is first selected by ISIS, and then a penalized regression step is carried out using the least absolute shrinkage and selection operator (LASSO), smoothly clipped absolute deviation (SCAD), or minimax concave penalty (MCP, Zhang (2010)) with the regularization parameter tuned using the BIC. These three methods are termed as ISIS-LASSO-BIC, ISIS-SCAD-BIC and ISIS-MCP-BIC. Among the Bayesian competitors, we consider two methods based on Bayesian credible region (BCR joint and BCR marginal, Bondell and Reich (2012)) and Bayesian shrinking and diffusing prior (BASAD, Narisetty and He (2014)). We have used R codes for all the methods. For ISIS, we have implemented codes from the R package `SIS`. The R codes for BCR are obtained from the first author's website, while the first author of Narisetty and He (2014) kindly shared the codes for BASAD with us. Further, we have implemented the approximate version of BASAD to reduce the computing time.

Simulation setup. We consider two values for n, namely, 50 and 100. For $n = 50$, we choose $p_n = 100$ and 500, while for $n = 100$ we choose $p_n = 500$, 1000 and 2000. The model $\mathbf{y}_n = \boldsymbol{\mu}_n + \mathbf{e}_n$ is considered as the true model, where $\boldsymbol{\mu}_n = X_{\gamma_c}\boldsymbol{\beta}_{\gamma_c}$. The vector $\boldsymbol{\beta}_{\gamma_c}$ is assumed to be sparse, *i.e.*, $p(\gamma_c) \ll p_n$, and these $p(\gamma_c)$ components are chosen randomly from the set of all covariates. When $p_n \leq 500$, we set $p(\gamma_c) = 5$, while $p(\gamma_c) = 10$ is set for higher values of p_n. All the $p(\gamma_c)$ values of $\boldsymbol{\beta}_{\gamma_c}$ are taken to be equal to 2.

Each data row $\mathbf{x}_i$ of the design matrix $X_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ is assumed to follow the Gaussian distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}_{p_n}$ for $i = 1, \ldots, n$. The covariance structure of $\boldsymbol{\Sigma}_{p_n} = ((\sigma_{ij}))$ for $1 \leq i, j \leq p_n$ is taken to be of the following four types:

Case 1. (*Identity*) $\boldsymbol{\Sigma}_{p_n} = \mathbf{I}$, *i.e.*, there is no correlation among the covariates.

Case 2. (*Block dependence*) $\boldsymbol{\Sigma}_{p_n}$ has a block covariance setting, where the *active* covariates have common correlation $\rho_1 = 0.25$, the *inactive* covariates have common correlation $\rho_2 = 0.75$ and each pair of active and inactive covariate has correlation $\rho_3 = 0.50$. This is an interesting co-variance structure as it attributes different correlations depending on whether the covariate is important, or not (also see Narisetty and He (2014)).

Case 3. (*Equi-correlation*) $\boldsymbol{\Sigma}_{p_n} = 0.5\mathbf{I} + 0.5\mathbf{1}\mathbf{1}'$, where $\mathbf{1}$ is the p_n-dimensional vector of ones. This exhibits a strong dependence structure uniformly among the covariates.

Case 4. (*Auto-regressive*) Here, we take $\sigma_{ii} = 1$ for $1 \leq i \leq p_n$, and $\sigma_{ij} = 0.9^{|i-j|}$ for

$1 \leq i \neq j \leq p_n$. With the increase in distance, the correlation decreases here.

Although theoretically we consider only Gaussian errors, in simulation studies, we consider two errors distributions, namely, the Gaussian and the heavy-tailed t distribution with 2 degrees of freedom. In the tables below, we report the proportion of times each method selects the true model in 100 random iterations. Additionally, we report the proportion of times our first-stage screening algorithm chooses a supermodel of the true model.

Simulation results. Tables 1 and 2 contain the results corresponding to $n = 50$ and $n = 100$, respectively. We notice that the covariance structure in Case 2 becomes singular for $p_n \geq 1000$, and therefore, we have restricted Case 2 to $p_n \leq 500$.

Table 1: Proportion of times true model is selected by each method for $n = 50$

Gaussian error								
Methods	Case 1		Case 2		Case 3		Case 4	
$\downarrow \quad p_n \rightarrow$	100	500	100	500	100	500	100	500
ISIS-SCAD-BIC	0.65	0.42	0.05	0.00	0.46	0.19	0.66	0.38
ISIS-MCP-BIC	0.44	0.23	0.02	0.00	0.12	0.04	0.50	0.24
BCR	0.26	0.00	0.45	0.00	0.15	0.00	0.22	0.00
BASAD	0.93	0.50	**0.82**	0.07	0.82	0.55	0.92	0.49
Proposed	**0.99**	**0.84**	0.72	**0.09**	**0.96**	**0.80**	**1.00**	**0.87**
Proposed (Step 1)	1.00	0.85	0.77	0.09	0.96	0.81	1.00	0.87
t_2 error								
	Case 1		Case 2		Case 3		Case 4	
Methods $\downarrow \quad p_n \rightarrow$	100	500	100	500	100	500	100	500
ISIS-SCAD-BIC	0.33	0.34	0.02	0.00	0.28	0.21	0.33	0.29
ISIS-MCP-BIC	0.26	0.26	0.02	0.00	0.20	0.17	0.27	0.26
BCR	0.15	0.01	0.29	0.00	0.12	0.00	0.20	0.00
BASAD	**0.69**	0.30	**0.55**	**0.09**	0.61	0.38	0.69	0.37
Proposed	**0.69**	**0.60**	0.54	0.08	**0.66**	**0.53**	**0.72**	**0.59**
Proposed (Step 1)	0.83	0.67	0.65	0.08	0.77	0.56	0.84	0.65

Among the three frequentist methods based on ISIS, we have reported the results for SCAD and MCP only, as ISIS-LASSO-BIC is outperformed by these two methods. For the other two methods, SCAD has shown uniformly better performance than MCP (see Table 1). For BCR, we observe that the joint version leads to singularity in several iterations in the simulation settings. Therefore, we have reported results for the more stable marginal version only. It is also clear from Tables 1 and 2 that ISIS is affected drastically when the dependence structure varies among the different sets of covariates. For example, for $n = 100$, ISIS-SCAD-BIC leads to the best performance under independence (Case 1) when $p_n = 2000$. However, it fails to identify the true model in a single instance under block-diagonal covariance structure (Case 2). This is due to the fact that ISIS relies on marginal information, and ignores the joint structure of the covariates.

Generally, the Bayesian methods turn out to be *more robust* than frequentist approaches. Among the Bayesian methods, BASAD and the proposed method clearly outperform BCR for all the cases. However, the performance of BASAD falls drastically for higher values of p_n. For example, when $p_n = 2000$, BASAD fails completely, irrespective of the underlying covariance structure. Note that BASAD needs to compute the inverse of

Table 2: Proportion of times true model is selected by each method for $n = 100$

Gaussian error										
Methods ↓	Case 1			Case 2	Case 3			Case 4		
$p_n \rightarrow$	500	1000	2000	500	500	1000	2000	500	1000	2000
ISIS-SCAD-BIC	0.85	0.39	0.28	0.00	0.64	0.18	0.02	0.84	0.43	0.25
ISIS-MCP-BIC	0.66	0.25	0.16	0.00	0.11	0.01	0.00	0.62	0.24	0.11
BCR	0.38	0.00	0.00	0.39	0.14	0.00	0.00	0.24	0.00	0.00
BASAD	0.93	0.19	0.00	0.92	0.93	0.36	0.00	0.98	0.27	0.00
Proposed	**0.98**	**0.95**	**0.66**	**0.97**	**1.00**	**0.92**	**0.31**	**1.00**	**0.92**	**0.27**
Proposed (Step 1)	1.00	0.96	0.66	0.97	1.00	0.92	0.31	1.00	0.93	0.57
t_2 error										
Methods ↓	Case 1			Case 2	Case 3			Case 4		
$p_n \rightarrow$	500	1000	2000	500	500	1000	2000	500	1000	2000
ISIS-SCAD-BIC	0.44	0.41	**0.32**	0.00	0.39	0.29	0.30	0.45	0.36	0.30
ISIS-MCP-BIC	0.38	0.39	0.29	0.00	0.24	0.23	0.28	0.40	0.33	0.27
BCR	0.26	0.00	0.00	0.23	0.09	0.00	0.00	0.21	0.00	0.00
BASAD	0.91	0.06	0.00	0.75	0.78	0.19	0.00	**0.88**	0.12	0.00
Proposed	**0.93**	**0.70**	0.21	**0.84**	**0.85**	**0.60**	**0.39**	0.78	**0.70**	**0.39**
Proposed (Step 1)	0.96	0.70	0.48	0.87	0.95	0.60	0.39	0.78	0.71	0.40

the covariance matrix for each model, which is computationally prohibitive for such high-dimensional data. To resolve this problem, they use a block covariance structure to simplify some of the matrix computations and this might be one of the reasons behind its poor performance. The strength of our proposed method is re-iterated from the simulation study, especially for higher values of p_n. Notably, there is a systematic improvement of the proposed method over BASAD when we move from $p_n = 100$ to $p_n \geq 500$, especially under cases 1, 3 and 4, for both the error distributions.

The performance of the first-stage screening algorithm is noteworthy. Except for the high-dimension-low-sample size situation with high correlation, *i.e.*, for $n = 50$, $p_n = 500$ in Case 2, this algorithm selects the true model for a high-proportion of times in all other cases.

To check the sensitivity of our method to the value of $\boldsymbol{\beta}_{\gamma_c}$, we perform a further simulation study. We consider Case 1 ($\boldsymbol{\Sigma}_{p_n} = \mathbf{I}$) with the Gaussian error distribution for $n = 100$; and two choices of $\boldsymbol{\beta}_{\gamma_c}$. First, a set of equi-spaced values of $\boldsymbol{\beta}_{\gamma_c}$ in the range $[1, 2]$ and next in the range $[2, 3]$. An increment of 0.2 is taken for $p_n = 500$ so that we have $p(\gamma_c) = 6$, and an increment of 0.1 is taken for $p_n = 1000$ and 2000 so that $p(\gamma_c) = 11$. The results are summarized in Table 3 below.

Table 3: Proportion of times true model is selected by each method for $n = 100$

Methods ↓	$\boldsymbol{\beta}_{\gamma_c} = (1.0, 1.2, \ldots, 2)'$			$\boldsymbol{\beta}_{\gamma_c} = (2.0, 2.1, \ldots, 3)'$		
$p_n \rightarrow$	500	1000	2000	500	1000	2000
ISIS-SCAD-BIC	0.66	0.40	0.24	0.82	0.47	0.33
ISIS-MCP-BIC	0.63	0.26	0.00	0.68	0.27	0.19
BCR	0.14	0.00	0.00	0.24	0.00	0.00
BASAD	**0.99**	0.14	0.00	0.98	0.28	0.00
Proposed	**1.00**	**0.93**	**0.87**	**1.00**	**0.94**	**0.76**
Proposed (Step 1)	1.00	0.93	0.87	1.00	0.94	0.76

Good performance of the proposed method is further re-iterated from the numerical results of Table 3. Also, it is observed that the method is not much sensitive to the level of signal strength, as long as the minimal signal strength is not negligible.

6. Real data analysis

6.1. Metabolic quantitative trait loci experiment

The first example is related to a metabolic quantitative trait loci experiment which links single nucleotide polymorphisms (SNPs) data to metabolomics data. The *predictors* come from a GWAS study of the candidate genes for alanine amino-transferase enzyme elevation in the liver along with the mass spectroscopy metabolomics data. A total of 10000 SNPs are pre-selected as candidate predictors, and the number of subjects included in the data set is 50. The genotype of each SNP is coded as 0, 1 and 2 for homozygous rare, heterozygous, and homozygous common allele, respectively. A particular metabolite bin that discriminates well between the disease status of the clinical trial's participants is selected as the *response variable*.

The SAM approach of Song and Liang (2015) selected two SNPs, rs17041311 and rs17392161. The first SNP has the same genotype as the SNP rs7896824, while the second SNP shares the same genotype with eleven other SNPs. We implement our proposed method by starting with $d_n = 5$ till $d_n = 50$ (which is the maximum possible value of d_n). From our analysis, the proposed method identifies all the SNPs (two from the first group, and all the twelve from the second group) from $d_n = 25$ onwards. We further observe that the proposed method consistently identifies a new set of SNPs consists of rs6704330 and rs12744386. This is a novel set of SNPs which *were not detected* in the earlier study, and further investigation may establish their association with the metabolite under study.

For the sake of comparison, we implement all the competing methods from our simulations in Section 5. We first fix a value of the model size (d_n), and then a model selection method is used to obtain a d_n-dimensional subset of the predictor variables. To assess the relative performance of these methods, we compute both the mean and the median square errors based on leave-one-out cross-validation (LOOCV). For all the methods, values of the mean square errors turn out to be quite high. Therefore, we use the median square errors for comparison. For increasing values of d_n, Figure 1 below gives us an idea about the overall performance of each of these methods. Clearly, BASAD yields the lowest median square of errors, while the performance for our proposal is the second best.

6.2. Polymerase chain reaction

This data is related to a polymerase chain reaction. A total of $n = 60$ samples, with 31 female and 29 male mice, are used to monitor the expression levels of $p_n = 22575$ genes. Some physiological phenotypes, including numbers of phosphoenolpyruvate carboxykinase, glycerol-3-phosphate acyltransferase, and stearoyl-CoA desaturase 1 are measured by quantitative real-time polymerase chain reaction. The relationship between the gene expression level (*perdictor*) and phosphoenolpyruvate carboxykinase (*response*) is of interest in this data. The gene expression data is standardized before the statistical analysis. To analyze this data, we repeat the same procedure as in Section 6.1 above.

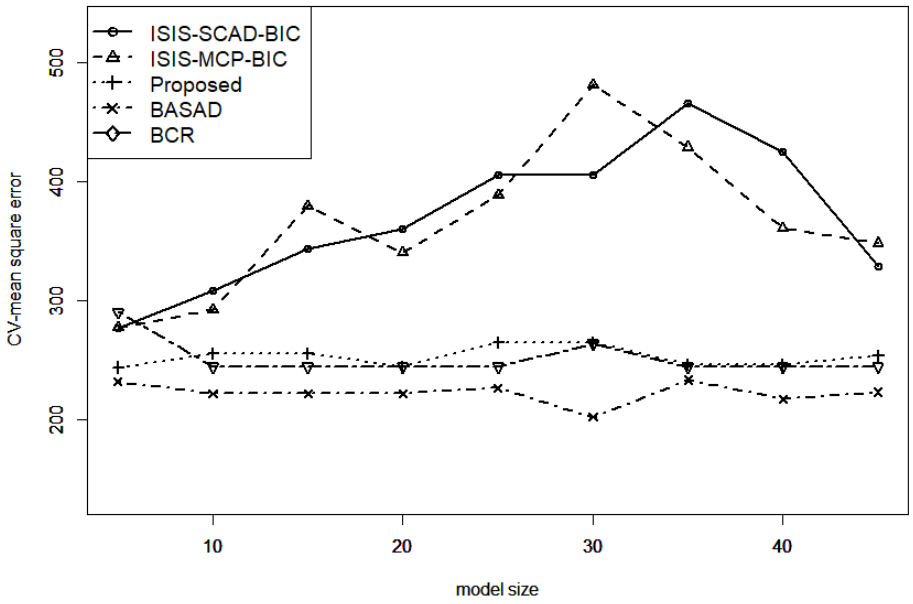


Figure 1: Comparison of the different methods using median square errors

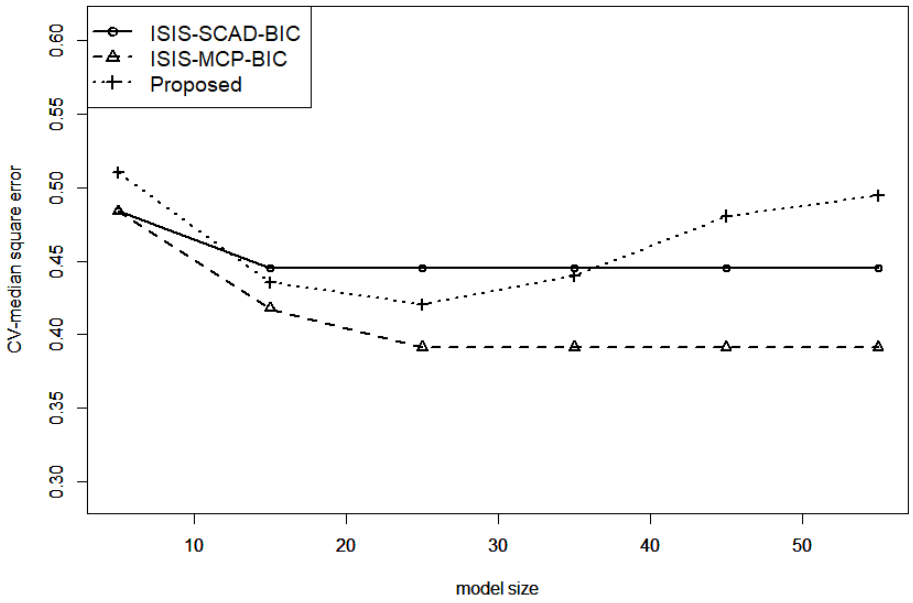


Figure 2: Comparison of the different methods using median square errors

Both BASAD and BCR could not be implemented for this data due to memory overflow for this data. Figure 2 gives us the overall picture of the performance of the other methods, and they all yield quite low median square errors. Clearly, ISIS-MCP leads to the lowest overall errors, and the proposed method performs marginally better than ISIS-SCAD

for $d_n = 15$ to 35. However, the maximum difference in errors of the proposed method with both methods based on ISIS is less than 0.11 over all values of d_n.

7. Concluding remarks

This paper addresses the variable selection problem in ultrahigh-dimensional linear regression settings. A new methodology for variable selection based on Zellner's g-prior is developed, taking into account the key features of the ultrahigh-dimensional regression settings, such as sparsity and multicollinearity, and adapting it accordingly. Variable selection in ultrahigh dimensions poses significant challenges due to the exponential growth of the model space with the number of covariates. Despite its various advantages, the predominant Bayesian variable selection procedure, the maximum a-posteriori (MAP) approach, becomes impractical in this context due to the vast model space. To address this problem, we propose a two-stepped model selection procedure that incorporates an initial screening.

While the idea of screening out unimportant covariates in the initial stage is not new, existing screening algorithms typically rely on marginal utilities and overlook the joint structure of the covariates. Our proposed screening algorithm takes the joint structure of the covariates into account, demonstrating greater efficiency and robustness across various correlation structures, as evidenced by our numerical results. In the second stage, we conduct a thorough model search within the class of submodels of the first-stage-selected model. Notably, we establish the strong selection consistency property of our two-stage algorithm theoretically under exponential growth of p_n with n. To our knowledge, this is the first selection consistency result addressing the exponential growth of p_n with n.

We conclude this section with some future directions. The effectiveness of our proposed two-stage procedure is heavily dependent on the sparsity assumption of the optimal model. While sparsity is commonly observed in high-dimensional regression, it is essential to expedite the search for the MAP model in denser cases as well. Relevantly, the choice of d_n is a critical factor in our method. A smaller d_n can enhance the speed and efficiency of both algorithms but may also lead to exclusion of important covariates. Thus, it is necessary to develop a mechanism for determining the optimal choice of d_n based on the data at hand.

Acknowledgements

We are thankful to the reviewer and the Editor for their constructive comments and helpful suggestions. The first author has been partially supported by the grant MTR/2023/001170, while the second author has been partially supported by the grant MTR/2023/000884.

Reproducibility

Codes for the proposed two-stage algorithm are available in the following link: https://github.com/mmukhopadhyay/Bayesian-Variable-Selection-for-Ultrahigh-dimensional-Sparse-Linear-Models.git.

References

Bondell, H. D. and Reich, B. J. (2012). Consistent high-dimensional Bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association*, **107**, 1610–1624.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer Series in Statistics. Springer, Heidelberg. Methods, theory and applications.

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, **43**, 1986–2018.

Chipman, H., George, E. I., and McCulloch, R. E. (2001). The practical implementation of Bayesian model selection. In *Model selection*, volume 38 of *IMS Lecture Notes-Monograph Series*, pages 65–134. Institute of Mathematical Statistics, Beachwood, OH.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, **106**, 544–557.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, **70**, 849–911.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, **38**, 3567–3604.

Fernández, C., Ley, E., and Steel, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, **100**, 381–427.

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, **87**, 731–747.

George, E. I. and Mcculloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, **33**, 730–773.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, **28**, 1302–1338.

Li, D., Dutta, S., and Roy, V. (2023). Model based screening embedded Bayesian variable selection for ultra-high dimensional settings. *Journal of Computational and Graphical Statistics*, **32**, 61–73.

Liang, F., Song, Q., and Yu, K. (2013). Bayesian subset modeling for high-dimensional generalized linear models. *Journal of the American Statistical Association*, **108**, 589–606.

Moreno, E., Girón, J., and Casella, G. (2015). Posterior model consistency in variable selection as the model dimension grows. *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, **30**, 228–241.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, **42**, 789–817.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, **103**, 681–686.

Shin, M., Bhattacharya, A., and Johnson, V. E. (2018). Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statistica Sinica*, **28**, 1053–1078.

Song, Q. and Liang, F. (2015). A split-and-merge Bayesian variable selection approach for ultrahigh dimensional regression. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, **77**, 947–972.

Song, R., Yi, F., and Zou, H. (2014). On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statistica Sinica*, **24**, 1735–1752.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Methodological*, **58**, 267–288.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, **104**, 1512–1524.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, **68**, 49–67.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti (P. K. Goel and A. Zellner, eds.)*, pages 233–243.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**, 894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, **67**, 301–320.

A. ANNEXURE

This section contains the proof of all the theorems. In all the proofs, the notation c is used as a generic symbol for constants. In many situations, the existence of a constant, rather than the value, is important. In such cases, the constant is denoted by c. Thus, all constants denoted by c are not necessarily the same.

A.1. Auxiliary results

In this section, we present auxiliary results which are used in proving the main results.

Lemma 1: Let M_γ, $\gamma \in \mathcal{G}$ and $M_{\gamma'}$ be two models with dimensions $p_n(\gamma)$ and $p_n(\gamma')$, where $p_n(\gamma), p_n(\gamma') \lesssim \log n$. Further, suppose $\tau_{\min} \leq \lambda_{\min}(A) \leq \lambda_{\max}(A) \leq n^2\tau_{\max}$ for some $\tau_{\min} > 0$ and $\tau_{\max} > 0$ (free of n) for both matrices $A = X_\gamma' X_\gamma$ and $A = X_{\gamma'}' X_{\gamma'}$. Then,

$$\frac{\left|I + g_n X_\gamma' X_\gamma\right|^{-1}}{\left|I + g_n X_{\gamma'}' X_{\gamma'}\right|^{-1}} = \frac{\left|I + g_n X_{\gamma'}' X_{\gamma'}\right|}{\left|I + g_n X_\gamma' X_\gamma\right|} \leq (1+\epsilon)\left(\frac{\tau_{\max}}{\tau_{\min} \wedge 1}\right)^{p_n(\gamma)\vee p_n(\gamma')} n^{2p_n(\gamma')} g_n^{p_n(\gamma')-p_n(\gamma)},$$

for any $\epsilon > 0$, when $g_n \gtrsim n$.

Proof: The j-th largest eigenvalue of any square matrix of the form $I + A$ are $1 + \lambda_j(A)$, where $\lambda_j(A)$ is the j-th largest eigenvalue of A. Further, both $X_\gamma' X_\gamma$ and $X_{\gamma'}' X_{\gamma'}$ are non-negative definite. Therefore, the highest eigenvalue of $I + g_n X_{\gamma'}' X_{\gamma'}$ is $1 + g_n n^2 \tau_{\max}$ and the lowest eigenvalue of $I + g_n X_\gamma' X_\gamma$ is $1 + \tau_{\min}$. By the trivial bound $\lambda_{\min}^d(A) \leq |A| \leq \lambda_{\max}^d(A)$, where d is the dimension of A, we get

$$\begin{aligned}
\frac{\left|I + g_n X_{\gamma'}' X_{\gamma'}\right|}{\left|I + g_n X_\gamma' X_\gamma\right|} &\leq \frac{(1 + g_n n^2 \tau_{\max})^{p_n(\gamma')}}{(1 + g_n \tau_{\min})^{p_n(\gamma)}} \\
&= n^{2p_n(\gamma')} g_n^{p_n(\gamma')-p_n(\gamma)} \tau_{\max}^{p_n(\gamma')} \tau_{\min}^{-p_n(\gamma)} \frac{\{1 + 1/(g_n n^2 \tau_{\max})\}^{p_n(\gamma')}}{\{1 + 1/(g_n \tau_{\min})\}^{p_n(\gamma)}} \\
&\leq (1+\epsilon)\left(\frac{\tau_{\max}}{\tau_{\min} \wedge 1}\right)^{p_n(\gamma)\vee p_n(\gamma')} n^{2p_n(\gamma')} g_n^{p_n(\gamma')-p_n(\gamma)},
\end{aligned}$$

for any $\epsilon > 0$ whenever $g_n \gtrsim n$. The last inequality is due to the fact that both terms $(1 + g_n n^2 \tau_{\max})^{p_n(\gamma')}$ and $(1 + g_n \tau_{\min})^{p_n(\gamma)}$ converges to one as $n \to \infty$ if $g_n \gtrsim n$. □

Lemma 2: Let M_γ be a full-rank model, $R_\gamma^{2\star} = \mathbf{y}_n' \left\{ I_n - X_\gamma \left(I_{p_n(\gamma)}/g_n + X_\gamma' X_\gamma \right)^{-1} X_\gamma' \right\} \mathbf{y}_n$, and $R_\gamma^2 = \mathbf{y}_n' \{I_n - P_n(\gamma)\} \mathbf{y}_n$, where $P_n(\gamma) = X_\gamma (X_\gamma' X_\gamma)^{-1} X_\gamma'$ is the projection matrix on the column space of X_γ. Then, under the assumptions (A2)-(A3), the following statements hold.

(a) $R_\gamma^{2\star} \geq R_\gamma^2$, and for any model M_γ satisfying (A3), $\sup_\gamma R_\gamma^{2\star} - R_\gamma^2 \leq cn/(1 + g_n \tau_{\min})$ for some appropriate constant $c > 0$ with probability at least $1 - \exp\{-n\}$,

(b) For any $\epsilon > 0$, there exists an appropriate constant $c > 0$ such that $R_{\gamma_c}^2 > n(1+\epsilon)\sigma^2$, and $R_{\gamma_c}^2 < n(1-\epsilon)\sigma^2$, with probability at least $1 - \exp\{-cn\}$.

Proof: Part(a). Observe that $I_{p_n(\gamma)}/g_n + X'_\gamma X_\gamma \geq X'_\gamma X_\gamma$, and so, $I_n - X_\gamma \left(I_{p_n(\gamma)}/g_n + X'_\gamma X_\gamma\right)^{-1} X'_\gamma \geq I_n - P_n(\gamma)$, which proves $R^2_\gamma \leq R^{2\star}_\gamma$.

To see the other side, observe that under (A3), and uniformly over any model M_γ

$$\begin{aligned} X_\gamma \left(I_{p_n(\gamma)}/g_n + X'_\gamma X_\gamma\right)^{-1} X'_\gamma &= X_\gamma \left(X'_\gamma X_\gamma\right)^{-1/2} \left[I_{p_n(\gamma)} + \left(X'_\gamma X_\gamma\right)^{-1}/g_n\right]^{-1} \left(X'_\gamma X_\gamma\right)^{-1/2} X'_\gamma \\ &\geq \{1 + 1/(g_n\tau_{\min})\}^{-1} P_n(\gamma) \end{aligned}$$

as $\lambda_{\max}\left(I_{p_n(\gamma)} + \left(X'_\gamma X_\gamma\right)^{-1}/g_n\right) \leq 1 + 1/(g_n\tau_{\min})$. Therefore,

$$\begin{aligned} \sup_{\gamma: p_n(\gamma) \lesssim \log n} R^{2\star}_\gamma - R^2_\gamma &\leq \sup_{\gamma: p_n(\gamma) \lesssim \log n} \mathbf{y}'_n \left[I_n - \{1 + 1/(g_n\tau_{\min})\}^{-1} P_n(\gamma) - I_n + P_n(\gamma)\right] \mathbf{y}_n \\ &= \sup_{\gamma: p_n(\gamma) \lesssim \log n} \frac{1}{1 + g_n\tau_{\min}} \mathbf{y}'_n P_n(\gamma) \mathbf{y}_n \leq \frac{1}{1 + g_n\tau_{\min}} \mathbf{y}'_n \mathbf{y}_n. \end{aligned}$$

Now, $\mathbf{y}'_n\mathbf{y}_n \leq 2\|\boldsymbol{\mu}_n\|^2 + 2\|\mathbf{e}_n\|^2$. By assumption (A2), $\|\boldsymbol{\mu}_n\|^2 = O(n)$ and as $\|\mathbf{e}_n\|^2 \sim \sigma^2\chi^2_n$, from Laurent and Massart (2000), we have $\|\mathbf{e}_n\|^2 \leq 6n\sigma^2$ with probability at least $1 - \exp\{-n\}$. Therefore, with probability at least $1 - \exp\{-n\}$, $R^{2\star}_\gamma - R^2_\gamma \leq cn/(1 + g_n\tau_{\min})$ for some appropriate constant $c > 0$.

Part(b). The random variable $\mathbf{e}'_n(I - P_n(\gamma_c))\mathbf{e}_n/\sigma^2$ follows a χ^2 distribution with $(n - p(\gamma_c))$ degrees of freedom. By (Laurent and Massart, 2000, Lemma 1), we have

$$\begin{aligned} P(R^2_{\gamma_c} > n(1+\epsilon)\sigma^2) &= P(\mathbf{y}'_n(I - P_n(\gamma_c))\mathbf{y}_n > n(1+\epsilon)\sigma^2) \\ &= P\left(\mathbf{e}'_n(I - P_n(\gamma_c))\mathbf{e}_n > n(1+\epsilon)\sigma^2\right) \\ &\leq \exp\left\{-c\frac{(n\epsilon + p(\gamma_c))^2}{(n - p(\gamma_c))}\right\} \leq \exp\{-cn\}, \end{aligned}$$

for some $c > 0$. Thus, the first part of the result follows. The proof of the second part follows similarly from (Laurent and Massart, 2000, Lemma 1). □

Lemma 3: Let $\mathbf{y}_n = \boldsymbol{\mu}_n + \mathbf{e}_n$ with $\mathbf{e}_n \sim N(\mathbf{0}, \sigma^2 I)$ and $\boldsymbol{\mu}'_n\boldsymbol{\mu}_n = O(n)$. For any $0.5 < k < 1$ and $\epsilon > 0$, there exists a constant $c > 0$ such that $n^{-k}|\boldsymbol{\mu}'_n\mathbf{e}_n| < \epsilon$ with probability at least $1 - \exp\{-cn^{2k-1}\}$

Proof: The random variable $\boldsymbol{\mu}'_n\mathbf{e}_n$ is distributed as a centered normal distribution with variance $\sigma^2\|\boldsymbol{\mu}_n\|^2$. Therefore, we get

$$P\left(|\boldsymbol{\mu}'_n\mathbf{e}_n| \geq \epsilon n^k\right) \leq \exp\{-cn^{2k}/\|\boldsymbol{\mu}_n\|^2\} \tag{4}$$

for an appropriate constant $c > 0$ depending on ϵ. By assumption (A2), $\|\boldsymbol{\mu}_n\|^2 = O(n)$. Therefore, the quantity on the right-hand side of the above expression is bounded above by $\exp\{-cn^{2k-1}\}$ for some $c > 0$. Thus, the result follows. □

A.2. Main results

A.2.1. Proof of Theorem 1

Proof: [**Part A.**] By (2), the ratio of posterior probabilities is

$$\begin{aligned}\sup_{\gamma_1\in\mathcal{G}_{1,d},\gamma_2\in\mathcal{G}_{2,d}}\frac{P(M_{\gamma_2}|\mathcal{G}_d,\mathbf{y}_n)}{P(M_{\gamma_1}|\mathcal{G}_d,\mathbf{y}_n)} &= \sup_{\gamma_1\in\mathcal{G}_{1,d},\gamma_2\in\mathcal{G}_{2,d}}\frac{P(M_{\gamma_2}|\mathbf{y}_n)}{P(M_{\gamma_1}|\mathbf{y}_n)}\\ &= \sup_{\gamma_1\in\mathcal{G}_{1,d},\gamma_2\in\mathcal{G}_{2,d}}\frac{\left|I+g_nX'_{\gamma_1}X_{\gamma_1}\right|^{1/2}}{\left|I+g_nX'_{\gamma_2}X_{\gamma_2}\right|^{1/2}}\left(\frac{R^{\star 2}_{\gamma_1}}{R^{\star 2}_{\gamma_2}}\right)^{n/2}. \qquad (5)\end{aligned}$$

By assumption (A3) and Lemma 1

$$\sup_{\gamma_1\in\mathcal{G}_{1,d},\gamma_2\in\mathcal{G}_{2,d}}\frac{\left|I+g_nX'_{\gamma_1}X_{\gamma_1}\right|^{1/2}}{\left|I+g_nX'_{\gamma_2}X_{\gamma_2}\right|^{1/2}}\leq 2\left(\frac{\tau_{\max}}{\tau_{\min}\wedge 1}\right)^{d_n/2}n^{d_n}. \qquad (6)$$

Next, we write the last part in the RHS of (5) as follows:

$$\begin{aligned}\sup_{\gamma_1\in\mathcal{G}_{1,d},\gamma_2\in\mathcal{G}_{2,d}}\left(\frac{R^{\star 2}_{\gamma_1}}{R^{\star 2}_{\gamma_2}}\right)^{n/2} \leq & \sup_{\gamma_1\in\mathcal{G}_{1,d}}\left(\frac{R^{\star 2}_{\gamma_1}}{R^2_{\gamma_1}}\right)^{n/2}\sup_{\gamma_1\in\mathcal{G}_{1,d}}\left(\frac{R^2_{\gamma_1}}{R^2_{\gamma_c}}\right)^{n/2}\\ & \sup_{\gamma_2\in\mathcal{G}_{2,d}}\left(\frac{R^2_{\gamma_c}}{R^2_{\gamma_2}}\right)^{n/2}\sup_{\gamma_2\in\mathcal{G}_{2,d}}\left(\frac{R^2_{\gamma_2}}{R^{\star 2}_{\gamma_2}}\right)^{n/2}. \qquad (7)\end{aligned}$$

We consider each term of the RHS of the above expression consecutively. By Lemma 2

$$\begin{aligned}\sup_{\gamma_1\in\mathcal{G}_{1,d}}\left(\frac{R^{\star 2}_{\gamma_1}}{R^2_{\gamma_1}}\right)^{n/2} &= \sup_{\gamma_1\in\mathcal{G}_{1,d}}\left(1+\frac{R^{\star 2}_{\gamma_1}-R^2_{\gamma_1}}{R^2_{\gamma_1}}\right)^{n/2}\leq\sup_{\gamma_1\in\mathcal{G}_{1,d}}\left(1+\frac{R^{\star 2}_{\gamma_1}-R^2_{\gamma_1}}{R^2_{\gamma_c}}\right)^{n/2}\\ &\leq \left(1+\frac{c}{1+g_n\tau_{\min}}\right)^{n/2}, \qquad (8)\end{aligned}$$

with probability at least $1-\exp\{-n\}$ for some $c>0$. Consider the second term of (7)

$$\sup_{\gamma_1\in\mathcal{G}_{1,d}}\left(\frac{R^2_{\gamma_1}}{R^2_{\gamma_c}}\right)^{n/2}=\sup_{\gamma_1\in\mathcal{G}_{1,d}}\left(1-\frac{R^2_{\gamma_c}-R_{\gamma_1}}{R_{\gamma_c}}\right)^{n/2}\leq 1,$$

by the fact that $R^2_{\gamma_c}-R_{\gamma_1}=\mathbf{y}'_n\left(P_n(\gamma_1)-P_n(\gamma_c)\right)\mathbf{y}_n\geq 0$ as $\gamma_c\subseteq\gamma_1$ and consequently, $P_n(\gamma_1)-P_n(\gamma_c)$ is non-negative definite matrix.

Next, consider the third expression of (7). The ratio

$$\inf_{\gamma_2\in\mathcal{G}_{2,d}}\left(\frac{R^2_{\gamma_2}}{R^2_{\gamma_c}}\right)^{n/2} = \inf_{\gamma_2\in\mathcal{G}_{2,d}}\left(1+\frac{R^2_{\gamma_2}-R^2_{\gamma_c}}{R^2_{\gamma_c}}\right)^{n/2}. \qquad (9)$$

Now, by assumption (A4)

$$\begin{aligned} R^2_{\gamma_2} - R^2_{\gamma_c} &= \boldsymbol{\mu}_n' \left\{I - P_n(\gamma_2)\right\} \boldsymbol{\mu}_n + \mathbf{e}_n' \left\{P_n(\gamma_c) - P_n(\gamma_2)\right\} \mathbf{e}_n + 2\boldsymbol{\mu}_n' \left\{I - P_n(\gamma_2)\right\} \mathbf{e}_n \\ &\geq \boldsymbol{\mu}_n' \left\{I - P_n(\gamma_2)\right\} \boldsymbol{\mu}_n - 2\boldsymbol{\mu}_n' P_n(\gamma_2) \mathbf{e}_n \\ &\geq \Delta_0 - 2\left|\boldsymbol{\mu}_n' \mathbf{e}_n\right|, \end{aligned}$$

uniformly over $\mathcal{G}_{2,d}$ as $\mathcal{G}_{2,d} \subseteq \mathcal{G}_0$, with probability one. By the choice of Δ_0 in (A4) and Lemma 3, we have $|\boldsymbol{\mu}_n' \mathbf{e}_n| = o(\Delta_0)$ with probability at least $1 - \exp\{-cn^\xi\}$ for $\xi > 0$ as in (A4) and some $c > 0$. Thus, from (9), by the above derivations,

$$\inf_{\gamma_2 \in \mathcal{G}_{2,d}} \left(1 + \frac{R^2_{\gamma_2} - R^2_{\gamma_c}}{R^2_{\gamma_c}}\right)^{n/2} \geq \left(1 + \frac{\Delta_0 \{1 + o(1)\}}{n\sigma^2(1+\epsilon)}\right)^{n/2} \gtrsim \exp\left\{\Delta_0(1-\epsilon)/(2\sigma^2)\right\},$$

for any $\epsilon > 0$. Finally, it can be verified by examining the definitions of R^2_γ and $R^{2\star}_{\gamma_c}$ that the last part of RHS of (7) is bounded above by 1. Thus, combining all the above facts we get, for any $\epsilon > 0$, and with probability at least $1 - \exp\{-cn^\xi\}$ for some $c > 0$,

$$\sup_{\gamma_1 \in \mathcal{G}_{1,d}, \gamma_2 \in \mathcal{G}_{2,d}} \left(\frac{R^{\star 2}_{\gamma_1}}{R^{\star 2}_{\gamma_2}}\right)^{n/2} \leq \left(1 + \frac{c}{1 + g_n \tau_{\min}}\right)^{n/2} \exp\{-\Delta_0(1-\epsilon)/(2\sigma^2)\}$$

and

$$\begin{aligned} &\sup_{\gamma_1 \in \mathcal{G}_{1,d}, \gamma_2 \in \mathcal{G}_{2,d}} \frac{P(M_{\gamma_2} | \mathcal{G}_d, \mathbf{y}_n)}{P(M_{\gamma_1} | \mathcal{G}_d, \mathbf{y}_n)} \\ &\quad \leq 2\left(\frac{\tau_{\max}}{\tau_{\min} \wedge 1}\right)^{d_n/2} n^{d_n} \left(1 + \frac{c}{1 + g_n \tau_{\min}}\right)^{n/2} \exp\{-\Delta_0(1-\epsilon)/(2\sigma^2)\} \\ &\quad \leq cn^{d_n} \exp\{-\Delta_0(1-\epsilon)/(2\sigma^2)\} \to 0, \end{aligned}$$

for an appropriate constant $c > 0$. This completes the proof of part A.

[Part B.] Observe that, by choice of δ_0 in (A4)

$$\begin{aligned} \frac{\sum_{\gamma_2 \in \mathcal{G}_{2,d}} P(M_{\gamma_2} | \mathcal{G}_d, \mathbf{y}_n)}{\sum_{\gamma_1 \in \mathcal{G}_{1,d}} P(M_{\gamma_1} | \mathcal{G}_d, \mathbf{y}_n)} &\leq \sup_{\gamma_1 \in \mathcal{G}_{1,d}, \gamma_2 \in \mathcal{G}_{2,d}} \frac{P(M_{\gamma_2} | \mathbf{y}_n)}{P(M_{\gamma_1} | \mathbf{y}_n)} \frac{|\mathcal{G}_{2,d}|}{|\mathcal{G}_{1,d}|} \\ &\leq cn^{d_n} \exp\{-\Delta_0(1-\epsilon)/(2\sigma^2)\} \frac{\binom{p_n}{d}}{\binom{p_n - p(\gamma_c)}{d - p(\gamma_c)}} \\ &\leq cn^{d_n} p_n^{p(\gamma_c)} \exp\{-2(1-\epsilon)p(\gamma_c) \log p_n\} \\ &\leq cn^{d_n} p_n^{-(1-2\epsilon)p(\gamma_c)} \end{aligned}$$

with probability at least $1 - \exp\{-cn^\xi\}$ for some $c > 0$, and for any $\epsilon > 0$.

[Part C.] Observe that $P(\gamma \in \mathcal{G}_d \mid \mathcal{G}_d, \mathbf{y}_n) = 1$. Therefore,

$$\begin{aligned} 1 &= P(\gamma \in \mathcal{G}_{1,d} \mid \mathcal{G}_d, \mathbf{y}_n) + P(\gamma \in \mathcal{G}_{2,d} \mid \mathcal{G}_d, \mathbf{y}_n) \\ &= P(\gamma \in \mathcal{G}_{1,d} \mid \mathcal{G}_d, \mathbf{y}_n) \left\{1 + \frac{P(\gamma \in \mathcal{G}_{2,d} \mid \mathcal{G}_d, \mathbf{y}_n)}{P(\gamma \in \mathcal{G}_{1,d} \mid \mathcal{G}_d, \mathbf{y}_n)}\right\} \\ &= P(\gamma \in \mathcal{G}_{1,d} \mid \mathcal{G}_d, \mathbf{y}_n) \left\{1 + \frac{\sum_{\gamma_2 \in \mathcal{G}_{2,d}} P(M_{\gamma_2} | \mathcal{G}_d, \mathbf{y}_n)}{\sum_{\gamma_1 \in \mathcal{G}_{1,d}} P(M_{\gamma_1} | \mathcal{G}_d, \mathbf{y}_n)}\right\} \\ &\leq P(\gamma \in \mathcal{G}_{1,d} \mid \mathcal{G}_d, \mathbf{y}_n) \left\{1 + cn^{d_n} p_n^{-(1-2\epsilon)p(\gamma_c)}\right\}, \end{aligned}$$

with probability at least $1-\exp\{-cn^{\xi}\}$ from part B. Observe that from (A1) and the choice of $d_n \sim \log n$, the sequence $n^{d_n} p_n^{-(1-2\epsilon)p(\gamma_c)} \to 0$, as $p_n \to \infty$. Further, as $n \to \infty$, the probability $1-\exp\{-cn^{\xi}\}$ converges to one. This completes the proof. □

A.3. Proof of Theorem 2

Proof: Recall that, $P\left(M_{\gamma_c} \mid \mathbf{y}_n, \mathcal{G}^{\star}\right)$ is the posterior probability of the model M_{γ_c}, restricted to the class $\mathcal{G}^{\star}$. We will first provide an uniform probabilistic upper bound to $P\left(M_{\gamma_c} \mid \mathbf{y}_n, \mathcal{G}^{\star}\right)$ for any fixed $\gamma^{\star}$ such that $\gamma_c \in \gamma^{\star}$. Observe that

$$P\left(M_{\gamma_c} \mid \mathcal{G}^{\star}, \mathbf{y}_n\right) = \left\{1 + \sum_{\gamma \subseteq \gamma^{\star}, \gamma \neq \gamma_c} \frac{P(M_{\gamma} \mid \mathbf{y}_n)}{P(M_{\gamma_c} \mid \mathbf{y}_n)}\right\}^{-1}.$$

The ratio of posterior probabilities of any model to the true model is given by

$$\frac{P(M_{\gamma}|\mathbf{y}_n)}{P(M_{\gamma_c}|\mathbf{y}_n)} = \left(\frac{1}{p_n-1}\right)^{p_n(\gamma)-p(\gamma_c)} \left(\frac{R^{2\star}_{\gamma_c}}{R^{2\star}_{\gamma}}\right)^{n/2} \frac{\left|I + g_n X'_{\gamma_c} X_{\gamma_c}\right|^{1/2}}{\left|I + g_n X'_{\gamma} X_{\gamma}\right|^{1/2}}. \tag{10}$$

We split $\mathcal{G}$ into two subclasses as follows:

(*i*) *Supermodel of the true model,* $\mathcal{G}_1^{\star} = \{\gamma : M_{\gamma_c} \subset M_{\gamma}\} \cap \mathcal{G}^{\star}$.

(*ii*) *Non-supermodels,* $\mathcal{G}_2^{\star} = \{\gamma : M_{\gamma_c} \nsubseteq M_{\gamma}\} \cap \mathcal{G}^{\star}$.

Case I: Super-models ($\gamma \in \mathcal{G}_1^{\star}$) First, we obtain a uniform upper bound for the ratio of the posterior probabilities of any model M_{γ} and M_{γ_c}, given in (10). Note that

$$\frac{R^{2\star}_{\gamma}}{R^{2\star}_{\gamma_c}} = \frac{R^{2\star}_{\gamma}}{R^{2}_{\gamma}} \frac{R^{2}_{\gamma}}{R^{2}_{\gamma_c}} \frac{R^{2}_{\gamma_c}}{R^{2\star}_{\gamma_c}} \geq \left(1 - \frac{\epsilon}{n(1+\epsilon)}\right) \frac{R^{2}_{\gamma}}{R^{2}_{\gamma_c}} \tag{11}$$

by Lemma 3 and $R^{2\star}_{\gamma} \geq R^{2}_{\gamma}$, and with probability at least $1-\exp\{-cn\}$ for some $c > 0$.

Next, consider that for any $\epsilon > 0$ and $R = 2(1+\epsilon)$, we have

$$\begin{aligned} &P\left[\sup_{\gamma^{\star} \in \mathcal{G}_{1,d}} \sup_{\gamma_c \subseteq \gamma \subseteq \gamma^{\star}} \left(R^2_{\gamma_c} - R^2_{\gamma}\right) < R\sigma^2\{p_n(\gamma) - p(\gamma_c)\} \log p_n\right] \\ &= P\left[\sup_{\{\gamma : \gamma_c \subseteq \gamma, |\gamma| \leq d_n\}} \left(R^2_{\gamma_c} - R^2_{\gamma}\right) < R\sigma^2\{p_n(\gamma) - p(\gamma_c)\} \log p_n\right]. \end{aligned} \tag{12}$$

The last equality holds due to the equality of the sets

$$\{\gamma : \gamma_c \subseteq \gamma, |\gamma| \leq d_n\} = \{\gamma : \gamma_c \subseteq \gamma \subseteq \gamma^{\star}, \gamma^{\star} \in \mathcal{G}_{1,d}\}.$$

Next, observe that the right-hand side (RHS) of (12) is bounded above by

$$\begin{aligned}
&\sum_{\{\gamma:\gamma_c\subseteq\gamma,|\gamma|\le d_n\}} P\left[\left(R^2_{\gamma_c}-R^2_{\gamma}\right) < R\sigma^2\{p_n(\gamma)-p(\gamma_c)\}\log p_n\right] \\
&\le \sum_{p_n(\gamma)-p(\gamma_c)=1}^{d_n-p(\gamma_c)} \binom{p_n-p(\gamma_c)}{p_n(\gamma)-p(\gamma_c)} \exp\left\{-R\{p_n(\gamma)-p(\gamma_c)\}\log p_n/2\right\} \\
&\le \sum_{p_n(\gamma)-p(\gamma_c)=1}^{d_n-p(\gamma_c)} (p_n-p(\gamma_c))^{p_n(\gamma)-p(\gamma_c)} p_n^{-R\{p_n(\gamma)-p(\gamma_c)\}/2} \\
&\le (d_n-p(\gamma_c))p_n^{-\epsilon}\to 0, \qquad (13)
\end{aligned}$$

where $\epsilon>0$ be any constant. Therefore, with probability at least $1-cp_n^{-\epsilon}$ for any $\epsilon>0$ and an appropriate $c>0$, the following holds uniformly over $\{\gamma:\gamma_c\subseteq\gamma\subseteq\gamma^\star,\gamma^\star\in\mathcal{G}_{1,d}\}$

$$\left(\frac{R^{2\star}_{\gamma_c}}{R^{2\star}_{\gamma}}\right)^{n/2} \le (1+\epsilon)\left(1-\frac{R(p_n(\gamma)-p(\gamma_c))\log p_n}{n(1-\epsilon)}\right)^{-n/2} \lesssim (1+\epsilon)p_n^{(1+\epsilon)(p_n(\gamma)-p(\gamma_c))}.$$

Again, by Lemma 1 and assumptions (A2)-(A3) we have

$$\frac{\left|I+g_nX'_\gamma X_\gamma\right|^{-1/2}}{\left|I+g_nX'_{\gamma_c}X_{\gamma_c}\right|^{-1/2}} \le cg_n^{-(p_n(\gamma)-p(\gamma_c))/2}n^{p(\gamma_c)},$$

where $c>0$ is some appropriate constant. Therefore, summing the ratio of posterior probabilities over $M_\gamma\in\mathcal{G}_1^\star$, we have

$$\begin{aligned}
\sum_{\gamma\in\mathcal{G}_1^\star}\frac{p(M_\gamma|\mathbf{y}_n)}{p(M_{\gamma_c}|\mathbf{y}_n)} &\le n^{p(\gamma_c)}\sum_{\gamma\in\mathcal{G}_1^\star}\frac{cp_n^{(1+\epsilon)(p_n(\gamma)-p(\gamma_c))}}{\{\sqrt{g_n}(p_n-1)\}^{p_n(\gamma)-p(\gamma_c)}} \\
&\le \sum_{p_n(\gamma)-p(\gamma_c)=1}^{d_n-p(\gamma_c)}\binom{d_n-p(\gamma_c)}{p_n(\gamma)-p(\gamma_c)}n^{p(\gamma_c)}c\left(\frac{p_n^{2\epsilon}}{g_n}\right)^{(p_n(\gamma)-p(\gamma_c))/2} \\
&\le c2^{d_n-p(\gamma_c)}n^{p(\gamma_c)}\left(\frac{p_n^{2\epsilon}}{g_n}\right)^{1/2} \\
&\le c2^{d_n-p(\gamma_c)}n^{p(\gamma_c)}\left(\frac{p_n^{2\epsilon}}{g_n}\right)^{1/2} \le cn^{p(\gamma_c)+1}\left(\frac{p_n^{2\epsilon}}{g_n}\right)^{1/2}
\end{aligned}$$

for any $\epsilon>0$ and a suitable choice of $c>0$. When we choose $\epsilon<\delta/3$, we get that the above expression converges to 0, as $p_n\to\infty$.

Case II: Non-super models ($\gamma\in\mathcal{G}_2^\star$) We split $R^{2\star}_\gamma/R^{2\star}_{\gamma_c}$ as before in (11). Observe that

$$\begin{aligned}
R^2_\gamma-R^2_{\gamma_c} &= \mathbf{y}'_n(P_n(\gamma_c)-P_n(\gamma))\mathbf{y}_n \\
&= \boldsymbol{\mu}'_n(P_n(\gamma_c)-P_n(\gamma))\boldsymbol{\mu}_n+2\boldsymbol{\mu}'_n(P_n(\gamma_c)-P_n(\gamma))\mathbf{e}_n+\mathbf{e}'_n(P_n(\gamma_c)-P_n(\gamma))\mathbf{e}_n \\
&\ge \boldsymbol{\mu}'_n(P_n(\gamma_c)-P_n(\gamma))\boldsymbol{\mu}_n-2|\boldsymbol{\mu}'_n\mathbf{e}_n|.
\end{aligned}$$

Note that $\boldsymbol{\mu}_n'(P_n(\gamma_c) - P_n(\gamma))\boldsymbol{\mu}_n = \boldsymbol{\mu}_n'(I - P_n(\gamma))\boldsymbol{\mu}_n > \Delta_0$ uniformly over the class of all small dimensional non-supermodels by assumption (A4). Further, by Lemma 3, we get $|\boldsymbol{\mu}_n'\mathbf{e}_n| = o(\Delta_0)$ with probability at least $1-\exp\{-cn^\xi\}$ for $\xi > 0$ as in (A4) and some $c > 0$. Combining all these facts and using (A4), we have with probability at least $1 - \exp\{-cn^\xi\}$

$$\sup_{\gamma^\star \in \mathcal{G}_{1,d}} \sup_{\gamma \in \mathcal{G}_2^\star} \left(\frac{R_\gamma^{2\star}}{R_{\gamma_c}^{2\star}}\right)^{-n/2} \leq (1+\epsilon)\left(1 + (1-\epsilon)\frac{\Delta_0}{n\sigma^2}\right)^{-n/2} \lesssim (1+\epsilon)\exp\left\{-(1-\epsilon)\Delta_0/2\sigma^2\right\}.$$

Further, from Lemma 1, the ratio of determinants in the last term of (10) is less than $c\left(n\sqrt{g_n\tau_{\max}}\right)^{p(\gamma_c)}$ for an appropriately chosen $c > 0$. Therefore,

$$\begin{aligned}\sup_{\gamma^\star \in \mathcal{G}_{1,d}} \sum_{\gamma \in \mathcal{G}_2^\star} \frac{p(M_\gamma|\mathbf{y}_n)}{p(M_{\gamma_c}|\mathbf{y}_n)} &\leq c\left(np_n\sqrt{g_n\tau_{\max}}\right)^{p(\gamma_c)} \exp\left\{-(1-\epsilon)\frac{\Delta_0}{2\sigma^2}\right\} \sum_{q=1}^{d_n} \binom{d_n}{q} \frac{1}{(p_n-1)^q} \\ &\leq c\left(\frac{n\sqrt{g_n}}{p_n^{1-2\epsilon}}\right)^{p(\gamma_c)}, \end{aligned} \tag{14}$$

for any $\epsilon > 0$, with probability at least $1-\exp\{-cn^\xi\}$ for $\xi > 0$ as in (A4), and uniformly over $\gamma^\star \in \mathcal{G}_{1,d}$. Combining the above facts, we get, with probability at least $1-cp_n^{-c_0}-\exp\{-cn^\xi\}$, where $c_0 \ll \delta/2$ and $\xi > 0$ is as in (A4),

$$\inf_{\gamma^\star \in \mathcal{G}_{1,d}} P\left(M_{\gamma_c} \mid \mathbf{y}_n, \mathcal{G}^\star\right) \geq \left[1 + cn^{p(\gamma_c)+1}\left(\frac{p_n^{2\epsilon}}{g_n}\right)^{1/2} + c\left(\frac{n\sqrt{g_n}}{p_n^{1-2\epsilon}}\right)^{p(\gamma_c)}\right]^{-1}$$

for some $\epsilon \ll \delta/2$, where δ is as in the choice of g_n. For the choice of g_n taken in Theorem 2, the above expression converges to 1 as $p_n \to \infty$. □

A.4. Proof of Theorem 3

Proof: Observe that

$$\begin{aligned} P_2(M_{\gamma_c}|\mathbf{y}_n) &= \sum_{\gamma^\star \in \mathcal{G}_d} P_2(M_{\gamma_c}, M_{\gamma^\star}|\mathbf{y}_n) = \sum_{\gamma^\star \in \mathcal{G}_d} P(M_{\gamma_c}|\mathbf{y}_n, \mathcal{G}^\star) P_1(M_{\gamma^\star}|\mathbf{y}_n) \\ &= \sum_{\gamma^\star \in \mathcal{G}_{1,d}} P(M_{\gamma_c}|\mathbf{y}_n, \mathcal{G}^\star) P_1(M_{\gamma^\star}|\mathbf{y}_n) + \sum_{\gamma^\star \in \mathcal{G}_{2,d}} P(M_{\gamma_c}|\mathbf{y}_n, \mathcal{G}^\star) P_1(M_{\gamma^\star}|\mathbf{y}_n), \end{aligned}$$

where $M_{\gamma^\star}$ is the model chosen in the first stage. Observe that $P(M_{\gamma_c}|\mathbf{y}_n, \mathcal{G}^\star) = 0$ if $\gamma^\star \in \mathcal{G}_{2,d}$, *i.e.*, if the model chosen in the first stage is a non-supermodel. Therefore,

$$\begin{aligned} P_2(M_{\gamma_c}|\mathbf{y}_n) &= \sum_{\gamma^\star \in \mathcal{G}_{1,d}} P(M_{\gamma_c}|\mathbf{y}_n, \mathcal{G}^\star) P_1(M_{\gamma^\star}|\mathbf{y}_n) \\ &\geq \inf_{\gamma^\star \in A_{1,d}} P\left(M_{\gamma_c} \mid \mathbf{y}_n, \mathcal{G}^\star\right) \left[1 + \frac{\sum_{\gamma^\star \in \mathcal{G}_{2,d}} P_1(M_{\gamma^\star}|\mathbf{y}_n)}{\sum_{\gamma^\star \in \mathcal{G}_{1,d}} P_1(M_{\gamma^\star}|\mathbf{y}_n)}\right]^{-1} \\ &\geq \left[1 + cn^{p(\gamma_c)+1}\left(\frac{p_n^{2\epsilon}}{g_n}\right)^{1/2} + c\left(\frac{n\sqrt{g_n}}{p_n^{1-2\epsilon}}\right)^{p(\gamma_c)}\right]^{-1} \left[1 + c_3 n^{d_n/2} p_n^{-(1-2\epsilon)p(\gamma_c)}\right]^{-1}, \end{aligned}$$

with a probability at least $1 - cp_n^{-c_0} - 2\exp\{-cn^\xi\}$, where $c_0 \ll \delta/2$ and $\xi > 0$ is as in (A4). Thus, $P_2(M_{\gamma_c}|\mathbf{y}_n) \to 1$ with probability tending to 1. □

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 509–533
http://www.ssca.org.in/journal

On High-Dimensional Modifications of the Nearest Neighbor Classifier

Annesha Ghosh[1], **Deep Ghoshal**[2], **Bilol Banerjee**[1] **and Anil K. Ghosh**[1]
[1] *Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata, India*
[2] *Department of Statistics, University of Illinois at Urbana-Champaign, USA.*

Received: 06 May 2024; Revised: 26 September 2024; Accepted: 30 September 2024

Abstract

Nearest neighbor classifier is arguably the most simple and popular nonparametric classifier available in the literature. However, due to the concentration of pairwise distances and the violation of the neighborhood structure, this classifier often suffers in high-dimension, low-sample size (HDLSS) situations, especially when the scale difference between the competing classes dominates their location difference. Several attempts have been made in the literature to take care of this problem. In this article, we discuss some of those existing methods and propose some new ones. We carry out some theoretical investigations in this regard and analyze several simulated and benchmark datasets to compare the empirical performances of our proposed methods with some of the existing ones.

Key words: Dimension reduction; Feature extraction; HDLSS asymptotics; Mixture distributions; Nearest neighbors.

AMS Subject Classifications: 62H30, 68T10

1. Introduction

In supervised classification, we use a training set of labeled observations from different competing classes to form a decision rule for classifying unlabeled test set observations as accurately as possible. Starting from Fisher (1936), Rao (1948) and Fix and Hodges (1951), several parametric as well as nonparametric classifiers have been developed for this purpose (see, *e.g.*, Duda *et al.*, 2007; Hastie *et al.*, 2009). Among them, the nearest neighbor classifier (see, *e.g.*, Cover and Hart, 1967) is perhaps the most popular one. The k-nearest neighbor classifier (k-NN) classifies an observation $\mathbf{x}$ to the class having the maximum number of representatives among the k nearest neighbors of $\mathbf{x}$. This classifier works well if the training sample size is large compared to the dimension of the data. For a suitable choice of k (which increases with the training sample size at an appropriate rate), under some mild regularity conditions, the misclassification rate of the k-NN classifier converges to the Bayes risk (*i.e.*, the misclassification rate of the Bayes classifier) as the training sample size grows to infinity (see, *e.g.* Devroye *et al.*, 2013; Hall *et al.*, 2008). However, like other nonparametric

Corresponding Author: Annesha Ghosh
Email: anneshaghosh659@gmail.com

methods, this classifier also suffers from the curse of dimensionality (see, *e.g.*, Carreira-Perpinan, 2009), especially when the dimension of the data is much larger than the training sample size. In such high-dimension, low-sample-size (HDLSS) situations, the concentration of pairwise distances (see, *e.g.*, Hall *et al.*, 2005; François *et al.*, 2007), presence of hubs and the violation of the neighborhood structure (see, *e.g.*, Radovanovic *et al.*, 2010; Pal *et al.*, 2016) often have adverse effects on the performance of the nearest neighbor classifier.

To demonstrate this, we consider some simple examples involving two d-dimensional normal distributions. Descriptions of these examples are given below.

Examples 1 - 3: *In these three examples, the first class has a normal distribution with the mean vector* $\mathbf{0}_d = (0, 0, \ldots, 0)^\top$ *and the dispersion matrix* $\mathbf{I}_d$ *(the* $d \times d$ *identity matrix), while the second class has the mean vector* $\mu\mathbf{1}_d = \mu(1, 1, \ldots, 1)^\top = (\mu, \mu, \ldots, \mu)^\top$ *and the dispersion matrix* $\sigma^2\mathbf{I}_d$*. In Example 1, we consider a location problem where we take* $\mu = 1$ *and* $\sigma = 1$*. Example 2 deals with a location-scale problem with* $\mu = 1$ *and* $\sigma = 2$*. As Example 3, we choose a scale problem, where* μ *and* σ *are taken as* 0 *and* 2*, respectively.*

In each of these examples, we carry out our experiment for 7 different choices of d ranging between 10 and 1000 (d= 10, 20, 50, 100, 200, 500 and 1000). In each case, taking an equal number of observations from the two competing classes, we form the training and test sets of size 50 and 500, respectively. This is done 100 times, and the average test set misclassification rates of the 1-NN classifier over these 100 trials are reported in Figure 1.

Note that in each of these examples, the distribution of each measurement variable differs in two competing classes. So, each of them contains information about class separability, and as a result, the separability between the two classes increases with the dimension. One can check that in each of these examples, the Bayes risk converges to 0 as the dimension grows. Therefore, the misclassification rate of any good classifier is also expected to go down as the dimension increases. We observed the same for the 1-NN classifier in Example 1 (location problem), but surprisingly, in the other two cases, its misclassification rates were close to 0.5 in high dimensions.

A careful investigation explains the reasons for this diametrically opposite behavior. Let $\{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_{n1}\}$ and $\{\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_{n_2}\}$ be the training samples from two competing classes (here we have $n_1 = n_2 = 25$) $N(\mathbf{0}_d, \mathbf{I}_d)$ and $N(\mu\mathbf{1}_d, \sigma^2\mathbf{I}_d)$, respectively. Now, for a test case $\mathbf{Z}$ from $N(\mathbf{0}_d, \mathbf{I}_d)$, one can show that for each $i = 1, 2, \ldots, n_1$, $\frac{1}{d}\|\mathbf{Z} - \mathbf{X}_i\|^2$, being the average of independent and identically distributed (i.i.d.) random variables, converges in probability to 2 as d increases to infinity. Similarly, it can be shown that for each $i = 1, 2, \ldots, n_2$, $\frac{1}{d}\|\mathbf{Z} - \mathbf{Y}_i\|^2 \xrightarrow{P} 1 + \mu^2 + \sigma^2$. So, $\mathbf{Z}$ is correctly classified by the 1-NN classifier (or any k-NN classifier with $k \leq \min\{n_1, n_2\}$) if $\mu^2 + \sigma^2 > 1$, Note that it was the case in all three examples. So, all observations from $N(\mathbf{0}_d, \mathbf{I}_d)$ were correctly classified. But for a test case $\mathbf{Z}'$ from $N(\mu\mathbf{1}_d, \sigma^2\mathbf{I}_d)$, we have $\frac{1}{d}\|\mathbf{Z}' - \mathbf{X}_i\|^2 \xrightarrow{P} 1 + \mu^2 + \sigma^2$ for $i = 1, 2, \ldots, n_1$ and $\frac{1}{d}\|\mathbf{Z}' - \mathbf{Y}_i\|^2 \xrightarrow{P} 2\sigma^2$ for $i = 1, 2, \ldots, n_2$. So, it is correctly classified if and only if $\sigma^2 < 1 + \mu^2$. This condition was satisfied in Example 1, but not in the other two cases. Because of this violation of the neighborhood structure (where observations from one class have all neighbors from other classes), in Examples 2 and 3, the 1-NN classifier misclassified all observations from $N(\mu\mathbf{1}_d, \sigma^2\mathbf{I}_d)$ and had misclassification rates close to 0.5.

This phenomenon of distance concentration in high dimension was observed by Hall *et al.* (2005) for Euclidean distances and François *et al.* (2007) for fractional distances. Hall *et al.* (2005) also studied the high dimensional behavior of some popular classifiers and observed this undesirable behavior of the nearest neighbor classifier. To take care of this problem, Chan and Hall (2009b) proposed an adjustment for the scale difference between the competing classes. They suggested to compute

$$\rho_1(\mathbf{Z}, \mathbf{X}_i) = \|\mathbf{Z} - \mathbf{X}_i\|^2 - \frac{1}{2}\binom{n_1}{2}^{-1} \sum_{s<t} \|\mathbf{X}_s - \mathbf{X}_t\|^2 \text{ for } i = 1, 2, \ldots, n_1,$$

$$\rho_2(\mathbf{Z}, \mathbf{Y}_i) = \|\mathbf{Z} - \mathbf{Y}_i\|^2 - \frac{1}{2}\binom{n_2}{2}^{-1} \sum_{s<t} \|\mathbf{Y}_s - \mathbf{Y}_t\|^2 \text{ for } i = 1, 2, \ldots, n_2$$

and classify $\mathbf{Z}$ to the first (respectively, second) class if $\min \rho_1(\mathbf{Z}, \mathbf{X}_i) < \min \rho_2(\mathbf{Z}, Y_i)$ (respectively, $\min \rho_1(\mathbf{Z}, \mathbf{X}_i) > \min \rho_2(\mathbf{Z}, Y_i)$). Note that without the scale adjustments (second terms on the right-hand side of the equations), it turns out to be the usual 1-NN classifier. Figure 1 also shows the performance of this classifier (we refer to it as the CH classifier) in Examples 1-3. In Example 1, it performed like the 1-NN classifier. Interestingly, in Example 2, while the 1-NN classifier failed, this scale adjustment led to improved performance by the CH classifier in high dimensions. But in Example 3, like the 1-NN classifier, it also misclassified almost 50% observations. Note that for any $\mathbf{Z}$ from $N(\mathbf{0}_d, \mathbf{I}_d)$, here $\rho_1(\mathbf{Z}, \mathbf{X}_i)/d \xrightarrow{P} 1$ $(i = 1, 2, \ldots, n_1)$ and $\rho_2(\mathbf{Z}, \mathbf{Y}_i)/d \xrightarrow{P} 1 + \mu^2$ $(i = 1, 2, \ldots, n_2)$ as d increases. So, it is correctly classified if $\mu^2 > 0$. Again, for any $\mathbf{Z}'$ from $N(\mu\mathbf{1}_d, \sigma^2\mathbf{I}_d)$, we have $\rho_1(\mathbf{Z}', \mathbf{X}_i)/d \xrightarrow{P} \mu^2 + \sigma^2$ for $i = 1, 2, \ldots, n_1$ and $\rho_2(\mathbf{Z}', \mathbf{Y}_i)/d \xrightarrow{P} \sigma^2$ for $i = 1, 2, \ldots, n_2$. So, here also, we need $\mu^2 > 0$ for correct classification. In Examples 1 and 2, we had $\mu^2 > 0$. So, the CH classifier performed well in those two examples for large values of d. But in Example 3, where we had $\mu^2 = 0$, it misclassified almost 50% observations. This example shows that the CH classifier may fail to discriminate between two high-dimensional distributions differing only in their scales.

However, if we slightly modify Chan and Hall (2009b)'s proposal of scale adjustment, we can take care of high dimensional scale problems as well. Our modified version (which

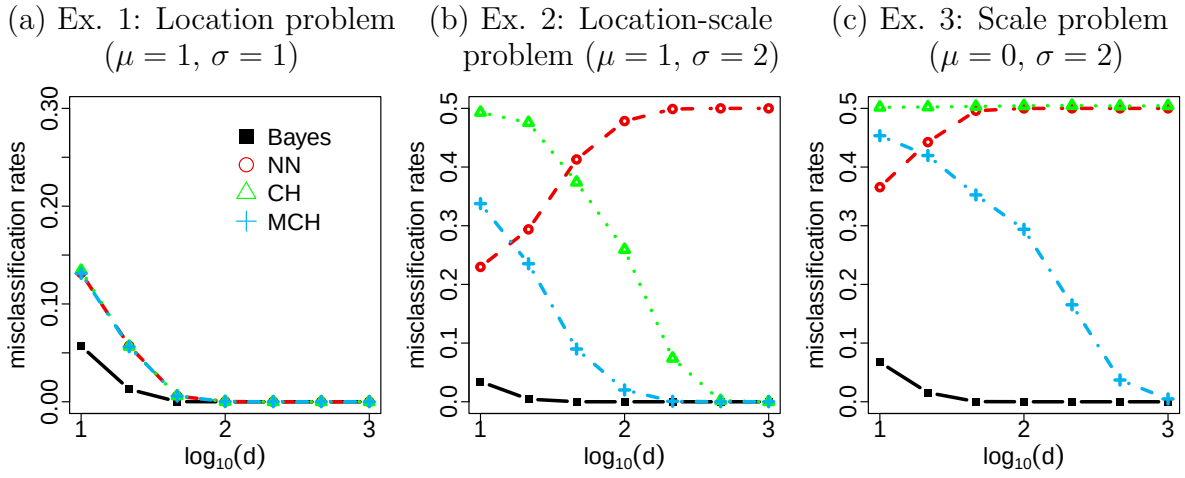


Figure 1: Misclassification rates of Bayes, NN, CH and MCH classifiers in Examples 1-3.

we refer to as the Modified Chan and Hall classifier or the MCH classifier) had excellent performance in all three examples (see Figure 1), especially for large values of d. In Section 2, we propose this modification and carry out some theoretical and numerical studies to understand the high-dimensional behavior of the resulting classifier.

An alternative strategy to deal with any high-dimensional problem is to reduce the dimension of the data and work on the reduced subspace. The simplest method of dimension reduction is to consider some random linear projections (see, *e.g.*, Fern and Brodley, 2003; Fradkin and Madigan, 2003; Vempala, 2005, and the references therein), and we can adopt that method for nearest neighbor classification as well. Another popular approach is to use projections based on principal component analysis (see, *e.g.*, Deegalla and Bostrom, 2006; Maciończyk *et al.*, 2023). But as pointed out in Dutta and Ghosh (2016), these methods often lead to poor performance in high-dimensional classification problems. For instance, from our description, it is quite clear that neither the principle component directions nor the random projections are meaningful in Example 3. Also getting consistent estimates of the principal components in high dimension is challenging (see, *e.g.*, Jung and Marron, 2009). Other approaches towards nearest neighbor classification of high-dimensional data include those based on mean absolute difference of distances (see,*e.g.*, Pal *et al.*, 2016; Roy *et al.*, 2022), hubness-based fuzzy measures (see, *e.g.*, Tomašev *et al.*, 2014) and distance metrics learning (see, *e.g.* Weinberger and Saul, 2009). Chan and Hall (2009a) proposed a robust version of the nearest neighbor method for classifying high-dimensional data, but their method can be used only for a specific type of two-class location problem. Instead of using random projections or principal components, Dutta and Ghosh (2016) suggested extracting some distance-based features from the data and performing nearest neighbor classification based on those features. They proposed two such methods, one using transformation based on average distances (TRAD) and the other using transformation based on inter-point distances (TRIPD). We briefly discuss these two methods in Section 3 and also propose some other methods for selecting distance-based features for nearest neighbor calssification of high dimensional data. A comparative discussion of these methods is also given in this section based on our analysis of some simulated data sets. Some benchmark data sets are analyzed in Section 4 to compare the performances of these methods with two popular state-of-the-art classifiers, support vector machines (see, *e.g.* Cristianini and Shawe-Taylor, 2003; Steinwart and Christmann, 2008; Scholkopf and Smola, 2018) and random forest (see, *e.g.*, Breiman, 2001; Genuer and Poggi, 2020), which are known to perform well for high dimensional data. Finally, a brief summary of the work and some concluding remarks are given in Section 5. All proofs and mathematical details are given in the Appendix.

2. Modified scale-adjusted nearest neighbor classifier

We have seen that while the 1-NN classifier failed in Examples 2 and 3, the scale-adjusted CH classifier worked well in Example 2 when the dimension was large. But, in Example 3, this scale adjustment could not improve the performance of the 1-NN classifier.

This motivates us to look for a modified scale adjustment. We define

$$\rho_1^*(\mathbf{Z},\mathbf{X}_i)=\|\mathbf{Z}-\mathbf{X}_i\|-\frac{1}{2}\binom{n_1}{2}^{-1}\sum_{s<t}\|\mathbf{X}_s-\mathbf{X}_t\| \text{ for } i=1,2,\ldots,n_1,$$

$$\rho_2^*(\mathbf{Z},\mathbf{Y}_i)=\|\mathbf{Z}-\mathbf{Y}_i\|-\frac{1}{2}\binom{n_2}{2}^{-1}\sum_{s<t}\|\mathbf{Y}_s-\mathbf{Y}_t\| \text{ for } i=1,2,\ldots,n_2$$

and classify a test set observation $\mathbf{Z}$ to the first (respectively, second) class if $\min \rho_1^*(\mathbf{Z},\mathbf{X}_i)$ is smaller (respectively, larger) than $\min\ \rho_2^*(\mathbf{Z},\mathbf{Y}_i)$. Figure 1 shows that this modified scale adjusted nearest neighbor classifier (henceforth referred to as the Modified Chan and Hall classifier or the MCH classifier) had excellent performance in high dimensions in all three examples. A small theoretical analysis explains the reasons for its superior performance.

Following our previous discussion on distance convergence, one can show that for a test set observation $\mathbf{Z}$ from $N(\mathbf{0}_d,\mathbf{I}_d)$, as d tends to infinity, we have $\rho_1^*(\mathbf{Z},\mathbf{X}_i)/\sqrt{d}\xrightarrow{P}1/\sqrt{2}$ for $i=1,2,\ldots,n_1$, while $\rho_2^*(\mathbf{Z},\mathbf{Y}_i)/\sqrt{d}\xrightarrow{P}\sqrt{1+\mu^2+\sigma^2}-\sigma/\sqrt{2}$ for $i=1,2,\ldots,n_2$. So, it is correctly classified if $\sqrt{1+\mu^2+\sigma^2}>(\sigma+1)/\sqrt{2}\Leftrightarrow 1+\mu^2+\sigma^2>(\sigma+1)^2/2\Leftrightarrow \mu^2+\frac{1}{2}(\sigma-1)^2>0$. Again for an observation $\mathbf{Z}'$ from $N(\mu\mathbf{1}_d,\sigma^2\mathbf{I}_d)$, as $d\to\infty$, we have $\rho_1^*(\mathbf{Z}',\mathbf{X}_i)/\sqrt{d}\xrightarrow{P}\sqrt{1+\mu^2+\sigma^2}-1/\sqrt{2}$ for $i=1,2,\ldots,n_1$ and $\rho_2^*(\mathbf{Z}',\mathbf{Y}_i)/\sqrt{d}\xrightarrow{P}\sigma/\sqrt{2}$ for $i=1,2,\ldots,n_2$. So, here also, $\mathbf{Z}'$ is correctly classified if $\mu^2+\frac{1}{2}(\sigma-1)^2>0$. This inequality holds in all three examples considered in Section 1. This was the reason for the excellent performance of the MCH classifier in high dimensions.

Like the usual nearest neighbor classifier, multi-class generalizations of CH and MCH classifiers are quite straightforward. If there are J competing classes $F_1,F_2,\ldots,F_J$ with the training samples $\{\mathbf{X}_{j1},\mathbf{X}_{j2},\ldots,\mathbf{X}_{jn_j}\}$ from the j-th class, $(j=1,2,\ldots,J)$, for classifying a test case $\mathbf{Z}$ by the CH classifier, we can compute

$$\rho_j(\mathbf{Z},\mathbf{X}_{ji})=\|\mathbf{Z}-\mathbf{X}_{ji}\|^2-\frac{1}{2}\binom{n_j}{2}^{-1}\sum_{s<t}\|\mathbf{X}_{js}-\mathbf{X}_{jt}\|^2 \text{ for } j=1,2,\ldots,J,\ i=1,2,\ldots,n_j$$

and assign $\mathbf{Z}$ to class j_0 if $\min\limits_{1\le i\le n_{j_0}}\rho_{j_0}(\mathbf{Z},\mathbf{X}_{j_0i})<\min\limits_{1\le i\le n_j}\rho_j(\mathbf{Z},\mathbf{X}_{ji})$ for all $j\neq j_0$. Similarly, for the MCH classifier, one can compute

$$\rho_j^*(\mathbf{Z},\mathbf{X}_{ji})=\|\mathbf{Z}-\mathbf{X}_{ji}\|-\frac{1}{2}\binom{n_j}{2}^{-1}\sum_{s<t}\|\mathbf{X}_{js}-\mathbf{X}_{jt}\| \text{ for } j=1,2,\ldots,J,\ i=1,2,\ldots,n_j$$

and assign $\mathbf{Z}$ to class j_0 if $\min\limits_{1\le i\le n_{j_0}}\rho_{j_0}^*(\mathbf{Z},\mathbf{X}_{j_0i})<\min\limits_{1\le i\le n_j}\rho_j^*(\mathbf{Z},\mathbf{X}_{ji})$ for all $j\neq j_0$.

For the sake of simplicity, in Examples 1-3, we considered binary classification problems involving two normal distributions each having i.i.d. measurement variables. Now, one may be curious to know how CH and MCH classifiers perform in high-dimensional multi-class classification problems involving more general class distributions with possibly dependent and non-identically distributed measurement variables. For this investigation, we consider the following assumptions.

(A1) In each of the J competing classes, the measurement variables have uniformly bounded fourth moments.

(A2) If $\mathbf{X} = (X_1, \ldots, X_d)^\top \sim F_j$ and $\mathbf{Y} = (Y_1, \ldots, Y_d)^\top \sim F_i$ $(1 \leq j, i \leq J)$ are independent, for $\mathbf{U} = \mathbf{X} - \mathbf{Y}$, $\sum_{r \neq s} |Corr(U_r^2, U_s^2)|$ is of the order $o(d^2)$.

(A3) Let $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ be the mean vector and the dispersion matrix of F_j $(1 \leq j \leq J)$. For each $j = 1, \ldots, J$, there exists a constant σ_j^2 such that $\text{trace}(\boldsymbol{\Sigma}_J)/d \to \sigma_j^2$ as $d \to \infty$. Also, for each $i \neq j$, there exists a constant ν_{ji}^2 such that $\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\|^2/d \to \nu_{ji}^2$ as $d \to \infty$.

Under (A1) and (A2), we have the weak law of large numbers (WLLN) (see, *e.g.*, Feller, 1991) for the sequence of possibly dependent and non-identically distributed random variables $\{U_q^2 : q \geq 1\}$,*i.e.*, $\left|\frac{1}{d}\|\mathbf{U}\|^2 - E\left(\frac{1}{d}\|\mathbf{U}\|^2\right)\right| \xrightarrow{P} 0$ as $d \to \infty$ (note that if the measurement variables are i.i.d., as they were in Examples 1-3, the WLLN holds under the second moment assumption, (A1) and (A2) are not needed there). Assumption (A3) gives the limiting value of $E\left(\frac{1}{d}\|\mathbf{U}\|^2\right)$ and hence that of $\frac{1}{d}\|\mathbf{U}\|^2 = \frac{1}{d}\|\mathbf{X} - \mathbf{Y}\|^2$ for $\mathbf{X} \sim F_j$ and $\mathbf{Y} \sim F_i$ $(1 \leq j, i \leq J)$. So, under (A1)-(A3), we have high-dimensional convergence of all pairwise distances and their limiting values (see Lemma 1 in Appendix). These assumptions are quite standard in the HDLSS literature. Hall *et al.* (2005) considered the d-dimensional observations as time series truncated at time d, and in addition to (A1) and (A3), they assumed the ρ-mixing property of the time series to study the high dimensional behavior of some popular classifier as d increases. Note that (A2) holds under that ρ-mixing condition. François *et al.* (2007) observed that for high-dimensional data with highly correlated or dependent measurement variables, pairwise distances are less concentrated than if all variables are independent. They claimed that the distance concentration phenomenon depends on the intrinsic dimension (see, *e.g.*, Levina and Bickel, 2004; Camastra and Staiano, 2016) of the data, instead of the dimension of the embedding space. So, in order to have distance concentration in high dimensions, one needs high intrinsic dimensionality of the data or weak dependence among the measurement variables. The assumption (A2) ensures that weak dependence. Some other similar relevant conditions for the convergence of pairwise distances can be found in (Ahn *et al.*, 2007; Jung and Marron, 2009; Sarkar and Ghosh, 2019; Yata and Aoshima, 2020; Banerjee and Ghosh, 2025). Under (A1)-(A3), we have the following theorem on the misclassification rates of the usual nearest neighbor, CH and MCH classifiers.

Theorem 1: If J competing classes satisfy assumptions (A1)-(A3), and there are at least two observations from each of them (*i.e*, $n_j \geq 2$ for all $j = 1, 2, \ldots, J$), then we have the following results.

(a) If $\nu_{ji}^2 > |\sigma_j^2 - \sigma_i^2|$ for all $j \neq i$, the misclassification probability of the k-nearest neighbor classifier with $k < \min\{n_1, \ldots, n_J\}$ converges to 0 as the dimension d grows to infinity. However, if $\nu_{ji}^2 < |\sigma_j^2 - \sigma_i^2|$ for some $j \neq i$, all observations from at least one class is misclassified with probability tending to 1 as d diverges to infinity.

(b) If $\nu_{ji}^2 > 0$ for $j \neq i$, the misclassification probability of the CH classifier converges to 0 as d grows to infinity.

(c) Suppose that for all $j \neq i$, either $\nu_{ji}^2 > 0$ or $\sigma_j^2 \neq \sigma_i^2$. Then, the misclassification probability of the MCH classifier converges to 0 as d grows to infinity.

Note that in Examples 1-3, we had $\nu_{12}^2 = \mu^2$, $\sigma_1^2 = 1$ and $\sigma_2^2 = \sigma^2$. The condition $\nu_{12}^2 > |\sigma_1^2 - \sigma_2^2|$ was violated in Examples 2 and 3, whereas the condition $\nu_{12}^2 > 0$ was also violated in Example 3. We had poor performance of NN and CH classifiers in these respective cases. But the condition $\nu_{12}^2 > 0$ or $\sigma_1^2 \neq \sigma_2^2$ was satisfied in all three examples. Consequently, the MCH classifier had good high-dimensional performance.

(a) Ex. 4: Two mixture normal distributions

(b) Ex. 5: Uniform vs. mixture of uniforms

(c) Ex. 6: Normal($\mathbf{0}_d, 3\mathbf{I}_d$) vs. standard t with 3 d.f.

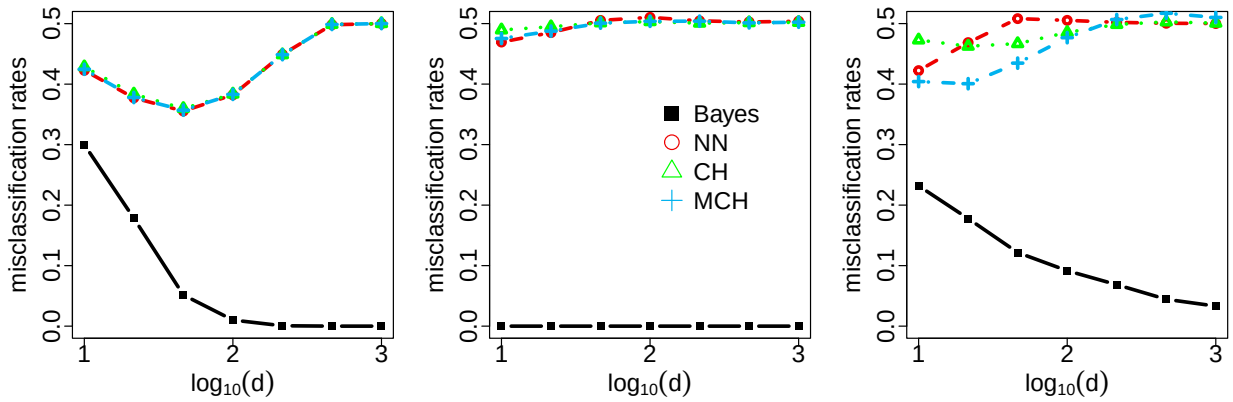


Figure 2: Misclassification rates of Bayes, NN, CH, and MCH classifiers in Examples 4-6.

Now, we consider three more examples (Examples 4-6) to investigate how the MCH classifier performs when at least one of the assumptions of Theorem 1 is violated.

Example 4: *Each of the two classes is an equal mixture of two normal distributions. While one class is a mixture of* $N(\mathbf{0}_d, \mathbf{I}_d)$ *and* $N(\mathbf{1}_d, 2\mathbf{I}_d)$*, the other one is a mixture of* $N(\boldsymbol{\alpha}_d, \mathbf{I}_d)$ *and* $N((\mathbf{1}_d - \boldsymbol{\alpha}_d), 2\mathbf{I}_d)$*, where* $\boldsymbol{\alpha}_d$ *is a* d*-dimensional vector with entries* 0 *and* 1 *at even and odd places, respectively.*

Example 5: *Here the two classes are* $U_d(1, 1.5)$ *and an equal mixture of* $U_d(0.5, 1)$ *and* $U_d(1.5, 2)$*, where* $U_d(a, b)$ *denotes the* d*-dimensional uniform distribution over the region* $\{\mathbf{x} \in \mathbb{R}^d : a \leq \|S^{1/2}x\| \leq b\}$ *for* $S = 0.5\mathbf{I}_d + 0.5\mathbf{1}_d\mathbf{1}_d^\top$.

Examples 4 and 5 are dealing with mixture distributions. Here, (A1)-(A3) hold for each of the four sub-classes, but (A2) is violated for both competing classes. We also consider the following example:

Example 6: *In this example, the two competing classes are* $N(\mathbf{0}_d, 3\mathbf{I}_d)$ *and the standard multivariate* t*-distribution with* 3 *degrees of freedom.*

In Example 6, (A2) is violated for the t-distribution. Moreover, since the two classes have the same mean vector and the same dispersion matrix, we have $\nu_{12}^2 = 0$ and $\sigma_1^2 = \sigma_2^2$. For each example, we consider different values of d ranging between 10 and 1000, and in each case, we form training and test samples of size 50 and 500, respectively, taking an equal number of observations from each class. Each experiment is repeated 100 times to compute the average test set misclassification rates of different classifiers, and they are reported in Figure 2. In these examples, NN, CH and MCH classifiers, all had poor performance, and they had misclassification rates close to 0.5 in high dimensions. These examples clearly

show the necessity to develop some new methods for high dimensional nearest neighbor classification, particularly for the examples involving mixture distributions. In the next section, we propose and discuss some methods for this purpose.

3. Nearest neighbor classification using distance-based features

In the previous sections, we have seen that for high dimensional classification based on nearest neighbors, the scale adjustment methods (CH and MCH) may not always be helpful. To take care of this problem, we suggest extracting some distance-based features from the data and constructing a suitable classifier on that feature space.

3.1. Classification based on minimum distances

Note that in a binary classification problem with training samples $\{\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1}\}$ and $\{\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2}\}$ from the two competing classes, for classification of a test case $\mathbf{Z}$, the 1-NN classifier computes its minimum distances $d_1(\mathbf{Z}) = \min_{1\le i\le n_1} \|\mathbf{Z} - \mathbf{X}_{1i}\|$ and $d_2(\mathbf{Z}) = \min_{1\le i\le n_2} \|\mathbf{Z} - \mathbf{X}_{2i}\|$ from Class-1 and Class-2, respectively. Then it classifies $\mathbf{Z}$ to the first class if $d_2(\mathbf{Z}) > d_1(\mathbf{Z})$ or $d_2^2(\mathbf{Z}) > d_1^2(\mathbf{Z})$. Like the 1-NN classifier, the CH classifier

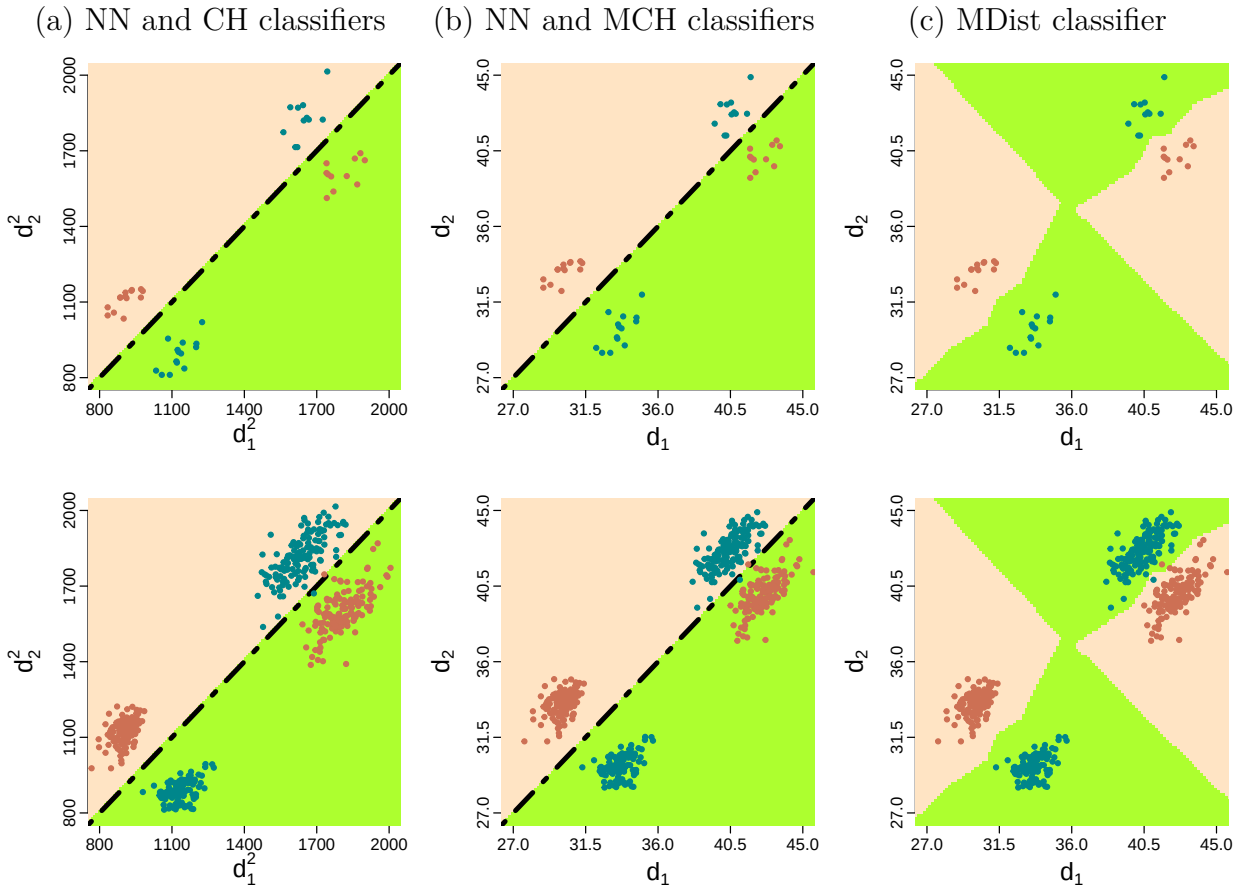


Figure 3: Scatter plots of training (top row) and test (bottom row) samples along with the class boundaries estimated by NN, CH, MCH, and MDist classifiers in Example 4.

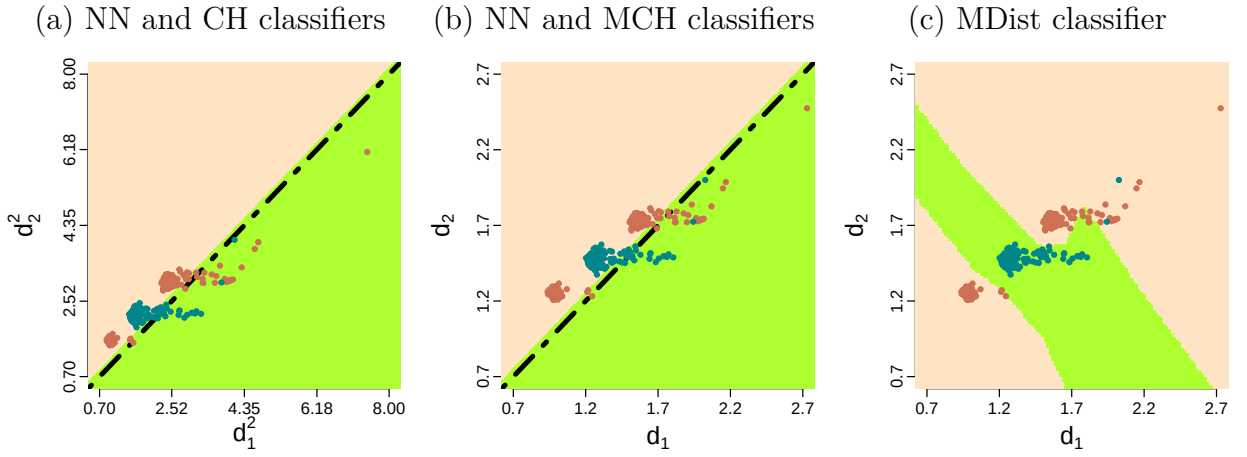


Figure 4: Scatter plots of the test samples and the class boundaries estimated by NN, CH, MCH and MDist classifiers in Example 5.

also leads to linear classification in the $d_1^2 - d_2^2$ space and classifies $\mathbf{Z}$ to the first class if $d_2^2(\mathbf{Z}) > d_1^2(\mathbf{Z}) + C_1$, where $C_1 = \frac{1}{2}\Big[\binom{n_2}{2}^{-1} \sum_{s<t} \|\mathbf{X}_{2s} - \mathbf{X}_{2t}\|^2 - \binom{n_1}{2}^{-1} \sum_{s<t} \|\mathbf{X}_{1s} - \mathbf{X}_{1t}\|^2\Big]$. Similarly, the MCH classifier leads to linear classification in the $d_1 - d_2$ space and classifies $\mathbf{Z}$ to the first class if $d_2(\mathbf{Z}) > d_1(\mathbf{Z}) + C_2$, where $C_2 = \frac{1}{2}\Big[\binom{n_2}{2}^{-1} \sum_{s<t} \|\mathbf{X}_{2s} - \mathbf{X}_{2t}\| - \binom{n_1}{2}^{-1} \sum_{s<t} \|\mathbf{X}_{1s} - \mathbf{X}_{1t}\|\Big]$. The first and the second columns in Figure 3 show the class boundaries estimated by these classifiers (the back line in the first and the second column shows the class boundary estimated by the 1-NN classifier) in Example 4 for dimension 500. They also show the scatter plots of $(d_1(\cdot), d_2(\cdot))$ $\Big($ or $\Big(d_1^2(\cdot), d_2^2(\cdot)\Big)\Big)$ for all training (top row) and test (bottom row) sample observations. For the training data points, the leave-one-out method (see, *e.g.*, Wong, 2015) is used to compute its minimum distances from the two classes. From this figure, it is quite evident that minimum distances (or squared minimum distances) from the two classes contain substantial information about class separability, but the resulting data clouds from the two classes are not linearly separable in that space. As a result, NN, CH, and MCH classifiers, all had poor performance. But we can overcome this problem if we use a suitable nonlinear classifier in that space. For instance one can use the 1-NN classifier in the $d_1 - d_2$ space. This classifier, which is referred to as the MDist classifier, performed well in this example. The last column in Figure 3 shows the class boundary estimated by the MDist classifier. Note that it correctly classified almost all observations.

We observed a similar phenomenon in Example 5 as well (see Figure 4). Like Example 4, here also the observations from different sub-classes form distinct clusters in the $d_1 - d_2$ space (or the $d_1^2 - d_2^2$ space). So, this feature space contains useful information about class separability, but the feature vectors from the two classes are not linearly separable. Therefore, while NN, CH and MCH classifiers had misclassification rates close to 50%, the MDist classifier had an excellent performance.

In Example 6, we have the convergence of pairwise distances for observations from the normal distribution, but not for observations from the multivariate t distribution. In this example, NN, CH and MCH classifiers classified almost all observations into a single class (see Figure 5), but the MDist classifier had much superior performance.

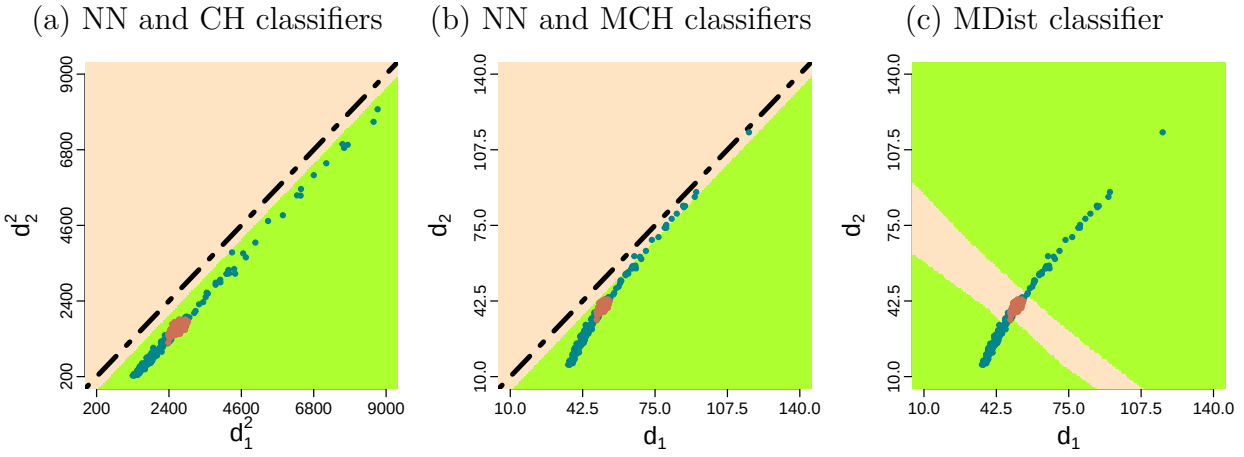


Figure 5: Scatter plots of the test samples and the class boundaries estimated by NN, CH, MCH and MDist classifiers in Example 6.

A similar idea of projecting the observations into a distance-based feature space and using the nearest neighbor classifier on that space was also considered in Dutta and Ghosh (2016), where the authors suggested using average distances $\bar{d}_1(\mathbf{Z}) = \text{avg}_i\|\mathbf{Z} - \mathbf{X}_{1i}\|$ and $\bar{d}_2(\mathbf{Z}) = \text{avg}_i\|\mathbf{Z} - \mathbf{X}_{2i}\|$ from the competing classes as features. Figure 6 shows these features for the test sample observations in Examples 4-6 and also the class boundaries estimated by the resulting classifier, called the TRAD classifier (see Dutta and Ghosh, 2016). Here also, for computing the feature vectors for the training sample observations, the leave-one-out method is used. Figure 6(a) shows that in Example 4, we have reasonable separability in the feature space, but the four distinct clusters are not as prominent as they were in Figure 3. Here we have some overlaps between the clusters corresponding to two competing classes. As a result, TRAD performed better than NN, CH and MCH classifiers, but not as good as the MDist classifier. This is also evident from Figure 7(a), which shows the average (over 100 replications) test set misclassification rates of these classifiers for various choices of d. In

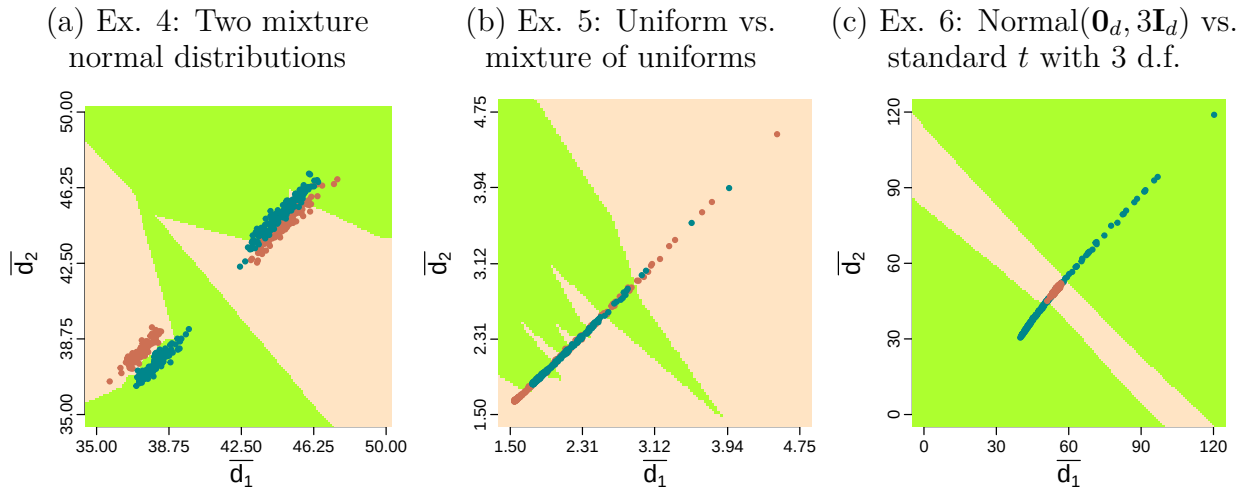


Figure 6: Scatter plots of the test samples and the class boundaries estimated by the TRAD classifier in Examples 4-6.

(a) Ex. 4: Two mixture normal distributions

(b) Ex. 5: Uniform vs. mixture of uniforms

(c) Ex. 6: Normal($\mathbf{0}_d, 3\mathbf{I}_d$) vs. standard t with 3 d.f.

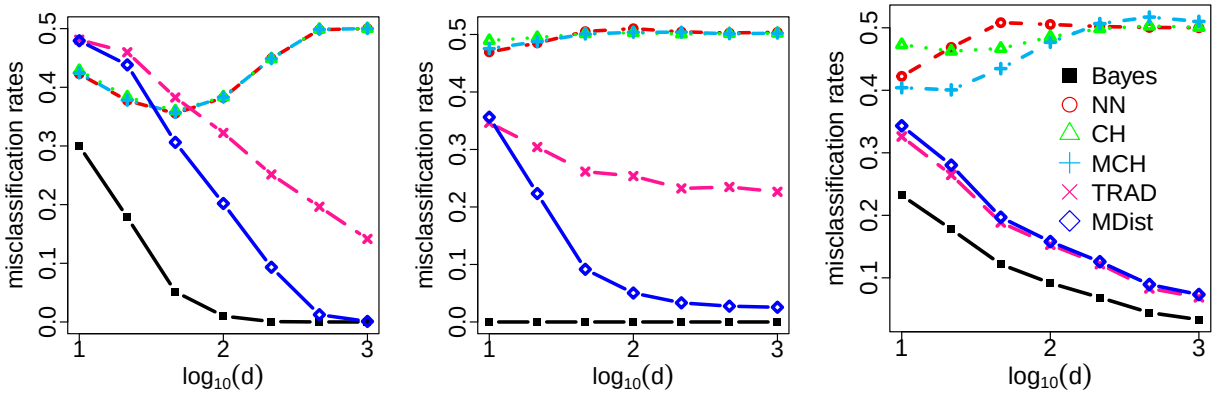


Figure 7: Misclassification rates of Bayes, CH, MCH, TRAD and MDist classifiers in Examples 4-6.

Example 5, the features based on average distances do not provide much separation between the two classes (see Figure 6(b)). So, as expected, TRAD had much higher misclassification rates (see 7(b). However, in Example 6, the $\bar{d}_1 - \bar{d}_2$ space shows almost the same degree of separation as in the $d_1 - d_2$ space (see Figure 6). So, the class boundaries estimated by TRAD and MDist classifiers were almost similar, and they had almost similar misclassification rates. (see Figures 7(c)).

The success of the MDist classifier motivates us to carry out some theoretical analysis to understand its high-dimensional behavior. For this investigation, we again consider assumptions (A1)-(A3), and prove the perfect classification property of the MDist classifier in high dimensions.

Theorem 2: Suppose that J competing classes satisfy assumptions (A1)-(A3), and from each of them, there are at least two observations (*i.e*, $n_j \geq 2$ for all $j = 1, 2, \ldots, J$). If for all $j \neq i$, $\nu_{ji}^2 > 0$ or $\sigma_j^2 \neq \sigma_i^2$, the misclassification rate of the MDist classifier converges to 0 as d grows to infinity.

However, as we have discussed before, if the competing classes are mixtures of several sub-classes, the assumptions (A1)-(A3) may hold for each of the sub-classes but none of the competing classes (as in Examples 4). We have seen that in such situations, CH and MCH classifiers often have poor performance in high dimensions. However, from the proof of Theorem 2 (see Appendix), it is clear that in such cases, for each of the sub-classes, the feature vectors of minimum distances converge to a point as the dimension increases. If these points are distinct for each sub-class, we get some distinct clusters in the feature space, and the MDist classifier leads to perfect classification. We have already seen that in Example 4. A theorem similar to Theorem 2 can be stated for these mixture distributions as well, but the conditions for perfect classification by the MDist classifier (*i.e.*, the conditions needed to ensure that for any two sub-classes from two competing classes, the feature vectors converge to two distinct points, one for each sub-class) becomes mathematically complicated to interpret. That is why we choose not to state that theorem here.

Now we consider two interesting examples (Example 7 and 8) involving binary classification, where each of the two competing classes satisfies assumptions (A1)-(A3), but we have $\nu_{12}^2 = 0$ and $\sigma_1^2 = \sigma_2^2$. So, in this case, the feature vectors $(d_1(\cdot), d_2(\cdot))$ corresponding to two competing classes converge to the same limiting value as d increases. One may be curious to know how the MDist classifier performs in such situations, and we investigate it here. Here also, we consider different values of d ranging between 10 and 1000, form the training and the test sets of size 50 and 500 by taking an equal number of observations from the two classes and repeat the experiment 100 times to compute the average test set misclassification rates of different classifiers.

Example 7: *We consider two normal distributions having the same mean vector* $\mathbf{0}_d$ *but different dispersion matrices* Λ_1=*diag*$(\lambda_{11}, \ldots, \lambda_{1d})$ *and* Λ_2=*diag*$(\lambda_{21}, \ldots, \lambda_{2d})$. *Here* $\lambda_{1i} = 1/2$ *and* $\lambda_{2i} = 2$ *for* $i \leq d/2$, *whereas* $\lambda_{1i} = 2$ *and* $\lambda_{2i} = 1/2$ *for* $i > d/2$.

Figure 8(a) show the scatter plots of the test set observations in the $d_1 - d_2$ space and the class boundary estimated by the MDist classifier for $d = 500$. It is clear that unlike previous examples, here the features based on minimum distances fail to discriminate between the two classes. As a result, the MDist classifier had much higher misclas-

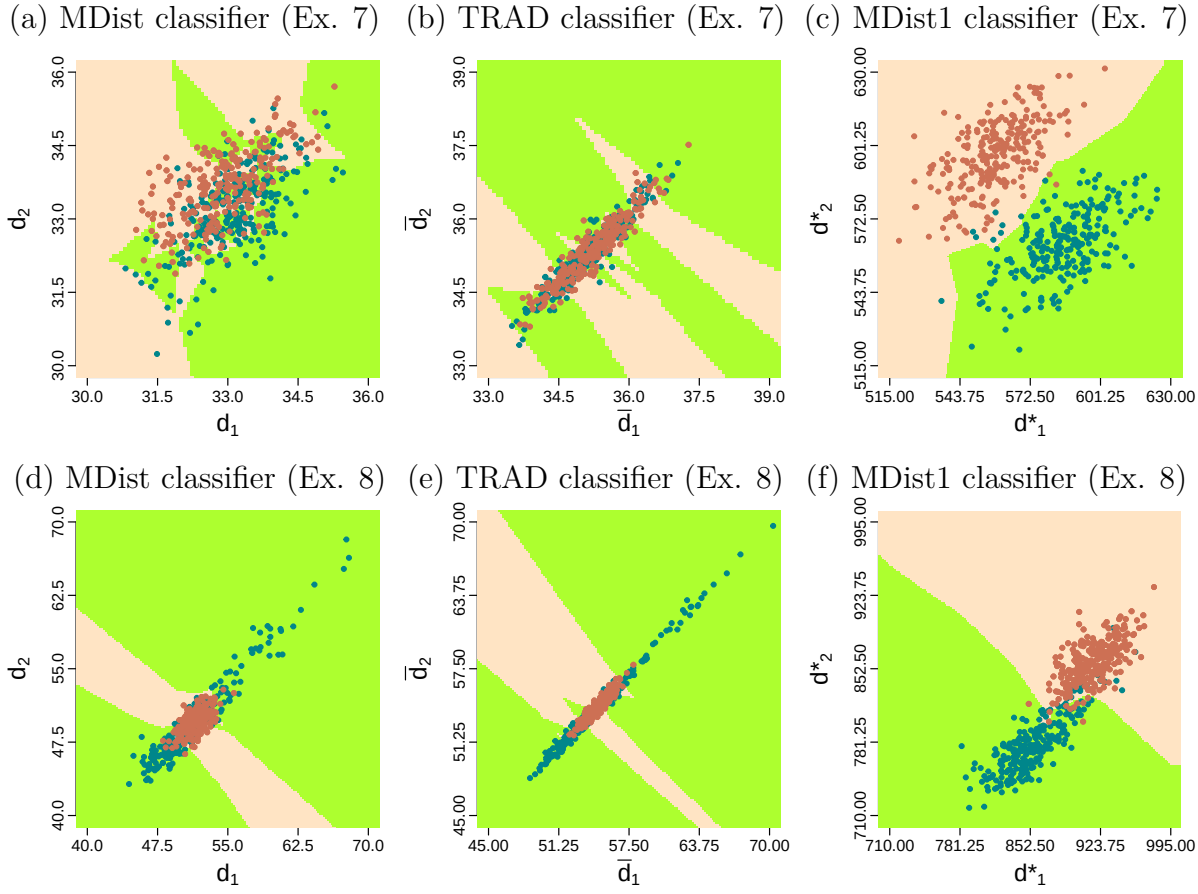


Figure 8: Scatter plots of the test samples and the class boundaries estimated by the MDist, TRAD and MDist1 classifiers in Example 7 (top row) and Example 8 (bottom row).

(a) Ex. 7: Two normals with different dispersion matrices having same trace

(b) Ex. 8: Product of $N(0,3)$ vs product of univariate t_3

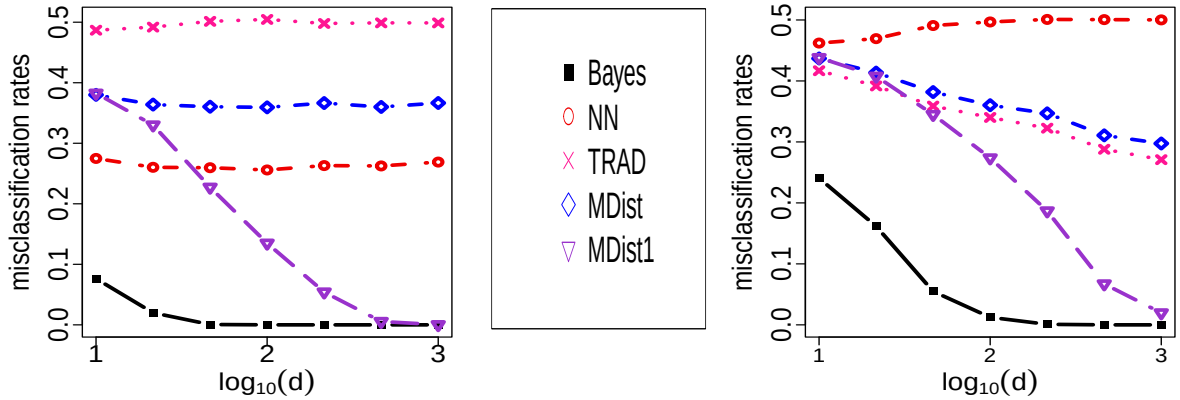


Figure 9: Misclassification rates of Bayes, NN, TRAD, MDist and MDist1 classifiers in Examples 7 and 8.

sification rates (see Figure 9(a)). The features based on average ℓ_2 distances also fail to provide any discriminatory information (see Figure 8(b)) in this case. The performance of the TRAD classifier was even worse. It misclassified almost half of the test set observations (see Figure 9(a)). Surprisingly, in this example, we get a good result if instead of ℓ_2 distance (Euclidean distance), we use ℓ_1 distance (Manhattan distance) for finding the neighbors. If $\{\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1}\}$ and $\{\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2}\}$ are training sample observations from two competing classes (here we have $n_1 = n_2 = 25$), for any $\mathbf{Z}$, we can use $d_1^*(\mathbf{Z}) = \min_{1\le i\le n_1} \|\mathbf{Z} - \mathbf{X}_{1i}\|_1$ and $d_2^*(\mathbf{Z}) = \min_{1\le i\le n_2} \|\mathbf{Z} - \mathbf{X}_{2i}\|_1$ as features and perform usual nearest neighbor classification on that feature space. Here also for computing d_1^* and d_2^* at the training data points, we use the leave-one-out method. Figure 8(c) shows the scatter plot of these features for all test set observations and the class boundary estimated by the 1-NN classifier on this feature space (we call it the MDist1 classifier). Here we have two distinct clusters in the feature space, one for each class. As a result, the MDist1 classifier had an excellent performance and correctly classified almost all observations. The average test set misclassification rate of this classifier (reported in Figure 9(a)) also tells us the same story. Now, let us consider the following example.

Example 8: *Here each of the two classes has i.i.d. measurement variables. In Class-1, they follow the* $N(0,3)$ *distribution, while in Class-2, they follow the standard t distribution with* 3 *degrees of freedom.*

Note that this is different from the multivariate t distribution considered in Example 6. In this example also, the features based on minimum ℓ_2 distances and those based on average ℓ_2 distances do not provide much separability between the two classes (see Figure 8(d) and (e)), but the features based on minimum ℓ_1 distances make the data clouds better separated (see Figure 8(f)). As a result, the MDist1 classifier outperformed TRAD and MDist classifiers (see Figure 9(b)).

To understand the high-dimensional behavior of the MDist1 classifier, we carry out some theoretical investigations under the following assumptions, which are similar to (A1)-(A3) stated before.

(A1°) In each of the J competing classes, the measurement variables have uniformly bounded second moments.

(A2°) If $\mathbf{X} = (X_1, \ldots, X_d)^\top \sim F_j$ and $\mathbf{Y} = (Y_1, \ldots, Y_d)^\top \sim F_i$ $(1 \leq j, i \leq J)$ are independent, for $\mathbf{U} = \mathbf{X} - \mathbf{Y}$, $\sum_{r \neq s} |Corr(|U_r|, |U_s|)|$ is of the order $o(d^2)$.

(A3°) For independent random vectors $\mathbf{X} \sim F_j$ and $\mathbf{Y} \sim F_i$ $(1 \leq j, i \leq J)$, $E\Big(\frac{1}{d}\|\mathbf{X} - \mathbf{Y}\|_1\Big) = \frac{1}{d}\sum_{q=1}^{d} E|X_q - Y_q|$ converges to a constant τ_{ji} as $d \to \infty$.

Under (A1°) and (A2°), we have the convergence of pairwise ℓ_1 distances. Following similar steps as used in the proof of Lemma 1, one can show that for $\mathbf{X} \sim F_j$ and $\mathbf{Y} \sim F_i$ $(1 \leq j, i \leq J)$, $\Big|\frac{1}{d}\|\mathbf{X} - \mathbf{Y}\|_1 - E\Big(\frac{1}{d}\|\mathbf{X} - \mathbf{Y}\|_1\Big)\Big| \xrightarrow{P} 0$ as $d \to \infty$.

For any $q = 1, 2, \ldots d$, define $e_{ji}^{(q)} = 2E|X_q - Y_q| - E|X_q - X_q'| - E|Y_q - Y_q'|$, where $\mathbf{X}, \mathbf{X}' \sim F_j$ and $\mathbf{Y}, \mathbf{Y}' \sim F_i$ $(j \neq i)$ are independent random vectors. This quantity $e_{ji}^{(q)}$ can be viewed as the energy distance (see, *e.g.*, Székely and Rizzo, 2023) between the q-th marginals of F_j and F_i ($F_j^{(q)}$ and $F_i^{(q)}$, say). Following Baringhaus and Franz (2004), one can show that $e_{ji}^{(q)}$ is non-negative, and it takes the value 0 if and only if $F_j^{(q)} = F_i^{(q)}$. So, for any fixed dimension d, we have $\frac{1}{d}\Big[2E\|\mathbf{X} - \mathbf{Y}\|_1 - E\|\mathbf{X} - \mathbf{X}'\|_1 - E\|\mathbf{Y} - \mathbf{Y}'\|_1\Big] = \frac{1}{d}\sum_{q=1}^{d} e_{ji}^{(q)} = \bar{e}_{ji}(d) \geq 0$, where the equality holds if and only if $F_j^{(q)} = F_i^{(q)}$ for $q = 1, 2, \ldots, d$. Therefore, it is somewhat reasonable to assume that $\mathcal{E}_{ji} = \lim_{d\to\infty} \bar{e}_{ji}(d) > 0$, which essentially says that the average coordinate-wise energy distance is asymptotically non-negligible. Under this assumption, we have the perfect separation property of the MDist1 classifier, which is asserted by the following theorem.

Theorem 3: Suppose that J competing classes satisfy assumptions (A1°)-(A3°), and from each of them, there are at least two observations (*i.e*, $n_j \geq 2$ for all $j = 1, 2, \ldots, J$). If the limiting value of the average coordinate-wise energy distance $\mathcal{E}_{ji} > 0$ for all $j \neq i$, the misclassification rate of the MDist1 classifier converges to 0 as d grows to infinity.

Note that while TRAD and MDist classifiers fail to discriminate between two distributions differing outside the first two moments, the MDist1 classifier can discriminate between them as long as they differ in their one-dimensional marginals. That is why it outperformed TRAD and MDist classifiers in Examples 7 and 8.

This classifier enjoys the perfect separation property in high dimensions even when the competing classes are mixtures of several sub-classes, and these sub-class distributions satisfy assumptions (A1°)-(A3°). It becomes clear from the proof of Theorem 3 (see Appendix) that in such cases, for each of the sub-classes, the feature vectors of minimum ℓ_1 distances (after appropriate scaling) converge to a point as the dimension increases. If these points are distinct for each sub-class, the MDist1 classifier leads to perfect classification. But here also writing the conditions for perfect classification becomes mathematically complicated to interpret. So, we decide not to state another theorem in this regard. However, our analysis of simulated data sets clearly demonstrates this. Figure 10 shows the misclassification rates of TRAD, MDist and MDist1 classifiers in Examples 1-6 along with those of 1-NN and Bayes

classifiers. In all these examples including Example 4 and 5, where we deal with mixture distributions, MDist and MDist1 classifiers had similar performance.

This figure shows another interesting phenomenon. In Examples 1-3, when the underlying distributions are unimodal, the TRAD classifier, which considers the average of distances from all observations, performed better than MDist and MDist1 classifiers, which consider the distance of one nearest neighbor only. In Example 6 also, TRAD had an edge over the other two classifiers. But, in the case of mixture distributions (see Examples 4 and 5), taking the average over all observations coming from different sub-classes does not seem to be a meaningful option. In those cases, MDist and MDist1 classifiers outperformed the TRAD classifier. These result shows that instead of always going for features based on a single nearest neighbor from each class, sometimes it is more meaningful to consider distances from multiple nearest neighbors. We can include these distances in the set of features and go for classification in the extended feature space. We consider such methods in the following subsection.

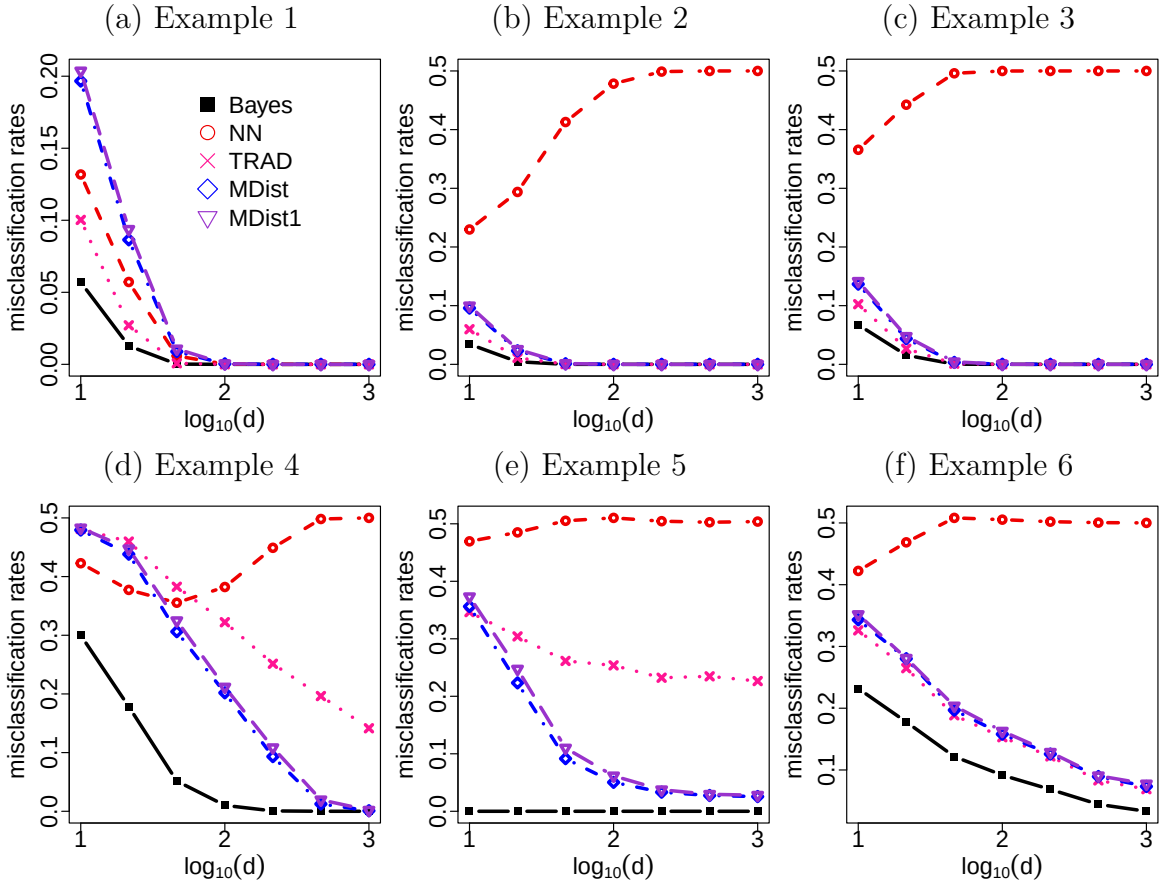


Figure 10: Misclassification rates of Bayes, NN, TRAD, MDist and MDist1 classifiers in Examples 1-6.

3.2. Clssification based on multiple neighbors

Instead of considering only the distance of the first neighbor from each class, here we consider the distances of the first r ($r \geq 1$) neighbors from each class and use them

as features. So, if there are J competing classes, we consider a total of Jr many features and use the 1-NN classifier on that feature space. Here also, we can use ℓ_2 distances or ℓ_1 distances as features, and the resulting classifier is referred to as rMDist and rMDist1 classifiers, respectively. One may also consider both ℓ_1 and ℓ_2 distances and deal with $2Jr$ many features simultaneously. We refer to the resulting classifier as the rMDistC classifier. In all these cases the value of r is chosen by minimizing the leave-one-out cross-validation

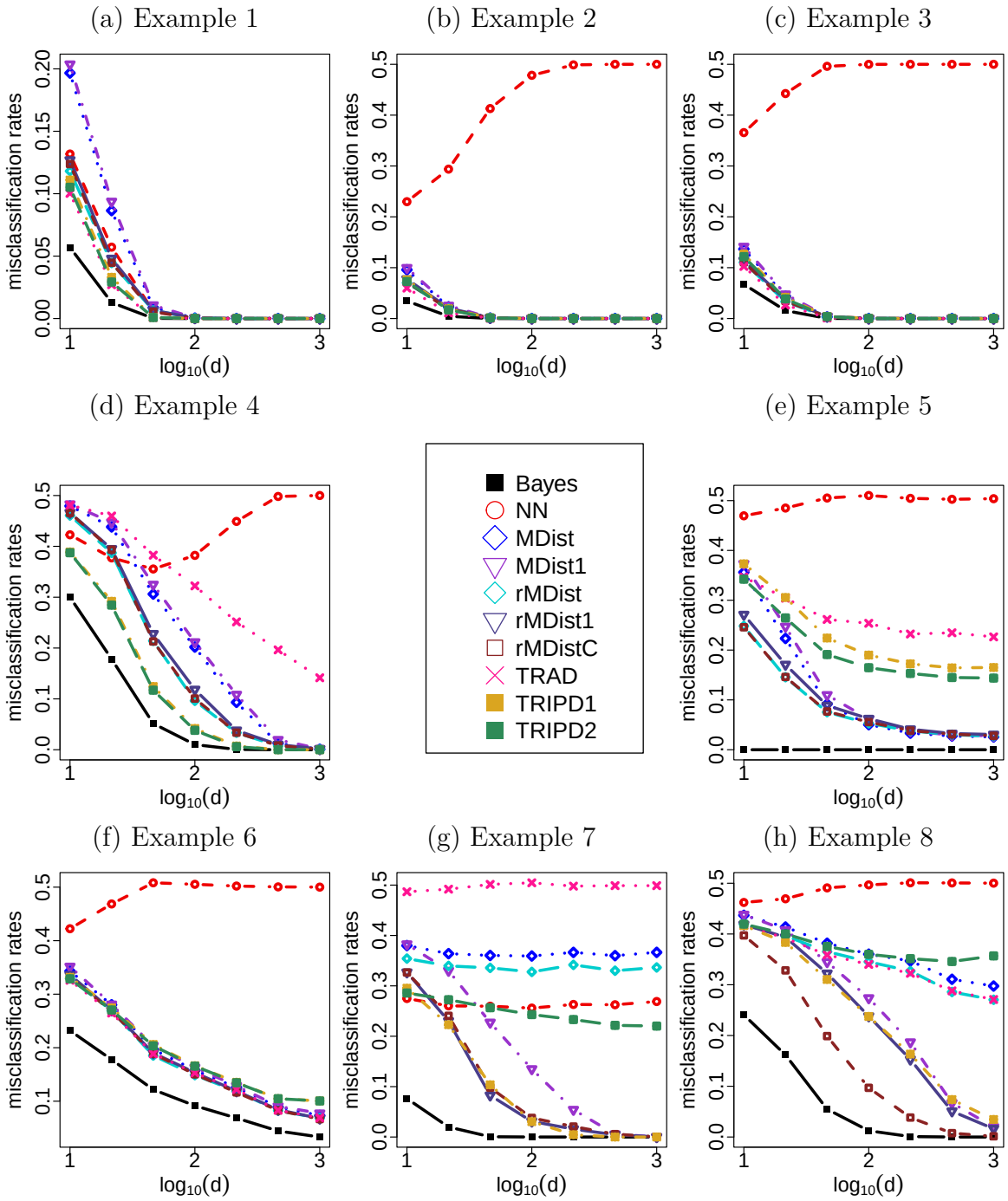


Figure 11: Misclassification rates of different classifiers in Examples 1-8.

estimate (see, *e.g.* Wong, 2015) of the misclassification rate. Figure 11 shows the average test set misclassification rates of these classifiers in Examples 1-8. One can see that in most of the examples, rMDist and rMDist1 classifiers performed better than MDist and MDist1 classifiers, respectively. The rMDistC classifier also performed well in almost all examples. In Examples 7 and 8, it outperformed rMDist and rMDist1 classifiers.

A somewhat similar classification method was considered in Dutta and Ghosh (2016), where the authors transformed each observation into an n-dimensional vector containing its distances from all training sample observations. They also considered ℓ_1 and ℓ_2 distances for transformation and called the resulting classifiers as TRIPD1 and TRIPD2. respectively. Misclassification rates of those two methods are also reported in Figure 11. In Examples 1-3 and 6, their misclassification rates were similar to our proposed methods. In Example 4, they had the lowest misclassification rates, but in Example 5, they were outperformed by our proposed classifiers. In Examples 7 and 8, the TRIPD2 classifier, which is based on ℓ_2 distances, had higher misclassification rates. In Example 7, the performance of the TRIPD1 classifier was comparable to rMDist1 and rMDistC classifiers, but in Example 8, the rMdistC classifier had a clear edge.

4. Results from the analysis of benchmark datasets

We analyze 10 benchmark datasets for further evaluation of the performance of the proposed and existing methods discussed in the previous two sections. For these benchmark datasets, since the true class distributions are not known, it is not possible to compute the Bayes risks. Therefore, to facilitate comparison, we report the misclassification rates of two popular classifiers, support vector machines (SVM) (see, *e.g.* Cristianini and Shawe-Taylor, 2003; Scholkopf and Smola, 2018) and random forest (RF) (see, *e.g.* Breiman, 2001; Genuer and Poggi, 2020), which are known to perform well for high dimensional data. Since the nearest neighbor classifiers are nonlinear, to make it fair, here we use the nonlinear SVM for comparison. For our numerical study, we use the radial basis function kernel, where all tuning parameters are chosen using the 5-fold cross-validation method (see, *e.g.*, Wong, 2015). We use the R package `caret` for this purpose. The same package is used for the random forest classifier as well, where we use default tuning parameters.

Out of these 10 datasets, Chowdary and Nutt datasets are taken from CompCancer dataset-Schliep lab. The rest of the datasets are taken from the UCR Time Series Classification Archive. Detailed descriptions of these datasets are available at these respective sources. The datasets taken from the UCR archive have specific training and test sets. We merge these two sets and divide the pooled dataset randomly into two parts to form the training and the test samples. Except for the Synthetic Control Chart data, in all other cases, the sizes of training and test samples are taken to be the same as they are in the data archive. Note that in all these cases, the size of the training sample is smaller than the dimension. In the case of Synthetic Control Chart data, instead of an equal partition (as in UCR achieve), we use training and test samples of size 60 and 540, respectively, so that the training sample size does not become larger than the dimension. The datasets from the CompCancer database do not have specific training and test samples. In these cases, we divide the data sets into equal halves to form the training and the test samples. Brief descriptions of these datasets are given in Table 1. In all these cases, we form the training and the test samples in such a way that the proportions of different classes in the two samples are

Table 1: Brief descriptions of the benchmark datasets.

dataset	d	J	Sample size		dataset	d	J	Sample size	
			Train	Test				Train	Test
Synthetic Control	60	6	60	540	Lightning7	319	7	70	73
Chowdary	182	2	52	52	Herring	512	2	64	64
Trace	275	4	100	100	Nutt	1070	2	14	14
Toe Segmentation1	277	2	40	228	Gordon	1628	2	90	91
Coffee	286	2	28	28	Colon Cancer	2000	2	31	31

Table 2: Average misclassification rates (in %) of different classifiers and their standard errors (reported inside the bracket) in benchmark datasets.

Dataset	Synth. Control	Chowdary	Trace	Toe Segment.1	Coffee	Lightning7	Herring	Nutt	Gordon	Colon Cancer
NN	18.78	4.83	20.33	38.62	2.00	37.97	51.33	34.00	2.96	26.10
	(0.28)	(0.29)	(0.37)	(0.36)	(0.31)	(0.43)	(0.59)	(0.92)	(0.16)	(0.65)
MDist	10.02	7.98	13.51	42.80	2.61	39.29	45.53	14.14	1.92	32.42
	(0.26)	(0.45)	(0.46)	(0.45)	(0.34)	(0.47)	(0.52)	(0.74)	(0.17)	(0.95)
MDist1	12.29	7.04	18.88	37.30	4.43	35.72	47.20	15.71	1.19	34.74
	(0.30)	(0.46)	(0.46)	(0.45)	(0.39)	(0.52)	(0.57)	(0.70)	(0.10)	(0.94)
rMDist	10.06	6.56	14.90	40.92	2.93	38.16	46.86	15.14	1.73	22.06
	(0.28)	(0.37)	(0.48)	(0.38)	(0.32)	(0.42)	(0.54)	(0.86)	(0.14)	(0.92)
rMDist1	10.12	5.35	19.61	35.64	4.50	36.00	46.06	15.93	1.27	27.03
	(0.27)	(0.35)	(0.44)	(0.47)	(0.39)	(0.44)	(0.63)	(0.68)	(0.10)	(1.03)
rMDistC	9.25	5.90	15.01	35.61	3.07	34.67	46.88	14.21	1.64	24.65
	(0.27)	(0.32)	(0.49)	(0.47)	(0.34)	(0.47)	(0.54)	(0.80)	(0.15)	(0.96)
TRAD	14.78	7.31	24.48	49.93	4.11	37.28	48.08	12.07	6.52	18.06
	(0.30)	(0.34)	(0.37)	(0.41)	(0.43)	(0.39)	(0.56)	(0.72)	(0.19)	(0.72)
TRIPD1	7.67	4.42	23.25	33.92	6.14	31.47	47.36	17.57	1.21	25.77
	(0.19)	(0.25)	(0.43)	(0.32)	(0.40)	(0.39)	(0.53)	(0.95)	(0.10)	(0.66)
TRIPD2	6.42	5.38	21.08	38.15	3.79	32.27	50.50	8.93	3.38	21.58
	(0.19)	(0.30)	(0.39)	(0.34)	(0.39)	(0.40)	(0.56)	(0.86)	(0.20)	(0.62)
Random	13.07	5.00	14.23	38.21	3.21	28.29	40.30	21.07	0.93	29.55
Forest	(0.24)	(0.26)	(0.48)	(0.41)	(0.45)	(0.47)	(0.47)	(1.09)	(0.11)	(0.64)
Nonlin.	9.50	7.63	10.48	45.02	4.25	35.77	39.11	11.21	2.53	20.90
SVM	(0.32)	(0.50)	(0.36)	(0.37)	(0.46)	(0.47)	(0.41)	(0.75)	(0.22)	(0.75)

as close as possible. In each case, this partitioning is carried out 100 times, and the average test set misclassification rates of different classifiers are reported in Table 2 along with their corresponding standard errors. Overall performances of CH and MCH classifiers (especially, that of the former one) were much inferior compared to all other classifiers considered here. Therefore, we do not report them in this section.

Though the 1-NN classifier had the lowest misclassification rate in the Coffee dataset and the second lowest misclassification rate in the Chowdary dataset, in many cases, its performance was far from the best one (see Table 2). For instance, in Nutt and Synthetic Control Chart datasets, its misclassification rates were much higher compared to all other classifiers considered here. Furthermore, it had the highest misclassification rate in the Her-

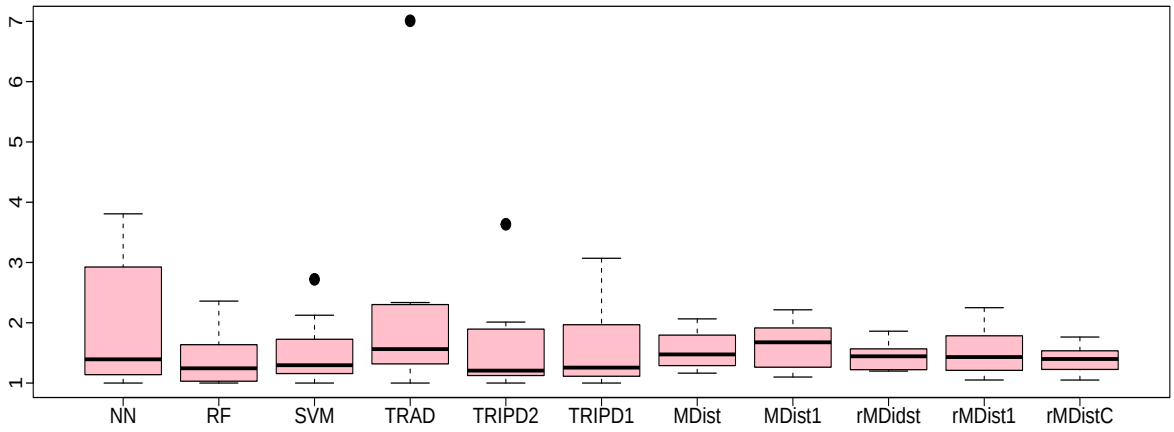


Figure 12: Boxplots showing the robustness of different classifiers in benchmark datasets.

ring data set. TRAD also had relatively higher misclassification rates in many examples (*e.g.*, Synthetic Control Chart, Trace, Toe Segmentation, and Gordon datasets). Only in the case of Colon Cancer data, it outperformed others. Our proposed classifiers had good performance in most of these examples. While the classifiers based on multiple neighbors (rMDist, rMDist1 and rMDistC) outperformed those based on single neighbors (MDist and MDist1) in Chowdary, Toe Segmentation, and Colon Cancer data sets, in all other cases, they had comparable performance. In Trace and Herring data sets, these classifiers performed better than TRAD, TRIPD1 and TRIPD2 classifiers. Table 2 clearly shows that the performances of our proposed classifiers, particularly for classifiers based on multiple neighbors, were comparable to nonlinear SVM and random forest.

To compare the overall performances of different classifiers concisely and comprehensively, we used the notion of robustness introduced in Friedman (1994). If there are T classifiers who have misclassification rates $e_1, e_2, \ldots, e_T$ in a particular data set, the robustness of the t-th classifier is computed as $R(t) = e_t/e_0$, where $e_0 = \min_{1\le t\le T} e_t$. So, in an example, the best classifier has $R(t) = 1$, while higher values of $R(t)$ indicate the lack of robustness of the t-th classifier. For each of these benchmark data sets, we computed these ratios for all classifiers, and they are graphically represented using box plots in Figure 12. This figure clearly shows that the overall performances of all other classifiers were somewhat better than the usual nearest neighbor classifier. It also shows that among our proposed classifiers, those based on multiple neighbors performed better than the corresponding classifiers based on a single nearest neighbor. While the rMDist classifier exhibited better robustness properties than TRIPD2, the rMDist1 classifier turned out to be more robust than the TRIPD1 classifier. The rMDistC classifier, which considers both ℓ_1 and ℓ_2 distances of the nearest neighbors, also had an excellent overall performance. If not better, the performances of our proposed classifiers were comparable to the popular classifiers like nonlinear SVM and random forest.

5. Concluding remarks

In this article, we have proposed some possible modifications to the nearest neighbor classifier for the classification of high-dimensional data. We have seen that if the location

difference among the competing classes gets masked by their scale difference, the usual nearest neighbor classifier performs poorly in high dimensions. The adjustment proposed by Chan and Hall (2009b) takes care of this problem, but the resulting classifier fails when the competing classes differ outside their first moments. The MCH classifier overcomes this limitation, and it can discriminate between two high-dimensional distributions differing either in their locations or in their scales. However, this method may not work well in many situations, especially when the class distributions are mixtures of several widely varying subclasses. The proposed classifiers based on minimum distances are helpful in such situations. The MDist1 classifier can even discriminate among competing classes differing outside the first two moments. Instead of considering only one neighbor from each class, sometimes it is helpful to consider the distances of the first r neighbors and perform nearest neighbor classification in that feature space. Analyzing several simulated and benchmark datasets, we have amply demonstrated that if not better, our proposed classifiers yield competitive performance in high dimensions.

In this article, we have used nearest neighbor classification on the feature space of ℓ_1 or ℓ_2 distances. Though we have seen some theoretical advantages of using the ℓ_1 distance, our analysis of benchmark datasets clearly shows that in practice, there is no clear winner. So, a user may wonder which of the two feature spaces to use in a given problem. One may also like to use ℓ_p-distances for other choices of p or features based on other generalized distance functions of the form $\varphi_{h,\psi}(\boldsymbol{x},\boldsymbol{y}) = h\Big\{\frac{1}{d}\sum_{i=1}^{d}\psi\Big(|x^{(i)}-y^{(i)}|^2\Big)\Big\}$ where $h:\mathbb{R}_+\to\mathbb{R}_+$ and $\psi:\mathbb{R}_+\to\mathbb{R}_+$ are continuous, strictly increasing functions with $h(0)=\psi(0)=0$ as introduced in Sarkar and Ghosh (2018). For suitable choices of h and ψ (*e.g.*, when $h(t)=t$ and $\psi(t)$ has non-constant completely monotone derivatives), it ensures the positivity of the energy distance (as discussed in the paragraph before Theorem 3) in any finite dimension. Moreover, if ψ is bounded, it also makes the resulting classifier robust against outliers. However, the choice of the optimal features is a challenging problem, and it would be helpful if a data-driven method can be developed for this purpose. Throughout this article, for our proposed methods, we have always used the 1-NN classifier in the feature space. This is mainly for a fair comparison with other competing nearest neighbor methods (*e.g.*, CH, TRAD, TRIPD1, and TRIPD2), where 1-NN classification is considered. However, in practice, one may use the k-NN classifier for other values of k as well. For constructing the rMDistC classifier, though we have considered the same number of ℓ_1 and ℓ_2 distances as features, it is possible to include r_1 many ℓ_1 distances and r_2 many ℓ_2 distance in the set of features. We avoid choosing different values of r_1 and r_2 to reduce the computing cost at the cross-validation step. In practice, distances from all of the first r neighbors may not always be important for classification. In such cases, a suitable feature selection criterion would be helpful. Instead of feature selection, one can also think about constructing an ensemble classifier (see, *e.g.* Dietterich, 2000; Zhang and Zhang, 2009; Kiziloz, 2021) like random forest, where we construct different classifiers based on different sets of features and judiciously aggregate them. These problems can be investigated in a separate work in future.

References

Ahn, J., Marron, J. S., Muller, K. M., and Chi, Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika*, **94**, 760–766.

Banerjee, B. and Ghosh, A. K. (2025). On high dimensional behaviour of some two-sample tests based on ball divergence. *Statistica Sinica*, **35**, To appear.

Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, **88**, 190–206.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5–32.

Camastra, F. and Staiano, A. (2016). Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, **328**, 26–41.

Carrerira-Perpinan, M. (2009). A review of dimension reduction techniques. Technical report. Department of Computer Science, University of Sheffield.

Chan, Y. B. and Hall, P. (2009a). Robust nearest neighbor methods for classifying high-dimensional data. *The Annals of Statistics*, **37**, 3186–3203.

Chan, Y.-B. and Hall, P. (2009b). Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika*, **96**, 469–478.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, **13**, 21–27.

Cristianini, N. and Shawe-Taylor, J. (2003). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Deegalla, S. and Bostrom, H. (2006). Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification. In *5th International Conference on Machine Learning and Applications*, pages 245–250. IEEE.

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A Probabilistic Theory of Pattern Recognition*. Springer Science & Business Media, Berlin.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15. Springer.

Duda, R. O., Hart, P. E., and Stork, D. G. (2007). *Pattern Classification*. Wiley, New York.

Dutta, S. and Ghosh, A. K. (2016). On some transformations of high dimension, low sample size data for nearest neighbor classification. *Machine Learning*, **102**, 57–83.

Feller, W. (1991). *An Introduction to Probability Theory and its Applications, Volume 1*. John Wiley & Sons.

Fern, X. Z. and Brodley, C. E. (2003). Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 186–193.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188.

Fix, E. and Hodges, J. L. (1951). Discriminatory analysis - nonparametric discrimination: Consistency properties. Project 21-49-004, Report 4, US Air Force School of Aviation Medicine, Randolph Field.

Fradkin, D. and Madigan, D. (2003). Experiments with random projections for machine learning. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 517–522.

François, D., Wertz, V., and Verleysen, M. (2007). The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, **19**, 873–886.

Friedman, J. H. (1994). Flexible metric nearest neighbor classification. Technical report, Department of Statistics, Stanford University.

Genuer, R. and Poggi, J.-M. (2020). *Random Forests*. Springer.

Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B*, **67**, 427–444.

Hall, P., Park, B. U., and Samworth, R. J. (2008). Choice of neighbor order in nearest-neighbor classification. *The Annals of Statistics*, **36**, 2135–2152.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *The Annals of Statistics*, **37**, 4104–4130.

Kiziloz, H. E. (2021). Classifier ensemble methods in feature selection. *Neurocomputing*, **419**, 97–107.

Levina, E. and Bickel, P. (2004). Maximum likelihood estimation of intrinsic dimension. *Advances in Neural Information Processing Systems*, **17**.

Maciończyk, R., Moryc, M., and Buchtyar, P. (2023). Analyzing the impact of principal component analysis on k-nearest neighbors and naive bayes classification algorithms. In *International Conference on Information and Software Technologies*, pages 247–263. Springer.

Pal, A. K., Mondal, P. K., and Ghosh, A. K. (2016). High dimensional nearest neighbor classification based on mean absolute differences of inter-point distances. *Pattern Recognition Letters*, **74**, 1–8.

Radovanovic, M., Nanopoulos, A., and Ivanovic, M. (2010). Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, **11**, 2487–2531.

Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society: Series B*, **10**, 159–203.

Roy, S., Sarkar, S., Dutta, S., and Ghosh, A. K. (2022). On generalizations of some distance based classifiers for HDLSS data. *Journal of Machine Learning Research*, **23**, 1–41.

Sarkar, S. and Ghosh, A. K. (2018). On some high-dimensional two-sample tests based on averages of inter-point distances. *Stat*, **7**, e187, 16.

Sarkar, S. and Ghosh, A. K. (2019). On perfect clustering of high dimension, low sample size data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, 2257–2272.

Scholkopf, B. and Smola, A. J. (2018). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer Science and Business Media.

Székely, G. J. and Rizzo, M. L. (2023). *The Energy of Data and Distance Correlation*. Chapman and Hall/CRC.

Tomašev, N., Radovanović, M., Mladenić, D., and Ivanović, M. (2014). Hubness-based fuzzy measures for high-dimensional k-nearest neighbor classification. *International Journal of Machine Learning and Cybernetics*, **5**, 445–458.

Vempala, S. S. (2005). *The Random Projection Method*. DIMACS - Series in Discrete Mathematics and Theoretical Computer Science, Volume 65, American Mathematical Society.

Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, **10**, 207–244.

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, **48**, 2839–2846.

Yata, K. and Aoshima, M. (2020). Geometric consistency of principal component scores for high-dimensional mixture models and its application. *Scandinavian Journal of Statistics*, **47**, 899–921.

Zhang, C.-X. and Zhang, J.-S. (2009). A novel method for constructing ensemble classifiers. *Statistics and Computing*, **19**, 317–327.

APPENDIX

Lemma 1: If J competing classes satisfy (A1)-(A3), for two independent random vectors $\mathbf{X} \sim F_j$ and $\mathbf{Y} \sim F_i$ $(1 \leq j, i \leq J)$, $\|\mathbf{X}-\mathbf{Y}\|^2/d \xrightarrow{P} \sigma_j^2 + \sigma_i^2 + \nu_{ji}^2$, where $\nu_{ji}^2 = 0$ for $j = i$.

Proof: Note that using Chebyshev's inequality, for any $\epsilon > 0$, we get

$$P\left(\left|\frac{1}{d}\|\mathbf{X}-\mathbf{Y}\|^2 - E\Big(\frac{1}{d}\|\mathbf{X}-\mathbf{Y}\|^2\Big)\right| \geq \epsilon\right) \leq \frac{1}{\epsilon^2} Var\left(\frac{1}{d}\|\mathbf{X}-\mathbf{Y}\|^2\right).$$

Now, $Var\left(\frac{1}{d}\|\mathbf{X}-\mathbf{Y}\|^2\right) = \frac{1}{d^2}\left[\sum_{s=1}^{d} Var((X_s-Y_s)^2) + \sum_{s=1}^{d}\sum_{t=1,t\neq s}^{d} Cov\Big((X_s-Y_s)^2, (X_t-Y_t)^2\Big)\right]$
Since the measurement variables from all classes have uniformly bounded fourth moments (see (A1)), we have $\sum_{s=1}^{d} Var((X_s-Y_s)^2) = O(d)$. Also, one can show that under assumptions (A1) and (A2), $\sum_{s=1}^{d}\sum_{t=1,t\neq s}^{d} Cov\Big((X_s-Y_s)^2, (X_t-Y_t)^2\Big) = o(d^2)$. So, $Var\left(\frac{1}{d}\|\mathbf{X}-\mathbf{Y}\|^2\right) \to 0$ and hence $\left|\frac{1}{d}\|\mathbf{X}-\mathbf{Y}\|^2 - E\Big(\frac{1}{d}\|\mathbf{X}-\mathbf{Y}\|^2\Big)\right| \xrightarrow{P} 0$ as $d \to \infty$.

Now, $E\Big(\frac{1}{d}\|\mathbf{X}-\mathbf{Y}\|^2\Big) = E\Big(\frac{1}{d}\big\|\big(\mathbf{X}-E(\mathbf{X})\big) - \big(\mathbf{Y}-E(\mathbf{Y})\big) + \big(E(\mathbf{X})-E(\mathbf{Y})\big)\big\|^2\Big) =$ $\frac{1}{d}\text{trace}(\mathbf{\Sigma}_j) + \frac{1}{d}\text{trace}(\mathbf{\Sigma}_i) + \frac{1}{d}\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\|^2 \to \sigma_j^2 + \sigma_i^2 + \nu_{ji}^2$ as $d \to \infty$. So, $\frac{1}{d}\|\mathbf{X}-\mathbf{Y}\|^2 \xrightarrow{P} \sigma_j^2 + \sigma_i^2 + \nu_{ji}^2$. Note that if $\mathbf{X}$ and $\mathbf{Y}$ follow the same distribution (*i.e.* $j = i$), we have $\nu_{ji}^2 = 0$. □

Proof of Theorem 1: (a) From Lemma 1, it is clear that for any test case $\mathbf{Z}$ from the j-th class $(j = 1, 2, \ldots, J)$, $\frac{1}{d}\|\mathbf{Z} - \mathbf{X}_{i\ell}\|^2 \xrightarrow{P} \sigma_j^2 + \sigma_i^2 + \nu_{ji}^2$ for $i = 1, 2, \ldots, J$ and $\ell = 1, 2, \ldots, n_i$. Therefore, for $k < \min\{n_1, n_2, \ldots, n_J\}$, the k-nearest neighbor classifier correctly classifies $\mathbf{Z}$ if $2\sigma_j^2 < \sigma_j^2 + \sigma_i^2 + \nu_{ji}^2$ for all $i \neq j$ or equivalently, $\nu_{ji}^2 > \sigma_j^2 - \sigma_i^2$ for all $i \neq j$. Similarly, for correct classification for a test case from the i-th class, we need $\nu_{ij}^2 > \sigma_i^2 - \sigma_j^2$ for all $j \neq i$. Combining these, we get $\nu_{ji}^2 > |\sigma_j^2 - \sigma_i^2|$ for all $j \neq i$.

If $\nu_{ji}^2 < |\sigma_j^2 - \sigma_i^2|$ for any pair (j, i), we have either $\nu_{ji}^2 < \sigma_j^2 - \sigma_i^2$ or $\nu_{ji}^2 < \sigma_i^2 - \sigma_j^2$. Without loss of generality, let us assume the first one. In that case, for any class-j observation, the distances of its neighbors from the i-th class turn out to be smaller than

those from j-th class with probability tending to 1 as d increases. So, all observations from the j-th class are misclassified with probability tending to 1.

(b) From Lemma 1, it is clear that for any test case $\mathbf{Z}$ from the j-th class, $\frac{1}{d}\rho_j(\mathbf{Z}, \mathbf{X}_{j\ell}) \xrightarrow{P} \sigma_j^2$ for $\ell = 1, 2, \ldots, n_j$ whereas for any $i \neq j$, $\frac{1}{d}\rho_i(\mathbf{Z}, \mathbf{X}_{i\ell}) \xrightarrow{P} \sigma_j^2 + \nu_{ji}^2$ for $\ell = 1, 2, \ldots, n_i$. So, $\mathbf{Z}$ is correctly classified if $\nu_{ji}^2 > 0$ for all $i \neq j$. Repeating this argument for $j = 1, 2, \ldots, J$, we get the result.

(c) Lemma 1 shows that for any test case $\mathbf{Z}$ from the j-th class, $\frac{1}{\sqrt{d}}\rho_j^*(\mathbf{Z}, \mathbf{X}_{j\ell}) \xrightarrow{P} \sigma_j/\sqrt{2}$ for $\ell = 1, 2, \ldots, n_j$ whereas for any $i \neq j$, $\frac{1}{\sqrt{d}}\rho_i^*(\mathbf{Z}, \mathbf{X}_{i\ell}) \xrightarrow{P} \sqrt{\sigma_j^2 + \sigma_i^2 + \nu_{ji}^2} - \sigma_i/\sqrt{2}$ for $\ell = 1, 2, \ldots, n_i$. So, $\mathbf{Z}$ is correctly classified if $\sqrt{\sigma_j^2 + \sigma_i^2 + \nu_{ji}^2} > (\sigma_j + \sigma_i)/\sqrt{2}$ or $\sigma_j^2 + \sigma_i^2 + \nu_{ji}^2 > (\sigma_j + \sigma_i)^2/2$ for all $i \neq j$. Note that $\sigma_j^2 + \sigma_i^2 + \nu_{ji}^2 - (\sigma_j + \sigma_i)^2/2 = \nu_{ji}^2 + (\sigma_j - \sigma_i)^2/2$, which is positive under the given condition. Now, the proof follows by the repetition of the same argument for $j = 1, 2, \ldots, J$. □

Proof of Theorem 2: For the sake of simplicity, let us prove it for $J = 2$. For $J > 2$, it can be proved similarly. From Lemma 1, for any training sample observation $\mathbf{X}_{1i}$ $(i = 1, 2, \ldots, n_1)$ from the first class, as d grows to infinity, we have

$$\left(\frac{1}{\sqrt{d}} \min_{1 \leq \ell (\neq i) \leq n_1} \|\mathbf{X}_{1i} - \mathbf{X}_{1\ell}\|, \frac{1}{\sqrt{d}} \min_{1 \leq \ell \leq n_2} \|\mathbf{X}_{1i} - \mathbf{X}_{2\ell}\| \right) \xrightarrow{P} (\sigma_1\sqrt{2}, \sqrt{\sigma_1^2 + \sigma_2^2 + \nu_{12}^2}) = \mathbf{a}_1, \text{ say.}$$

Similarly, for a training sample observation $\mathbf{X}_{2i}$ $(i = 1, 2, \ldots, n_2)$ from the second class, as d tends to infinity, we have

$$\left(\frac{1}{\sqrt{d}} \min_{1 \leq \ell \leq n_1} \|\mathbf{X}_{2i} - \mathbf{X}_{1\ell}\|, \frac{1}{\sqrt{d}} \min_{1 \leq \ell (\neq i) \leq n_2} \|\mathbf{X}_{2i} - \mathbf{X}_{2\ell}\| \right) \xrightarrow{P} (\sqrt{\sigma_1^2 + \sigma_2^2 + \nu_{12}^2}, \sigma_2\sqrt{2}) = \mathbf{a}_2, \text{ say.}$$

So, if $\mathbf{a}_1 \neq \mathbf{a}_2$, the feature vectors obtained from two sets of training sample observations converge to two distinct points $\mathbf{a}_1$ and $\mathbf{a}_2$, respectively. Now, for any test case $\mathbf{Z}$, using Lemma 1, it can be shown that as d grows to infinity, $(\frac{1}{\sqrt{d}} \min_{1 \leq i \leq n_1} \|\mathbf{Z} - \mathbf{X}_{1i}\|, \frac{1}{\sqrt{d}} \min_{1 \leq i \leq n_2} \|\mathbf{Z} - \mathbf{X}_{2i}\|)$ converges in probability to $\mathbf{a}_1$ and $\mathbf{a}_2$ for $\mathbf{Z} \sim F_1$ and $\mathbf{Z} \sim F_2$, respectively.

Therefore, for any $\mathbf{Z} \sim F_1$ (respectively, F_2), in the $d_1 - d_2$ space, while the scaled versions of its distances from the feature vectors from Class-1 (respectively, Class-2) converge to 0, those from the feature vectors from Class-2 (respectively, Class-1) converge to $\|\mathbf{a}_1 - \mathbf{a}_2\|$ as d tends to infinity. So, it is correctly classified with probability tending to 1. Therefore, for perfect classification by the MDist classifier, we need $\mathbf{a}_1$ and $\mathbf{a}_2$ to be distinct,*i.e.*, $2\sigma_1^2$, $2\sigma_2^2$ and $\sigma_1^2 + \sigma_2^2 + \nu_{12}^2$ cannot be all equal. Note that these three quantities are equal if and only if $\nu_{12}^2 = 0$ and $\sigma_1^2 = \sigma_2^2$, which cannot happen under the assumptions of Theorem 2. □

Lemma 2: Suppose that $\mathbf{X}, \mathbf{X}' \sim F_1$ and $\mathbf{Y}, \mathbf{Y}' \sim F_2$ are four independent d-dimensional random vectors with finite first moments. Then, we have

$$2E\|\mathbf{X} - \mathbf{Y}\|_1 - E\|\mathbf{X} - \mathbf{X}'\|_1 - E\|\mathbf{Y} - \mathbf{Y}'\|_1 \geq 0$$

where the equality holds if and only if F_1 and F_2 have identical one-dimensional marginals.

Proof: First note that

$$2E\|\mathbf{X}_1-\mathbf{Y}_1\|_1-E\|\mathbf{X}_1-\mathbf{X}_2\|_1-E\|\mathbf{Y}_1-\mathbf{Y}_2\|_1=\sum_{q=1}^{d}\Big[2E|X_q-Y_q|-E|X_q-X_q'|-E|Y_q-Y_q'|\Big].$$

Now, from Baringhaus and Franz (2004), we get

$$2E|X_q-Y_q|-E|X_q-X_q'|-E|Y_q-Y_q'|=2\int_{-\infty}^{\infty}\Big(F_1^{(q)}(t)-F_2^{(q)}(t)\Big)^2dt,$$

where $F_1^{(q)}$ and $F_2^{(q)}$ are the distribution functions of X_q and Y_q, respectively. Clearly, it is non-negative, and it takes the value 0 if and only if $F_1^{(q)}=F_2^{(q)}$,*i.e.*, X_q and Y_q have the same distribution. This shows that $2E\|\mathbf{X}_1-\mathbf{Y}_1\|_1-E\|\mathbf{X}_1-\mathbf{X}_2\|_1-E\|\mathbf{Y}_1-\mathbf{Y}_2\|_1\geq 0$, where the equality holds if and only if X_q and Y_q have the same distribution for all $q=1,,2\ldots,d$. □

Proof of Theorem 3: As in the proof of Theorem 2, here also, for the sake of simplicity, we prove the result for $J=2$. For $J>2$, it can be proved similarly.

Consider two random vectors $\mathbf{X}\sim F_j$ and $\mathbf{Y}\sim F_i$ $(1\leq j,i\leq 2)$. Under (A1°) and (A2°), we have $\Big|\frac{1}{d}\|\mathbf{X}-\mathbf{Y}\|_1-E\Big(\frac{1}{d}\|\mathbf{X}-\mathbf{Y}\|_1\Big)\Big|\xrightarrow{P}0$ as $d\to\infty$, and under (A3°), we have $\lim_{d\to\infty}E\Big(\frac{1}{d}\|\mathbf{X}-\mathbf{Y}\|_1\Big)=\tau_{ji}$. Lemma 2 shows that $2\tau_{12}-\tau_{11}-\tau_{22}\geq 0$ and under the assumption $\mathcal{E}_{12}>0$, the equality is ruled out. So, here we have $2\tau_{12}-\tau_{11}-\tau_{22}>0$, which implies that τ_{11},τ_{12} and τ_{22} cannot be equal.

Now note that for any training sample observation $\mathbf{X}_{1i}$ $(i=1,2,\ldots,n_1)$ from the first class, as d grows to infinity,

$$\left(\frac{1}{d}\min_{1\leq\ell(\neq i)\leq n_1}\|\mathbf{X}_{1i}-\mathbf{X}_{1\ell}\|_1,\frac{1}{d}\min_{1\leq\ell\leq n_2}\|\mathbf{X}_{1i}-\mathbf{X}_{2\ell}\|_1\right)\xrightarrow{P}(\tau_{11},\tau_{12})=\mathbf{a}_1^{\circ},\ \text{say.}$$

Similarly, for a training sample observation $\mathbf{X}_{2i}$ $(i=1,2,\ldots,n_2)$ from the second class, as d tends to infinity,

$$\left(\frac{1}{d}\min_{1\leq\ell\leq n_1}\|\mathbf{X}_{2i}-\mathbf{X}_{1\ell}\|_1,\frac{1}{d}\min_{1\leq\ell(\neq i)\leq n_2}\|\mathbf{X}_{2i}-\mathbf{X}_{2\ell}\|_1\right)\xrightarrow{P}(\tau_{12},\tau_{22}),=\mathbf{a}_2^{\circ},\ \text{say..}$$

Since τ_{11}, τ_{12} and τ_{22} are not equal, we have $\mathbf{a}_1^{\circ}\neq\mathbf{a}_2^{\circ}$. So, the feature vectors obtained from two sets of training sample observations converge to two distinct points $\mathbf{a}_1^{\circ}$ and $\mathbf{a}_2^{\circ}$, respectively. For a test case $\mathbf{Z}$, as d grows to infinity, $(\frac{1}{d}\min_{1\leq i\leq n_1}\|\mathbf{Z}-\mathbf{X}_{1i}\|_1,\frac{1}{d}\min_{1\leq i\leq n_2}\|\mathbf{Z}-\mathbf{X}_{2i}\|_1)$ converges in probability to $\mathbf{a}_1^{\circ}$ and $\mathbf{a}_2^{\circ}$ for $\mathbf{Z}\sim F_1$ and $\mathbf{Z}\sim F_2$, respectively.

Therefore, for any $\mathbf{Z}\sim F_1$ (respectively, F_2), in the $d_1^*-d_2^*$ space, while the scaled versions of its distances from the feature vectors from Class-1 (respectively, Class-2) converge to 0, those from the feature vectors from Class-2 (respectively, Class-1) converge to $\|\mathbf{a}_1^{\circ}-\mathbf{a}_2^{\circ}\|$. So, it is correctly classified with probability tending to 1. □

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 535–554
http://www.ssca.org.in/journal

Access Structure Hiding Verifiable Tensor Designs

Anandarup Roy[1], Bimal Kumar Roy[1], Kouichi Sakurai[2] and Suprita Talnikar[3]
[1]*Applied Statistics Unit, Indian Statistical Institute, Kolkata, India*
[2]*Department of Computer Science, Kyushu University, Japan*
[3]*Digital Security, Radboud University, The Netherlands*

Received: 30 April 2024; Revised: 24 September 2024; Accepted: 30 September 2024

Abstract

The field of verifiable secret sharing schemes was introduced by Verheul *et al.* and has evolved over time, including well-known examples by Feldman and Pedersen. Stinson made advancements in combinatorial design-based secret sharing schemes in 2004. Desmedt *et al.* introduced the concept of frameproofness in 2021, while recent research by Sehrawat *et al.* in 2021 focuses on LWE-based access structure hiding verifiable secret sharing with malicious-majority settings. Furthermore, Roy *et al.* combined the concepts of reparable threshold schemes by Stinson *et al.* and frameproofness by Desmedt *et al.* in 2023, to develop extendable tensor designs built from balanced incomplete block designs, and also presented a frameproof version of their design. This paper explores ramp-type verifiable secret sharing schemes, and the application of hidden access structures in such cryptographic protocols. Inspired by Sehrawat *et al.*'s access structure hiding scheme, we develop an ϵ-almost access structure hiding scheme, which is verifiable as well as frameproof. We detail how the concept ϵ-almost hiding is important for incorporating ramp schemes, thus making a fundamental generalisation of this concept.

Key words: Combinatorial secret sharing; Tensor designs; Ramp schemes; Access structure hiding; Verifiability; Frameproofness.

1. Introduction

A verifiable secret sharing scheme Verheul and van Tilborg (1997); Peng (2012); Hofmeister *et al.* (2000); Pedersen (1991); Dehkordi *et al.* (2024) is a cryptographic protocol that allows a dealer to distribute shares of a secret to a group of parties in such a way that (i) the secret remains confidential and cannot be determined by any unauthorized collection of parties, (ii) the secret can be reconstructed correctly by the authorized collection of parties when they combine their shares, (iii) there is a mechanism for parties to verify the correctness of the shares they receive and for the reconstruction process, and (iv) the scheme can withstand malicious behavior from both the dealer and the parties, thus ensuring the security and integrity of the secret sharing process.

Corresponding Author: Anandarup Roy
Email: ananda1101@gmail.com

Repairable Threshold Schemes (RTSs) Stinson and Wei (2018); Laing and Stinson (2018) are cryptographic schemes that allow for the reconstruction of lost or corrupted shares in a threshold scheme without the need for the dealer who initially set up the scheme to be involved in the repair process. In RTSs, a subset of authorized parties can collaboratively reconstruct the lost share, ensuring the integrity and availability of the shared secret. Roy and Roy (2023) explores the concept of repairable ramp schemes for secret sharing and various applications, including cloud storage, sensor-based IoTs, and electronic identification cards. It proposes a protocol for extending schemes that allow for the retrieval of shares through collaborative efforts in case of loss or corruption, thereby enhancing data security and privacy. Roy and Roy (2023) also introduces the concept of tensor products of balanced incomplete block designs (BIBDs), which help securely combine individual secrets from various systems, enabling multi-level or multi-system secret sharing schemes in a robust and efficient manner. Desmedt *et al.* (2021) introduced the concept of frameproofness of secret sharing schemes, which ensures the security and integrity of shared secrets and analyses the resistance of a scheme to attempts of falsely implicating (framing) a (set of) player(s) in the unauthorized disclosure of secret information. Roy and Roy (2023) establishes a theoretical framework for frameproofness within its extension protocol, and ensures that its extended scheme upholds the principles of frameproofness by leveraging concepts from combinatorial design theory.

Sehrawat *et al.* (2021) provides a detailed discussion on how secret sharing can be achieved with hidden access structures, allowing for a wide range of access policies to be enforced in the secret sharing process. The scheme is designed to support verifiability even when a majority of the parties are malicious, and its verification procedure does not incur any communication overhead, making it "free" in terms of computational resources. The scheme provides a maximum share size formula that allows for efficient sharing of secrets while maintaining security guarantees. The share size is optimized to balance security and efficiency considerations. It also includes mechanisms to detect and identify malicious behavior during the secret sharing process.

1.1. Our contribution

This motivation clearly begs the question of verifiability of secret sharing schemes constructed as the extended tensor designs from Roy and Roy (2023), and how frameproofness applies to the resulting composition. Our approach results in a fundamental generalisation of the novel access structure hiding technique introduced by Sehrawat *et al.* (2021) to incorporate ramp schemes, thus allowing for a wider range of secret sharing schemes to use this technique. We provide detailed explanations for how our generalised ϵ-almost access structure hiding ramp-type tensor design satisfies all properties of an almost-verifiable secret sharing scheme, as well as almost fully hides its access structure, and has a frameproof version that does not lose any original information.

1.2. Organisation of the paper

Beginning with the introduction of various important types of secret sharing schemes such as VSS schemes, RTSs, BIBDs and access structure hiding schemes in Section 1, we define various notations, definitions and other preliminaries in Section 2. We introduce our modified concept of ϵ-almost access structure hiding ramp-type tensor designs in section 3,

where we provide a background of the existing theory of extending tensor designs by Roy *et al.* Roy and Roy (2023), as well as demonstrate various secret sharing properties (such as correctness, ϵ-correctness and computational secrecy for their tensor design schemes. We also recall the concept of frameproof tensor designs through an example and show that it is also applicable to our scheme, and detail an algorithm for access structure token generation according to our requirements. In section 4, we state the mains results of this paper in the form of Theorems 3, 4, 5 and 6. Sections 5 and 6 present detailed proofs of these theorems. In Section 7, we enumerate a few applications of our results in the real world, and then conclude in Section 8.

2. Preliminaries

Given a collection $\mathbf{P} = \{P_1, \ldots, P_\ell\}$ of (say) players in a secret sharing scheme, we denote the power set of $\mathbf{P}$, *i.e.* the set of all subsets of $\mathbf{P}$, by $2^{\mathbf{P}}$. The closure of a subset $\mathbf{A} \in 2^{\mathbf{P}}$ is the set $cl(\mathbf{A}) \coloneqq \{\mathbf{C} : \mathbf{C}^* \subseteq \mathbf{C} \subseteq \mathbf{P} \text{ for some } \mathbf{C}^* \in \mathbf{A}\}$. Given a security parameter ω, a function $\delta(\omega)$ is called *negligible* if for all $c > 0$, there exists an ω_0 such that $(\omega) < 1/\omega_c$ for all $\omega > \omega_0$. Given a probability distribution X, the notation $\Pr[t \leftarrow X]$ denotes a sampling of t by the distribution X.

Definition 1: Let $X = \{X_\lambda\}_{\lambda\in\mathbb{N}}$ and $Y = \{Y_\lambda\}_{\lambda\in\mathbb{N}}$ be collections of probability distributions (or *ensembles*) X_λ and Y_λ over $\{0,1\}^{\kappa(\lambda)}$ for some polynomial $\kappa(\lambda)$. These two ensembles are *polynomially or computationally indistinguishable* if for every (probabilistic) polynomial-time algorithm $\mathbf{D}$, for all $\lambda \in \mathbb{N}$, and a negligible function δ,

$$|\Pr[t \leftarrow X_\lambda : \mathbf{D}(t) = 1] - \Pr[t \leftarrow Y_\lambda : \mathbf{D}(t) = 1]| \leq \delta(\lambda).$$

Assume that there exist positive integers θ, Θ and ℓ, where $\theta < \Theta \leq \ell$. A (θ, Θ, ℓ)-*ramp scheme* Paterson and Stinson (2013) involves a dealer selecting a secret and then distributing a share to each of ℓ players in a manner that fulfills the following criteria:

Reconstruction: Any subset of Θ players has the ability to collectively determine the secret using the shares they possess.

Secrecy: No subset of θ players is able to deduce any details regarding the secret.

The terms θ and Θ are referred to as the lower and upper thresholds of the scheme, respectively. For the sake of convenience, we shall refer to collections of players $\mathbf{C} \in 2^{\mathbf{P}}$ such that $\theta < |\mathbf{C}| < \Theta$ by the term *ramp collection*. In the event where $\Theta = \theta + 1$, the scheme is recognized as a (Θ, ℓ)-threshold scheme. In the context of such a Θ-threshold scheme, the problem of *share repairability* pertains to the identification of a secure protocol for restoring the lost share of a specific player ($P_{\mathfrak{i}} \in \mathbf{P}$). This process involves a certain subset of d players (excluding $P_{\mathfrak{i}} \in \mathbf{P}$) engaging in message exchange amongst themselves and with $P_{\mathfrak{i}} \in \mathbf{P}$, with the objective of successfully repairing its share. The smallest integer d required to accomplish this task is known as the *repairing degree* of the scheme. If an honest-but-curious coalition of no more than $\Theta - 1$ players of a (Θ, ℓ)-threshold scheme combines all the information it holds (this includes their shares, as well as all messages that they send or receive during the protocol) and still obtains no information about the secret, then we say that it is a (Θ, ℓ, d)-*repairable threshold scheme*, or a (Θ, ℓ, d)-RTS.

Definition 2: Suppose $2 \leq k < v$. A (b, v, k, r, λ)-balanced incomplete block design or a (b, v, k, r, λ)-BIBD is a design $(X, \mathcal{B})$ such that:

1. $|X| = v$;
2. each block $B \in \mathcal{B}$ contains exactly k points;
3. every pair of distinct points from X is contained in exactly λ blocks.

Observe that if each point occurs in exactly r blocks, then the parameters b, v, k, r, λ of a BIBD satisfy the following relations Stinson (2004):

(i) $bk = vr$;

(ii) $\lambda(v-1) = r(k-1)$;

(iii) $b \geq v$ (and hence $r > k$).

We sometimes refer to a (b, v, k, r, λ)-BIBD as simply a (v, k, λ)-BIBD.

Definition 3: Let $\mathbf{P} = \{P_1, \ldots, P_\ell\}$ be a set of parties or players. A collection $\Gamma \subseteq 2^{\mathbf{P}}$ is monotone if $\mathbf{A} \in \Gamma$ and $\mathbf{A} \subseteq \mathbf{B}$ imply that $\mathbf{B} \in \Gamma$. An *access structure* $\Gamma \subseteq 2^{\mathbf{P}}$ is a monotone collection of non-empty subsets of $\mathbf{P}$. Sets in γ are called *authorized*, and sets not in Γ are called *unauthorized*.

Definition 4: For an access structure Γ, $\Gamma_0 = \{\mathbf{A} \in \Gamma : \mathbf{B} \not\subset \mathbf{A} \text{ for all } \mathbf{B} \in \Gamma \setminus \mathbf{A}\}$ is the family of *minimal authorized subsets* in Γ.

Definition 5: A *computational secret sharing scheme* with respect to an access structure Γ, security parameter ω, a set of ℓ polynomial-time parties or players $\mathbf{P} = \{P_1, \ldots, P_\ell\}$, and a set of secrets $\mathbf{K}$, consists of a pair of polynomial-time algorithms $(\mathsf{Share}, \mathsf{Recon})$, where:

- Share is a randomized algorithm that gets a secret $k \in \mathbf{K}$ and access structure Γ as inputs, and outputs ℓ shares, $\{s_1^{(k)}, \ldots, s_\ell^{(k)}\}$, of k, and
- Recon is a deterministic algorithm that gets as input the shares of a subset $\mathbf{A} \subseteq \mathbf{P}$, denoted by $\left\{s_i^{(k)}\right\}_{i \in \mathbf{A}}$, and outputs a string in $\mathbf{K}$,

such that the following two requirements are satisfied:

1. *(Perfect Correctness)* for all secrets $k \in \mathbf{K}$ and every authorized collection $\mathbf{A} \in \Gamma$, it holds that: $\Pr\left[\mathsf{Recon}\left(\left\{s_i^{(k)}\right\}_{i \in \mathbf{A}}, \mathbf{A}\right) = k\right] = 1$,
2. *(Computational Secrecy)* for every unauthorized collection $\mathbf{B} \notin \Gamma$ and all distinct secrets $k_1, k_2 \in \mathbf{K}$, it holds that the distributions $\left\{s_i^{(k_1)}\right\}_{i \in \mathbf{A}}$ and $\left\{s_i^{(k_2)}\right\}_{i \in \mathbf{A}} \in \mathbf{B}$ are computationally indistinguishable (with respect to ω).

Traditionally, secret sharing relies on honest participants. However, a *verifiable secret sharing (VSS) scheme* is also required to withstand active attacks, specifically:

- a dealer sending inconsistent or incorrect shares to some of the participants during the distribution protocol, and
- participants submitting incorrect shares during the reconstruction protocol.

VSS schemes were first introduced by Verheul and van Tilborg (1997). Clearly, Shamir's threshold scheme is not a VSS scheme, since it does not exclude either of these attacks. Well-known examples of VSS schemes are Feldman's VSS scheme Hofmeister *et al.* (2000) and Pedersen's VSS scheme Pedersen (1991).

The access structure hiding verifiable (computational) secret sharing scheme of Sehrawat *et al.* (2021) defined below guarantees a relaxed definition of verifiability of shares of authorised collections of players even when a majority of the parties are malicious. Their scheme supports all monotone access structures, and its security — in particular, verifiability — relies on the hardness of the LWE problem.

Definition 6: An *access structure hiding verifiable (computational) secret sharing scheme* with respect to an access structure Γ, security parameter ω, a set of ℓ polynomial-time parties or players $\mathbf{P} = \{P_1, \ldots, P_\ell\}$, and a set of secrets $\mathbf{K}$, consists of two sets of polynomial-time algorithms, $(\mathsf{HsGen}, \mathsf{HsVer})$ and $(\mathsf{VerShr}, \mathsf{Recon}, \mathsf{Ver})$, which are defined as follows:

- VerShr is a randomized algorithm that gets a secret $k \in \mathbf{K}$ and access structure Γ as inputs, and outputs ℓ shares, $\{s_1^{(k)}, \ldots, s_\ell^{(k)}\}$, of k,
- Recon is a deterministic algorithm that gets as input the shares of a subset $\mathbf{A} \subseteq \mathbf{P}$, denoted by $\left\{s_i^{(k)}\right\}_{i \in \mathbf{A}}$, and outputs a string in $\mathbf{K}$, and
- Ver is a deterministic Boolean algorithm that gets $\left\{s_i^{(k)}\right\}_{i \in \mathbf{A}}$ and a secret k' in $\mathbf{K}$ as inputs, and outputs $b \in \{0, 1\}$,

such that the following three requirements are satisfied:

1. *(Perfect Correctness)* for all secrets $k \in \mathbf{K}$ and every authorized collection $\mathbf{A} \in \Gamma$, it holds that: $\Pr\left[\mathsf{Recon}\left(\left\{s_i^{(k)}\right\}_{i \in \mathbf{A}}, \mathbf{A}\right) = k\right] = 1$.
2. *(Computational Secrecy)* for every unauthorized collection $\mathbf{B} \notin \Gamma$ and all distinct secrets $k_1, k_2 \in \mathbf{K}$, it holds that the distributions $\left\{s_i^{(k_1)}\right\}_{i \in \mathbf{A}}$ and $\left\{s_i^{(k_2)}\right\}_{i \in \mathbf{A}} \in \mathbf{B}$ are computationally indistinguishable (with respect to ω).
3. *(Computational Verifiability)* Every authorized collection $\mathbf{A} \in \Gamma$ can use Ver to verify whether its set of shares $\left\{s_i^{(k)}\right\}_{i \in \mathbf{A}}$ is consistent with a given secret $k \in \mathcal{K}$. Formally, for a negligible function δ, it holds that:

- If all shares $s_i^{(k)} \in \left\{s_i^{(k)}\right\}_{i\in\mathbf{A}}$ are consistent with the secret k, then

$$\Pr\left[\mathsf{Ver}\left(k, \left\{s_i^{(k)}\right\}_{i\in\mathbf{A}}\right) = 1\right] = 1 - \delta(\omega)$$

- If any share $s_i^{(k)} \in \left\{s_i^{(k)}\right\}_{i\in\mathbf{A}}$ is inconsistent with the secret k, then

$$\Pr\left[\mathsf{Ver}\left(k, \left\{s_i^{(k)}\right\}_{i\in\mathbf{A}}\right) = 0\right] = 1 - \delta(\omega).$$

- HsGen is a randomized algorithm that gets $\mathbf{P}$ and Γ as inputs, and outputs ℓ *access structure tokens* $\left\{\mho_1^{(\Gamma)}, \ldots, \mho_\ell^{(\Gamma)}\right\}$, and
- HsVer is a deterministic algorithm that gets as input the access structure tokens of a subset $\mathbf{A} \subseteq \mathbf{P}$ $\left(\text{denoted } \left\{\mho_i^{(\Gamma)}\right\}_{i\in\mathbf{A}}\right)$, and outputs $b \in \{0, 1\}$,

such that the following three requirements are satisfied:

1. *(Perfect completeness)* Every authorized collection of parties $\mathbf{A} \in \Gamma$ can identify itself as a member of the access structure Γ, *i.e.* $\Pr\left[\mathsf{HsVer}\left(\left\{\mho_i^{(\Gamma)}\right\}_{i\in\mathbf{A}}\right) = 1\right] = 1$.
2. *(Perfect soundness)* Every unauthorized collection of parties $\mathbf{B} \notin \Gamma$ can identify itself to be outside of the access structure Γ, *i.e.* $\Pr\left[\mathsf{HsVer}\left(\left\{\mho_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}\right) = 0\right] = 1$.
3. *(Statistical hiding)* For all access structures $\Gamma, \Gamma' \subseteq 2^{\mathbf{P}}$ where $\Gamma \neq \Gamma'$, and for all unauthorised collections $\mathbf{B} \notin \Gamma, \Gamma'$,

$$\left|\Pr\left[\Gamma \mid \left\{\mho_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{B}}\right] - \Pr\left[\Gamma' \mid \left\{\mho_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{B}}\right]\right| = 2^{-\omega}.$$

3. ϵ-Almost access structure hiding ramp-type tensor designs

We incorporate the novel access structure hiding technique of Sehrawat *et al.* (2021) in the tensor design obtained by extending BIBDs as introduced in the work of Roy and Roy (2023). Since the scheme of Roy and Roy (2023) is a ramp scheme for both variants (non-frameproof and frameproof, defined below) of the tensor design, we introduce the new concept of an ϵ-almost access structure hiding ramp scheme.

Definition 7: Consider a (θ, Θ, ℓ)-ramp scheme, so that its access structure Γ is characterised by the ramp bounds (θ, Θ). For $\epsilon = (\epsilon_{\text{Corr}}, \epsilon_1, \epsilon_2, \epsilon_3)$, an *$\epsilon$-almost access structure hiding (θ, Θ, ℓ)-ramp scheme* with respect to a security parameter ω, a set of ℓ polynomial-time parties or players $\mathbf{P} = \{P_1, \ldots, P_\ell\}$, and a set of secrets $\mathbf{K}$, consists of two sets of polynomial-time algorithms, $(\mathsf{HsGen}, \mathsf{HsVer})$ and $(\mathsf{VerShr}, \mathsf{Recon}, \mathsf{Ver})$, which are defined as follows:

- VerShr is a randomized algorithm that gets a secret $k \in \mathbf{K}$ and the bounds θ, Θ as inputs, and outputs ℓ shares, $\{s_1^{(k)}, \ldots, s_\ell^{(k)}\}$, of k,
- Recon is a deterministic algorithm that gets as input the shares of a subset $\mathbf{A} \subseteq \mathbf{P}$, denoted by $\left\{s_i^{(k)}\right\}_{i\in\mathbf{A}}$, and outputs a string in $\mathbf{K}$, and
- Ver is a deterministic Boolean algorithm that gets $\left\{s_i^{(k)}\right\}_{i\in\mathbf{A}}$ and a secret $k' \in \mathbf{K}$ as inputs, and outputs $b \in \{0, 1\}$,

such that the following four requirements are satisfied:

1. *(Perfect Correctness)* for all secrets $k \in \mathbf{K}$ and every authorized collection $\mathbf{A}$ such that $|\mathbf{A}| \geq \Theta$, it holds that: $\Pr\left[\mathsf{Recon}\left(\left\{s_i^{(k)}\right\}_{i\in\mathbf{A}}, \mathbf{A}\right) = k\right] = 1$.
2. *(ϵ_{corr}-Correctness)* for all secrets $k \in \mathbf{K}$ and every ramp collection $\mathbf{C}$ such that $\theta < |\mathbf{C}| < \Theta$, there exists $\epsilon_{\text{corr}} > 0$ such that: $\Pr\left[\mathsf{Recon}\left(\left\{s_i^{(k)}\right\}_{i\in\mathbf{A}}, \mathbf{A}\right) = k\right] = \epsilon_{\text{corr}}$.
3. *(Computational Secrecy)* for every unauthorized collection $\mathbf{B}$ with $|\mathbf{B}| \leq \theta$ and all distinct secrets $k_1, k_2 \in \mathbf{K}$, it holds that the distributions $\left\{s_i^{(k_1)}\right\}_{i\in\mathbf{A}}$ and $\left\{s_i^{(k_2)}\right\}_{i\in\mathbf{A}} \in \mathbf{B}$ are computationally indistinguishable (with respect to ω).
4. *(Computational Verifiability)* Every authorized collection $\mathbf{A}$ such that $|\mathbf{A}| \geq \Theta$ can use Ver to verify whether its set of shares $\left\{s_i^{(k)}\right\}_{i\in\mathbf{A}}$ is consistent with a given secret $k \in \mathcal{K}$. Formally, for a negligible function δ, it holds that:
   - If all shares $s_i^{(k)} \in \left\{s_i^{(k)}\right\}_{i\in\mathbf{A}}$ are consistent with the secret k, then
   $$\Pr\left[\mathsf{Ver}\left(k, \left\{s_i^{(k)}\right\}_{i\in\mathbf{A}}\right) = 1\right] = 1 - \delta(\omega)$$
   - If any share $s_i^{(k)} \in \left\{s_i^{(k)}\right\}_{i\in\mathbf{A}}$ is inconsistent with the secret k, then
   $$\Pr\left[\mathsf{Ver}\left(k, \left\{s_i^{(k)}\right\}_{i\in\mathbf{A}}\right) = 0\right] = 1 - \delta(\omega).$$

- HsGen is a randomized algorithm that gets $\mathbf{P}$, θ and Θ as inputs, and outputs ℓ *access structure tokens* $\left\{\mathfrak{V}_1^{(\Gamma)}, \ldots, \mathfrak{V}_\ell^{(\Gamma)}\right\}$, and
- HsVer is a deterministic algorithm that gets as input the access structure tokens of a subset $\mathbf{A} \subseteq \mathbf{P}$ $\left(\text{denoted } \left\{\mathfrak{V}_i^{(\Gamma)}\right\}_{i\in\mathbf{A}}\right)$, and outputs $b \in \{0, 1\}$,

such that the following six requirements are satisfied:

1. *(Perfect completeness)* Every authorized collection of parties $\mathbf{A}$ such that $|\mathbf{A}| \geq \Theta$ can identify itself as a member of the access structure Γ, *i.e.* $\Pr\left[\mathsf{HsVer}\left(\left\{\mathfrak{V}_i^{(\Gamma)}\right\}_{i\in\mathbf{A}}\right) = 1\right] = 1$.

2. *(ϵ_1-Completeness)* Every ramp collection of parties $\mathbf{C}$ (where $\theta < |\mathbf{C}| < \Theta$) can almost always identify itself as a member of the access structure Γ),
i.e. $\Pr\left[\mathsf{HsVer}\left(\left\{\mathfrak{V}_i^{(\Gamma)}\right\}_{i\in\mathbf{A}}\right) = 1\right] = 1 - \epsilon_1$.

3. *(Perfect soundness)* Every unauthorized collection of parties $\mathbf{B}$ with $|\mathbf{B}| \leq \theta$ can identify itself to be outside of the access structure Γ, *i.e.* $\Pr\left[\mathsf{HsVer}\left(\left\{\mathfrak{V}_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}\right) = 0\right] = 1$.

4. *(ϵ_2-Soundness)* Every ramp collection of parties $\mathbf{C}$ (where $\theta < |\mathbf{C}| < \Theta$) can almost always identify itself to be outside of the access structure Γ, *i.e.*
$\Pr\left[\mathsf{HsVer}\left(\left\{\mathfrak{V}_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}\right) = 0\right] = 1 - \epsilon_2$.

5. *(Statistical hiding)* For all ramp access structures $\Gamma \neq \Gamma'$ and for all unauthorised collections $\mathbf{B}$ with $|\mathbf{B}| \leq \theta, \theta'$,

$$\left|\Pr\left[\Gamma \mid \left\{\mathfrak{V}_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{B}}\right] - \Pr\left[\Gamma' \mid \left\{\mathfrak{V}_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{B}}\right]\right| = 2^{-\omega}.$$

6. *(ϵ_3-Statistical Hiding)* For all ramp access structures $\Gamma, \Gamma' \subseteq 2^{\mathbf{P}}$ where $\Gamma \neq \Gamma'$, and for all ramp collections $\mathbf{C}$ such that $\theta < |\mathbf{C}| < \Theta$,

$$\left|\Pr\left[\Gamma \mid \left\{\mathfrak{V}_i^{(\Gamma)}\right\}_{i\in\mathbf{C}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{C}}\right] - \Pr\left[\Gamma' \mid \left\{\mathfrak{V}_i^{(\Gamma)}\right\}_{i\in\mathbf{C}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{C}}\right]\right| \leq \epsilon_3(\omega).$$

3.1. Tensor design

Let $\mathcal{A}$ and $\mathcal{B}$ be the share matrices generated by ramp schemes with respectively b_1 and b_2 blocks having shares of sizes k_1 and k_2. Suppose $\mathcal{A}$ and $\mathcal{B}$ also denote the $b_1 \times k_1$ and $b_2 \times k_2$ matrices corresponding to the two schemes. The Krönecker product of $\mathcal{A} \otimes \mathcal{B}$ is therefore

$$M = \begin{pmatrix} \mathfrak{a}_{11}\mathcal{B} & \mathfrak{a}_{12}\mathcal{B} & \dots & \mathfrak{a}_{1k_1}\mathcal{B} \\ \mathfrak{a}_{21}\mathcal{B} & \mathfrak{a}_{22}\mathcal{B} & \dots & \mathfrak{a}_{2k_1}\mathcal{B} \\ \vdots & & & \\ \mathfrak{a}_{b_1 1}\mathcal{B} & \mathfrak{a}_{b_1 2}\mathcal{B} & \dots & \mathfrak{a}_{b_1 k_1}\mathcal{B} \end{pmatrix}. \tag{1}$$

If the share matrix $\mathcal{A}$ is defined over the field $\mathbb{F}_{p_1}$ and $\mathcal{B}$ over the field $\mathbb{F}_{p_2}$ for some primes p_1 and p_2, then we define the scalar multiplication as the simple integer multiplication:

$$\begin{aligned} \mathbb{F}_{p_1} \times \mathbb{F}_{p_2} &\to \mathbb{Z} \\ \text{such that } (x_1, x_2) &\mapsto x_1 \cdot x_2. \end{aligned}$$

The reason behind taking such a multiplication is that the product elements are not distinguishable from integers. Therefore, M is a matrix over the integer ring $\mathbb{Z}$.

Theorem 1 (Reconstruction from Tensor Designs, Roy and Roy (2023))**:** Consider a $(v_1, k_1, \lambda_1, b_1, r_1)$-BIBD $\mathcal{A}$ and a $(v_2, k_2, \lambda_2, b_2, r_2)$-BIBD $\mathcal{B}$.

1. The matrix $\mathcal{A} \otimes \mathcal{B}_d$ produces a tensor design (over the integer ring $\mathbb{Z}$) for a (public) integer d such that there are no multiplicative collisions of the type $x_i(y_j + d) = x_k(y_l + d)$ for $(i, j) \neq (k, l)$.
2. - If $\gcd(x_1, x_2, \ldots, x_{v_1}) = 1$;
   - if $\gcd(y_1, y_2, \ldots, y_{v_2}) = 1$;

   then $\mathcal{A}$ and $\mathcal{B}$ can be reproduced from a collection of players in the new scheme $\mathcal{A} \otimes \mathcal{B}_d$, hence enabling share repair and secret reconstruction.

For the purpose of real-world implementation, we consider a prime power q, which is computed from p_1, p_2 and d such that it is sufficiently greater than all the elements in $\mathcal{A} \otimes \mathcal{B}_d$.

3.2. Secret sharing properties of $\mathcal{A} \otimes \mathcal{B}_d$

Since $\mathcal{A} \otimes \mathcal{B}_d$ is a (θ, Θ, ℓ)-ramp scheme, it clearly satisfies the following properties of Definition 7:

Perfect Correctness: From Lemmas 4–9 of Roy and Roy (2023), it is clear that $\mathcal{A} \otimes \mathcal{B}_d$ is a (θ, Θ, ℓ)-ramp scheme, for $\theta = (\tau_1 - 1)(\tau_2 - 1) + 1$ and $\Theta = \min\{(\tau_1 - 1)b_2 + 1, (\tau_2 - 1)b_1 + 1\}$. Hence, any $\mathbf{A}$ with $|\mathbf{A}| \geq \Theta$ can reconstruct the secret with probability 1, *i.e.* $\Pr\left[\mathsf{Recon}\left(\left\{s_i^{(k)}\right\}_{i \in \mathbf{A}}, \mathbf{A}\right) = k\right] = 1$.

ϵ_{corr}-Correctness: Suppose $\theta < |\mathbf{C}| < \Theta$ and $\mathbf{C}$ gets partial information about $\mathcal{A} \otimes \mathcal{B}_d$, *i.e.* it can reconstruct exactly one of $\mathcal{A}$ and $\mathcal{B}_d$, say $\mathcal{A}$ (respectively $\mathcal{B}_d$). Then it must guess the secret of the other factor, *i.e.* $\mathcal{B}_d$ (respectively $\mathcal{A}$) uniformly at random at best, ie. with probability $\frac{1}{p_2}$ (respectively $\frac{1}{p_1}$). Therefore, for all secrets $k \in \mathbf{K}$ and such a ramp collection $\mathbf{C}$, we denote $\epsilon_{\text{corr}} := \max\left\{\frac{1}{p_1}, \frac{1}{p_2}\right\}$. Therefore, $\Pr\left[\mathsf{Recon}\left(\left\{s_i^{(k)}\right\}_{i \in \mathbf{A}}, \mathbf{A}\right) = k\right] \leq \epsilon_{\text{corr}}$.

Computational Secrecy: Consider an unauthorised collection $\mathbf{B}$, with $|\mathbf{B}| \leq \theta$ or $\theta < |\mathbf{B}| < \Theta$. Thus, $\mathbf{B}$ gets no information about the secret, which means it must guess (at best) uniformly at random, the secrets of both the factors $\mathcal{A}$ and $\mathcal{B}_d$ of $\mathcal{A} \otimes \mathcal{B}_d$. Hence, given the access structure Γ, it holds for every unauthorised collection $\mathbf{B} \notin \Gamma$ and every pair of different secrets $k_1 \neq k_2$ in $\mathcal{K}$ that the distributions $\left\{s_i^{(k_1)}\right\}_{i \in \mathbf{B}}$ and $\left\{s_i^{(k_2)}\right\}_{i \in \mathbf{B}}$ are computationally indistinguishable w.r.t. the parameter $\delta := \frac{1}{p_1 p_2}$, according to Definition 1.

3.3. Frameproofness

The concept of *framing* a player (or a collection of players), and subsequently the property of frameproofness of a secret sharing scheme was introduced by Desmedt *et al.* in

Desmedt *et al.* (2021). Sehrawat *et al.* (2021) proposes an access structure hiding verifiable secret sharing scheme, where it establishes indistinguishability of authorisation of any collection of players by use of *access structure tokens*. For the collection $\mathbf{P}$ of all players in the scheme, they make the following claim regarding its frameproofness:

"...the share of each party P_i is sealed as a PRIM-LWE instance such that the lattice basis, A_i, used to generate it is known only to P_i. Since A_i is required to generate P_i's share, it is infeasible for any coalition of polynomial-time parties $\mathbf{A} \subset \mathbf{P}$ to compute the share of $P_i \in \mathbf{P} \setminus \mathbf{A}$ without solving the LWE problem."

Furthermore, Roy and Roy (2023) shows that for the tensor design in Equation (1), only two players — one from the $r_1 - 1$ players possessing $\mathfrak{a}_{11}\mathfrak{b}_{11}$ and one from the $b_2 - 1$ players possessing $\frac{\mathfrak{a}_{12}}{\mathfrak{a}_{11}}, \frac{\mathfrak{a}_{13}}{\mathfrak{a}_{11}}, \ldots$ — can reconstruct the entire share of player P_1, and hence, frame this player. They address this problem by reducing the repetitive nature of shares of the participants — by decreasing the size of each share, while retaining all the information that a player had in the previous construction. In fact, the secret reconstruction for the modified scheme is then shown to require at $\tau_1 + \tau_2$ players. Additionally, Theorem 2 below ensures that $\mathcal{F}(\mathcal{A}, \mathcal{B})$ is simply a Θ-threshold scheme for $\Theta = \tau_1 + \tau_2$ (and not a ramp scheme like (AoB).

Example

Consider an example, where matrix $\mathcal{A}$ represents a 2-$(4, 3, 2)$-BIBD and $\mathcal{B}$ a 2-$(5, 4, 3)$-BIBD over the points $\{1, 2, 3, 4\}$ and $\{1, 2, 3, 4, 5\}$, respectively (note that $r_1 = 3, r_2 = 4$), and $d = 21$. The Krönecker product tensor design obtained from these two matrices is represented by the matrix $\mathcal{A} \otimes \mathcal{B}_d$ as defined in Roy and Roy (2023):

$$\begin{pmatrix}
22 & 23 & 24 & 25 & 44 & 46 & 48 & 50 & 66 & 69 & 72 & 75 \\
23 & 24 & 25 & 26 & 46 & 48 & 50 & 52 & 69 & 72 & 75 & 78 \\
24 & 25 & 26 & 22 & 48 & 50 & 52 & 44 & 72 & 75 & 78 & 66 \\
25 & 26 & 22 & 23 & 50 & 52 & 44 & 46 & 75 & 78 & 66 & 69 \\
26 & 22 & 23 & 24 & 52 & 44 & 46 & 48 & 78 & 66 & 69 & 72 \\
44 & 46 & 48 & 50 & 66 & 69 & 72 & 75 & 88 & 92 & 96 & 100 \\
46 & 48 & 50 & 52 & 69 & 72 & 75 & 78 & 92 & 96 & 100 & 104 \\
48 & 50 & 52 & 44 & 72 & 75 & 78 & 66 & 96 & 100 & 104 & 88 \\
50 & 52 & 44 & 46 & 75 & 78 & 66 & 69 & 100 & 104 & 88 & 92 \\
52 & 44 & 46 & 48 & 78 & 66 & 69 & 72 & 104 & 88 & 92 & 96 \\
66 & 69 & 72 & 75 & 88 & 92 & 96 & 100 & 22 & 23 & 24 & 25 \\
69 & 72 & 75 & 78 & 92 & 96 & 100 & 104 & 23 & 24 & 25 & 26 \\
72 & 75 & 78 & 66 & 96 & 100 & 104 & 88 & 24 & 25 & 26 & 22 \\
75 & 78 & 66 & 69 & 100 & 104 & 88 & 92 & 25 & 26 & 22 & 23 \\
78 & 66 & 69 & 72 & 104 & 88 & 92 & 96 & 26 & 22 & 23 & 24 \\
88 & 92 & 96 & 100 & 22 & 23 & 24 & 25 & 44 & 46 & 48 & 50 \\
92 & 96 & 100 & 104 & 23 & 24 & 25 & 26 & 46 & 48 & 50 & 52 \\
96 & 100 & 104 & 88 & 25 & 26 & 22 & 23 & 48 & 50 & 52 & 44 \\
100 & 104 & 88 & 92 & 25 & 26 & 22 & 23 & 50 & 52 & 44 & 46 \\
104 & 88 & 92 & 96 & 26 & 22 & 23 & 24 & 52 & 44 & 46 & 48
\end{pmatrix}$$

On applying certain permutations on each block of $\mathcal{A} \otimes \mathcal{B}_d$ (and removing zeroes), we obtain a scheme that extends the BIBDs $\mathcal{A}$ and $\mathcal{B}$, where it is no longer possible to reconstruct the secret from just two players. The full algorithm may be found in Roy and Roy (2023). The shares of players in this version, which we shall denote here by $\mathcal{F}(\mathcal{A}, \mathcal{B})$, are:

$$\begin{pmatrix}
22 & 50 & 72 \\
23 & 46 & 78 \\
25 & 48 & 72 \\
22 & 52 & 75 \\
24 & 46 & 66 \\
50 & 72 & 88 \\
46 & 78 & 92 \\
48 & 72 & 100 \\
52 & 75 & 88 \\
46 & 66 & 96 \\
72 & 88 & 25 \\
78 & 92 & 23 \\
72 & 100 & 24 \\
75 & 104 & 26 \\
66 & 92 & 23 \\
88 & 25 & 48 \\
92 & 23 & 52 \\
100 & 25 & 48 \\
88 & 26 & 50 \\
96 & 23 & 44
\end{pmatrix}$$

3.4. Secret sharing properties of $\mathcal{F}(\mathcal{A}, \mathcal{B})$

From Theorem 2 stated below, it is clear that $\mathcal{F}(\mathcal{A}, \mathcal{B})$ is a (θ, Θ, ℓ)-ramp scheme, for $\theta = \tau_1 + \tau_2$ and $\Theta = \min\{(\tau_1 - 1)b_2 + 1, (\tau_2 - 1)b_1 + 1\}$. Therefore, it clearly satisfies the following properties of perfect correctness for all authorised collections of players of size greater than Θ, ϵ_{corr}-correctness for ramp collections of players that are authorised, and computational secrecy for all unauthorised collections of players (irrespective of size), from Definition 7.

A complete explanation is very similar to that for $\mathcal{A} \otimes \mathcal{B}_d$ given in Section 3.2.

3.5. Graphical representation

Definition 8: A bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is said to *induce* a tensor design $\mathcal{B}$ if

- the vertex set $\mathcal{V} = \mathbf{P} \sqcup \mathbf{V}$ the disjoint union of the set of players $\mathbf{P} = \{P_1, \ldots, P_b\}$ and the set of points $\mathbf{V} = \{x_1, \ldots, x_v\}$ of $\mathcal{B}$, and
- the edge set is the collection $\bigcup_{\substack{i \in [b] \\ j \in [v]}} \{(P_i, x_j) : x_j \in \text{ share of } P_i\}$.

Theorem 2: Given a bipartite graph $\mathcal{G}$ inducing a tensor design $\mathcal{B}$, and given subsets $\delta(P_i) \subseteq N(P_i)$ of size s,

(i) If $\bigcup_{i\in[b]} \delta(P_i) = \mathbf{V}$, then reconstruction of the modified scheme $\mathcal{F}(\mathcal{A}, \mathcal{B})$ is possible.

(ii) If $s \geq 1$, then (i) holds.

3.6. Defining access structure tokens

Consider first, the Krönecker product tensor design $\mathcal{A} \otimes \mathcal{B}_d$ as defined in Equation (1).

Let $\mathfrak{a}_1, \ldots, \mathfrak{a}_{v_1} \in \mathbb{F}_{p_1}$ be the elements in $\mathcal{A}$ and $\mathfrak{b}_1, \ldots, \mathfrak{b}_{v_2} \in \mathbb{F}_{p_2}$ be the elements in $\mathcal{B}$. The access structure tokens for the share of each player are elements of $\in \mathbb{Z}_2^{v_1} \times \in \mathbb{Z}_2^{v_1}$, computed according to Algorithm 1.

Algorithm 1 HsGen: Access structure tokens for the tensor designs $\mathcal{A} \otimes \mathcal{B}_d$ and $\mathcal{F}(\mathcal{A}, \mathcal{B})$

$\gamma \xleftarrow{\$} \mathsf{Perm}(\{0,1\}^{v_1} \times \{0,1\}^{v_2})$.
for $1 \leq i \leq b_1 b_2$ **do**: *//player* P_i
  for $1 \leq j \leq v_1$ **do**: *//element* $\mathfrak{a}_j$
    $\hat{\mho}_i^{(1,\Gamma)} \leftarrow (\omega_1, \ldots, \omega_{v_1})$ such that $\omega_j = 1$ if and only if element $\mathfrak{a}_j$ of $\mathcal{A}$ occurs as a product $\mathfrak{a}_j\mathfrak{b}_l$ in the share of P_i.
  end for
  for $1 \leq l \leq v_2$ **do**: *//element* $\mathfrak{b}_l$
    $\hat{\mho}_i^{(2,\Gamma)} \leftarrow (\omega_1, \ldots, \omega_{v_2})$ such that $\omega_l = 1$ if and only if element $\mathfrak{b}_l$ of $\mathcal{B}$ occurs as a product $\mathfrak{a}_j\mathfrak{b}_l$ in the share of P_i.
  end for
  $\left(\mho_1^{(\Gamma)}, \ldots, \mho_{b_1 b_2}^{(\Gamma)}\right) \leftarrow \gamma\left(\hat{\mho}_1^{(1,\Gamma)} \| \hat{\mho}_1^{(2,\Gamma)}, \ldots, \hat{\mho}_{b_1}^{(1,\Gamma)} \| \hat{\mho}_{b_2}^{(2,\Gamma)}\right)$. *//permutation*
end for

Logical condition

From Algorithm 1, it is clear that the authorisation of a collection of players $\mathbf{B}$ can be determined directly from the intermediate vectors $\hat{\mho}_i^{(1,\Gamma)}$ and $\hat{\mho}_i^{(2,\Gamma)}$ used to compute their access structure tokens. Consider the two logical statements P and Q:

$$\begin{aligned} P &: \mathbf{B} \in \Gamma \\ Q &: \left(\bigvee_{i\in\mathbf{B}} \hat{\mho}_i^{(1,\Gamma)} \text{ has Hamming weight } \geq \tau_1\right) \wedge \left(\bigvee_{i\in\mathbf{B}} \hat{\mho}_i^{(2,\Gamma)} \text{ has Hamming weight } \geq \tau_2\right). \end{aligned} \tag{2}$$

Then from the definition of $\hat{\mho}_i^{(1,\Gamma)}$ and $\hat{\mho}_i^{(2,\Gamma)}$, it is clear that $P \leftrightarrow Q$. The proceeding lemma easily follows from this observation:

Lemma 1: Let Γ denote the access structure for the tensor design $\mathcal{A} \otimes \mathcal{B}_d$. Then there exist parameters θ and Θ such that Γ is fully characterised by the following three conditions on any collection of players $\mathbf{B} \in 2^{\mathbf{P}}$:

1. If $|\mathbf{B}| < \theta$, then $\mathbf{B} \notin \Gamma$.
2. If $\theta \leq |\mathbf{B}| < \Theta$, then $\mathbf{B}$ may or may not belong to Γ, *i.e.* it may or may not be authorised.

3. If $|\mathbf{B}| \geq \Theta$, then $\mathbf{B} \in \Gamma$.

Proof: The proof follows by checking which collections of players satisfy the condition Q. If τ_1 and τ_2 are the reconstruction numbers of $\mathcal{A}$ and $\mathcal{B}$, respectively. Then from Lemmas 4 and 7 of Roy and Roy (2023), $\theta = (\tau_1 - 1)(\tau_2 - 1) + 1$. Also, from Lemmas 5, 6, 8 and 9 of Roy and Roy (2023), $\Theta = \min\{(\tau_1 - 1)b_2 + 1, (\tau_2 - 1)b_1 + 1\}$. □

Further observe that the permutation γ in Algorithm 1 ensures that a collection of players $\mathbf{B}$ of size $t < \Theta$ cannot simply examine their tokens and conclude (with probability 1) whether or not it is authorised.

4. Main results

Theorem 3: Given a positive integer d that satisfies Theorem 1, consider the tensor designs $\mathcal{A} \otimes \mathcal{B}_d$ with ramp structure (θ, Θ, ℓ), for a secret k, and shares $s_i^{(k)}$ for each player $P_i \in \mathbf{P}$. Then there exists an access structure token generation algorithm that makes $\mathcal{A} \otimes \mathcal{B}_d$ an ϵ-almost access structure hiding (θ, Θ, ℓ)-ramp tensor design.

Theorem 4: Given a positive integer d that satisfies Theorem 1, consider the tensor designs $\mathcal{F}(\mathcal{A}, \mathcal{B})$ with ramp structure (θ, Θ, ℓ), for a secret k, and shares $s_i^{(k)}$ for each player $P_i \in \mathbf{P}$. Then there exists an access structure token generation algorithm that makes $\mathcal{F}(\mathcal{A}, \mathcal{B})$ an ϵ-almost access structure hiding (θ, Θ, ℓ)-ramp tensor design.

Theorem 5: The access structure hiding tensor design $\mathcal{A} \otimes \mathcal{B}_d$ is verifiable.

Theorem 6: The access structure hiding tensor design $\mathcal{F}(\mathcal{A}, \mathcal{B})$ is verifiable.

5. Proof of Theorems 3 and 4

Proof: [Proof of Theorem 3.] This is easily seen as the scheme $\mathcal{A} \otimes \mathcal{B}_d$ satisfies the six properties enumerated in Definition 7.

Completeness and ϵ_1-completeness:

Case 1: $|\mathbf{A}| \geq \Theta$. Since the access structure tokens of any collection of size at least Θ always satisfy the logical condition (2), $\mathbf{A}$ can simply check this condition and output 1. Therefore,

$$\Pr\left[\mathsf{HsVer}\left(\left\{\mho_i^{(\Gamma)}\right\}_{i\in\mathbf{A}}\right) = 1\right] = 1.$$

Case 2: $\theta < |\mathbf{C}| < \Theta$, and $\mathbf{C}$ is authorised. Let $|\mathbf{C}| = T$, such that $\theta < T < \Theta$ and $\mathbf{C}$ is an authorised collection of players.

$$\begin{aligned} \text{Number of permutations that fix the access structure tokens of } \mathbf{C} &= (\ell - T)! \\ \text{Total number of permutations on all } \ell \text{ access structure tokens} &= \ell! \end{aligned}$$

As there is a uniformly random distribution on the access structure tokens, $\mathbf{C}$ can make a uniformly random guess from $\{0, 1\}$ about its authorisation status. Therefore,

the probability that any collection of size T can identify itself as authorised can be bounded above by the summation

$$\sum_{\substack{\mathbf{C}\in\Gamma \\ \text{with } |\mathbf{C}|=T}} \frac{(\ell-T)!}{\ell!} \leq \frac{1}{\binom{\ell}{T}},$$

$$\text{and thus, } \Pr\left[\mathsf{HsVer}\left(\left\{\mathfrak{V}_i^{(\Gamma)}\right\}_{i\in\mathbf{C}}\right)=1\right] \leq \sum_{\theta<T<\Theta} \frac{1}{\binom{\ell}{T}}. \tag{3}$$

Denoting $\epsilon_1 \coloneqq \sum_{\theta<T<\Theta} \frac{1}{\binom{\ell}{T}}$, we then have

$$\Pr\left[\mathsf{HsVer}\left(\left\{\mathfrak{V}_i^{(\Gamma)}\right\}_{i\in\mathbf{C}}\right)=1\right] \geq 1-\epsilon_1.$$

Soundness and ϵ_2-soundness:

Case 1: $|\mathbf{B}| \leq \theta$. Since the access structure tokens of any collection of size at most θ never satisfy the logical condition (2), $\mathbf{B}$ can simply check this condition and output 0. Therefore,

$$\Pr\left[\mathsf{HsVer}\left(\left\{\mathfrak{V}_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}\right)=0\right]=1.$$

Case 2: $\theta < |\mathbf{C}| < \theta$, and $\mathbf{C}$ is unauthorised. Let $|\mathbf{C}| = T$, such that $\theta < T < \Theta$ and $\mathbf{C}$ is an unauthorised collection of players. We arrive at the upper bound $\epsilon_2 \coloneqq \sum_{\theta<T<\Theta} \frac{1}{\binom{\ell}{T}}$ as in Equation (3), by the same argument as for ϵ_1-completeness above. Hence,

$$\Pr\left[\mathsf{HsVer}\left(\left\{\mathfrak{V}_i^{(\Gamma)}\right\}_{i\in\mathbf{C}}\right)=0\right] \geq 1-\epsilon_2.$$

Statistical hiding and ϵ_3-statistical hiding: As $\mathcal{A}\otimes\mathcal{B}_d$ is a (θ,Θ,ℓ)-ramp scheme, any non-ramp collection of parties can simply count the access structure tokens of all its players and determine its authorisation.

Case 1: $|\mathbf{B}| \leq \theta$. By definition of the access structure tokens, $\bigvee_{i\in\mathbf{B}} \hat{\mathfrak{V}}_i^{(1,\Gamma)} < \tau_1$ and $\bigvee_{i\in\mathbf{B}} \hat{\mathfrak{V}}_i^{(2,\Gamma)} < \tau_2$.

Thus, for any such collection and for any access structure $\Gamma' \subseteq 2^{\mathbf{P}}$ characterised by the ramp bounds (θ,Θ) such that $\mathbf{B}\notin\Gamma'$, $\Gamma' \mid \left\{\mathfrak{V}_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}$ follows the uniform distribution. Hence,

$$\Pr\left[\Gamma' \mid \left\{\mathfrak{V}_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}\right] = \frac{2}{\ell(\ell-3)} = \frac{2}{2^{b_1b_2}(2^{b_1b_2}-3)}.$$

And therefore, $\left|\Pr\left[\Gamma \mid \left\{\mathfrak{V}_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{B}}\right] - \Pr\left[\Gamma' \mid \left\{\mathfrak{V}_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{B}}\right]\right| = 0.$

If Γ' is any other type of access structure (which does not characterise a ramp scheme), then $\Pr\left[\Gamma' \mid \left\{\mho_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{B}}\right] = 0$.

And therefore, $\left|\Pr\left[\Gamma \mid \left\{\mho_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{B}}\right] - \Pr\left[\Gamma' \mid \left\{\mho_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{B}}\right]\right|$
$= \dfrac{2}{2^{b_1 b_2}(2^{b_1 b_2} - 3)}$.

Case 2(a): $\theta < |\mathbf{C}| < \Theta$ and $\mathbf{C}$ is unauthorised. Since $\mathbf{C}$ is an unauthorised collection of parties, it knows no information about either factor, $\mathcal{A}$, $\mathcal{B}_d$, of $\mathcal{A} \otimes \mathcal{B}_d$. Therefore, by the same arguments as for Case 1,

$$\left|\Pr\left[\Gamma \mid \left\{\mho_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{B}}\right] - \Pr\left[\Gamma' \mid \left\{\mho_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{B}}\right]\right| = \frac{2}{2^{b_1 b_2}(2^{b_1 b_2} - 3)}.$$

Case 2(b): $\theta < |\mathbf{C}| < \Theta$ and $\mathbf{C}$ has partial information about the secret. Let us assume $\mathbf{C}$ knows the secret of the factor $\mathcal{A}$ of $\mathcal{A} \otimes \mathcal{B}_d$. Then it must guess the shares of players of $\mathcal{B}_d$ at best uniformly at random. So, a similar computation as in Case 1 allows us to arrive at the bound

$$\left|\Pr\left[\Gamma \mid \left\{\mho_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{B}}\right] - \Pr\left[\Gamma' \mid \left\{\mho_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{B}}\right]\right| \leq \frac{2}{2^{b_2}(2^{b_2} - 3)}.$$

On the other hand, if $\mathbf{C}$ knows the secret of the factor $\mathcal{B}_d$ of $\mathcal{A} \otimes \mathcal{B}_d$, then the bound becomes

$$\left|\Pr\left[\Gamma \mid \left\{\mho_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{B}}\right] - \Pr\left[\Gamma' \mid \left\{\mho_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{B}}\right]\right| \leq \frac{2}{2^{b_1}(2^{b_1} - 3)}.$$

The equality in the two previous inequalities can be achieved when Γ' is not a ramp type scheme even when $\mathbf{C}$ has information about one threshold scheme. To sum it up, the required value for the parameter ϵ_3 is therefore the maximum of these two bounds. Without loss of generality, we have assumed that $b_1 \leq b_2$ and hence among the three expressions on the right side, the last one is the largest.

Thus,

$$\left|\Pr\left[\Gamma \mid \left\{\mho_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{B}}\right] - \Pr\left[\Gamma' \mid \left\{\mho_i^{(\Gamma)}\right\}_{i\in\mathbf{B}}, \left\{s_i^{(k)}\right\}_{i\in\mathbf{B}}\right]\right| \leq \frac{2}{2^{b_1}(2^{b_1} - 3)}.$$

□

The proof of Theorem 4 is exactly similar to the proof above.

6. Proof of theorems 5 and 6

Proof: If $\mathbf{A}$ is an authorised collection of parties (irrespective of its size), then clearly,

$$\Pr\left[\mathsf{Ver}\left(k, \left\{s_i^{(k)}\right\}_{i\in\mathbf{A}}\right) = 1\right] = 1$$

as $\mathbf{A}$ can reconstruct the secret perfectly.

Recall the definition of the prime power q from Section 3.1. For an unauthorised collection of parties $\mathbf{A}$ such that $\mathbf{A}$ cannot compute all elements of even one of $\mathcal{A}$ or $\mathcal{B}_d$,

$$\begin{aligned} \Pr\left[\mathsf{Ver}\left(k, \left\{s_i^{(k)}\right\}_{i\in\mathbf{A}}\right) = 1\right] &\leq \frac{1}{q} \\ \text{and therefore, } \Pr\left[\mathsf{Ver}\left(k, \left\{s_i^{(k)}\right\}_{i\in\mathbf{A}}\right) = 0\right] &\geq 1 - \frac{1}{q}. \end{aligned} \tag{4}$$

For a ramp collection of parties $\mathbf{A}$ such that $\theta < |\mathbf{A}| < \Theta$, *i.e.* $\mathbf{A}$ can compute all elements of exactly one of $\mathcal{A}$ or $\mathcal{B}_d$,

$$\begin{aligned} \Pr\left[\mathsf{Ver}\left(k, \left\{s_i^{(k)}\right\}_{i\in\mathbf{A}}\right) = 1\right] &\leq \max\left\{\frac{1}{p_1}, \frac{1}{p_2}\right\} \\ \text{and therefore, } \Pr\left[\mathsf{Ver}\left(k, \left\{s_i^{(k)}\right\}_{i\in\mathbf{A}}\right) = 0\right] &\geq 1 - \max\left\{\frac{1}{p_1}, \frac{1}{p_2}\right\}. \end{aligned} \tag{5}$$

The bounds in Equations (4) and (5) are simply because $\mathcal{A}$ and $\mathcal{B}_d$ are τ_1- and τ_2-threshold schemes based on Shamir schemes Shamir (1979), which means any collection of players that cannot reconstruct the entire secret cannot obtain any information about the secret. □

The proof of Theorem 6 is exactly similar to the proof above.

7. Applications

Our technique has real-world applications in a very wide range of domains, including secure multiparty computation Chaum (1989); Andrychowicz *et al.* (2016); Smart *et al.* (2024), secure distributed storage Garay *et al.* (1997); Rajasekaran and Duraipandian (2024), attribute-based encryption Nali *et al.* (2005); Ibraimi *et al.* (2009); Saidi *et al.* (2024); Asaithambi *et al.* (2024), access control mechanisms Eland (1978); di Vimercati (2011); Gondara (2011); Nour *et al.* (2022), secure cloud computing Xu *et al.* (2009); Cui and Yi (2024), e-voting systems Rabia *et al.* (2023), secure data sharing in blockchain technology Zhang and Lin (2018); Alshehri *et al.* (2023); Wang *et al.* (2023), and privacy-preserving machine learning algorithms Çatak (2015); Xu *et al.* (2015); Qin *et al.* (2024); Mestari *et al.* (2024), to name a few.

For example in cloud storage systems Shin *et al.* (2017), our technique can enhance data integrity and availability by enabling authorized parties to reconstruct lost or corrupted shares without involving the initial dealer, avoiding framing of various parties, and computationally easy verification of shares against malicious adversary interactions.

Within sensor-based IoT systems Sikder *et al.* (2018), repairable ramp schemes safeguard the confidentiality and integrity of sensitive information exchanged among devices. The ability to repair lost or corrupted shares while maintaining frameproofness, and verifiability of these shares, along with the ability to ensure their completeness and soundness without the need to actually access the shares ensures uninterrupted operation and security, critical for IoT applications.

Furthermore, repairable ramp schemes are instrumental in multi-level security systems Gao and Xiao (2011); Wagner (1997), such as those employed by government agencies and financial institutions. Our techniques would only improve their guarantees of security, while maintaining accessibility of critical information. They would also enable secure collaborative data sharing in environments where multiple parties require access to confidential data.

8. Conclusion and future work

In this paper, we discuss verifiability and frameproofness of access structure hiding ramp-type tensor designs. We do this through the introduction of a new type of secret sharing scheme, called an ϵ-almost access structure hiding (θ, Θ, ℓ)-ramp tensor design, thus making an essential generalisation of the existing novel design introduced by Sehrawat *et al.*. We explore ways of enhancing data security and privacy, especially Roy *et al.*'s concept of extending repairable threshold schemes, using tensor products of balanced incomplete block designs. This concept provides a fundamental generalization of existing designs, and thus plays an important role in enhancing the security and verifiability of secret sharing schemes by providing a mechanism for parties to verify the correctness of the shares they receive and ensuring that the reconstruction process is accurate. By incorporating ramp schemes, the construction becomes more robust against malicious behavior and unauthorized access, thus strengthening the overall security and integrity of the secret sharing process. We also list a few real-world applications where our techniques could be utilised for improved security.

While we demonstrate our concept of ϵ-almost access structure hiding for only extendable combinatorial tensor designs, it opens up a wide range of possibilities for any ramp-type scheme to incorporate this technique for further improvement of confidentiality, secrecy and verifiability.

Acknowledgements

The authors express their thanks to the Editors for their guidance and counsel. The authors are also grateful to the reviewer for valuable comments and suggestions of generously listing many useful references.

Bibliography

Alshehri, S., Bamasag, O., Alghazzawi, D. M., and Jamjoom, A. (2023). Dynamic secure access control and data sharing through trusted delegation and revocation in a blockchain-enabled cloud-iot environment. *IEEE Internet Things*, **10**, 4239–4256.

Andrychowicz, M., Dziembowski, S., Malinowski, D., and Mazurek, L. (2016). Secure multiparty computations on bitcoin. *Communications of the ACM*, **59**, 76–84.

Asaithambi, S., Ravi, L., Devarajan, M., Selvalakshmi, A., Almaktoom, A. T., Almazyad, A. S., Xiong, G., and Mohamed, A. W. (2024). Blockchain-assisted hierarchical attribute-based encryption scheme for secure information sharing in industrial internet of things. *IEEE Access*, **12**, 12586–12601.

Çatak, F. Ö. (2015). Secure multi-party computation based privacy preserving extreme learning machine algorithm over vertically distributed data. In Arik, S., Huang, T., Lai, W. K., and Liu, Q., editors, *Neural Information Processing - 22nd International Conference, ICONIP 2015, Istanbul, Turkey, November 9-12, 2015, Proceedings, Part II*, volume 9490 of *Lecture Notes in Computer Science*, pages 337–345. Springer.

Chaum, D. (1989). The spymasters double-agent problem: Multiparty computations secure unconditionally from minorities and cryptographically from majorities. In Brassard, G., editor, *Advances in Cryptology - CRYPTO '89, 9th Annual International Cryptology Conference, Santa Barbara, California, USA, August 20-24, 1989, Proceedings*, volume 435 of *Lecture Notes in Computer Science*, pages 591–602. Springer.

Cui, H. and Yi, X. (2024). Secure internet of things in cloud computing via puncturable attribute-based encryption with user revocation. *IEEE Internet Things*, **11**, 3662–3670.

Dehkordi, M. H., Farahi, S. T., and Mashhadi, S. (2024). Lwe-based verifiable essential secret image sharing scheme ((t , s , k , n) $({t,s,k,n})$ - VESIS). *IET Image Process*, **18**, 1053–1072.

Desmedt, Y., Mo, S., and Slinko, A. M. (2021). Framing in secret sharing. *IEEE Transactions on Information Forensics and Security*, **16**, 2836–2842.

di Vimercati, S. D. C. (2011). Access control policies, models, and mechanisms. In van Tilborg, H. C. A. and Jajodia, S., editors, *Encyclopedia of Cryptography and Security, 2nd Ed*, pages 13–14. Springer.

Eland, N. (1978). *Language-Based Access Control Mechanisms for Shared Databases*. PhD thesis, Cornell University, USA.

Gao, C. and Xiao, C. (2011). A security model for information systems with multi-level security. In Wang, Y., Cheung, Y., Guo, P., and Wei, Y., editors, *Seventh International Conference on Computational Intelligence and Security, CIS 2011, Sanya, Hainan, China, December 3-4, 2011*, pages 620–624. IEEE Computer Society.

Garay, J. A., Gennaro, R., Jutla, C. S., and Rabin, T. (1997). Secure distributed storage and retrieval. In Mavronicolas, M. and Tsigas, P., editors, *Distributed Algorithms, 11th International Workshop, WDAG '97, Saarbrücken, Germany, September 24-26, 1997, Proceedings*, volume 1320 of *Lecture Notes in Computer Science*, pages 275–289. Springer.

Gondara, M. K. (2011). Access control mechanisms for semantic web services-a discussion on requirements & future directions. *Clinical Orthopaedics and Related Research*, **abs/1105.0141**.

Hofmeister, T., Krause, M., and Simon, H. U. (2000). Contrast-optimal k out of n secret sharing schemes in visual cryptography. *Theoretical Computer Science*, **240**, 471–485.

Ibraimi, L., Tang, Q., Hartel, P. H., and Jonker, W. (2009). Efficient and provable secure ciphertext-policy attribute-based encryption schemes. In Bao, F., Li, H., and Wang, G., editors, *Information Security Practice and Experience, 5th International Conference, ISPEC 2009, Xi'an, China, April 13-15, 2009, Proceedings*, volume 5451 of *Lecture Notes in Computer Science*, pages 1–12. Springer.

Laing, T. M. and Stinson, D. R. (2018). A survey and refinement of repairable threshold schemes. *Journal of Mathematical Cryptology*, **12**, 57–81.

Mestari, S. Z. E., Lenzini, G., and Demirci, H. (2024). Preserving data privacy in machine learning systems. *Computer Security*, **137**, 103605.

Nali, D., Adams, C. M., and Miri, A. (2005). Using threshold attribute-based encryption for practical biometric-based access control. *International Journal of Network Security*, **1**, 173–182.

Nour, B., Khelifi, H., Hussain, R., Mastorakis, S., and Moungla, H. (2022). Access control mechanisms in named data networks: A comprehensive survey. *ACM Computing Surveys*, **54**, 1–35.

Paterson, M. B. and Stinson, D. R. (2013). A simple combinatorial treatment of constructions and threshold gaps of ramp schemes. *Cryptography and Communications*, **5**, 229–240.

Pedersen, T. P. (1991). Non-interactive and information-theoretic secure verifiable secret sharing. In Feigenbaum, J., editor, *Advances in Cryptology - CRYPTO '91, 11th Annual International Cryptology Conference, Santa Barbara, California, USA, August 11-15, 1991, Proceedings*, volume 576 of *Lecture Notes in Computer Science*, pages 129–140. Springer.

Peng, K. (2012). Critical survey of existing publicly verifiable secret sharing schemes. *IET Information Security*, **6**, 249–257.

Qin, H., He, D., Feng, Q., Khan, M. K., Luo, M., and Choo, K. R. (2024). Cryptographic primitives in privacy-preserving machine learning: A survey. *IEEE Transactions on Knowledge and Data Engineering*, **36**, 1919–1934.

Rabia, F., Arezki, S., and Gadi, T. (2023). A review of blockchain-based e-voting systems: Comparative analysis and findings. *International Journal of Interactive Mobile Technologies*, **17**, 49–67.

Rajasekaran, P. and Duraipandian, M. (2024). Secure cloud storage for iot based distributed healthcare environment using blockchain orchestrated and deep learning model. *Journal of Intelligent and Fuzzy Systems*, **46**, 1069–1084.

Roy, B. K. and Roy, A. (2023). Iot-applicable generalized frameproof combinatorial designs. *IoT*, **4**, 466–485.

Saidi, A., Amira, A., and Nouali, O. (2024). A secure multi-authority attribute based encryption approach for robust smart grids. *Concurrency and Computation: Practice and Experience*, **36**.

Sehrawat, V. S., Yeo, F. Y., and Desmedt, Y. (2021). Extremal set theory and LWE based access structure hiding verifiable secret sharing with malicious-majority and free verification. *Theoretical Computer Science*, **886**, 106–138.

Shamir, A. (1979). How to share a secret. *Communications of the ACM*, **22**, 612–613.

Shin, Y., Koo, D., and Hur, J. (2017). A survey of secure data deduplication schemes for cloud storage systems. *ACM Computing Surveys*, **49**, 1–38.

Sikder, A. K., Petracca, G., Aksu, H., Jaeger, T., and Uluagac, A. S. (2018). A survey on sensor-based threats to internet-of-things (iot) devices and applications. *Clinical Orthopaedics and Related Research*, **abs/1802.02041**.

Smart, N., Baron, J. W., Saravanan, S., Brandt, J., and Mashatan, A. (2024). Multiparty computation: To secure privacy, do the math: A discussion with nigel smart, joshua w. baron, sanjay saravanan, jordan brandt, and atefeh mashatan. *ACM Queue*, **21**, 78–100.

Stinson, D. R. (2004). *Combinatorial Designs - Constructions and Analysis*. Springer.

Stinson, D. R. and Wei, R. (2018). Combinatorial repairability for threshold schemes. *Designs Codes and Cryptography*, **86**, 195–210.

Verheul, E. R. and van Tilborg, H. C. A. (1997). Constructions and properties of k out of n visual secret sharing schemes. *Designs Codes and Cryptography*, **11**, 179–196.

Wagner, G. (1997). Multi-level security in multiagent systems. In Kandzia, P. and Klusch, M., editors, *Cooperative Information Agents, First International Workshop, CIA' 97, Kiel, Germany, February 26-28, 1997, Proceedings*, volume 1202 of *Lecture Notes in Computer Science*, pages 272–285. Springer.

Wang, N., Fu, J., Zhang, S., Zhang, Z., Qiao, J., Liu, J., and Bhargava, B. K. (2023). Secure and distributed iot data storage in clouds based on secret sharing and collaborative blockchain. *IEEE/ACM Transactions on Networking*, **31**, 1550–1565.

Xu, J., Huang, R., Huang, W., and Yang, G. (2009). Secure document service for cloud computing. In Jaatun, M. G., Zhao, G., and Rong, C., editors, *Cloud Computing, First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. Proceedings*, volume 5931 of *Lecture Notes in Computer Science*, pages 541–546. Springer.

Xu, K., Yue, H., Guo, L., Guo, Y., and Fang, Y. (2015). Privacy-preserving machine learning algorithms for big data systems. In *35th IEEE International Conference on Distributed Computing Systems, ICDCS 2015, Columbus, OH, USA, June 29 - July 2, 2015*, pages 318–327. IEEE Computer Society.

Zhang, A. and Lin, X. (2018). Towards secure and privacy-preserving data sharing in e-health systems via consortium blockchain. *Journal of Medical Systems*, **42**, 1–18.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 555–573
http://www.ssca.org.in/journal

Analysis of Spatial and Temporal Patterns in Deaths of Despair in the Appalachian Region of the United States

Vishal Deo[1,2†], Raanan Gurewitsch[3†], Saurav Guha[2,4], Meghana Ray[2] and Saumyadipta Pyne[2,5]

[1]*National Institute for Research in Digital Health and Data Science, ICMR New Delhi, India*
[2]*Health Analytics Network, Pittsburgh, PA, USA*
[3]*University of Pittsburgh, Pittsburgh, PA, USA*
[4]*Department of Statistics, Mathematics & Computer Application, Bihar Agricultural University, Bhagalpur, Bihar, India*
[5]*Department of Statistics and Applied Probability, University of California Santa Barbara, Santa Barbara, CA, USA*

[†]These authors contributed equally.

Received: 02 August 2024; Revised: 28 September 2024; Accepted: 04 October 2024

Abstract

Mortality data in the United States (U.S.) revealed a precipitous rise in deaths among non-Hispanic white populations starting in the later part of 1990s due to such causes as Suicide, Alcohol consumption, and Drug accidental overdose. In particular, opioid-related deaths have increased dramatically across the U.S. during this period. For a systemic analysis of the temporal and spatial patterns of this critical phenomenon from the perspective of public health, we studied it in the context of Appalachian Region (AR), which spans across 13 U.S. states and is home to more than 8 percent of the country's population. We identified 8 spatial and temporal metaclusters of AR counties with relatively high rates of deaths due to the above-mentioned causes over the period 1979-2017 based on U.S. county- and cause-specific mortality data. Thus, we analyzed the mortality trends for each of the metaclusters, which were characterized based on their respective demographic and socioeconomic changes.

Key words: Deaths of despair; Mortality data; Spatial clustering, Temporal patterns; Opioid epidemic; Appalachian Region.

AMS Subject Classifications: 62H11, 62H30

Corresponding Author: S. Pyne
Email: spyne@ucsb.edu

For successful policy-making,
a government needs good statistics
as well as good statisticians.
One is not a substitute for the other.

Calyampudi Radhakrishna (C.R.) Rao

1. Introduction

Advancements in science, medicine and technology have extended life expectancy and improved health outcomes over the past half-century in the U.S., but not everyone has benefited equally. Researchers studying mortality have noted a precipitous rise in deaths between 1999 and 2005 due to such causes as Suicide, Alcohol consumption and Drug accidental overdose. Here, these causes of deaths are collectively denoted by the acronym: *SAD*. This phenomenon was characterized initially among non-Hispanic white Americans – mostly without a four-year college degree – as "deaths of despair" by Anne Case and Angus Deaton (Case and Deaton, 2015). They noted that white working-class lives over the last half century were affected by long-term labor market declines (Case and Deaton, 2020) which, in turn, led to a decline of families and relationships, limited access to high-quality healthcare, increased social isolation and loneliness, and a general loss of hope for the future (George *et al.*, 2021). Simultaneously, increases in ease of access to handguns, inexpensive alcohol, and prescription or non-prescription drugs including opioids have played their role in these excess, premature, and sometimes preventable, SAD deaths (George *et al.*, 2021; Shiels *et al.*, 2020).

Notably, while the historic deaths of despair were largely observed among non-Hispanic whites in rural America, more recently, around 2015, they increased across all races (Hedegaard *et al.*, 2018b,a). In 2017, there were 158,000 documented despair-related deaths that contributed to the longest sustained decline in life expectancy since 1915 (Woolf *et al.*, 2018). In the period between 2000 and 2017, there were 1,446,177 drug poisoning, suicide, and alcohol-induced premature deaths in the U.S., that included 563,765 drug poisoning deaths (17.6 per 100,000 person-years), 517,679 suicides (15.8 per 100,000 person-years), and 364,733 alcohol related deaths (10.5 per 100,000 person-years). These amounted to 451,596 excess deaths than those expected based on the rates of 2000 (Shiels *et al.*, 2020). For instance, alcohol related deaths increased by 77% from 2000 (19,627) to 2016 (34,857). Alcohol deaths have risen in all races/ethnicities and across all age groups in both men and women between 2000 to 2016 (Spillane *et al.*, 2020).

Opioid-related deaths have increased dramatically in the past two decades in the U.S. In 2016 alone, there were 45,838 opioid related deaths, and in 2017, the U.S. Department of Health and Human Services declared the opioid epidemic as a "public health emergency" (U.S. Department of Health and Human Services, 2023). Increase in mortality due to drug overdose rose by 15%, while alcohol-related and suicide deaths have increased yearly by 4.1% and 1.5% respectively (Shiels *et al.*, 2020). Trajectories in the former rates of mortality have also varied by geographical regions – high in some predominantly rural states such as Maine, Kentucky and West Virginia, with the lowest in other largely rural states such as

Nebraska and Iowa (Rigg *et al.*, 2018). The distribution of local patterns of the phenomenon is complex, and national-level analyses (*e.g.*, Jalal *et al.* (2018)) seldom consider myriad key factors that may influence the data. These include different types of opioids (say, cheaper synthetic ones, *e.g.*, Fentanyl), societal stigma (which in turn influences treatment seeking behaviors), differences in reporting or physician training, accuracy of determination of the cause of death, effectiveness of different interventions in addressing the specific types of opioid use, *etc.*

Discussions that focus on opioid mortality often overlook the intersectionality among its various social determinants including educational attainment or employment dynamics of a population. The effects of such factors are both spatial as well as temporal in nature. Taking cognizance of the fact that despair is, in general, a complex psychosocial phenomenon, we chose to focus our study specifically on SAD as the notified causes of death (as given by the ICD-10 codes) in the mortality database. Attempts were made in the past to broadly study national level data on socioeconomic disparities, mortality statistics, or the opioid epidemic (Wallace *et al.*, 2019; Case and Deaton, 2021; Jalal *et al.*, 2018). However, for gaining key insights into SAD deaths, we think it is more effective to concentrate on the occurrence (or recurrence) of mortality patterns in a particular highly-affected geographical region, which could then be partitioned into spatial and temporal subregions for systematic investigation. Towards this, in the present study, we focused on the Appalachian Region (AR) that is known for high rates of poverty and mortality due to despair as well as being among those parts of the U.S. that were seriously affected by drug overdose deaths over the past few decades (Rigg *et al.*, 2018; NACo and ARC, 2019).

Notably, AR is a 205,000-square-mile region that spans the Appalachian Mountains from southern New York to northern Mississippi. It includes all of West Virginia (WV) and parts of 12 other states: Alabama (AL), Georgia (GA), Kentucky (KY), Maryland (MD), Mississippi (MS), New York (NY), North Carolina (NC), Ohio (OH), Pennsylvania (PA), South Carolina (SC), Tennessee (TN), and Virginia (VA). AR includes 423 counties (around 13% of all counties in the U.S.) and 8 independent cities in 13 states, and has a population of approximately 25 million people. It is divided into 5 sub-regions: Northern, North Central, Central, South Central, and Southern. As per the Appalachian Regional Commission (ARC), the region has overall low educational attainment, increasing unemployment and poverty. In 2019, 24.9% and 24.6% of eligible individuals earned Bachelor's degrees in mining and non-mining counties respectively compared to 32.8% in the rest of the U.S. (Bowen *et al.*, 2020). Between 2005 and 2020, employment in the coal industry fell by around 54% (Bowen *et al.*, 2020). Between 2013 and 2017, poverty rates in Appalachia averaged 16.3% compared to 14.6% for rest of the U.S. A regional analysis indicates that poverty rates ranged between 6.5% to 41% with poverty mostly concentrated in central Appalachia encompassing Eastern KY and WV (Appalachian Regional Commission, 2019).

AR has seen a steeper rise in deaths of despair since around 1998, especially among the middle-aged population, as compared to the non-Appalachian regions of the country (Meit *et al.*, 2017). While the region is home to around 32.5% of the U.S. population, it accounted for 49.6% of excess deaths in U.S. caused by the increase in midlife mortality during 2010-2017 (Meit *et al.*, 2019; Woolf *et al.*, 2019). Further, Meit *et al.* (2019) have noted that this disparity in deaths of despair is more evident in the Central and North Central Appalachian sub-regions. Accidental drug overdoses were identified as a major contributor towards the

rising deaths of despair in AR between 1999 and 2017, sustained by the easy and abundant availability of prescription opioids and heroin in the region (Woolf *et al.*, 2019; Monnat, 2020). In rural Appalachia, women have been found to be at a higher risk of committing suicide than men (Christine *et al.*, 2020). Declining manufacturing and mining industries, persistent poverty, rurality, social isolation, and physically demanding and injury-prone manual labor jobs have been studied as some of the possible socio-economic determinants of distress in AR, *e.g.*, George *et al.* (2021), Rigg *et al.* (2018), Meit *et al.* (2017), Meit *et al.* (2019), Woolf *et al.* (2019), and Monnat (2020).

In this study, our aim is to identify sub-regions of AR with high prevalence of SAD deaths at multiple time-periods, and to investigate the association of SAD mortality trends in these sub-regions with their economic and demographic characteristics. Given the dynamic nature of such characteristics at local (county) levels, we divided the overall study time-period of 1979-2017 into eight five-year periods, and identified flexibly-shaped spatial clusters of AR counties based on high SAD mortality rates for each period. Further, 8 metaclusters were constructed by combining spatially contiguous counties that had multiple occurrences among the clusters identified in different time-periods. These 8 metaclusters represent sub-regions with persistent prevalence of high SAD mortality in AR, which were then characterized based on relevant covariates. The metaclusters were compared with respect to temporal trends in various demographic and economic parameters such as annual average overall employment rate, industry-specific employment rates (mining and manufacturing), population size, median age, and median household income. After description of the methods and results in the subsequent sections, we end with an overall discussion of the analysis, its findings and limitations.

2. Data

We obtained the time-series data of age-adjusted mortality rates due to the three SAD causes (based on the corresponding ICD 10 codes) for each county in AR from the publicly accessible Mortality Information and Research Analytics System (MOIRA) of University of Pittsburgh (www.moira.pitt.edu). MOIRA data is sourced from the Centers for Disease Control and Prevention (CDC) National Center for Health Statistics (NCHS), and the U.S. Census Bureau. The MOIRA system facilitates extraction and visualization of U.S. mortality and population data in a standardized format and categorized by causes of death given by International Classification of Diseases (ICD 10) codes. Data was collected for the period 1979-2017, which also contained mortality rates grouped by sex and race. Additional data such as on employment, wages, population size, median age, and median household income was obtained from the official websites of the U.S. Bureau of Labor Statistics (www.bls.gov) and the U.S. Census Bureau (www.data.census.gov).

3. Methods

For each of the eight five-year periods, we performed spatial clustering of the 423 Appalachian counties based on county-level SAD Age-Adjusted mortality Rates (AAR) using a flexibly shaped spatial scan statistic due to Tango and Takahashi (2005) implemented with a restricted likelihood ratio in the R package rflexscan (Otani and Takahashi, 2021). The counties appearing in these spatial clusters were identified and ranked by their number of occurrences over the eight time-periods. Recurrent counties, *i.e.*, counties with more

than one occurrence in the time-period-specific clusters, were combined using the K (=1) nearest-neighbor strategy to obtain the final spatiotemporal metaclusters. We characterized the metaclusters using known Socio-Economic Status (SES) and race-based county labels (Wallace *et al.*, 2019). Smooth log-transformed trends of SAD AAR were obtained using the MortlitySmooth package in R (Camarda, 2012) and were plotted by age, sex, race, and SAD causes of death for each of the 8 metaclusters. SAD AARs were also predicted using the same library beyond 2017 for each metacluster until 2020, to avoid conflation with the Covid-19-associated mortality rates of the same areas thereafter.

Trends of annual employment rate in the mining and manufacturing industries were also plotted for the metaclusters. In addition, distributions of average annual unemployment rates in the metaclusters in five-year periods were visualized using boxplots. Percentage changes in population size and median age of the population of the metaclusters from 1980 to 2020 were also evaluated for the metaclusters. Total population of metaclusters in a year was calculated as the sum of census population of the counties. Median age of the metaclusters for a given year was calculated as the median of the county-wise median age. In addition, median household income of the metaclusters was calculated as the median of the county-wise median household income. Since the county-wise estimates of median household income were available for the years 1979 and 2021 (which are based on the 1980 and 2020 census, respectively), the percentage change was calculated from 1979 to 2021. All the three metacluster-wise percentage changes were obtained as the median of the percentage changes calculated for the respective counties during the mentioned period.

4. Results

Spatially-flexible scan statistics identified clusters comprising of the counties in AR with relatively higher SAD age-adjusted mortality rates (AAR) for each of the 8 five-year time-periods (Figure 2). Although the compositions of the clusters vary across time-periods, some counties appeared recurrently in the identified clusters over time. Such counties are mostly concentrated around the South Central, Central, and North Central Appalachia, with a few in Northern and Southern regions. Assuming that a higher number of recurrences of any county in the clusters over the eight time periods would indicate a longer prevalence of SAD mortality therein, we used the K (=1) nearest-neighbor strategy for selecting nearby counties that have multiple such recurrences, and then combining them to form a *metacluster*. The neighborhood of (two or more) counties is determined by their sharing of common boundary lines or points even if they lie across different states.

The above procedure led to the construction of 8 metaclusters of counties spanning across AR that are both spatially contiguous and temporally recurrent hotspots of SAD deaths (Figure 3). Five of these eight metaclusters are located around the western regions of the Central and South-Central Appalachia, with one cluster extending to the lower region of the North Central Appalachia. Among the remaining three clusters, which are relatively smaller in size, two are located in the Northern Appalachia and one in the Southern Appalachia. Details of the identified metaclusters are provided in Table 1. A closer look at the physical map of the Appalachian region (Figure 3 inset) reveals the location and the underlying landform patterns of these metaclusters. Evidently, the metaclusters are located mostly along the Valley & Ridge region, and the southern part of the Blue Ridge Mountains.

Figure 1: Details of spatiotemporal metaclusters

Cluster No.	Cluster Name	State(s) of counties	No. of counties	County-FIPS	Appalachian Subregion (no. of counties)
1	Alabama (AL)	Alabama	18	37005, 37067, 37171, 37197, 51021, 51027, 51051, 51063, 51077, 51089, 51105, 51141, 51155, 51167, 51169, 51185, 51195, 51197	Central (7), and South Central (11)
2	Eastern PA	Pennsylvania	15	13241, 37021, 37027, 37043, 37087, 37089, 37099, 37113, 37149, 37161, 37173, 37175, 45021, 45073, 45083	South Central (11), and Southern (4)
3	Kentucky (KY)	Kentucky	12	21013, 21025, 21071, 21095, 21115, 21119, 21131, 21133, 21189, 21193, 21195, 21235	Central (12)
4	South NC + SC + GA	North Carolina, South Carolina, and Georgia	12	47013, 47001, 47025, 47049, 47057, 47059, 47063, 47093, 47145, 47151, 47163, 47173	Central (7), and South Central (5)
5	Tennessee (TN)	Tennessee	10	54005, 54019, 54039, 54045, 54047, 54055, 54059, 54081, 54089, 54109	North Central (5), and Central (5)
6	VA + North NC	Virginia and North Carolina	6	42003, 42007, 42051, 42125, 54009, 54029	Northern (6)
7	Western PA	Pennsylvania	5	42069, 42079, 42089, 42107, 42113	Northern (5)
8	West Virginia (WV)	West Virginia	4	1009, 1055, 1073, 1095	Southern (4)

To illustrate if spatiotemporal patterns of socioeconomic status (SES) have any association with the identified metaclusters of high SAD mortality rates, we used the SES class labels due to Wallace *et al.* (2019). The labels 1 & 8 represent high SES, 2 mid/low SES, and 4 mid SES; where 1 & 2 are semi-urban, 8 is rural, and 4 mostly-rural counties. Thus, Figure 4 provides us with a nuanced characterization of each metacluster. We observe that 6 out of 8 metaclusters have a majority of their counties falling in low SES categories. However, the two metaclusters in PA – Eastern PA and Western PA – appear to be distinctive from the rest as they have a more balanced distribution of both semi-urban and rural counties with high as well as mid SES. We discuss about this point further below. Here, we note that the categorization by Wallace *et al.* (2019), which is based on relatively recent SES of the counties, may not capture the full dynamics of SES over the entire time-period of this study.

Historical trends of SAD AAR (in logarithmic scale) in each of the 8 metaclusters from 1979 to 2017 for different causes of deaths of despair, age groups, races and sexes are presented in the Figures 5, 6, 7 and 8, respectively. The changing dynamics of the cause-specific SAD AARs in the 8 metaclusters is visualized in Figure 5. While the contribution of drug overdose deaths to SAD AAR was almost negligible in the early 1990s, it has grown

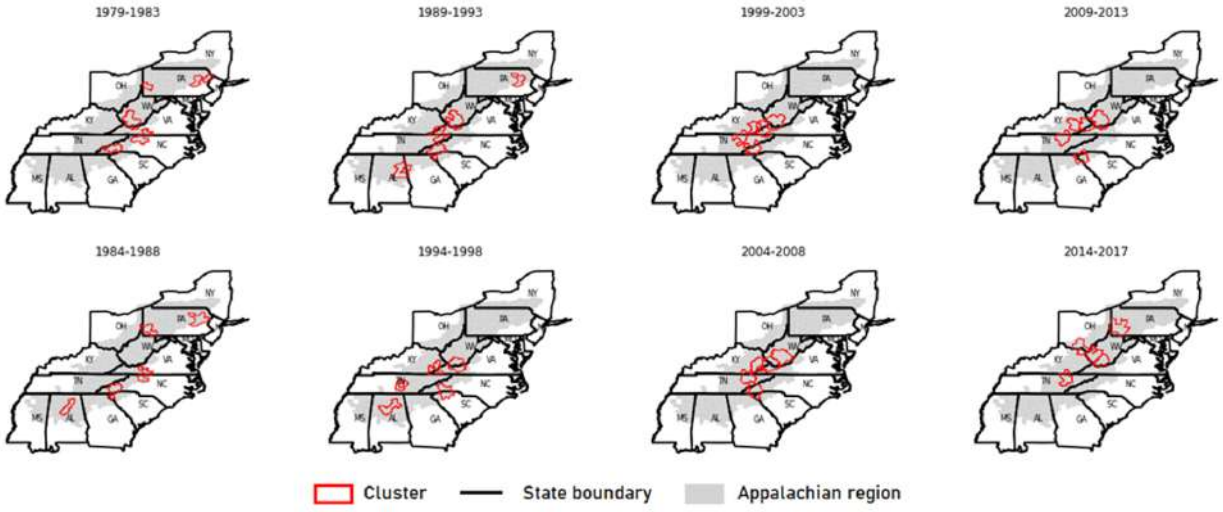


Figure 2: Spatial clustering of the AR counties in terms of their SAD AAR in each of the 8 successive 5-year time-periods between 1979 and 2017. The identified clusters' boundaries are shown with red lines. AR (grey area) and the state boundaries (black lines) are included for visual reference.

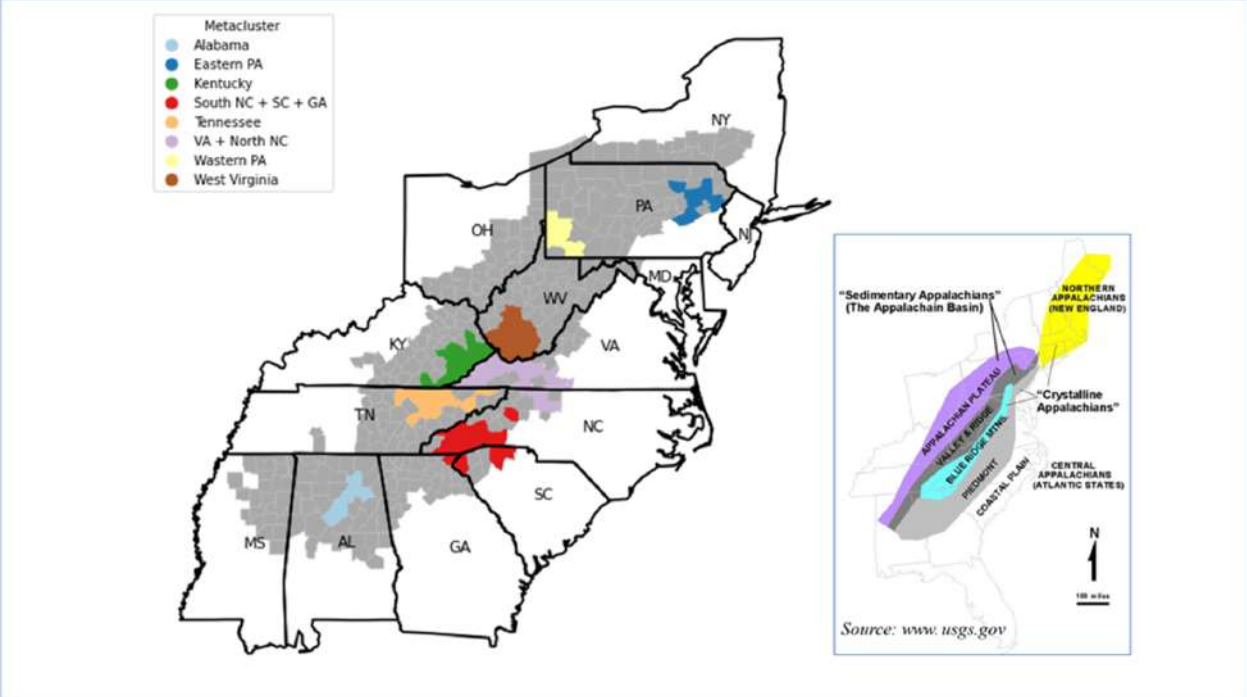


Figure 3: The 8 spatial metaclusters of SAD deaths in AR produced from the spatial clusters over time-periods between 1979 and 2017. These are shown in distinct colors and their labels in the legend. For visual reference, AR (grey area) and the state boundaries (black lines) are included along with an inset physical map of AR (due to www.usgs.gov).

continuously at a fast pace since then, and has become comparable to those due to suicides and alcohol related deaths. In fact, in some metaclusters (TN and VA + North NC), drug

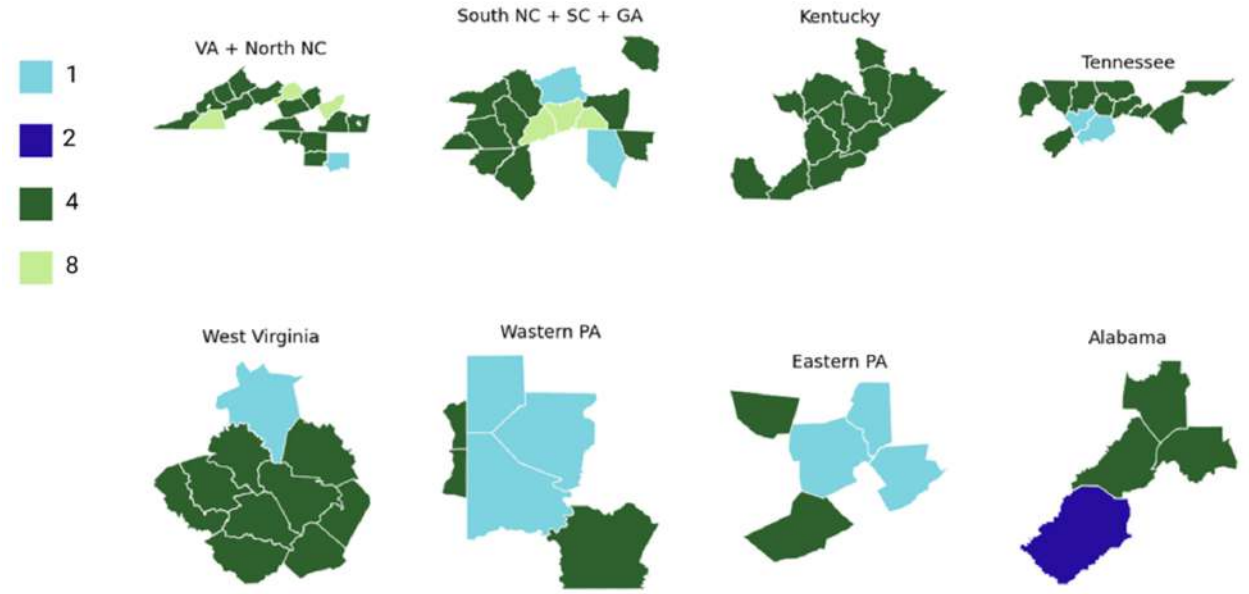


Figure 4: Characterization of the spatial metaclusters based on the known SES classification (given in the color-key) of their constituent counties.

overdose mortality AAR has even surpassed that of alcohol-related deaths and suicides. A slightly improved scenario on drug-overdose deaths could be noted in metacluster 3 (KY) by comparing the trends in Figures 5 and 6 during towards the end of the study period. Notably, the AAR trends of suicides and alcohol related deaths have remained high, and in fact, more or less constant, over the study period in all metaclusters.

Notably, the most alarming pattern appears in Figure 6 in which the SAD mortality AARs for both the younger and older age groups have increased over the years in all metaclusters; but the rate of change is markedly higher for the former age group of < 45 years. SAD AAR for the older age group (≥ 45 years) has been high since 1979, with occasional plateauing around the last decade of the 20th century, before it started increasing again, albeit at a slower pace compared to the younger age group. The trend for the younger age group is present in every metacluster, rising from around 12-20 per 100,000 in the year 1979 to around 33-55 per 100,000 in the year 2017, *i.e.*, at the end of the study-period. Interestingly, while the SAD AAR trends have continued to rise over the decades, Figure 7 shows little racial difference therein between whites and non-whites, except for marginally higher rates for the whites in the last two decades. Overall, the gradual and continual rise in the SAD AARs among the female and the younger populations – vis-a-vis the traditional trends of the male and the older populations – are clearly visible from Figures 6 and 8 respectively.

Moreover, the younger age group's SAD AAR has been continuously increasing in all metaclusters, except for metacluster 3 (KY) where a downward trend was observed in the last decade. In all other metaclusters, the gap between the SAD AAR of the younger and the older populations has continued to get narrower. Metacluster 6 (VA + North NC) has seen the sharpest rise in SAD mortality rate of the younger age group in recent years, with its value reaching almost 90 per 100,000. This phenomenon of alarmingly increasing SAD AAR among the younger age group is prevalent among all races, with almost similar trends for the non-Hispanic white and the other race groups (Figure 7). Although the SAD AAR

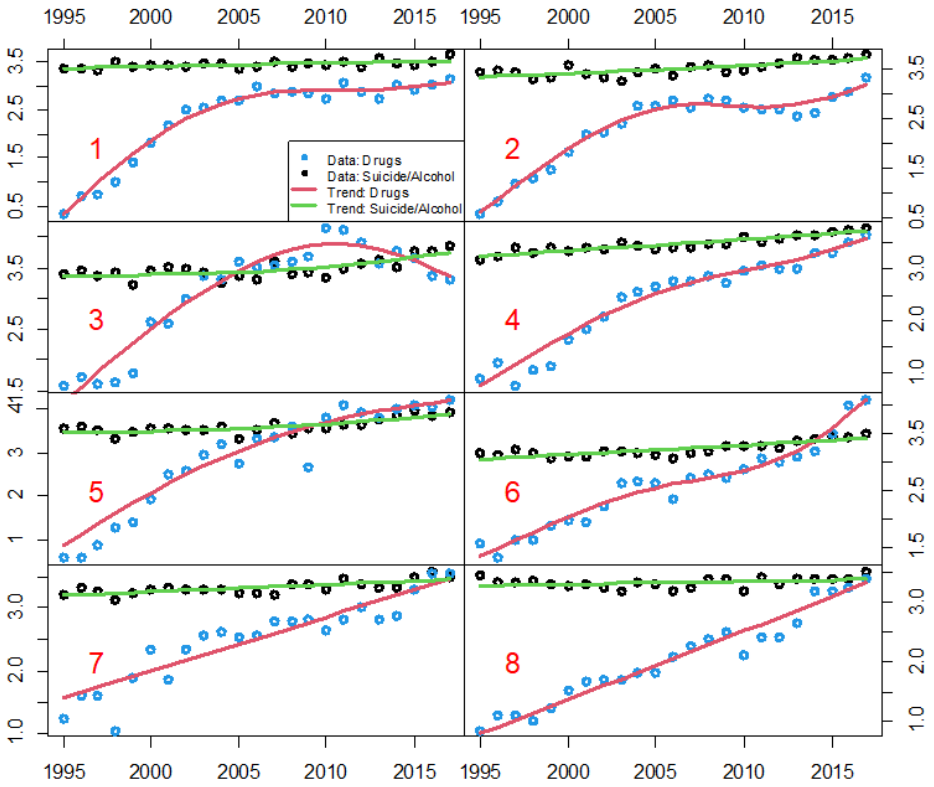


Figure 5: Longitudinal trends of SAD AAR for different causes of deaths of despair in the 8 spatial metaclusters of AR. The historically more common causes, *i.e.*, suicides and alcoholism (SA: green curves) are compared with the drug overdoses (D: red curves).

for younger age groups has been consistently high for the male population as compared to the female population, the rate of increase of the SAD AAR has been evidently higher for the female population in all metaclusters (see Figure 8).

For a visual comparison of the different metaclusters, Figure 9 overlays their trends of SAD AAR from 1979 to 2020 along with that of the remaining counties in AR (shown as a bold grey curve). Naturally, the latter, denoted by "Rest", has lower SAD AAR than every metacluster for every observed time-period while following a similarly increasing trend over time. In addition to this baseline trend, we also included the projected trend of SAD AAR for the time-period beyond 2017, up to 2020 (*i.e.*, the pre-pandemic years), as shown to the right of the dotted line. Overall, it is evident that the SAD AAR has been increasing in all metaclusters ever since 1979, and with a higher pace since around 2000. The projections provide clear insights into the temporal patterns for the metaclusters. For instance, the distinctive decline in SAD mortality of metacluster 3 (KY) in the last decade stands out among all of these trends. Metaclusters 1 (AL) and 5 (TN) started with very similar trends but went on to stray – during the 1990s – the most apart from each other. In fact, the latter is projected to have the highest SAD AAR among all the metaclusters exactly when the former is supposed to have AAR even lower than the Rest. Most importantly, several metaclusters {1, 2, 7, 8}, spanning various sub-regions of AR, that had exhibited different trends in the first decade, seemed to converge towards the Rest by the end of the study-period.

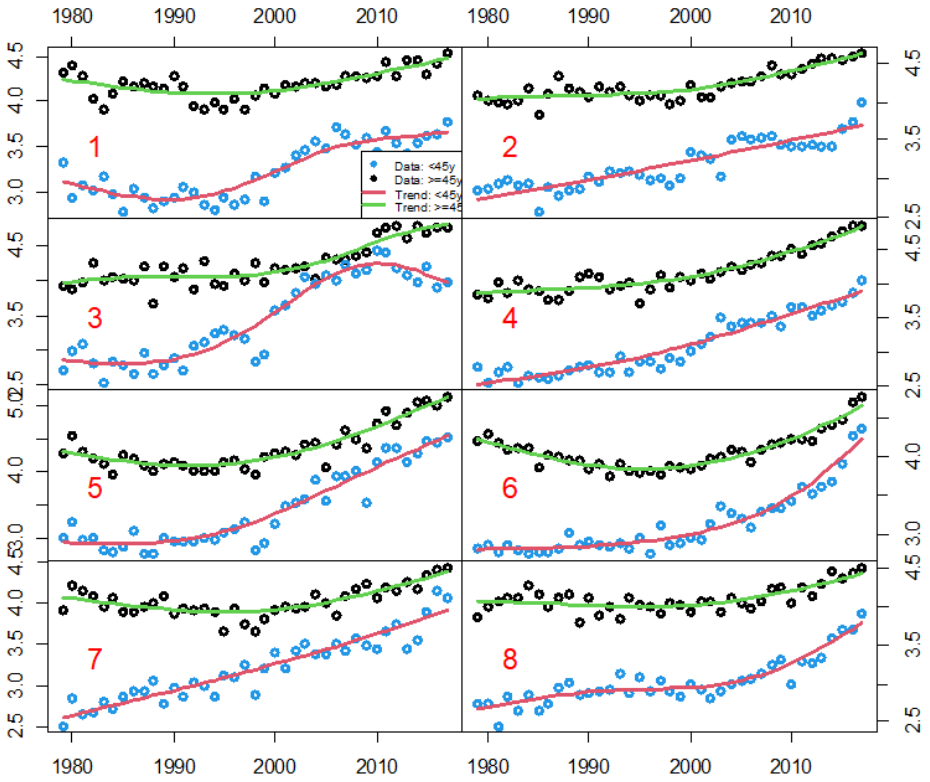


Figure 6: Longitudinal trends of SAD AAR for different age-groups in the 8 spatial metaclusters of AR. The older groups ($\geq$ 45 years: green curves) are compared with the younger groups (<45 years: red curves).

To gain insights into the socioeconomic conditions of each metacluster, the unemployment rates of the counties in the eight metaclusters for each 5-year period from 1990 to 2020, are presented as boxplots in Figure 10. Although their scale and range of variation differ across the metaclusters, the trends are similar *e.g.*, the lowering of unemployment rates between 2000 and 2005, and then again around 2015. Comparing these trends with those of the SAD AAR (Figure 9), we can clearly observe sharp increases in the gradient of the SAD AAR around the years of higher unemployment rate. For instance, in the year 1995, unemployment rates of metaclusters 3 (KY), 4 (South NC + SC + GA), 5 (TN), and 7 (Western PA) were very high, and around the same time, sharp upward shift in the trend of SAD AAR can be observed for these metaclusters in Figure 10. However, there are no apparent reduction in the gradient of the trends of SAD mortality rate during periods of lower unemployment rates, possibly hinting at deeper structural reasons for despair that may not be fully mitigated by employment alone.

The heightened problem of unemployment in the AR over the past few decades is generally associated with the steady decline in mining and manufacturing industries in the region. Historical trends of annual average number of jobs in these industries in the 8 metaclusters are plotted against the national average in Figure 11. Clearly, the decline in such jobs has continued over the past five decades for every metacluster, despite a brief recovery in mining during 2010-2015. This could be attributed to a wide array of factors ranging from demographic changes in terms of aging and migration to economic drivers such

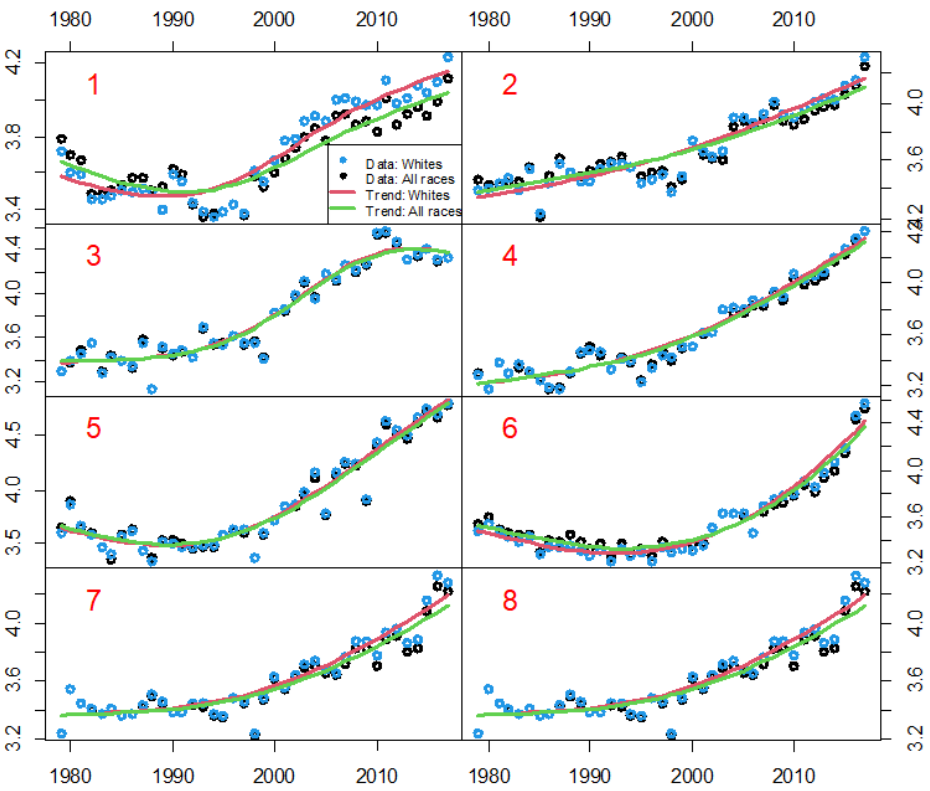


Figure 7: Longitudinal trends of SAD AAR for different races in the 8 spatial metaclusters of AR. The whites (red curves) are compared with the other races (green curves).

as outsourcing and global trade. Towards this, metacluster-wise median percentage change in population size, median age, and median household income, between 1980 and 2020 (Table 2) provide key insights into the shifts in the demographic and economic characteristics of each metacluster over the study-period. Interestingly, in 5 metaclusters, the population sizes have decreased between 5% and 29%. The same metaclusters, except for Western PA, are also characterized by higher rise in the median age of their population and much lower rise in their median household income as compared to the corresponding national change figures. The TN metacluster which has reported the highest SAD AAR in the last decade, has also seen the highest decline in population with the highest rise in median age and a decline in median household income.

Now we compare two metaclusters that are both from the same state, PA. The population of the Western PA metacluster decreased by 5% and its median age rose although lower than that of the U.S. Yet, its median household income has risen, in fact, by a greater percentage than the national average. This is likely to be driven by the urban sectors of the economy owing to Pittsburgh, the most prominent city in AR, in contrast to the ageing populations with limited opportunities for income generation among most of the other metaclusters that also have shrinking populations. Interestingly, it is also distinct from Eastern PA, which is one of the three metaclusters that have seen their populations rise. This metacluster has seen a 61% jump in population size and 51% increase in median household income, both of which are much higher than the national increments. These results not only

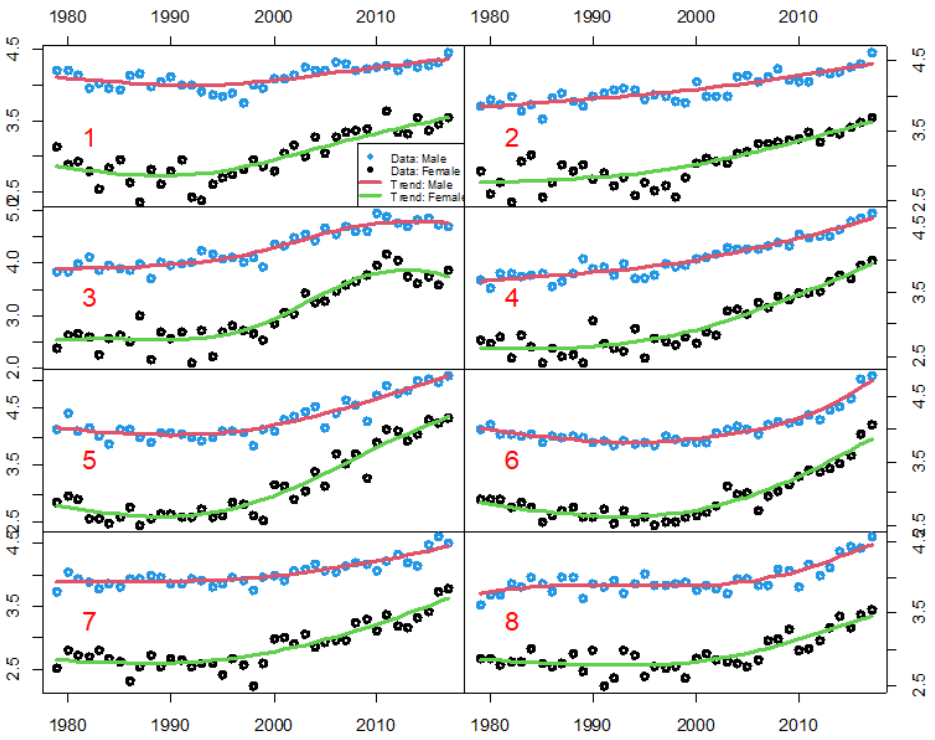


Figure 8: Longitudinal trends of SAD AAR for females (green curves) and males (red curves) in the 8 spatial metaclusters of AR.

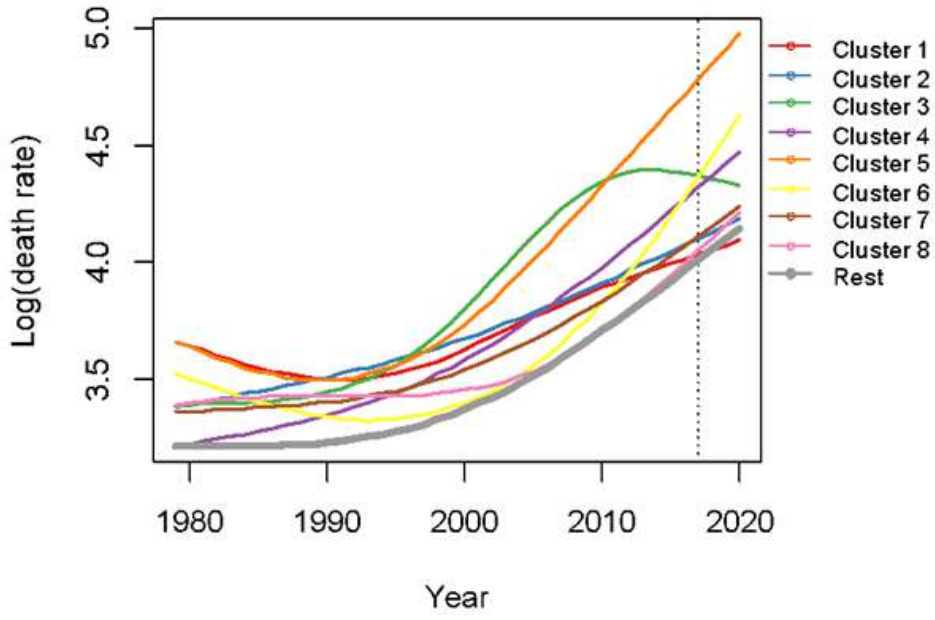


Figure 9: Trends of SAD AAR from 1979 to 2020 between 8 spatial metaclusters (as curves of different colors) are compared with the "Rest" of the Appalachian counties (bold grey curve). The predicted death rates for the time-period beyond 2017 are shown to the right of the dotted line.

showcase the distinct characteristics of these metaclusters even if they are from the same

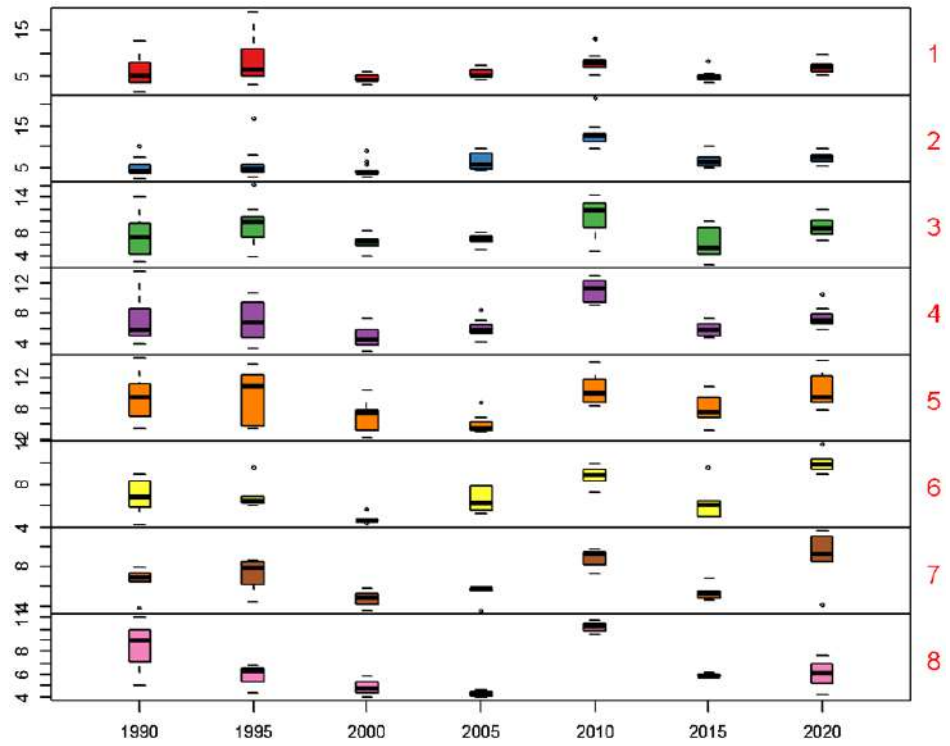


Figure 10: Unemployment rates (y-axis) over time (x-axis) are compared for the 8 spatial metaclusters (as labeled on the right) with boxplots for counties therein.

state, but also underscores how the complex problem of SAD deaths may not be adequately addressed merely by reducing poverty.

More granular insights can be derived from the county-level scatterplot of such changes in the demographic and economic parameters, upon grouping by the metaclusters, as shown in Figure 12. The inter-metacluster variation is more prominent in terms of the changes in population size. In general, counties with decline in population and steeper rise in median age have seen minimal rise in median household expenditure (smaller dots). Those with positive increases in population (appearing to the right-hand side of the dotted line) have witnessed rise in median household income (larger dots). As expected, higher rise in population is associated with lower rise in median age, but with some metaclusters that serve as notable exceptions. Moreover, we can observe intra-metacluster heterogeneity in terms of the changes in the demographic parameters and their interplay with household income. For example, some metaclusters have counties with moderate to high increases in their median age but notable rises in median household incomes. Such heterogeneities, within and across the metaclusters, may indicate the complexity underlying the phenomenon of SAD mortality, and underscore the need for investigating its social determinants at local community levels.

5. Discussion

Certain classic texts such as *The Other America* by Michael Harrington and *Night Comes to the Cumberlands* by Harry Caudill introduced Appalachian poverty to Americans during the early 1960s. The intense deprivation and hardships in AR led the then President

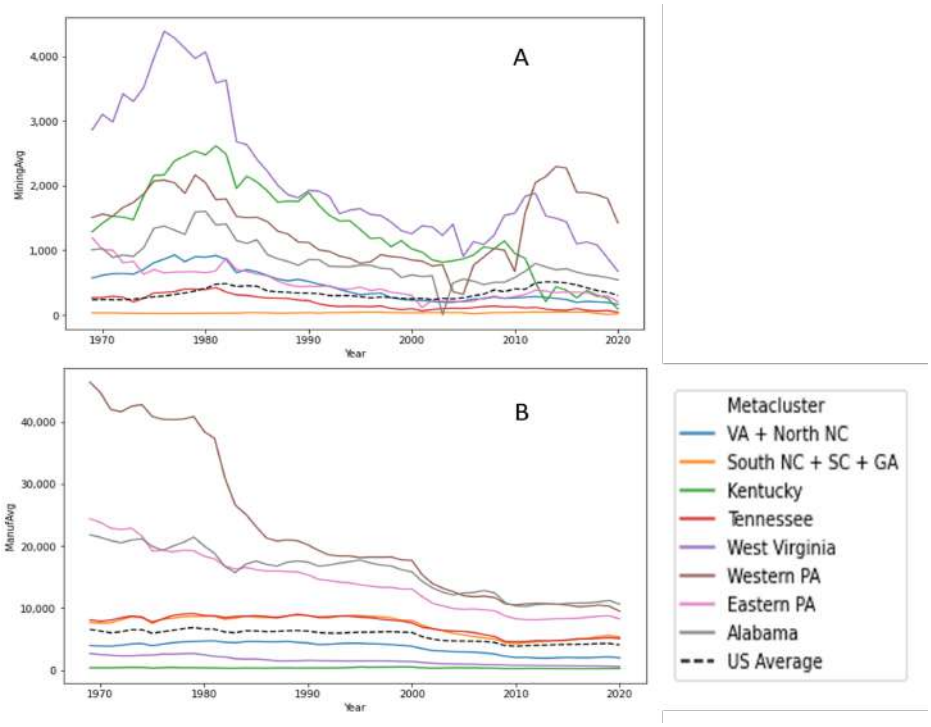


Figure 11: Longitudinal data on the average number of (A) mining and (B) manufacturing jobs in the 8 spatial metaclusters are shown (in different colors) against the national average (dotted).

John F. Kennedy to establish the President's Appalachian Regional Commission (PARC) in 1963. In its report, PARC categorically noted that "Appalachia is a region apart – both geographically and statistically" (PARC, 1964). In particular, AR "lags behind the rest of the Nation in its economic growth and that its people have not shared properly in the Nation's prosperity." In the subsequent decades, many steps have been taken towards reduction of the abject poverty in AR using various mechanisms, *e.g.*, the Appalachian Regional Development Act of 1965 (ARDA), which designated AR as a special economic zone and provided spending of more than $23 billion. Six decades later, the observable and compelling phenomenon of SAD deaths makes it vital for the researchers to analyze patterns of such dire yet disparate outcomes that have persisted in certain areas – and even evolved during the opioid epidemic – against the complex socioeconomic background of AR.

In rural communities, residents are more likely to work in physically demanding and injury-prone job sectors such as farms, factories, and mines, as compared to their urban counterparts. These place workers at increased risks for chronic pain and disability (Keyes *et al.*, 2014). Between 2015 and 2019, the share of Appalachian residents who reported a disability was 16.2% compared to 12.6% for the U.S. (Pollard and Jacobsen, 2021). Indeed, the prevalence of midlife pain epidemic in the U.S., which was highlighted by Case and Deaton (2015), exacerbated by the surge in the use of prescription (or otherwise) painkillers since the mid-1990s, has well-documented links to both addiction and SAD deaths in AR (Quinones, 2015). Not surprisingly, therefore, many Appalachian communities that are mining-dependent became targets for heavy marketing of Oxycodone and other strong prescription opioids much earlier than the rest of the country (Rigg *et al.*, 2018). Detailed patterns of such substance

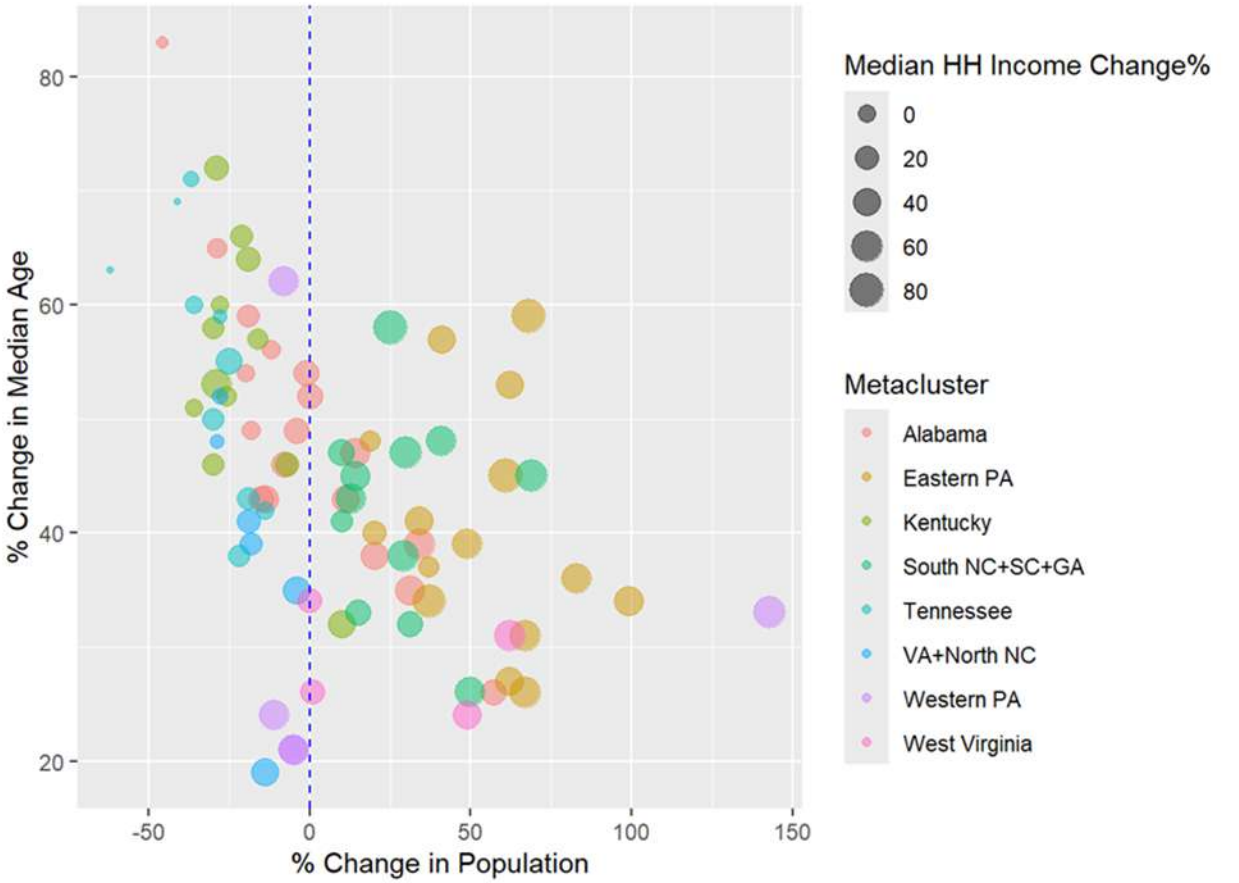


Figure 12: Percentage change in median age of the counties (grouped by their metacluster-specific colors) plotted against percentage change in their total population from 1980 to 2020. Size of the points is proportional to the percentage increase in the median household income of the counties.

and polysubstance uses among different population groups in the U.S. over the past 5 decades were identified by our previous studies based on NSDUH population surveys on substance use (Ray *et al.*, 2022).

In the present study, we identified 8 metaclusters of high rates of SAD deaths in AR over the period 1979-2017 based on U.S. county- and cause-specific mortality data. We observed patterns for each metacluster such as the dynamics for SAD mortality due to drug-overdoses, and its rising trends among the younger age group and women. We also noted heterogeneity among the metaclusters not only in terms of SES and rural/urban compositions but also their demographic and socioeconomic dynamics over the study-period. The patterns from our analysis showcase the need for further dissection of the data and covariates to detect the contributions of local vulnerabilities within each metacluster. For instance, the roles of such social determinants as poor public transport infrastructure that limits access to better jobs and healthcare, or sub-standard schools that may not prepare students for newer careers, cannot be ignored (George *et al.*, 2021). Based on our past use of subcounty-scale characteristics such as CDC Social Vulnerability Index, and techniques such as Small Area Estimation, we think further disaggregation of the SAD mortality data can lead to

a more nuanced understanding of despair in the diverse communities of AR, and thus aid public health and policy-making (Stacy *et al.*, 2023). Structural solutions at local levels can address issues involving strategies that may go beyond even poverty and unemployment.

We understand that our study has certain limitations. While different approaches of space-time clustering are known (*e.g.*, Knox test), the one that we adopted here is based on our intention to avoid the identified clusters from being necessarily temporally contiguous. Therefore, to allow for the occasional "ups and downs" in the SAD mortality rates within a cluster, we first clustered the AR counties in each successive 5-year window, and then used their recurrence for constructing the metaclusters. Further, to allow flexibly shaped clusters, we decided not to use scan statistics based on a circular window (say, due to Kulldorff), which have difficulty in correctly detecting irregularly shaped clusters that are more realistic. Instead, we applied the flexible spatial scan statistic of Tango and Takahashi (2005), which is able to detect a cluster of any shape reasonably well as its relative risk increases during the Monte Carlo simulation used in this approach.

Figure 13: Metacluster-wise median % change in population size, median age, and median household income (in USD). Data Source: U.S. Census Bureau.

Metacluster No.	Metacluster Name	Total population			Median Age of Population			Median Household Income		
		1980	2020	Median % Change	1980	2020	Median % Change	1979*	2021	Median % Change
1	Alabama (AL)	746676	850344	-6%	31.4	46.9	48%	38229	45519	29%
2	Eastern PA	789917	1230264	61%	32.5	47.1	39%	37790	51817	51%
3	Kentucky (KY)	383919	297485	-27%	27.8	42.9	55%	30189	34474	14%
4	South NC + SC + GA	805216	1057295	27%	30.8	44.2	44%	32730	48986	53%
5	Tennessee (TN)	670198	485884	-29%	29.1	44.6	57%	39926	40591	-1%
6	VA + North NC	2103297	1808600	-19%	32.8	46.0	40%	52830	58447	17%
7	Western PA	807375	858706	-5%	35.2	43.4	24%	39372	58645	54%
8	West Virginia (WV)	876509	934903	25%	31.5	40.2	29%	36554	53034	35%
US		226,542,250	331,449,281	46%	30	38.8	29%	47396	69,717	47%

*Converted to 2021 USD

Dedication

We dedicate this paper to the memory of the legendary statistician, the late Professor C.R. Rao (1920-2023). Following his retirement from the Indian Statistical Institute, Dr. Rao had had a second illustrious academic career in the U.S.; in particular, at the University of Pittsburgh and the Pennsylvania State University. Incidentally, both of these institutions are located in AR, the region of focus in the present study. As the quote at the beginning of the paper underscores, Dr. Rao had a profound interest in the use of statistics for policy-making and public health (Rao *et al.*, 2017a,b). While his work was extended by us to address some recent public health problems (*e.g.*, Guha *et al.* (2022)), his longevity provided

us with a direct historical connection to the classical past of statistics and its luminaries such as R.A. Fisher and P.C. Mahalanobis. One of the authors (SP) had the privilege of having Dr. Rao as a colleague at his eponymous institute in Hyderabad, India.

As a tribute to this trailblazing statistical scientist and an outstanding and prolific author as well as a wise and witty mentor to many, we echo the sentiment expressed in the Proceedings of the (U.S.) National Academy of Sciences earlier this year (DasGupta, 2024), "Goodbye, Dr. Rao. Thank you for your inspiration and guidance. We will remember you."

Acknowledgements

Initial parts of the work was conducted with support from the Public Health Dynamics Laboratory at University of Pittsburgh. The authors declare no conflicts of interest.

References

Appalachian Regional Commission (2019). Poverty rates in Appalachia 2013-2017. https://www.arc.gov/map/poverty-rates-in-appalachia-2013-2017/.

Bowen, E., Christiadi, D. J., and Lego, B. (2020). An overview of coal and the economy in Appalachia. Appalachian Regional Commission.

Camarda, C. G. (2012). Mortalitysmooth: An R package for smoothing poisson counts with P-splines. *Journal of Statistical Software*, **50**, 1–24.

Case, A. and Deaton, A. (2015). Rising morbidity and mortality in midlife among white non-Hispanic Americans in the 21st century. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 15078–15083.

Case, A. and Deaton, A. (2020). *Deaths of Despair and the Future of Capitalism*. Princeton University Press.

Case, A. and Deaton, A. (2021). Life expectancy in adulthood is falling for those without a BA degree, but as educational gaps have widened, racial gaps have narrowed. *Proceedings of the National Academy of Sciences of the United States of America*, **118**, e2024777118.

Christine, A. S., Lancaster, K. E., Gaynes, B. N., Wang, V., Pence, B. W., Miller, W. C., and Go, V. F. (2020). The opioid and related drug epidemics in rural Appalachia: A systematic review of populations affected, risk factors, and infectious diseases. *Substance Abuse*, **41**, 69–69.

DasGupta, A. (2024). C. R. Rao: Paramount statistical scientist (1920 to 2023). *Proceedings of the National Academy of Sciences*, **121**, e2321318121.

George, D., Snyder, B., Van Scoy, L., Brignone, E., and et al (2021). Perceptions of diseases of despair by members of rural and urban high-prevalence communities. *JAMA Network Open*, **4**, e2118134.

Guha, S., Petersen, A., Ray, S., and Pyne, S. (2022). On Rao's weighted distributions for modeling the dynamics of wildfires and air pollution. In Bapat, R. B., Karantha, M. P., Kirkland, S. J., Neogy, S. K., Pati, S., and Puntanen, S., editors, *Applied Linear Algebra, Probability and Statistics*, Indian Statistical Institute Series, pages 379–394. Springer.

Hedegaard, H., Curtin, S. C., and M, W. (2018a). Suicide mortality in the United States, 1999-2017. *NCHS Data Brief*, **330**, 1–8.

Hedegaard, H., Miniño, A. M., and Warner, M. (2018b). Drug overdose deaths in the United States, 1999-2017. *NCHS Data Brief*, **329**, 1–8.

Jalal, H., Buchanich, J. M., Roberts, M. S., Balmert, L. C., Zhang, K., and Burke, D. S. (2018). Changing dynamics of the drug overdose epidemic in the United States from 1979 through 2016. *Science*, **361**, aau1184.

Keyes, K. M., Cerda, M., Brady, J. E., Havens, J. R., and S, G. (2014). Understanding the rural-urban differences in nonmedical prescription opioid use and abuse in the United States. *American Journal of Public Health*, **104**, E52–E59.

Meit, M., Heffernan, M., and Tanenbaum, E. (2019). Investigating the impact of the diseases of despair in Appalachia. *Journal of Appalachian Health*, **1**, 1–18.

Meit, M., Heffernan, M., Tanenbaum, E., and Hoffman, T. (2017). Appalachian diseases of despair. The Walsh Center for Rural Health Analysis at NORC at the University of Chicago. https://www.arc.gov/wp-content/uploads/2020/06/AppalachianDiseasesofDespairAugust2017.pdf.

Monnat, S. M. (2020). Working-age non-Hispanic white mortality: Rural–urban and within-rural differences. *Population Research and Policy Review*, **39**, 805–834.

Otani, T. and Takahashi, K. (2021). Flexible scan statistics for detecting spatial disease clusters: The rflexscan R package. *Journal of Statistical Software*, **99**, 1–29.

PARC (1964). Appalachia: A report by the president's Appalachian commision. https://www.govinfo.gov/content/pkg/GOVPU-Y3_AP4_2-PURL-LPS99948/pdf/GOVPUB-Y3_AP4_2-PURL-LPS99948.pdf.

Pollard, K. and Jacobsen, L. A. (2021). The Appalachian region: A data overview from the 2015-2019 American community survey chartbook. The Appalachian Regional Commission. https://www.arc.gov/report/the-appalachian-region-a-data-overview-from-the-2015-2019-american-community-survey/.

Quinones, S. (2015). *Dreamland: The True Tale of America's Opiate Epidemic*. Bloomsbury Press.

Rao, A. S., Pyne, S., and Rao, C. R. (2017a). *Handbook of Statistics: Disease Modelling and Public Health, Part A*, volume 36. Elsevier.

Rao, A. S., Pyne, S., and Rao, C. R. (2017b). *Handbook of Statistics: Disease Modelling and Public Health, Part B*, volume 37. Elsevier.

Ray, S., Desai, M., and Pyne, S. (2022). Identifying patterns of association in the use of addictive substances over five decades in the United States. *Computers in Biology and Medicine*, **151**, 106175.

Rigg, K., Monnat, S. M., and Chavez, M. N. (2018). Opioid-related mortality in rural America: Geographic heterogeneity and intervention strategies. *International Journal of Drug Policy*, **57**, 119–129.

Shiels, M. S., Tatalovich, Z., Chen, Y., and et al (2020). Trends in mortality from drug poisonings, suicide, and alcohol-induced deaths in the United States from 2000 to 2017. *JAMA Network Open*, **3**, e2016217.

Spillane, S., Shiels, M. S., Best, A. F., Haozous, E. A., and et al (2020). Trends in alcohol related deaths in the United States, 2000-2016. *JAMA Network Open*, **3**, e1921451.

Stacy, S., Chandra, H., Guha, S., Gurewitsch, R., Brink, L., Robertson, L., Wilson, D., Yuan, J.-M., and Pyne, S. (2023). Re-scaling and small area estimation of behavioral risk survey guided by social vulnerability data. *BMC Public Health*, **23**, 184.

Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, **4**, 1–15.

NACo and ARC (2019). Opioids in Appalachia. the role of counties in reversing a regional epidemic. https://www.naco.org/sites/default/files/documents/opioids-full.pdf.

U.S. Department of Health and Human Services (2023). What is the opioid epidemic? https://www.hhs.gov/opioids/about-the-epidemic/index.html.

Wallace, M., Sharfstein, J. M., Kaminsky, J., and Lessler, J. (2019). Comparison of U.S. county-level public health performance rankings with county cluster and national rankings: Assessment based on prevalence rates of smoking and obesity and motor vehicle crash death rates. *JAMA Network Open*, **2**, e186816.

Woolf, S. H., Chapman, D. A., M, B. J., Bobby, K. J., Zimmerman, E. B., and Blackburn, S. M. (2018). Changes in midlife death rates across racial and ethnic groups in the United States: systematic analysis of vital statistics. *BMJ*, **362**, k3096.

Woolf, S. H., Schoomaker, H., Hill, L., and Orndahl, C. M. (2019). The social determinants of health and the decline in U.S. life expectancy: implications for Appalachia. *Journal of Appalachian Health*, **1**, 6–14.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 575–594
http://www.ssca.org.in/journal

A New Unit Root Test for an Autoregressive Model Subject to Measurement Errors

Weerapat Rattanachadjan[1], Jiraphan Suntornchost[1] and Partha Lahiri[2]
[1]*Department of Mathematics and Computer Science,*
Faculty of Science, Chulalongkorn University, Thailand
[2]*Joint Program in Survey Methodology, and Department of Mathematics,*
University of Maryland, College Park, USA.

Received: 13 July 2024; Revised: 26 September 2024; Accepted: 8 October 2024

Abstract

The unit root test – a test of the null hypothesis that a first-order autoregressive model is a random walk model against the alternative hypothesis that the model is a stationary model - has played a significant role in time series literature. The benchmark unit root test is the well-known Dickey-Fuller test widely extended to cover a variety of applications. However, to the best of our knowledge, all available unit root tests assume no measurement errors in the observed data. In this paper, we first investigate the effects of sampling errors, alternatively called as measurement errors, on the biases of the commonly used estimators of autocorrelation coefficient and the Dickey-Fuller test statistics. We then propose alternative estimators for the autocorrelation coefficient and the Dickey-Fuller test statistics to reduce such biases due to sampling errors. In our study, we prove that the adjusted estimators of the autocorrelation coefficient and the test statistics have the same asymptotic distributions as that of the Dickey-Fuller test statistics. Moreover, we conduct Monte Carlo simulation studies to investigate the performance of our proposed test statistics in terms of unbiasedness, the probability of Type-I error, and power of the test. Our simulation results demonstrate that the proposed estimators can reduce bias due to sampling errors. Finally, we apply the proposed test statistics to the Current Population Survey (CPS) data on unemployment of the United States during the period 1990 - 2013.

Key words: Unit root; Autoregressive coefficient; Sampling errors; Measurement errors; Likelihood ratio.

AMS Subject Classifications: 62F10, 62F12 , 62H20, 62M10

1. Introduction

Measurement errors in time series data occur in different applications of ecology, economics, finance, repeated surveys, and other disciplines. In ecological research, Shenk *et al.* (1998) introduced the concept of sampling errors in the form of measurement errors. Specifically, they investigated the effects of sampling variances on the first-order autoregressive

Corresponding Author: Jiraphan Suntornchost
Email: Jiraphan.s@chula.ac.th

population models in order to estimate population abundance. The concept was then studied in the context of time series population models such as the ones given in De Valpine and Hastings (2002), Dennis *et al.* (2006), Buonaccorsi and Staudenmayer (2009).

In Economics and Finance, Walters and Ludwig (1981) studied effects of measurement errors on the estimation of stock-recruitment relationships. Moreover, they obtained estimates of measurement errors. Besides the applications in stock markets, the measurement errors in time series data were also considered in other applications such as the U.K. GDP (Smith *et al.*, 1998) and the U.S. GDP (Aruoba *et al.*, 2016).

Time series data with measurement errors also occur in the context of repeated surveys where the actual characteristics of interest are usually not observed but are estimated by survey direct estimates. The problem was first considered in Scott and Smith (1974) where the authors considered an autoregressive time series model with sampling errors. The study was then further pursued by many researchers, such as Scott *et al.* (1977), Bell and Hillmer (1990) Ludwig and Walters (1981), Bell and Hillmer (1990), Staudenmayer and Buonaccorsi (2005), Rossi and Santucci de Magistris (2018).

Beside parameter estimation, one crucial tool for autoregessive time series analysis is the test of unit root. The benchmark unit root test was introduced by Dickey and Fuller (1979), where they obtained the test statistic and derived the asymptotic distribution of their test statistic under the null hypothesis of unit root. The test has been widely extended to higher order time series models and applied in many contexts during the last few decades. However, the test statistic was originally designed for real-time series data without accounting for sampling errors commonly found in repeated survey data. Ignoring sampling errors could cause biases to the test statistic and lead to a wrong conclusion of the unit root test in the presence of sampling errors. Therefore, to avoid such biases, effects of sampling errors to the unit root test deserve investigation and an effective adjustment to the test statistics is required. However, to the best of our knowledge, there is no unit root test for time series data with measurement errors available in literature.

In this paper, we investigate the effect of sampling errors on the unit root test of Dickey and Fuller (1979). Our study suggests that ignoring sampling errors could cause biases in the estimation of autocorrelation coefficient and the Dickey-Fuller unit root test statistics. Thus, we propose a modification of the Dickey-Fuller test that is bias-corrected for sampling errors. We derive its asymptotic properties, and conduct Monte Carlo simulation studies to investigate the performance of our proposed method by considering the unbiasedness, the probability of Type-I error, and the power of the test. Moreover, we apply the proposed test statistics to the Current Population Survey (CPS) data on unemployment of the United States during the period 1990 to 2013. The numerical results demonstrate that the new test can reduce the bias of the original Dickey-Fuller test when there is a present of sampling errors.

The organization of this paper is as follows. In Section 2, we review the Dickey-Fuller unit root test statistic for the first-order autoregressive model. In Section 3, we propose an adjusted estimate of the Dickey-Fuller unit root in the presence of sampling errors. In Section 4, we demonstrate Monte Carlo simulations to study the performance of the proposed test statistic in different aspects such as bias, probability of Type-I error, and power of the test. In Section 5, we apply the proposed test statistic to the Current Population Survey (CPS)

data on unemployment of the United States during the period 1990 to 2013. In Section 6, we offer some concluding remarks. Finally the proofs of theoretical properties of the proposed test statistic and important lemmas are provided in Section 7.

2. Unit root test for AR(1) model

Consider the first order autoregressive model for the time series $\{Y_t : t = 1, 2, \ldots, T\}$, defined as

$$Y_t = \rho Y_{t-1} + e_t, \tag{1}$$

where ρ is the regression coefficient and $\{e_t\}$ is a sequence of independent normal random variables with mean zero and unknown variance σ_e^2.
The least squares estimate $\hat{\rho}_Y$ of the autocorrelation coefficient ρ is defined as

$$\hat{\rho}_Y = \frac{S_{Y,T}(1)}{S_{Y,T}(0)}, \tag{2}$$

where $S_{Y,T}(k) = \sum_{t=2}^{T} Y_{t-1} Y_{t+k-1}$.

Dickey and Fuller (1979) constructed the unit root test statistic under the null hypothesis that $\rho = 1$ as

$$\hat{\tau} = \frac{(\hat{\rho}_Y - 1)\sqrt{\sum_{t=1}^{T} Y_t^2}}{\sqrt{\hat{\sigma}^2}}, \tag{3}$$

where

$$\hat{\sigma}^2 = \frac{1}{T-2} \sum_{t=2}^{T} (Y_t - \hat{\rho}_Y Y_{t-1})^2.$$

Moreover, they obtained the asymptotic distribution of $\hat{\rho}_Y$ as

$$T(\hat{\rho}_Y - 1) \xrightarrow{d} \frac{\left(\sum_{i=1}^{\infty} \sqrt{2}\gamma_i Z_i\right)^2 - 1}{2\sum_{i=1}^{\infty} \gamma_i^2 Z_i^2},$$

where $Z_i \overset{iid}{\sim} N(0,1)$ and $\gamma_i = (-1)^{i+1} \dfrac{2}{(2i-1)\pi}$.

Consequently, the asymptotic distribution of the test statistic $\hat{\tau}$ is obtained as

$$\hat{\tau} \xrightarrow{d} \frac{\left(\sum_{i=1}^{\infty} \sqrt{2}\gamma_i Z_i\right)^2 - 1}{2\sqrt{\sum_{i=1}^{\infty} \gamma_i^2 Z_i^2}}. \tag{4}$$

3. Unit root test for AR(1) with measurement errors

In this section, we consider the model in (1) when the actual time series $\{Y_t : t = 1, 2, \ldots, T\}$ is unobserved but its predicted value from a survey $\{W_t : t = 1, 2, \ldots, T\}$ can be obtained. Specifically, the model considered in this section consists of two sub-models: the autoregressive model for the actual time series defined in (1) and the sampling model assuming that the observed value can be written as a sum of the actual value and a sampling error. In particular, the sampling model is

$$W_t = Y_t + u_t, \tag{5}$$

where $\{W_t : t = 1, 2, \ldots, T\}$ is the sequence of observed variables with $W_0 = 0$ and $\{u_t : t = 1, 2, \ldots, T\}$ is the sequence of sampling errors assumed to be independently normally distributed with mean zero and known variances σ^2_{ut}. The assumption of known sampling variances σ^2_{ut} often follows from the asymptotic variances of transformed direct designed-based estimates such as in Efron and Morris (1975), Carter and Rolph (1974), Lahiri and Suntornchost (2015), and Marhuenda García *et al.* (2016).

To construct an adjustment of the unit root test, we first investigate the effect of ignoring the sampling errors to the estimations of the autocorrelation coefficient and the Dickey- Fuller unit root test statistic. By substituting Y_t with the survey estimate W_t in (2), the naive estimate of the autocorrelation coefficient is

$$\hat{\rho}_W = \frac{S_{W,T}(1)}{S_{W,T}(0)}$$

and the naive test statistic is

$$\hat{\tau}_{naive} = \frac{(\hat{\rho}_W - 1)\sqrt{S_{W,T}(0)}}{\sqrt{\hat{\sigma}^2_{W,e}}}, \tag{6}$$

where

$$\hat{\sigma}^2_{W,e} = \frac{1}{T-2}\sum_{t=2}^{T}(W_t - \hat{\rho}_W W_{t-1})^2.$$

Applying the conditional expectation, we found that

$$\mathbb{E}\left(S_{W,T}(0)|Y_t\right) = \sum_{t=2}^{T} Y^2_{t-1} + \sum_{t=2}^{T}\sigma^2_{u,t-1},$$
$$\mathbb{E}\left(S_{W,T}(1)|Y_t\right) = \sum_{t=2}^{T} Y_t Y_{t-1}.$$

Therefore, by applying the first order Taylor series approximation, we can show that the naive estimator of the autocorrelation coefficient, $\hat{\rho}_W$, is asymptotically biased and then the estimator is not reliable. Hence, following Lahiri and Suntornchost (2015), we propose an adjustment to each component in $\hat{\rho}_W$ by removing the biases of $S_{W,T}(0)$ and $S_{W,T}(1)$. Therefore, the proposed estimate of the autoregressive coefficient ρ is defined as

$$\hat{\rho}_{Adj} = \frac{S_{W,T}(1)}{\widetilde{S}_{W,T}(0)},$$

where $\widetilde{S}_{W,T}(0) = S_{W,T}(0) - S_{\sigma_u}(0)$, and $S_{\sigma_u}(0) = \sum_{t=2}^{T} \sigma^2_{u,t-1}$. Applying the first order Taylor series approximation, we prove in Theorem 1 that

$$\hat{\rho}_{Adj} - \hat{\rho}_Y = o_p(1), \tag{7}$$

under the assumption $\rho = 1$.
Moreover, we show in Theorem 2 that $T(\hat{\rho}_{Adj} - 1)$ has the same asymptotic distribution as $T(\hat{\rho}_Y - 1)$. In particular,

$$T(\hat{\rho}_{Adj} - 1) \xrightarrow{d} \frac{\left(\sum_{i=1}^{\infty} \sqrt{2}\gamma_i Z_i\right)^2 - 1}{2\sum_{i=1}^{\infty} \gamma_i^2 Z_i^2},$$

where $Z_i \overset{iid}{\sim} N(0,1)$ and $\gamma_i = (-1)^{i+1} \dfrac{2}{(2i-1)\pi}$.

Furthermore, we construct an adjusted estimate for σ^2 subject to sampling errors, defined as

$$\hat{\sigma}^2_{Adj,e} = |\hat{\sigma}^2_{W,e,1} - \hat{\sigma}^2_{W,e,2}|, \tag{8}$$

where

$$\hat{\sigma}^2_{W,e,1} = \frac{1}{T-2} \sum_{t=2}^{T} (W_t - \hat{\rho}_{Adj} W_{t-1})^2,$$

and

$$\hat{\sigma}^2_{W,e,2} = \frac{1}{T-2} \sum_{t=2}^{T} \left(\sigma^2_{u,t} + \hat{\rho}^2_{Adj} \sigma^2_{u,t-1}\right).$$

Then, we propose an adjusted test statistic for the unit root test of the first order autoregressive model subject to measurement errors defined as

$$\hat{\tau}_{Adj} = \frac{(\hat{\rho}_{Adj} - 1)\sqrt{\widetilde{S}_{W,T}(0)}}{\sqrt{\hat{\sigma}^2_{Adj,e}}}. \tag{9}$$

Moreover, we prove in Theorem 3 that the proposed test statistic has the same asymptotic distribution as the true estimate $\hat{\tau}_Y$. In particular,

$$\hat{\tau}_{Adj} \xrightarrow{d} \frac{\left(\sum_{i=1}^{\infty} \sqrt{2}\gamma_i Z_i\right)^2 - 1}{2\sqrt{\sum_{i=1}^{\infty} \gamma_i^2 Z_i^2}}, \tag{10}$$

where $\gamma_i = (-1)^{i+1} \frac{2}{(2i-1)\pi}$ and $Z_i \overset{iid}{\sim} N(0,1)$.

4. Monte carlo simulations

In this section, we conduct Monte Carlo simulations to study the performance of the proposed test statistic compared to the naive test that ignores sampling errors. For our simulation experiment, we set the true sampling variances of u_t, σ_{ut}^2 in model (5), by using estimated variances of 288 monthly survey-weighted direct estimates of the number of unemployed workers obtained from the U.S. Current Population Survey (CPS) conducted during the period 1990 - 2013. There are 12 simulation settings based on four selected states with different ranges of sampling standard deviations and three different values of the regression standard deviation σ_e of the autoregressive model (1). The values of σ_e are specified by the ratio $k = \dfrac{\bar{\sigma}_u}{\sigma_e}$ where $\bar{\sigma}_u$ is the average of sampling standard deviations defined as $\bar{\sigma}_u = T^{-1}\sum_{t=1}^{T}\sigma_{ut}$. The three values of k considered are $0.75, 1$, and 1.25 representing the cases where the average of standard deviations of sampling errors is smaller than, equal to, and larger than the regression standard deviation, respectively. In addition, we consider four different lengths (T) of time series, $T \in \{25, 50, 100, 250\}$, to study asymptotic behaviours of the test statistics. Each setting is repeated for 20,000 simulation runs. In particular,the steps of simulation are as follows.

1. For each combination of state and k, calculate the regression variance σ_e^2 from $\sigma_e = \dfrac{\bar{\sigma}_u}{k}$.
2. For each simulation setting and each $l = 1, 2, \ldots, 20,000$,
   (a) generate the variance components and sampling errors $\{(u_t^{(l)}, e_t^{(l)}) : t = 1, 2, \ldots, 250\}$,
   (b) calculate the time series $\{Y_t^{(l)} : t = 1, 2, \ldots, 250\}$, from model (1) with $\rho = 1$,
   (c) generate $\{W_t^{(l)} : t = 1, 2, \ldots, 250\}$ from model (5),
   (d) calculate $\hat{\tau}_{true}^{(l)}$, $\hat{\tau}_{naive}^{(l)}$, and $\hat{\tau}_{Adj}^{(l)}$ from the fomula in (3), (6), and (9), respectively.

To study the performances of the test statistics, we first consider different percentiles of the estimated test statistics and the estimated values of the probability of Type-I error. The estimates of the test statistics in different percentiles by using data from one selected state, State 3, are presented in Tables 1 - 3, respectively for the cases of $k = 0.75, 1$, and 1.25.

From Tables 1 - 3, we can see that the percentiles of the true test statistics and the proposed test statistic are close together, particularly those values between the 10th and 90th percentiles. In contrast, the naive test statistics are much lower than the true estimates in all cases. These results suggest that the naive estimator of the Dickey-Fuller test statistic underestimates the true test statistic, while the proposed estimator can reduce such underestimation.

Next, we consider the accuracy of the estimated probability of Type-I error, computed as the portion of the number of replications in which the unit root hypothesis is rejected when the actual time series is generated from the true autoregressive model (1) with $\rho = 1$. In particular, the estimated probability of Type-I error is computed as

$$\hat{\alpha} = \frac{1}{L}\sum_{l=1}^{L} \mathbb{1}_{\{\hat{\tau}_\cdot^{(l)} \text{ reject } H_0\}},$$

Table 1: The empirical percentiles of the different test statistics for $k = 0.75$

Length (T)	Statistics	Percentiles						
		1	10	25	50	75	90	99
$T = 25$	$\hat{\tau}_{true}$	-2.58	-1.61	-1.06	-0.51	0.21	0.87	2.28
	$\hat{\tau}_{naive}$	-3.88	-2.40	-1.65	-0.98	-0.34	0.25	1.24
	$\hat{\tau}_{Adj}$	-3.41	-1.78	-1.07	-0.46	0.32	1.24	4.16
$T = 50$	$\hat{\tau}_{true}$	-2.60	-1.68	-1.11	-0.53	0.22	0.90	2.08
	$\hat{\tau}_{naive}$	-3.91	-2.58	-1.82	-1.06	-0.36	0.20	1.10
	$\hat{\tau}_{Adj}$	-3.10	-1.78	-1.11	-0.47	0.31	1.17	3.21
$T = 100$	$\hat{\tau}_{true}$	-2.65	-1.61	-1.09	-0.54	0.23	0.86	2.06
	$\hat{\tau}_{naive}$	-3.87	-2.48	-1.76	-1.05	-0.37	0.19	1.13
	$\hat{\tau}_{Adj}$	-2.64	-1.65	-1.08	-0.48	0.27	0.97	2.41
$T = 250$	τ_{true}	-2.69	-1.62	-1.12	-0.55	0.19	0.87	2.16
	$\hat{\tau}_{naive}$	-3.88	-2.46	-1.78	-1.09	-0.38	0.22	1.12
	$\hat{\tau}_{Adj}$	-2.56	-1.64	-1.10	-0.54	0.21	0.91	2.20

Table 2: The empirical percentiles of the different test statistics for $k = 1$

Length (T)	Statistics	Percentiles						
		1	10	25	50	75	90	99
$T = 25$	$\hat{\tau}_{true}$	-2.68	-1.65	-1.09	-0.54	0.16	0.92	2.16
	$\hat{\tau}_{naive}$	-4.25	-2.81	-2.04	-1.29	-0.59	0.03	0.94
	$\hat{\tau}_{Adj}$	-3.87	-2.04	-1.25	-0.54	0.27	1.26	5.48
$T = 50$	$\hat{\tau}_{true}$	-2.63	-1.70	-1.15	-0.56	0.18	0.84	2.22
	$\hat{\tau}_{naive}$	-4.59	-3.05	-2.25	-1.41	-0.68	-0.09	0.77
	$\hat{\tau}_{Adj}$	-3.50	-1.85	-1.16	-0.51	0.29	1.20	4.94
$T = 100$	$\hat{\tau}_{true}$	-2.52	-1.67	-1.13	-0.57	0.15	0.84	1.88
	$\hat{\tau}_{naive}$	-4.49	-3.02	-2.19	-1.40	-0.67	-0.11	0.67
	$\hat{\tau}_{Adj}$	-2.95	-1.71	-1.10	-0.50	0.23	1.02	3.23
$T = 250$	τ_{true}	-2.58	-1.62	-1.11	-0.51	0.21	0.86	2.04
	$\hat{\tau}_{naive}$	-4.46	-2.95	-2.15	-1.34	-0.64	-0.07	0.86
	$\hat{\tau}_{Adj}$	-2.75	-1.65	-1.09	-0.49	0.24	0.97	2.68

where $\mathbb{1}_{\{\hat{\tau}_\cdot^{(l)} \text{ reject } H_0\}}$ is equal to 1 if the specific test statistic $\hat{\tau}_\cdot^{(l)} \in \{\hat{\tau}_{true}, \hat{\tau}_{Adj}, \hat{\tau}_{naive}\}$ rejects $\rho = 1$, and is equal to 0 for otherwise. The results for the tests with significance level 0.05 are presented in Table 4 as follows. From Table 4, we can see that the estimated probabilities of Type-I error of the true test statistic $\hat{\tau}_{true}$ and the proposed test statistic $\hat{\tau}_{adj}$ are approximately 0.05 in all cases. In contrast, the naive test statistic $\hat{\tau}_{naive}$ produces estimated probabilities of Type-I error different from 0.05 for all cases. Specifically, the values are approximately 0.2, 0.3, and 0.4 for the cases corresponding to $k = 0.75$, 1, and 1.25, respectively. This result suggests that the bias of the estimated probability of Type-I error obtained from the naive test statistic is higher when the sampling variance is higher. Moreover, the naive test statistic gives different conclusions from the actual test statistic. In contrast, our proposed test provides the same conclusion as the true test even with the large values of sampling variances.

Finally, we investigate the performance of the proposed test regarding the estimation

Table 3: The empirical percentiles of the different test statistics for $k = 1.25$

Length (T)	Statistics	Percentiles						
		1	10	25	50	75	90	99
$T = 25$	$\hat{\tau}_{true}$	-2.58	-1.62	-1.06	-0.47	0.23	0.86	2.17
	$\hat{\tau}_{naive}$	-4.09	-2.45	-1.71	-1.00	-0.35	0.25	1.24
	$\hat{\tau}_{Adj}$	-3.58	-1.76	-1.05	-0.44	0.34	1.33	4.53
$T = 50$	$\hat{\tau}_{true}$	-2.64	-1.61	-1.10	-0.49	0.18	0.87	1.89
	$\hat{\tau}_{naive}$	-4.05	-2.50	-1.81	-1.08	-0.44	0.10	0.93
	$\hat{\tau}_{Adj}$	-3.14	-1.68	-1.06	-0.44	0.31	1.14	3.16
$T = 100$	$\hat{\tau}_{true}$	-2.59	-1.59	-1.08	-0.50	0.20	0.87	2.00
	$\hat{\tau}_{naive}$	-3.84	-2.51	-1.76	-1.06	-0.42	0.16	0.91
	$\hat{\tau}_{Adj}$	-2.65	-1.59	-1.07	-0.46	0.20	1.03	2.34
$T = 250$	$\hat{\tau}_{true}$	-2.62	-1.59	-1.10	-0.50	0.22	0.90	1.98
	$\hat{\tau}_{naive}$	-3.81	-2.50	-1.80	-1.07	-0.37	0.19	0.99
	$\hat{\tau}_{Adj}$	-2.59	-1.61	-1.09	-0.49	0.22	0.96	2.21

Table 4: The empirical estimates of Type-I error

	Values of the ratio k								
	$k = 0.75$			$k = 1$			$k = 1.25$		
	$\hat{\tau}_{true}$	$\hat{\tau}_{naive}$	$\hat{\tau}_{Adj}$	$\hat{\tau}_{true}$	$\hat{\tau}_{naive}$	$\hat{\tau}_{Adj}$	$\hat{\tau}_{true}$	$\hat{\tau}_{naive}$	$\hat{\tau}_{Adj}$
State 1	0.0490	0.2090	0.0495	0.0450	0.3092	0.0485	0.0422	0.4078	0.0492
State 2	0.0450	0.1955	0.0470	0.0445	0.2895	0.0410	0.0511	0.4099	0.0656
State 3	0.0480	0.2010	0.0465	0.0485	0.3210	0.0535	0.0532	0.4104	0.0572
State 4	0.0475	0.1955	0.0455	0.0550	0.2915	0.0565	0.0473	0.4031	0.0488

of the power of the test for different values of the autocorrelation coefficient ρ, varying in the set $\{0.85, 0.9,\ 0.95, 0.975,\ 0.99, 0.995\}$. The simulation setting in this post is the same as previous algorithm except in the step 2(b), instead of using the data with a unit root, the time series $\{Y_t^{(l)} : t = 1, 2, \ldots, 250\}$, is generated from model (1) with specific $\rho = \rho_0$, where $\rho_0 \in \{0.85, 0.9,\ 0.95, 0.975,\ 0.99, 0.995\}$. The numerical results of the estimated power functions of the true test statistic $\hat{\tau}_{true}$ and the proposed test statistic $\hat{\tau}_{Adj}$ for $k = 0.75, 1, 1.25$ are presented in Figures 1-3, respectively.

From Figures 1-3, we can see that the estimated powers of the two tests are lower when the true value of ρ gets closer to one. The powers of the proposed test are close to the powers of the true test. These results suggest that the proposed test performs well in terms of the power of the test.

5. Applications

In this section, we apply the proposed test statistic to the CPS survey data of the four selected states, comparing with the naive test statistic ignoring sampling errors. Numerical results including the test statistics with their associated probabilities of Type-I errors are presented in Table 5.

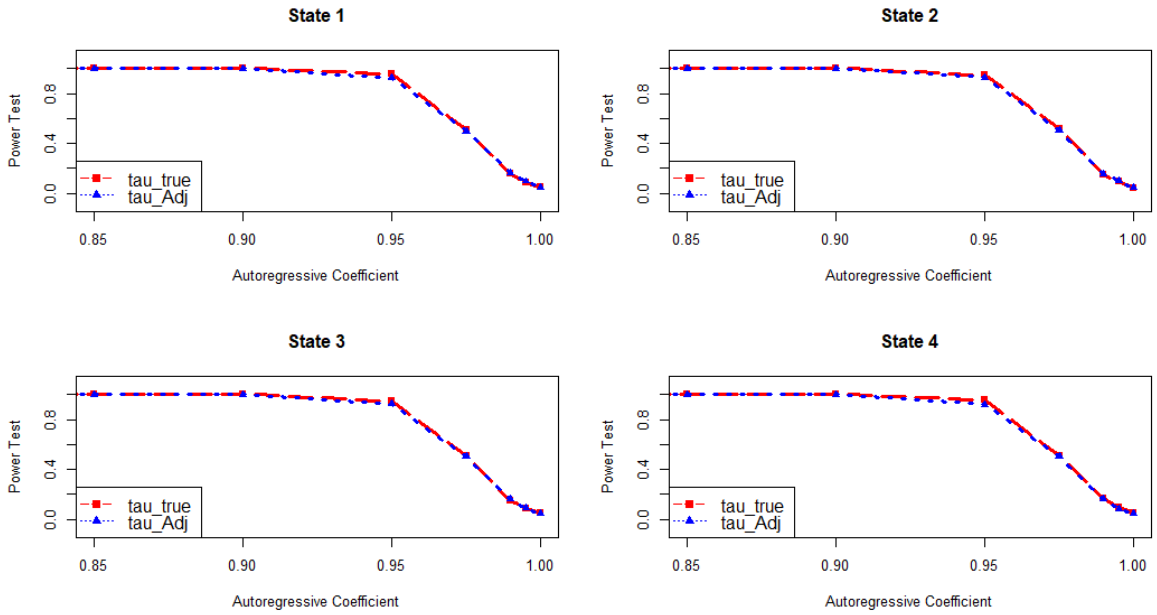


Figure 1: Empirical estimates of the power for $k = 0.75$

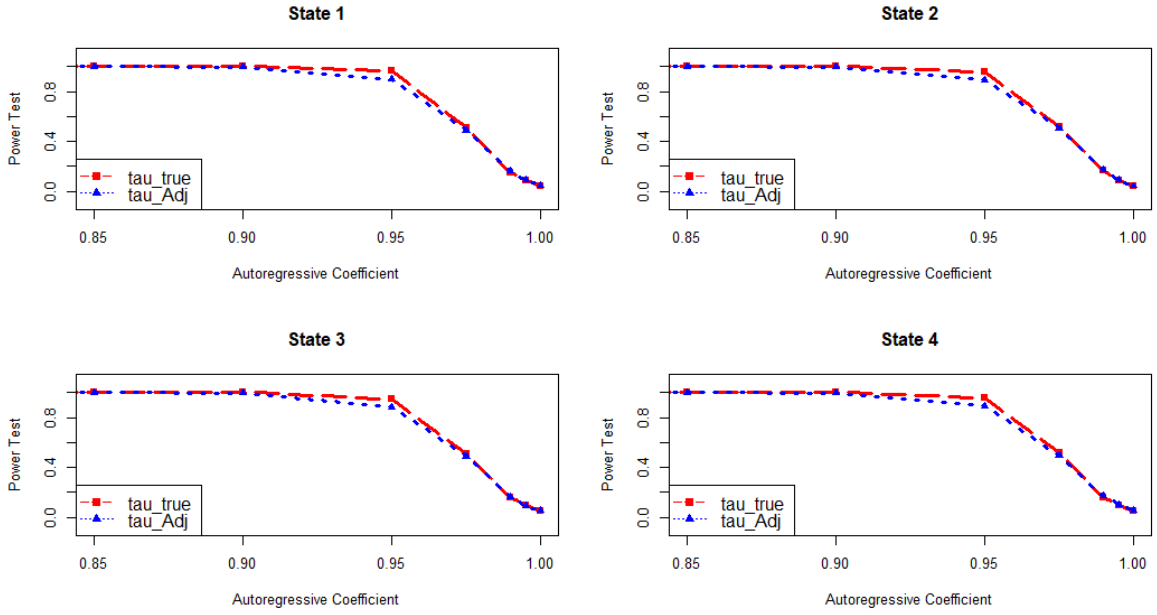


Figure 2: Empirical estimates of the power for $k = 1$

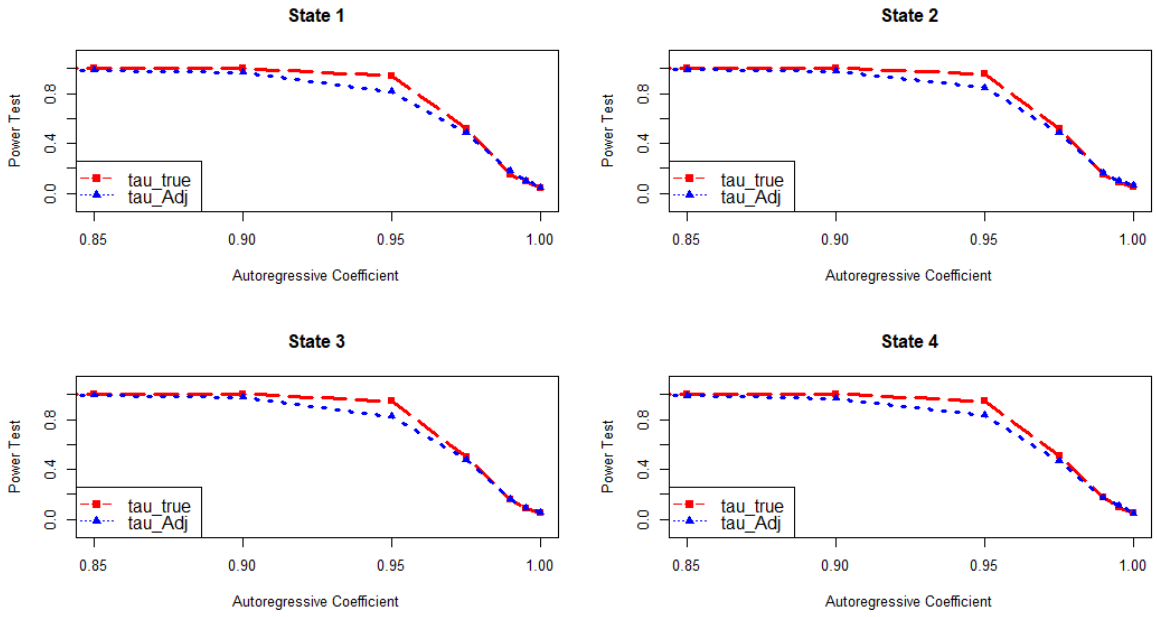


Figure 3: Empirical estimates of the power for $k = 1.25$

Table 5: The estimated test statistics and the corresponding p-values for four selected states

	$\hat{\tau}_{naive}$		$\hat{\tau}_{Adj}$	
	Calculated test Statistic	p-value	Calculated test Statistic	p-value
State 1	-6.59	$< 1 \times 10^{-4}$	-1.32	0.17
State 2	-4.89	$< 1 \times 10^{-4}$	-0.86	0.35
State 3	-7.90	$< 1 \times 10^{-4}$	-1.51	0.12
State 4	-4.18	$< 1 \times 10^{-4}$	-0.76	0.39

From Table 5, we observe the same behavior of the two estimates as the simulation results presented in Tables 1 – 3. In particular, the naive test provides much lower values of the test statistic than the proposed test statistics. The naive test statistics for the four states reject the null hypothesis and conclude that the time series are stationary. In contrast, the proposed test provides larger values of the p-values than 0.01 in all cases. Therefore, the proposed test suggests that the actual time series have a unit root at the significant level 0.01.

6. Conclusions and discussions

In this paper, we investigated the effects of sampling errors on the commonly used autocorrelation coefficient estimator and the well-known Dickey-Fuller unit root test statistic. We found that ignoring sampling errors could cause biases in the estimations of the correlation coefficient and the test statistic. This will lead to a wrong conclusion of the unit root test. Therefore, in our study, we introduced a new autocorrelation coefficient estimator and a unit root test statistic in order to reduce biases caused by sampling errors. Moreover, we obtained asymptotic distributions of our proposed estimator $\hat{\rho}_{Adj}$ and the proposed test

statistic $\hat{\tau}_{Adj}$ and showed that the two estimators have the same asymptotic distributions as of the estimators without measurement errors. Furthermore, we conducted simulation studies and applied the proposed method to real data. Numerical results suggested that our proposed method have good performances in terms of bias reduction, the accuracies of the estimated probability of Type-I error and the estimated power of the unit root test.

Acknowledgements

The authors would like to thank the editor and the referees for all valuable comments that improve the quality of the manuscript. The first author is supported by Development and Promotion of Science and Technology Talents (DPST) Project, administered by The Institute for the Promotion of Teaching Science and Technology (IPST).

References

Aruoba, S. B., Diebold, F. X., Nalewaik, J., Schorfheide, F., and Song, D. (2016). Improving GDP measurement: A measurement-error perspective. *Journal of Econometrics*, **191**, 384–397.

Bell, W. R. and Hillmer, S. C. (1990). The time series approach to estimation for repeated surveys. *Survey Methodology*, **16**, 195–215.

Buonaccorsi, J. P. and Staudenmayer, J. (2009). Statistical methods to correct for observation error in a density-independent population model. *Ecological Monographs*, **79**, 299–324.

Carter, G. M. and Rolph, J. E. (1974). Empirical Bayes methods applied to estimating fire alarm probabilities. *Journal of the American Statistical Association*, **69**, 880–885.

De Valpine, P. and Hastings, A. (2002). Fitting population models incorporating process noise and observation error. *Ecological Monographs*, **72**, 57–76.

Dennis, B., Ponciano, J. M., Lele, S. R., Taper, M. L., and Staples, D. F. (2006). Estimating density dependence, process noise, and observation error. *Ecological Monographs*, **76**, 323–341.

Dickey, D. A. (1976). *Estimation and Hypothesis Testing in Nonstationary Time Series.* Ph.D. thesis, Iowa State University.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, **74**, 427–431.

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, **70**, 311–319.

Fuller, W. A. (1976). *Introduction to Statistical Time Series*. John Wiley & Sons, New York.

Lahiri, P. and Suntornchost, J. (2015). Variable selection for linear mixed models with applications in small area estimation. *Sankhya B*, **77**, 312–320.

Ludwig, D. and Walters, C. J. (1981). Measurement errors and uncertainty in parameter estimates for stock and recruitment. *Canadian Journal of Fisheries and Aquatic Sciences*, **38**, 711–720.

Marhuenda García, Y., Morales, D., and Pardo Llorente, M. d. C. (2016). Tests for the variance parameter in the Fay-Herriot model. *Statistics*, **50**, 27–42.

Rossi, E. and Santucci de Magistris, P. (2018). Indirect inference with time series observed with error. *Journal of Applied Econometrics*, **33**, 874–897.

Scott, A. J. and Smith, T. M. F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, **69**, 674–678.

Scott, A. J., Smith, T. M. F., and Jones, R. G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review*, **45**, 13–28.

Shenk, T. M., White, G. C., and Burnham, K. P. (1998). Sampling-variance effects on detecting density dependence from temporal trends in natural populations. *Ecological Monographs*, **68**, 445–463.

Smith, R. J., Weale, M. R., and Satchell, S. E. (1998). Measurement error with accounting constraints: Point and interval estimation for latent data with an application to U.K. gross domestic product. *Review of Economic Studies*, **65**, 109–134.

Staudenmayer, J. and Buonaccorsi, J. P. (2005). Measurement error in linear autoregressive models. *Journal of the American Statistical Association*, **100**, 841–852.

Walters, C. J. and Ludwig, D. (1981). Effects of measurement errors on the assessment of stock–recruitment relationships. *Canadian Journal of Fisheries and Aquatic Sciences*, **38**, 704–710.

APPENDIX

A. Appendix: theoretical properties

In this section, we prove asymptotic properties of the adjusted estimators of the correlation coefficient and the unit root test statistic discussed in Section 3. We first obtain some important moment properties in Lemma 1 and then prove the three main results respectively in Theorem 1, Theorem 2, and Theorem 3.

Lemma 1: Under the assumption that $\rho = 1$, we have

1. $\mathbb{E}\left(S_{Y,T}(0)\right) = \frac{1}{2}T(T-1)\sigma_e^2$;
2. $\mathbb{E}\left(S_{Y,T}(1)\right) = \frac{1}{2}T(T-1)\sigma_e^2$;
3. $\mathrm{Var}\left(S_{Y,T}(0)\right) = \frac{1}{3}T(T-1)(T^2-T+1)\sigma_e^4$;
4. $\mathrm{Var}\left(S_{Y,T}(1)\right) = \frac{1}{3}T(T-1)(T^2-T+1)\sigma_e^4$;
5. for any positive integer k, $\mathbb{E}(S_{Y,T}^{-k}(0)) = O(T^{-2k})$; and
6. for any positive integers l and k, $\mathbb{E}(S_{Y,T}^{-k}(0)S_{Y,T}^{l}(1)) = O(T^{2(l-k)})$.

Proof:

1. Given that $Y_0 = 0$,

$$S_{Y,T}(0) = \sum_{t=1}^{T-1} \left(\sum_{j=1}^{t} e_j \right)^2$$
$$= \sum_{i=1}^{T-1} (T-i) e_i^2 + \sum_{i=2}^{T-1} \sum_{j=1}^{i-1} (T-i) e_i e_j. \tag{11}$$

   By the property that $\{e_i\}_{i \geq 1}$ is a sequence of independent random variables with zero mean and variance σ_e^2,

$$\mathbb{E}\left(S_{Y,T}(0)\right) = \sum_{i=1}^{T-1} (T-i)\sigma_e^2 = \frac{1}{2} T(T-1)\sigma_e^2.$$

2. Note that

$$S_{Y,T}(1) = S_{Y,T}(0) + \sum_{t=2}^{T} e_t Y_{t-1}.$$

   Since $\mathbb{E}(e_i) = 0$ and e_i and Y_{i-1} are independent, $\mathbb{E}\left(S_{Y,T}(1)\right) = \mathbb{E}\left(S_{Y,T}(0)\right)$.

3. Since $\{e_i\}_{i \geq 1}$ is a sequence of independent random variables with zero mean and variance σ_e^2, $\{e_i^2\}$ and $\{e_i e_j\}$ are uncorrelated sequences of uncorrelated random variables such that $\mathrm{Var}(e_i^2) = 2\sigma_e^4$ and $\mathrm{Var}(e_i e_j) = \sigma_e^4$ for $i \neq j$. From (11),

$$\mathrm{Var}\left(S_{Y,T}(0)\right) = \sum_{i=1}^{T-1} (T-i)^2 \,\mathrm{Var}\left(e_i^2\right) + \sum_{i=2}^{T-1} \sum_{j=1}^{i-1} (T-i)^2 \,\mathrm{Var}\left(e_i e_j\right)$$
$$= T(T-1)(T^2 - T + 1)\sigma_e^4.$$

4. Note that

$$\mathrm{Var}\left(\sum_{t=2}^{T} e_t Y_{t-1} \right) = \sum_{t=2}^{T} \mathrm{Var}(e_t Y_{t-1}) + 2 \sum_{2 \leq i < j \leq T} \mathrm{Cov}(e_i Y_{i-1}, e_j Y_{j-1})$$
$$= \frac{1}{2} T(T-1)\sigma_e^4,$$

   and

$$\mathrm{Cov}\left(S_{Y,T}(0), \sum_{t=2}^{T} e_t Y_{t-1} \right) = \mathrm{Cov}\left(\sum_{t=2}^{T} Y_{t-1}^2, \sum_{t=2}^{T} e_t Y_{t-1} \right)$$
$$= \sum_{t=2}^{T} \mathrm{Cov}\left(Y_{t-1}^2, e_t Y_{t-1} \right) + \sum_{t=2}^{T} \sum_{s=2}^{t-1} \mathrm{Cov}\left(Y_{t-1}^2, e_s Y_{s-1} \right)$$
$$+ \sum_{t=2}^{T} \sum_{s=t+1}^{T} \mathrm{Cov}\left(Y_{t-1}^2, e_s Y_{s-1} \right)$$

$$= \frac{1}{3}T(T-1)(T-2)\sigma_e^4.$$

Then,

$$\begin{aligned}\mathrm{Var}\left(S_{Y,T}(1)\right) &= \mathrm{Var}\left(S_{Y,T}(0)\right) + \mathrm{Var}\left(\sum_{t=2}^{T} e_t Y_{t-1}\right) + 2\,\mathrm{Cov}\left(S_{Y,T}(0), \sum_{t=2}^{T} e_t Y_{t-1}\right)\\ &= \frac{1}{3}T(T-1)(T^2-T+1)\sigma_e^4 + \frac{1}{2}T(T-1)\sigma_e^4 + \frac{2}{3}T(T-1)(T-2)\sigma_e^4\\ &= \frac{1}{6}T(T-1)(2T^2+2T-3)\sigma_e^4.\end{aligned}$$

5. To find the order of $\mathbb{E}(S_{Y,T}(0)^{-k})$, we apply the second order Taylor approximation to the function $f(x) = x^{-k}$ about $\mu = \mathbb{E}(S_{Y,T}(0))$ as follows.

$$\begin{aligned}\mathbb{E}(S_{Y,T}(0)^{-k}) &= \frac{1}{\mathbb{E}^k(S_{Y,T}(0))} + \frac{k(k+1)}{2}\frac{\mathrm{Var}(S_{Y,T}(0))}{\mathbb{E}^{k+2}(S_{Y,T}(0))} + O(T^{-2k})\\ &= O(T^{-2k}) + O(T^{-2(k+2)})O(T^4) + O(T^{-2k})\\ &= O(T^{-2k}).\end{aligned}$$

6. Similarly, we apply the second order Taylor approximation to the function $f(x,y) = y^{-k}x^l$ about $\mu = (\mathbb{E}(S_{Y,T}(1)), \mathbb{E}(S_{Y,T}(0)))$ to find the order of $\mathbb{E}(S_{Y,T}(0)^{-k}S_{Y,T}(1)^l)$ as follows.

$$\begin{aligned}\left|\mathbb{E}\left(\frac{S_{Y,T}(1)^l}{S_{Y,T}(0)^k}\right)\right| &\le \left|\frac{\mathbb{E}^l(S_{Y,T}(1))}{\mathbb{E}^k(S_{Y,T}(0))}\right| + \left|\frac{l(l-1)}{2}\frac{\mathbb{E}^{l-2}(S_{Y,T}(1))}{\mathbb{E}^k(S_{Y,T}(0))}\,\mathrm{Var}(S_{Y,T}(1))\right|\\ &\quad + \left|\frac{k(k+1)}{2}\frac{\mathbb{E}^l(S_{Y,T}(1))}{\mathbb{E}^{k+2}(S_{Y,T}(0))}\,\mathrm{Var}(S_{Y,T}(0))\right|\\ &\quad + \left|2kl\frac{\mathbb{E}^{l-1}(S_{Y,T}(1))}{\mathbb{E}^{k+1}(S_{Y,T}(0))}\,\mathrm{Cov}(S_{Y,T}(1), S_{Y,T}(0))\right| + O(T^{-2(l-k)})\\ &\le O(T^{2(l-k)}) + O(T^{2(l-k)}) + O(T^{2(l-k)}) + O(T^{2(l-k)}) + O(T^{-2(l-k)})\\ &= O(T^{2(l-k)}).\end{aligned}$$

□

Theorem 1: Under the assumption that $\rho = 1$,

$$\hat{\rho}_{Adj} - \hat{\rho}_Y = o_p(1) \qquad \text{as } T \text{ goes to infinity.}$$

Moreover,

$$\hat{\rho}_{Adj} - \rho = o_p(1) \qquad \text{as } T \text{ goes to infinity.}$$

Proof: To prove the theorem, we will show that $\mathbb{E}(\hat{\rho}_{Adj} - \hat{\rho}_Y)^2 = O(T^{-2})$ by proving the following statements:

(1) $\mathbb{E}(\hat{\rho}_{Adj} - \hat{\rho}_Y) = O(T^{-2})$,

(2) $\mathrm{Var}(\hat{\rho}_{Adj} - \hat{\rho}_Y) = O(T^{-2})$.

To prove (1), apply the second order Taylor series expansion to the function $f(x, y) = \frac{x}{y}$ around $(S_{Y,T}(1), S_{Y,T}(0))$ as follows.

$$\begin{aligned}\hat{\rho}_{Adj} - \hat{\rho}_Y = &\frac{1}{S_{Y,T}(0)}(S_{W,T}(1) - S_{Y,T}(1)) - \frac{S_{Y,T}(1)}{S^2_{Y,T}(0)}(\widetilde{S}_{W,T}(0) - S_{Y,T}(0)) \\ &+ \frac{S_{Y,T}(1)}{S^3_{Y,T}(0)}(\widetilde{S}_{W,T}(0) - S_{Y,T}(0))^2 \\ &- \frac{1}{S^2_{Y,T}(0)}(S_{W,T}(1) - S_{Y,T}(1))(\widetilde{S}_{W,T}(0) - S_{Y,T}(0)) + O_p(T^{-2}).\end{aligned}$$

Then, apply the conditional expectation given $\boldsymbol{Y}$, we have

$$\begin{aligned}\mathbb{E}\left(\hat{\rho}_{Adj} - \hat{\rho}_Y|\boldsymbol{Y}\right) &= \frac{S_{Y,T}(1)}{S^3_{Y,T}(0)}\,\mathrm{Var}\left(\widetilde{S}_{W,T}(0)\middle|\boldsymbol{Y}\right) - \frac{1}{S^2_{Y,T}(0)}\,\mathrm{Cov}\left(S_{W,T}(1), \widetilde{S}_{W,T}(0)\middle|\boldsymbol{Y}\right) \\ &= \frac{S_{Y,T}(1)}{S^3_{Y,T}(0)}\left(2\sum_{t=2}^{T}\sigma^4_{u,t-1} + 4\sum_{t=2}^{T}Y^2_{t-1}\sigma^2_{u,t-1}\right) \\ &\quad - \frac{2}{S^2_{Y,T}(0)}\sum_{t=2}^{T}\left(Y_tY_{t-1} + Y_{t-1}Y_{t-2}\right)\sigma^2_{u,t-1} + O_p(T^{-2}).\end{aligned}$$

Let $\sigma^2_u = \max\limits_{1\le t\le T}\sigma^2_{u,t}$. We can show that

$$\left|\mathbb{E}\left(\hat{\rho}_{Adj} - \hat{\rho}_Y|\boldsymbol{Y}\right)\right| \le \frac{2|S_{Y,T}(1)|}{S^3_{Y,T}(0)}T\sigma^4_u + \frac{4|S_{Y,T}(1)|}{S^2_{Y,T}(0)}\sigma^2_u + \frac{5}{S_{Y,T}(0)}\sigma^2_u + O_p(T^{-2}). \tag{12}$$

From Lemma 1, we can show that

$$\begin{aligned}\mathbb{E}\left(\frac{|S_{Y,T}(1)|}{S^3_{Y,T}(0)}\right) &= O(T^{-4}), \\ \mathbb{E}\left(\frac{|S_{Y,T}(1)|}{S^2_{Y,T}(0)}\right) &= O(T^{-2}), \\ \mathbb{E}\left(\frac{1}{S_{Y,T}(0)}\right) &= O(T^{-2}).\end{aligned}$$

Therefore, $|\mathbb{E}(\hat{\rho}_{Adj} - \hat{\rho}_Y)| = O(T^{-2})$.

To prove (2), we note that

$$\begin{aligned}\mathrm{Var}(\hat{\rho}_{Adj} - \hat{\rho}_Y) &= \mathbb{E}\left(\mathrm{Var}(\hat{\rho}_{Adj} - \hat{\rho}_Y|\boldsymbol{Y})\right) + \mathrm{Var}\left(\mathbb{E}(\hat{\rho}_{Adj} - \hat{\rho}_Y|\boldsymbol{Y})\right) \\ &\le \mathbb{E}\left(\mathrm{Var}(\hat{\rho}_{Adj}|\boldsymbol{Y})\right) + \mathbb{E}\left(\mathbb{E}^2(\hat{\rho}_{Adj} - \hat{\rho}_Y|\boldsymbol{Y})\right).\end{aligned} \tag{13}$$

To bound the first term of (13), we apply the first order Taylor approximation to the function $f(x,y) = \frac{x}{y}$ around the point $(S_{Y,T}(1), S_{Y,T}(0))$ as follows.

$$\frac{S_{W,T}(1)}{\widetilde{S}_{W,T}(0)} = \frac{S_{Y,T}(1)}{S_{Y,T}(0)} + \frac{1}{S_{Y,T}(0)}(S_{W,T}(1) - S_{Y,T}(1)) - \frac{S_{Y,T}(1)}{S^2_{Y,T}(0)}(\widetilde{S}_{W,T}(0) - S_{Y,T}(0)) + O_p(T^{-2}).$$

Therefore,

$$\begin{aligned}\operatorname{Var}\left(\frac{S_{W,T}(1)}{\widetilde{S}_{W,T}(0)}\middle|\boldsymbol{Y}\right) &= \frac{1}{S^2_{Y,T}(0)}\operatorname{Var}\left(S_{W,T}(1)|\boldsymbol{Y}\right) + \frac{S^2_{Y,T}(1)}{S^4_{Y,T}(0)}\operatorname{Var}\left(\widetilde{S}_{W,T}(0)\middle|\boldsymbol{Y}\right) \\ &\quad - \frac{2S_{Y,T}(1)}{S^3_{Y,T}(0)}\operatorname{Cov}\left(S_{W,T}(1), \widetilde{S}_{W,T}(0)\middle|\boldsymbol{Y}\right) + O(T^{-2}) \\ &:= A_1 + A_2 + A_3 + O_p(T^{-2}).\end{aligned}$$

To bound $\mathbb{E}(A_1)$, we notice that

$$\begin{aligned}\operatorname{Var}\left(S_{W,T}(1)|\boldsymbol{Y}\right) &= \sum_{t=2}^{T}(Y_t^2\sigma^2_{u,t-1} + Y^2_{t-1}\sigma^2_{u,t} + \sigma^2_{u,t}\sigma^2_{u,t-1} + 2Y_tY_{t-2}\sigma^2_{u,t-1}) \\ &\leq \sum_{t=2}^{T}(2Y_t^2 + Y^2_{t-1} + Y^2_{t-2})\sigma^2_u + T\sigma^4_u \\ &\leq 6S_{Y,T}(0)\sigma^2_u + T\sigma^4_u.\end{aligned}$$

From Lemma 1, we have $\mathbb{E}(A_1) = \mathbb{E}\left(\frac{6\sigma^2_u}{S_{Y,T}(0)} + \frac{T\sigma^4_u}{S^2_{Y,T}(0)}\right) = O(T^{-2})$.

For the term A_2, we have

$$\operatorname{Var}\left(\widetilde{S}_{W,T}(0)\middle|\boldsymbol{Y}\right) = 2\sum_{t=2}^{T}\sigma^4_{u,t-1} + 4\sum_{t=2}^{T}Y^2_{t-1}\sigma^2_{u,t-1} \leq 2T\sigma^4_u + 4\sigma^2_uS_{Y,T}(0).$$

From Lemma 1, $\mathbb{E}(A_2) = \mathbb{E}\left(\frac{2S^2_{Y,T}(1)}{S^4_{Y,T}(0)}T\sigma^4_u + \frac{4S^2_{Y,T}(1)}{S^3_{Y,T}(0)}\sigma^2_u\right) = O(T^{-2})$. For the last term A_3, we notice that

$$\operatorname{Cov}\left(S_{W,T}(1), \widetilde{S}_{W,T}(0)\middle|\boldsymbol{Y}\right) = 2\sum_{t=2}^{T}(Y_tY_{t-1} + Y_{t-1}Y_{t-2})\sigma^2_{u,t-1} \leq 10\sigma^2_uS_{Y,T}(0).$$

Hence, $\mathbb{E}(A_3) = \mathbb{E}\left(\frac{20\sigma^2_uS_{Y,T}(1)}{S^2_{Y,T}(0)}\right) = O(T^{-2})$. This implies that $\mathbb{E}\left(\operatorname{Var}(\hat{\rho}_{Adj}|\boldsymbol{Y})\right) = O(T^{-2})$. To consider $\mathbb{E}\left(\mathbb{E}^2(\hat{\rho}_{Adj} - \hat{\rho}_Y|\boldsymbol{Y})\right)$, we apply (12) and Cauchy-Schwartz inequality to obtain

$$\begin{aligned}\mathbb{E}^2\left(\hat{\rho}_{Adj} - \hat{\rho}_Y|\boldsymbol{Y}\right) &\leq 3\left(\frac{2|S_{Y,T}(1)|}{S^3_{Y,T}(0)}T\sigma^4_u\right)^2 + 3\left(\frac{4|S_{Y,T}(1)|}{S^2_{Y,T}(0)}\sigma^2_u\right)^2 + 3\left(\frac{5}{S_{Y,T}(0)}\sigma^2_u\right)^2 \\ &= \frac{12S^2_{Y,T}(1)}{S^6_{Y,T}(0)}T^2\sigma^8_u + \frac{48S^2_{Y,T}(1)}{S^4_{Y,T}(0)}\sigma^4_u + \frac{75}{S^2_{Y,T}(0)}\sigma^4_u.\end{aligned}$$

From Lemma 1, $\mathbb{E}\left(\mathbb{E}^2(\hat{\rho}_{Adj}-\hat{\rho}_Y|\boldsymbol{Y})\right)=O(T^{-4})$. Hence, from (13), $\mathrm{Var}(\hat{\rho}_{Adj}-\hat{\rho}_Y)=O(T^{-2})$.

From (1) and (2), we have $\mathbb{E}(\hat{\rho}_{Adj}-\hat{\rho}_Y)^2=O(T^{-2})$. Therefore, $\hat{\rho}_{Adj}-\hat{\rho}_Y=o_p(1)$ as T goes to infinity. Moreover, since $\hat{\rho}_Y-\rho=o_p(1)$, we have $\hat{\rho}_{Adj}-\rho=o_p(1)$ as T goes to infinity. □

Having proved the asymptotic property of $\hat{\rho}_{Adj}$, we will prove the asymptotic distribution of the test statistics $\hat{\tau}_{Adj}$ by first obtaining some important lemmas as follows.

Lemma 2: Under the assumption that $\rho=1$,

$$\frac{1}{T^2}\widetilde{S}_{W,T}(0)-\sum_{i=1}^{\infty}\gamma_i^2Z_i^{*2}=o_p(1)$$

as T goes to infinity, where $\gamma_i=(-1)^{i+1}\dfrac{2}{(2i-1)\pi}$ and $Z_i^*\overset{iid}{\sim}N(0,\sigma_e^2)$.

Proof: We know from Dickey (1976) that

$$\frac{1}{T^2}S_{Y,T}(0)-\sum_{i=1}^{\infty}\gamma_i^2Z_i^{*2}=o_p(1),$$

as T goes to infinity. To prove this lemma, we will show that

$$\frac{1}{T^2}\widetilde{S}_{W,T}(0)-\frac{1}{T^2}S_{Y,T}(0)=o_p(1) \tag{14}$$

as T goes to infinity.
First, we notice that

$$\frac{\widetilde{S}_{W,T}(0)}{T^2}-\frac{S_{Y,T}(0)}{T^2}=\frac{1}{T^2}\sum_{t=2}^{T}2Y_{t-1}u_{t-1}+\frac{1}{T^2}\sum_{t=2}^{T}(u_{t-1}^2-\sigma_{u,t-1}^2).$$

Since $\mathbb{E}(Y_tu_t)$ and $\mathbb{E}\left(u_t^2-\sigma_{u,t}^2\right)$ are equal to zero for all t,

$$\mathbb{E}\left(\frac{\widetilde{S}_{W,T}(0)}{T^2}-\frac{S_{Y,T}(0)}{T^2}\right)=\frac{1}{T^2}\sum_{t=2}^{T}2\,\mathbb{E}\left(Y_{t-1}u_{t-1}\right)+\frac{1}{T^2}\sum_{t=2}^{T}\mathbb{E}(u_{t-1}^2-\sigma_{u,t-1}^2)=0. \tag{15}$$

Since $\{Y_tu_t\}_{1\le t\le T}$ and $\{u_t^2-\sigma_{u,t}^2\}_{1\le t\le T}$ are uncorrelated random sequences,

$$\begin{aligned}\mathrm{Var}\left(\frac{\widetilde{S}_{W,T}(0)}{T^2}-\frac{S_{Y,T}(0)}{T^2}\right)&=\frac{1}{T^4}\sum_{t=2}^{T}4\,\mathrm{Var}(Y_{t-1}u_{t-1})+\frac{1}{T^4}\sum_{t=2}^{T}\mathrm{Var}(u_{t-1}^2-\sigma_{u,t-1}^2)\\&\le\frac{1}{T^4}\sigma_e^2\sigma_u^2\cdot\frac{1}{2}T(T-1)+\frac{2}{T^4}T\sigma_u^4\\&=O(T^{-2}).\end{aligned} \tag{16}$$

Hence, from (15) and (16), (14) is proved. Consequently,

$$\frac{1}{T^2}\widetilde{S}_{W,T}(0)-\sum_{i=1}^{\infty}\gamma_i^2Z_i^{*2}=o_p(1),$$

as T goes to infinity. □

Theorem 2: Under the assumption that $\rho = 1$, the statistics $T(\hat{\rho}_{adj} - 1)$ has the same limiting distribution as $T(\hat{\rho}_Y - 1)$ as T goes to infinity. In a particular,

$$T(\hat{\rho}_{Adj} - 1) \xrightarrow{d} \frac{\left(\sum_{i=1}^{\infty} \sqrt{2}\gamma_i Z_i\right)^2 - 1}{2\sqrt{\sum_{i=1}^{\infty} \gamma_i^2 Z_i^2}},$$

where $\gamma_i = (-1)^{i+1}\frac{2}{(2i-1)\pi}$ and $Z_i \overset{iid}{\sim} N(0,1)$.

Proof: From the definition of $\hat{\rho}_{Adj}$, $T(\hat{\rho}_{Adj} - 1)$ can be simplified as

$$T(\hat{\rho}_{Adj} - 1) = T\left(\frac{S_{W,T}(1) - \widetilde{S}_{W,T}(0)}{\widetilde{S}_{W,T}(0)}\right) = \left(\frac{1}{T^2}\widetilde{S}_{W,T}(0)\right)^{-1}\left(\frac{1}{T}\left(S_{W,T}(1) - \widetilde{S}_{W,T}(0)\right)\right). \quad (17)$$

From (1) and (5), we have

$$\begin{aligned}
\frac{1}{T}\left(S_{W,T}(1) - \widetilde{S}_{W,T}(0)\right) &= \frac{1}{T}\sum_{t=2}^{T}\left((Y_{t-1} + u_{t-1})(Y_t + u_t - Y_{t-1} - u_{t-1}) + \sigma^2_{u,t-1}\right) \\
&= \frac{1}{T}\sum_{t=2}^{T}\left((Y_{t-1} + u_{t-1})(e_t + u_t - u_{t-1}) + \sigma^2_{u,t-1}\right) \\
&= \frac{1}{T}\sum_{t=2}^{T} Y_{t-1}e_t + \frac{1}{T}\sum_{t=1}^{T-1} e_t u_T - \frac{Y_1 u_1}{T} - \frac{1}{T}\sum_{t=2}^{T-1} e_t u_{t-1} \\
&\quad + \frac{1}{T}\sum_{t=2}^{T} e_t u_{t-1} + \frac{1}{T}\sum_{t=2}^{T} u_t u_{t-1} - \frac{1}{T}\sum_{t=2}^{T}(u^2_{t-1} - \sigma^2_{u,t-1}). \qquad (18)
\end{aligned}$$

Notice that each of the terms in (18) except $\frac{1}{T}\sum_{t=2}^{T} Y_{t-1}e_t$ is a sum of uncorrelated random variables with zero means and finite variances. Therefore, by the law of large number, each of those terms converges in probability to zero.

Following the results of Fuller (1976) that

$$\frac{1}{T}\sum_{t=2}^{T} Y_{t-1}e_t \xrightarrow{d} \frac{1}{2}\left(\sum_{i=1}^{\infty}\sqrt{2}\gamma_i Z_i^*\right)^2 - \frac{\sigma_e^2}{2},$$

where $\gamma_i = (-1)^{i+1}\dfrac{2}{(2i-1)\pi}$ and $Z_i^* \overset{iid}{\sim} N(0,\sigma_e^2)$, we can show that

$$\frac{1}{T}\left(S_{W,T}(1) - \widetilde{S}_{W,T}(0)\right) \xrightarrow{d} \frac{1}{2}\left(\sum_{i=1}^{\infty}\sqrt{2}\gamma_i Z_i^*\right)^2 - \frac{\sigma_e^2}{2}. \quad (19)$$

From Lemma 2, (17), and (19),

$$T(\hat{\rho}_{Adj} - 1) \xrightarrow{d} \frac{\left(\sum_{i=1}^{\infty} \sqrt{2}\gamma_i Z_i\right)^2 - 1}{2\sqrt{\sum_{i=1}^{\infty} \gamma_i^2 Z_i^2}}, \quad (20)$$

where $\gamma_i = (-1)^{i+1}\dfrac{2}{(2i-1)\pi}$ and $Z_i \overset{iid}{\sim} N(0,1)$. □

Lemma 3: Define the statistic $\hat{\sigma}^2_{Adj,e}$ as

$$\hat{\sigma}^2_{Adj,e} = |\hat{\sigma}^2_{W,e,1} - \hat{\sigma}^2_{W,e,2}|,$$

where

$$\hat{\sigma}^2_{W,e,1} = \frac{1}{T-2}\sum_{t=2}^{T}(W_t - \hat{\rho}_{Adj}W_{t-1})^2,$$

and

$$\hat{\sigma}^2_{W,e,2} = \frac{1}{T-2}\sum_{t=2}^{T}\left(\sigma^2_{u,t} + \hat{\rho}^2_{Adj}\sigma^2_{u,t-1}\right).$$

Then, under the assumption that $\rho = 1$,

$$\hat{\sigma}^2_{Adj,e} - \hat{\sigma}^2_e = o_p(1).$$

In particular, $\hat{\sigma}^2_{Adj,e} - \sigma^2_e = o_p(1)$.

Proof: Notice that

$$\begin{aligned}
(T-2)\hat{\sigma}^2_{W,e,1} &= \sum_{t=2}^{T}(Y_t - \hat{\rho}_Y Y_{t-1} + (\hat{\rho}_Y - \hat{\rho}_{Adj})Y_{t-1} + u_t - \hat{\rho}_{Adj}u_{t-1})^2 \\
&= (T-2)\hat{\sigma}^2_e + (\hat{\rho}_Y - \hat{\rho}_{Adj})^2 S_{Y,T}(0) + \sum_{t=2}^{T}(u_t - \hat{\rho}_{Adj}u_{t-1})^2 \\
&\quad + 2(\hat{\rho}_Y - \hat{\rho}_{Adj})\sum_{t=2}^{T}Y_{t-1}(u_t - \hat{\rho}_{Adj}u_{t-1}) + 2\sum_{t=2}^{T}(Y_t - \hat{\rho}_Y Y_{t-1})(u_t - \hat{\rho}_{Adj}u_{t-1}) \\
&= (T-2)\hat{\sigma}^2_e + (\hat{\rho}_Y - \hat{\rho}_{Adj})^2 S_{Y,T}(0) + \sum_{t=2}^{T}(u_t - \hat{\rho}_{Adj}u_{t-1})^2 \\
&\quad + 2\sum_{t=2}^{T}\left(e_t + (\rho - \hat{\rho}_{Adj})Y_{t-1}\right)(u_t - \hat{\rho}_{Adj}u_{t-1}).
\end{aligned}$$

Then,

$$\begin{aligned}
(T-2)(\hat{\sigma}^2_{W,e,1} - \hat{\sigma}^2_{W,e,2} - \hat{\sigma}^2_e) &= (\hat{\rho}_Y - \hat{\rho}_{Adj})^2 S_{Y,T}(0) + \sum_{t=2}^{T}(u_t - \hat{\rho}_{Adj}u_{t-1})^2 \\
&\quad + 2\sum_{t=2}^{T}\left(e_t + (\rho - \hat{\rho}_{Adj})Y_{t-1}\right)(u_t - \hat{\rho}_{Adj}u_{t-1}) \\
&\quad - \sum_{t=2}^{T}\left(\sigma^2_{u,t} + \hat{\rho}^2_{Adj}\sigma^2_{u,t-1}\right) \\
&= (\hat{\rho}_Y - \hat{\rho}_{Adj})^2 S_{Y,T}(0) + \sum_{t=2}^{T}(u_t^2 - \sigma^2_{u,t}) + \hat{\rho}^2_{Adj}\sum_{t=2}^{T}(u^2_{t-1} - \sigma^2_{u,t-1}) \\
&\quad - 2\hat{\rho}_{Adj}\sum_{t=2}^{T}u_t u_{t-1} + 2\sum_{t=2}^{T}e_t u_t + 2(\rho_Y - \hat{\rho}_{Adj})\sum_{t=2}^{T}Y_{t-1}u_t
\end{aligned}$$

$$- 2\hat{\rho}_{Adj}\sum_{t=2}^{T} e_t u_{t-1} - 2\hat{\rho}_{Adj}(\rho_Y - \hat{\rho}_{Adj})\sum_{t=2}^{T} Y_{t-1}u_{t-1}$$
$$= o_p(T),$$

where we use Theorem 1, Lemma 2, and the weak law of large number to obtain the last equation. Therefore, $\hat{\sigma}^2_{W,e,1} - \hat{\sigma}^2_{W,e,2} - \hat{\sigma}^2_e = o_p(1)$. Consequently, $\hat{\sigma}^2_{Adj,e,1} - \sigma^2_e = o_p(1)$. □

Applying Lemma 2 - Lemma 3, we obtain the asymptotic distribution of the proposed statistic $\hat{\tau}_{Adj}$ in the following theorem.

Theorem 3: Let $\hat{\tau}_{Adj}$ be a statistic defined by

$$\hat{\tau}_{Adj} = \frac{(\hat{\rho}_{Adj} - 1)\sqrt{\widetilde{S}_{W,T}(0)}}{\sqrt{\hat{\sigma}^2_{Adj,e}}}.$$

Then $\hat{\tau}_{Adj}$ has the same asymptotic distribution as $\hat{\tau}$ in (4). That is

$$\hat{\tau}_{Adj} \xrightarrow{d} \frac{\left(\sum_{i=1}^{\infty}\sqrt{2}\gamma_i Z_i\right)^2 - 1}{2\sqrt{\sum_{i=1}^{\infty}\gamma_i^2 Z_i^2}},$$

where $\gamma_i = (-1)^{i+1}\dfrac{2}{(2i-1)\pi}$ and $Z_i \overset{iid}{\sim} N(0,1)$.

Proof: From Lemma 2 and Lemma 3, we have

$$\frac{1}{T^2}\widetilde{S}_{W,T}(0)\cdot\frac{1}{\hat{\sigma}^2_{Adj,e}} \xrightarrow{p} \sum_{i=1}^{\infty}\gamma_i^2\frac{Z_i^{*2}}{\sigma_e^2},$$

where $\gamma_i = (-1)^{i+1}\dfrac{2}{(2i-1)\pi}$ and $Z_i^* \overset{iid}{\sim} N(0,\sigma_e^2)$.

Then,

$$\sqrt{\frac{1}{T^2}\widetilde{S}_{W,T}(0)\cdot\frac{1}{\hat{\sigma}^2_{Adj,e}}} \xrightarrow{p} \sqrt{\sum_{i=1}^{\infty}\gamma_i^2 Z_i^2}, \tag{21}$$

where $Z_i \overset{iid}{\sim} N(0,1)$.

From (20) and (21), we can conclude that

$$\hat{\tau}_{Adj} = T(\hat{\rho}_{Adj} - 1)\cdot\sqrt{\frac{1}{T^2}\widetilde{S}_{W,T}(0)\cdot\frac{1}{\hat{\sigma}^2_{Adj,e}}} \xrightarrow{d} \frac{\left(\sum_{i=1}^{\infty}\sqrt{2}\gamma_i Z_i\right)^2 - 1}{2\sqrt{\sum_{i=1}^{\infty}\gamma_i^2 Z_i^2}}.$$

□

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 595–609
http://www.ssca.org.in/journal

Tests of Contrasts for Mean Vectors with Large Dimensions

Rauf Ahmad
Department of Statistics, Uppsala University, Sweden

Received: 17 April 2024; Revised: 28 September 2024; Accepted: 10 October 2024

Abstract

A test statistic for a fixed contrast comparison of high-dimensional mean vectors is introduced. The statistic can be used when the dimension of the vectors exceeds the sample size, and the data may not necessarily follow a multivariate normal distribution. The components of the test statistics are defined as U-statistics with optimal properties, where the same estimators are given equivalent, computationally highly efficient, formulation for practical applications. The properties of the statistic are studied under a general multivariate model and certain mild assumptions. Through simulations, the statistic is shown to have an accurate size control and high power properties. An extension of a set of fixed orthogonal contrasts is also discussed.

Key words: High-dimensional tests; Multivariate inference; Contrast comparisons; U-statistics.

AMS Subject Classifications: 62H11, 62H30

1. Introduction

Let $\mathbf{X}_{ik} = (X_{ik1}, \ldots, X_{ikp})^{\mathrm{T}} \sim \mathcal{F}_i$, $k = 1, \ldots, n_i$, be a random sample of n_i vectors from ith non-degenerate p-variate distribution, denoted $\mathcal{F}_i$, which need not necessarily be multivariate normal, $i = 1, \ldots, g \geq 2$. Further, the g populations are assumed to be independent, with $\mathrm{E}(\mathbf{X}_{ik}) = \boldsymbol{\mu}_i \in \mathbb{R}^p$ and $\mathrm{Cov}(\mathbf{X}_{ik}) = \boldsymbol{\Sigma}_i \in \mathbb{R}^{p\times p}$, and $\boldsymbol{\Sigma}_i > 0$, $\forall\ i$.

Most of the testing problems in multivariate theory pertain to the two basic parameters, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$; *e.g*, single- and multi-sample hypotheses for $\boldsymbol{\mu}_i$, such as $\boldsymbol{\mu}_i = \mathbf{0}$, $\boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_g$ $(g \geq 2)$. These hypotheses are termed global hypotheses, and their rejection often implies further exploration to sort out potential contributors to the rejection. For example, for $g = 1$, a level profile analysis is carried out to test if all components of $\boldsymbol{\mu}$ are same, *i.e.* if $\mu_1 = \ldots = \mu_p$.

In practice, however, situations exist, mainly in multi-sample cases, where certain specific contrast comparisons among $\boldsymbol{\mu}_i$ are of interest. For example, for $g = 3$, it might be of interest to test if $\boldsymbol{\mu}_1 - 2\boldsymbol{\mu}_2 + \boldsymbol{\mu}_3 = \mathbf{0}$. In general, such a *contrast hypothesis* is formulated

Corresponding Author: Rauf Ahmad
Email: rauf.ahmad@statistik.uu.se

as

$$\mathrm{H}_0 : \sum_{i=1}^{g} c_i \boldsymbol{\mu}_i = \mathbf{0} \quad \text{vs.} \quad \mathrm{H}_1 : \text{Not } \mathrm{H}_0, \tag{1}$$

with $\sum_{i=1}^{g} c_i = 0$, a condition which is an inevitable component of the definition of a contrast. In the aforementioned example, $(c_1, c_2, c_3) = (1, -2, 1)$.

Note that, for $g = 2$, the condition implies $c_2 = -c_1$ which, without loss of generality, can be taken as $c_1 = 1 \Rightarrow c_2 = -1$, so that H_0 reduces to the usual two-sample hypothesis $\mathrm{H}_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. Although, it is also a special form of contrast, the main advantage of contrast testing is apparent for the case of more than two populations.

Our objective in this article is to construct tests for H_0 in (1) when the dimension p may be large, and possibly larger than the sample sizes, *i.e.*, $p \gg n_i$, the $\mathcal{F}_i$ may be non-normal, and $\boldsymbol{\Sigma}_i$ may be unequal. For the classical case, *i.e.*, $p < n_i$, with $\mathcal{F}_i$ assumed multivariate normal, and often $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}\ \forall\ i$ (homoscedasticity assumption), the multivariate theory offers likelihood-ratio tests leading to Wilks' Λ criterion, which is further related to an F-statistic, and for moderately large sample sizes, follows an approximate χ^2-distribution; see *e.g* Anderson (2003).

As the likelihood-ratio testing framework collapses for high-dimensional data, particularly when $p \gg n_i$, new testing strategies are needed to cope with this issue. In this context, we are interested to introduce tests of (1) for $p \gg n_i$ under multivariate Behrens-Fisher setting, additionally relaxing normality assumption which is replaced with alternative mild assumptions stated below.

The test statistics are composed of estimators defined as U-statistics with optimality properties. The same estimators are alternatively also defined as simple functions of empirical covariance estimators, which makes them computationally very efficient. The U-statistics version, however, helps study their theoretical properties, including limiting distribution, conveniently, where the efficient formulation is useful for practical applications.

Whereas high-dimensional mean testing has generally attracted huge attraction in the recent past (see a list of references in Ahmad, 2019b), problems like contrast comparison have mostly been dealt with under the general rubric of multiple testing theory. For a related work in the classical case, *i.e.*, $n > p$, see Hayter (2014) and the references cited therein. A general, comprehensive reference for multiple testing problems, including for large data, containing abundant further references, is Dickhaus (2014).

Section 2 introduces test statistic for a single contrast hypothesis in (1), with an extension to a set of orthogonal contrasts in Section 3. Evaluation of the proposed tests through simulations is given in Section 4. Some technical results are deferred to the Appendix.

2. Test of a single contrast

Given the data set up in Sec. 1, let $\mathbf{X}_i = (\mathbf{X}_{i1}^{\mathrm{T}}, \ldots, \mathbf{X}_{in_i}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{n_i \times p}$ be the data matrix corresponding to the ith sample, so that the unbiased estimators of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are defined as

$$\overline{\mathbf{X}}_i = \frac{1}{n_i} \mathbf{X}_i^{\mathrm{T}} \mathbf{1}_{n_i}, \quad \widehat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i - 1} \mathbf{X}_i^{\mathrm{T}} \mathbf{C}_{n_i} \mathbf{X}_i, \tag{2}$$

respectively, where $\mathbf{C}_{n_i} = \mathbf{I}_{n_i} - \mathbf{J}_{n_i}/n_i$ is the centering matrix with $\mathbf{I}_{n_i}$ as identity matrix and $\mathbf{J}_{n_i} = \mathbf{1}_{n_i}\mathbf{1}_{n_i}^{\mathrm{T}}$ with $\mathbf{1}_{n_i}$ a vector of 1s. All vectors are column vectors by default.

Further, we denote vector inner product of $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$ as $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^{\mathrm{T}}\mathbf{b} : \mathbb{R}^p \mapsto \mathbb{R}$, so that $\|\mathbf{a}\|^2 = \mathbf{a}^{\mathrm{T}}\mathbf{a}$ is the (squared) norm of $\mathbf{a}$, and $\|\mathbf{A}\|^2 = \mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{A}) : \mathbb{R}^{q\times p} \mapsto \mathbb{R}$ is the Frobenius norm of $\mathbf{A} \in \mathbb{R}^{q\times p}$. Moreover, $\otimes$ and $\oplus$ are Kronecker product and sum, respectively.

To consider a test statistic for H_0 in (1), we can logically begin with the point estimator, $\sum_{i=1}^{g} c_i\overline{\mathbf{X}}_i = \overline{\mathbf{X}}_0$, and note that, under independence,

$$\mathrm{E}(\overline{\mathbf{X}}_0) = \boldsymbol{\mu}_0 = \sum_{i=1}^{g} c_i\boldsymbol{\mu}_i \quad \text{and} \quad \mathrm{Cov}(\overline{\mathbf{X}}_0) = \boldsymbol{\Sigma}_0 = \sum_{i=1}^{g} c_i^2\frac{\boldsymbol{\Sigma}_i}{n_i}. \tag{3}$$

In the classical setting, assuming normality and homoscedasticity, a test for (1) can be defined as $\mathrm{T}^2 = c_0^{-1}\overline{\mathbf{X}}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\overline{\mathbf{X}}_0$ with $c_0 = \sum_{i=1}^{g} c_i^2/n_i^2$, where $\mathbf{S}_0 = \sum_{i=1}^{g}(n_i-1)\widehat{\boldsymbol{\Sigma}}_i/(n-g)$ is the pooled estimator of $\boldsymbol{\Sigma}_0$ and $n = \sum_{i=1}^{g} n_i$. The T^2 statistic has optimality properties under the aforementioned assumptions, but its validity rests on the invertibility of $\mathbf{S}_0$ which, in turn, holds if and only if $n - g > p$. As this condition is not satisfied for high-dimensional data, and definitely not when $p \gg n_i$, T^2 collapses in this case and needs a modification.

The test statistic that we intend to propose for (1) is based on a modification of T^2-type statistics for testing different hypotheses on location parameters (see *e.g* Ahmad, 2014, 2019b). To see how this modification may work for the present case, first assume, tentatively, that $\boldsymbol{\Sigma}_i$ are known and, to avoid singularity issue of their empirical estimators at a later stage, consider the criterion

$$\mathrm{A} = \frac{\mathrm{A}_1}{\mathrm{tr}(\boldsymbol{\Sigma}_0)}, \tag{4}$$

with $\mathrm{A}_1 = \|\overline{\mathbf{X}}_0\|^2$, $\overline{\mathbf{X}}_0$, $\boldsymbol{\Sigma}_0$ as in (3), and $\mathrm{tr}(\cdot)$ is the trace operator. It follows that

$$\|\overline{\mathbf{X}}_0\|^2 = \left(\sum_{i=1}^{g} c_i\overline{\mathbf{X}}_i\right)^{\mathrm{T}}\left(\sum_{i=1}^{g} c_i\overline{\mathbf{X}}_i\right) = \sum_{i=1}^{g} c_i^2\|\overline{\mathbf{X}}_i\|^2 + \sum_{\substack{i=1}}^{g}\sum_{\substack{j=1\\ i\neq j}}^{g} c_ic_j\langle\overline{\mathbf{X}}_i, \overline{\mathbf{X}}_j\rangle.$$

Partitioning $\|\overline{\mathbf{X}}_i\|^2$ as

$$\|\overline{\mathbf{X}}_i\|^2 = \frac{1}{n_i^2}\sum_{k=1}^{n_i}\|\mathbf{X}_{ik}\|^2 + \frac{1}{n_i^2}\sum_{\substack{k=1\\ k\neq r}}^{n_i}\sum_{r=1}^{n_i}\langle\mathbf{X}_{ik}, \mathbf{X}_{ir}\rangle = \frac{1}{n_i}\,\mathrm{E}_i + \frac{n_i-1}{n_i}U_i = Q_i + U_i,$$

we can further write A_1 as

$$\mathrm{A}_1 = \|\overline{\mathbf{X}}_0\|^2 = \sum_{i=1}^{g} c_i^2Q_i + \sum_{i=1}^{g} c_i^2U_i + 2\sum_{\substack{i=1}}^{g}\sum_{\substack{j=1\\ i<j}}^{g} c_ic_jU_{ij} \;=\; \mathrm{A}_{11} + \mathrm{A}_{12}, \tag{5}$$

with $\mathrm{A}_{11} = \sum_{i=1}^{g} c_i^2Q_i$, where $Q_i = (\mathrm{E}_i - U_i)/n_i$, $\mathrm{E}_i = \sum_{k=1}^{n_i}\|\mathbf{X}_{ik}\|^2/n_i$. Moreover

$$U_i = \frac{1}{n_i(n_i-1)}\sum_{\substack{k=1\\ k\neq r}}^{n_i}\sum_{r=1}^{n_i}\langle\mathbf{X}_{ik}, \mathbf{X}_{ir}\rangle \quad \text{and} \quad U_{ij} = \langle\overline{\mathbf{X}}_i, \overline{\mathbf{X}}_j\rangle = \frac{1}{n_in_j}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j}\langle\mathbf{X}_{ik}, \mathbf{X}_{jl}\rangle \tag{6}$$

are, one- and two-sample U-statistics with symmetric kernels, $h(\overline{\mathbf{X}}_{ik}, \overline{\mathbf{X}}_{ir}) = \langle \overline{\mathbf{X}}_{ik}, \overline{\mathbf{X}}_{ir} \rangle$ and $h(\overline{\mathbf{X}}_{ik}, \overline{\mathbf{X}}_{jl}) = \langle \overline{\mathbf{X}}_{ik}, \overline{\mathbf{X}}_{jl} \rangle$, respectively. The motivation behind this decomposition becomes clear from the moments of the components of A_1 as summarized in the following theorem, proved in Appendix B.1.

Theorem 1: Given the partition of A_1 in (5) with

$$\mathrm{A}_{11} = \sum_{i=1}^{g} c_i^2 Q_i, \quad \mathrm{A}_{12} = \sum_{i=1}^{g} c_i^2 U_i + 2 \sum_{\substack{i=1 \\ i<j}}^{g} \sum_{j=1}^{g} c_i c_j U_{ij}, \tag{7}$$

we have $\mathrm{E}(\mathrm{A}_{11}) = \mathrm{tr}(\boldsymbol{\Sigma}_0)$, $\mathrm{E}(\mathrm{A}_{12}) = \|\boldsymbol{\mu}_0\|^2$ and $\mathrm{Var}(\mathrm{A}_{12}) = 2\|\boldsymbol{\Sigma}_0\|^2 + \mathrm{R}$, where

$$\begin{aligned} \mathrm{R} \;=\; & 4 \sum_{i=1}^{g} (c_i \boldsymbol{\mu}_i)^{\mathrm{T}} \frac{c_i^2 \boldsymbol{\Sigma}_i}{n_i} (c_i \boldsymbol{\mu}_i) + 4 \left(\sum_{\substack{i=1 \\ i<j}}^{g} \sum_{j=1}^{g} (c_i \boldsymbol{\mu}_i)^{\mathrm{T}} \frac{c_j^2 \boldsymbol{\Sigma}_j}{n_j} (c_i \boldsymbol{\mu}_i) + \sum_{\substack{i=1 \\ i<j}}^{g} \sum_{j=1}^{g} (c_j \boldsymbol{\mu}_j)^{\mathrm{T}} \frac{c_i^2 \boldsymbol{\Sigma}_i}{n_i} (c_j \boldsymbol{\mu}_j) \right) \\ & + 8 \left(\sum_{\substack{i=1 \\ i<j}}^{g} \sum_{j=1}^{g} (c_i \boldsymbol{\mu}_i)^{\mathrm{T}} \frac{c_j^2 \boldsymbol{\Sigma}_j}{n_j} (c_j \boldsymbol{\mu}_j) + \sum_{\substack{i=1 \\ i<j}}^{g} \sum_{j=1}^{g} (c_j \boldsymbol{\mu}_j)^{\mathrm{T}} \frac{c_i^2 \boldsymbol{\Sigma}_i}{n_i} (c_i \boldsymbol{\mu}_i) \right) \\ & + 8 \left(\sum_{\substack{i=1 \\ i<j<j'}}^{g} \sum_{j=1}^{g} \sum_{j'=1}^{g} (c_i \boldsymbol{\mu}_i)^{\mathrm{T}} \frac{c_j^2 \boldsymbol{\Sigma}_j}{n_j} (c_{i'} \boldsymbol{\mu}_{i'}) + \sum_{\substack{i=1 \\ i<i'<j}}^{g} \sum_{i'=1}^{g} \sum_{j=1}^{g} (c_j \boldsymbol{\mu}_j)^{\mathrm{T}} \frac{c_i^2 \boldsymbol{\Sigma}_i}{n_i} (c_{j'} \boldsymbol{\mu}_{j'}) \right) \end{aligned}$$

Under H_0, $\mathrm{E}(\mathrm{A}_{12}) = 0$, $\mathrm{Var}(\mathrm{A}_{12}) = 2\|\boldsymbol{\Sigma}_0\|^2$, where $\mathrm{E}(\mathrm{A}_{11})$, $\mathrm{Var}(\mathrm{A}_{11})$ remain same.

We observe that, $\mathrm{E}(\mathrm{A}_{11})$ is independent of $\boldsymbol{\mu}_i$, hence of $\boldsymbol{\mu}_0$, and $\mathrm{E}(\mathrm{A}_{12})$ is independent of $\boldsymbol{\Sigma}_i$, hence of $\boldsymbol{\Sigma}_0$. Further, under H_0, $\mathrm{E}(\mathrm{A}_{12}) = 0$ and $\mathrm{Var}(\mathrm{A}_{12}) = 2\|\boldsymbol{\Sigma}_0\|^2$, so that

$$\mathrm{E}(\mathrm{A}) \;=\; 1 + \frac{\|\boldsymbol{\mu}_0\|^2}{\mathrm{tr}(\boldsymbol{\Sigma}_0)} = 1 \tag{8}$$

$$\mathrm{Var}(\mathrm{A}) \;=\; \frac{2\|\boldsymbol{\Sigma}_0\|^2}{[\mathrm{tr}(\boldsymbol{\Sigma}_0)]^2}. \tag{9}$$

From the proof in Appendix B.1, we note that we use a slight approximation for $\mathrm{Var}(\mathrm{A}_{12})$ since the first term in $2\|\boldsymbol{\Sigma}_0\|^2$ has denominator $n_i(n_i - 1)$, not n_i^2, which, precisely, gives $\mathrm{Var}(\mathrm{A}_{12}) = 2\|\boldsymbol{\Sigma}_0\|^2[1+o(1)]$ and $\mathrm{Var}(\mathrm{A}) = [2\|\boldsymbol{\Sigma}_0\|^2/[\mathrm{tr}(\boldsymbol{\Sigma}_0)]^2][1+o(1)]$. As $(n_i-1)/n_i$ makes no difference for the final limit as $n_i \to \infty$, we skip $o(1)$ term when the context is clear.

Note also that, $\mathrm{Var}(\mathrm{A}_{11})$ is not reported in Theorem 1. It will be a part of main theorem, Theorem 2, where it is shown that A_{11} is a simple plug-in, consistent estimator of $\mathrm{E}(\mathrm{A}_{11}) = \mathrm{tr}(\boldsymbol{\Sigma}_0)$ for $n_i, p \to \infty$, in the sense that $\mathrm{A}_1 \,/\, \mathrm{tr}(\boldsymbol{\Sigma}_0)$ and $[\mathrm{A}_1 \,/\, \mathrm{tr}(\boldsymbol{\Sigma}_0)][\mathrm{tr}(\boldsymbol{\Sigma}_0)/\, \mathrm{A}_{11}]$ have essentially the same limit. We can thus consider the following test statistic for H_0

$$\mathrm{T} = \frac{\mathrm{A}_1}{\mathrm{A}_{11}} = 1 + \frac{\mathrm{A}_{12}}{\mathrm{A}_{11}}. \tag{10}$$

With normality assumption relaxed, we replace it with a general multivariate model. Given $\mathbf{X}_{ik} \in \mathbb{R}^p$, let $\mathbf{Y}_{ik} = \mathbf{X}_{ik} - \boldsymbol{\mu}_i$, and define

$$\mathbf{Y}_{ik} = \boldsymbol{\Gamma}_i \mathbf{Z}_{ik}, \quad k = 1, \ldots, n_i, \ i = 1, \ldots, g, \tag{11}$$

with $\boldsymbol{\Gamma}_i = \boldsymbol{\Sigma}_i^{1/2}$, $\mathbf{Z}_{ik} \in \mathbb{R}^p$, $\mathbf{Z}_{ik} \sim \mathcal{F}_i$, where $\mathrm{E}(\mathbf{Z}_{ik}) = \mathbf{0}_p$ and $\mathrm{Cov}(\mathbf{z}_{ik}) = \mathbf{I}_p \ \forall \ i$. Here, $\mathbf{0}_p$ is a vector of zeros and $\mathbf{I}_p$ denotes the identity matrix. We supplement Model (11) with the following assumptions, where $\nu_{is} = \lambda_{is}/p$ and λ_{is}, $s = 1, \ldots, p$, denote the eigenvalues of $\boldsymbol{\Sigma}_i$.

Assumption 1: $\mathrm{E}(Y_{iks}^4) = \gamma_{is} \leq \gamma < \infty \ \forall \ s = 1, \ldots, p, \ \forall \ i = 1, \ldots, g, \ \gamma \in \mathbb{R}^+$.

Assumption 2: $\lim_{p\to\infty} \sum_{s=1}^{p} \nu_{is} = \nu_{i0} \leq \nu \in \mathbb{R}^+, \ \forall \ i = 1, \ldots, g$.

Assumption 3: $\lim_{n_i, p\to\infty} p/n_i = \xi_i \leq \xi = O(1), \ \forall \ i = 1, \ldots, g$.

Assumption 4: $\lim_{n_i\to\infty} n_i/n = \rho_i \leq \rho = O(1), \ \forall \ i = 1, \ldots, g, \ n = \sum_{i=1}^{g} n_i$.

Assumption 5: $\lim_{p\to\infty} \boldsymbol{\mu}_i^{\mathrm{T}} \boldsymbol{\Sigma}_k \boldsymbol{\mu}_j / p = \phi_{ijk} \leq \phi = O(1), \ \forall \ i, j, k = 1, \ldots, g$.

Assumption 1 helps deal with moments of quadratic forms under Model (11). Assumption 2 is often used in high-dimensional inference. Assumptions 3-4 ensure a non-degenerate limit by controlling simultaneous rates of convergence among sample sizes and in relation to dimension. Assumption 5 is only needed under the alternative. Using Theorem 1 and the probability convergence of A_{11} (see Appendix B.2), we write

$$\mathrm{T} - 1 = \frac{\mathrm{A}_{12}}{\mathrm{tr}(\boldsymbol{\Sigma}_0)}[1 + o_P(1)],$$

with A_{12} as in (7), $\mathrm{E}(\mathrm{T} - 1) = \|\boldsymbol{\mu}_0\|^2 / \mathrm{tr}(\boldsymbol{\Sigma}_0)$ and

$$\sigma_1^2 = \frac{2\|\boldsymbol{\Sigma}_0\|^2 + \mathrm{R}}{[\mathrm{tr}(\boldsymbol{\Sigma}_0)]^2},$$

where, under H_0, $\mathrm{E}(\mathrm{T} - 1) = 0$, $\sigma_0^2 = 2\|\boldsymbol{\Sigma}_0\|^2 / [\mathrm{tr}(\boldsymbol{\Sigma}_0)]^2$. Theorem 2 gives the distribution of $\widetilde{\mathrm{T}} = (\mathrm{T} - \mathrm{E}(\mathrm{T}))/\sigma_{\mathrm{T}}$ with $\widetilde{\mathrm{T}}_0$ as its value under H_0. For proof, see Appendix B.2.

Theorem 2: Let $\widetilde{\mathrm{T}}$ be as defined above. Under Model (11) and Assumptions 1-5, $\widetilde{\mathrm{T}} \xrightarrow{\mathcal{D}} N(0, 1)$, as $n_i, p \to \infty$. In particular, under H_0, $\widetilde{\mathrm{T}}_0 \xrightarrow{\mathcal{D}} N(0, 1)$.

For power of $\widetilde{\mathrm{T}}$, let Z_α be the quantile of $Z \sim N(0, 1)$, and $\widetilde{\mathrm{T}}$, $\widetilde{\mathrm{T}}_0$ be as in Theorem 2. For any n_i and p, $P(\widetilde{\mathrm{T}}_0 \geq Z_\alpha) = \alpha$ and $P(\widetilde{\mathrm{T}} \geq -\delta + \tau Z_\alpha) = 1 - \beta$ define the size and power of the test, respectively, where $\delta = \|\boldsymbol{\mu}_0\|^2 / \sqrt{2\|\boldsymbol{\Sigma}_0\|^2 + \mathrm{R}}$ and $\tau = \sigma_0/\sigma_1$, with σ_0^2 and σ_1^2 as $\mathrm{Var}(\mathrm{T})$ under the null and alternative, respectively, as in Theorem 2. It follows, under the assumptions, that $\tau \to [2 + \xi^{-1}]^{-1/2} O(1) = O(1)$ and $\delta = n_i [2 + \xi^{-1}]^{-1/2} O(1) = O(n_i)$, so that $1 - \beta = 1$ - $P[\widetilde{\mathrm{T}} \leq -(n_i + Z_\alpha) O(1)] \Rightarrow 1$, as $n_i, p \to \infty$.

We need to estimate $\mathrm{Var}(\mathrm{T})$. As $\mathrm{A}_{11} \xrightarrow{\mathcal{P}} \sum_{i=1}^{g} \sum_{s=1}^{\infty} c_i^2 \xi_i \nu_{i0}$, $\mathrm{Var}(\mathrm{T})$ basically follows from $\mathrm{Var}(\mathrm{A}_{12})$ which is composed of $\|\boldsymbol{\Sigma}_i\|^2$ and $\|\boldsymbol{\Gamma}_i \boldsymbol{\Gamma}_j\|^2$, where $\boldsymbol{\Gamma}_i = \boldsymbol{\Sigma}_i^{1/2}$, since, under H_0,

$$\|\boldsymbol{\Sigma}_0\|^2 = \sum_{i=1}^{g} \frac{c_i^4}{n_i} \|\boldsymbol{\Sigma}_i\|^2 + 2 \sum_{\substack{i=1 \\ i<j}}^{g} \sum_{i=1}^{g} \frac{c_i^2 c_j^2}{n_i n_j} \|\boldsymbol{\Gamma}_i \boldsymbol{\Gamma}_j\|^2. \tag{12}$$

The estimators are defined as below, where $\widehat{\mathbf{\Gamma}}_i = \widehat{\mathbf{\Sigma}}_i^{1/2}$.

Definition 1: Estimators of $\|\mathbf{\Gamma}_i\mathbf{\Gamma}_j\|^2$, $\|\mathbf{\Sigma}_i\|^2$, under Model (11), are defined as below, where $\nu_i = (n_i-1)/n_i(n_i-2)(n_i-3)$, $\mathrm{Q}_i = \sum_{k=1}^{n_i}\|\widetilde{\mathbf{X}}_{ik}\|^2$, $\widetilde{\mathbf{X}}_{ik} = \mathbf{X}_{ik} - \overline{\mathbf{X}}_i$, $i = 1, \ldots, g$.

$$\mathrm{E}_{ij} = \|\widehat{\mathbf{\Gamma}}_i\widehat{\mathbf{\Gamma}}_j\|^2, \tag{13}$$

$$\mathrm{E}_i = \nu_i\left[(n_i-1)(n_i-2)\|\widehat{\mathbf{\Sigma}}_i\|^2 + [\|\widehat{\mathbf{\Gamma}}_i\|^2]^2 - n\,\mathrm{Q}_i\right], \tag{14}$$

As functions of empirical $\widehat{\mathbf{\Sigma}}_i$, the estimators are computationally very efficient. They are unbiased and high-dimensional consistent. To prove these properties, however, an alternative formulation of the same estimators, in terms of U-statistics, is very helpful.

Given Model (11), let $\mathbf{D}_{ikr} = \mathbf{Y}_{ik} - \mathbf{Y}_{ir}$ with $\mathrm{E}(\mathbf{D}_{ikr}) = \mathbf{0}$, $\mathrm{Cov}(\mathbf{D}_{ikr}) = 2\mathbf{\Sigma}_i = 2\|\mathbf{\Gamma}_i\|^2$, and $\widehat{\mathbf{\Sigma}}_i$ can be written as U-statistic with symmetric kernel $h(\mathbf{X}_{ik}, \mathbf{X}_{ir}) = \mathbf{D}_{ikr}\mathbf{D}_{ikr}^{\mathrm{T}}/2$, *i.e.*,

$$\widehat{\mathbf{\Sigma}}_i = \frac{1}{Q(n_i)}\sum_{\substack{k=1\\k\neq r}}^{n_i}\sum_{r=1}^{n_i}\frac{1}{2}\mathbf{D}_{ikr}\mathbf{D}_{ikr}^{\mathrm{T}}$$

where $Q(n_i) = n_i(n_i-1)$. Denote further $\mathrm{A}_{ijkrls} = \mathbf{D}_{ikr}^{\mathrm{T}}\mathbf{D}_{jls}$ and $\mathrm{A}_{ikrls} = \mathbf{D}_{ikr}^{\mathrm{T}}\mathbf{D}_{ils}$ with $\mathrm{E}(\mathrm{A}_{ijklrs}^2) = 4\|\mathbf{\Gamma}_i\mathbf{\Gamma}_j\|^2$, $\mathrm{E}(\mathrm{A}_{ikrls}^2) = 4\|\mathbf{\Sigma}_i\|^2$. The U-statistics forms of E_{ij} and E_i follow as

$$\mathrm{E}_{ij} = \frac{1}{Q(n_i)Q(n_j)}\sum_{\substack{k=1\\\pi(k,r)}}^{n_i}\sum_{r=1}^{n_i}\sum_{\substack{l=1\\\pi(l,s)}}^{n_j}\sum_{s=1}^{n_j}\frac{1}{4}\,\mathrm{A}_{ijklrs}^2 \tag{15}$$

$$\mathrm{E}_i = \frac{1}{P(n_i)}\sum_{\substack{k=1\\\pi(k,r,l,s)}}^{n_i}\sum_{r=1}^{n_i}\sum_{l=1}^{n_i}\sum_{s=1}^{n_i}\mathrm{B}_{ikrls}, \tag{16}$$

where $P(n_i) = n_i(n_i-1)(n_i-2)(n_i-3)$, $\mathrm{B}_{ikrls} = \mathrm{A}_{ikrls}^2 + \mathrm{A}_{iksrl}^2 + \mathrm{A}_{ilrsk}^2$ and $\pi(\cdot)$ implies all involved indices pairwise unequal. Note that, E_i and E_{ij} are one- and two-sample U-statistics with symmetric kernels $\mathrm{B}_{ikrls}/4$ and $\mathrm{A}_{ijklrs}^2/4$, respectively; see *e.g* Koroljuk and Borovskich (1994). The following theorem summarizes the properties of estimators. The proof of this theorem is a tedious computational exercise of projection properties of U-statistics and is omitted for simplicity; see *e.g* Ahmad (2017).

Theorem 3: Given Model (11), Assumption 1, and E_{ij}, E_i as in (15)-(16). Then, $\mathrm{E}(\mathrm{E}_{ij}) = \|\mathbf{\Gamma}_i\mathbf{\Gamma}_j\|^2$ and $\mathrm{E}(\mathrm{E}_i) = \|\mathbf{\Sigma}_i\|^2$. Further,

$$\mathrm{Var}(\mathrm{E}_{ij}) = \frac{2}{(n_i-1)(n_j-1)}\left[(n_i+n_j-1)\|\mathbf{\Sigma}_i\mathbf{\Sigma}_j\|^2 + \left\{\|\mathbf{\Gamma}_i\mathbf{\Gamma}_j\|^2\right\}^2 + M_2O(n) + M_3O(1)\right],$$

$$\mathrm{Var}(\mathrm{E}_i) = \frac{4}{P(n_i)}\left[a(n_i)\|\mathbf{\Sigma}_i^2\|^2 + b(n_i)\left\{\|\mathbf{\Sigma}_i\|^2\right\}^2 + M_2O(n_i^3) + M_3O(n_i^2)\right],$$

$$\mathrm{Cov}(\mathrm{E}_{ij}, \mathrm{E}_i) = \frac{4}{Q(n_i)}\left[n_i\,\mathrm{tr}(\mathbf{\Sigma}_i^3\mathbf{\Sigma}_j) + M_2O(n_i)\right],$$

where $a(n_i) = 2n_i^3 - 9n_i^2 + 9n_i - 16$, $b(n_i) = n_i^2 - 3n_i + 8$, $P(n_i) = n_i(n_i-1)(n_i-2)(n_i-3)$, $Q(n_i) = n_i(n_i-1)$, and M_2, M_3 are given in Lemma 1.

Note that, less emphasis on terms involving M_2, M_3 *etc.* is due to the fact that they eventually vanish, exactly under normality, and asymptotically under Model (11) and the assumptions. For the rest, Theorem 3 yields $\mathrm{Var}(\mathrm{E}_i \,/\, \mathrm{E}(\mathrm{E}_i)) \leq O(1/n_i)$, $\mathrm{Var}(\mathrm{E}_{ij} \,/\, \mathrm{E}(\mathrm{E}_{ij})) \leq O(1/n_i{+}1/n_j)$, $\mathrm{Cov}(\mathrm{E}_i \,/\, \mathrm{E}(\mathrm{E}_i), \mathrm{E}_{ij} \,/\, \mathrm{E}(\mathrm{E}_{ij})) \leq O(1/n_i)$, *i.e.*, the ratios are uniformly bounded in p. This, in particular, implies that p does not influence the non-degenerate limit of $\widetilde{\mathrm{T}}$ in Theorem 2. Following corollary can now replace Theorem 2 for practical applications.

Corollary 3.1: Theorem 2 remains valid if $\mathrm{Var}(\widetilde{\mathrm{T}})$ is replaced with $\widehat{\mathrm{Var}}(\widetilde{\mathrm{T}})$ obtained by substituting E_{ij} for $\|\boldsymbol{\Gamma}_i\boldsymbol{\Gamma}_j\|^2$ and E_i for $\|\boldsymbol{\Sigma}_i\|^2$ in (12).

3. Test of a set of orthogonal contrasts

Often, the researcher is interested to simultaneously test a set of multiple contrasts. In principal, this set can be of any cardinality, but only a set of orthogonal contrasts makes sense since any contrast beyond orthogonal set will carry redundant information. For g populations, an orthogonal set consists of $m = g - 1$ contrasts. We are thus interested in simultaneous testing of a set of m contrasts, *i.e.*,

$$\mathrm{H}_{0q} : \sum_{i=1}^{g} c_{iq}\boldsymbol{\mu}_{iq} = \mathbf{0} \;\; \text{vs.} \;\; \mathrm{H}_{1q} : \text{Not } \mathrm{H}_{0q}, \quad q = 1, \ldots, m, \tag{17}$$

where $\sum_{i=1}^{g} c_{iq} = 0$, as before, with additional orthogonality constraint, $\sum_{i=1}^{g} c_{iq}c_{iq'} = 0$, $q \neq q'$. Extending the notations in Sec. 2, we can re-write the set of hypotheses in (17) as

$$\mathrm{H}_{0s} : \boldsymbol{\Xi}_s = \mathbf{0} \;\; \text{vs.} \;\; \mathrm{H}_{1s} : \text{Not } \mathrm{H}_{0s}, \tag{18}$$

where s refers to the *set* of contrasts, with $\boldsymbol{\Xi}_s = (\boldsymbol{\mu}_{01}^{\mathrm{T}}, \ldots, \boldsymbol{\mu}_{0m}^{\mathrm{T}})$, $\boldsymbol{\mu}_{0q} = \sum_{i=1}^{g} c_{iq}\boldsymbol{\mu}_{iq}$. Letting $\overline{\mathbf{X}}_{0q} = \sum_{i=1}^{g} c_{iq}\overline{\mathbf{X}}_{iq}$ estimate $\boldsymbol{\mu}_{0q}$, an estimator of $\boldsymbol{\Xi}_s \in \mathbb{R}^{m\times p}$ follows as

$$\mathbf{M}_s = (\overline{\mathbf{X}}_{01}^{\mathrm{T}}, \ldots, \overline{\mathbf{X}}_{0m}^{\mathrm{T}}) \in \mathbb{R}^{m\times p}.$$

Denoting $\boldsymbol{\Sigma}_{0q} = \sum_{i=1}^{g} c_{iq}^2\boldsymbol{\Sigma}_{iq}/n_i$ and using $\mathrm{Cov}(\overline{\mathbf{X}}_{0q}, \overline{\mathbf{X}}_{0q'}) = \mathbf{0}$ for $q \neq q'$„ we get

$$\mathrm{E}(\mathbf{M}_s) = \boldsymbol{\Xi}_s \;\; \text{and} \;\; \mathrm{Cov}(\mathbf{M}_s) = \boldsymbol{\Sigma}_s = \mathrm{diag}(\boldsymbol{\Sigma}_{01}, \ldots, \boldsymbol{\Sigma}_{0m}) = \oplus_{q=1}^{m}\boldsymbol{\Sigma}_{0q},$$

It is obvious then that the theory for m orthogonal contrasts extends straightforwardly from that of one contrast in Sec. 2, where the orthogonality condition particularly simplifies the computations. Thus, partitioning $\|\overline{\mathbf{X}}_{0q}\|^2$ similarly as $\|\overline{\mathbf{X}}_0\|^2$ in Sec. 2, we have

$$\|\overline{\mathbf{X}}_{0q}\|^2 = \sum_{i=1}^{g} c_{iq}^2 Q_{iq} + \sum_{i=1}^{g} c_{iq}^2 U_{iq} + 2\sum_{\substack{i=1 \\ i<j}}^{g}\sum_{j=1}^{g} c_{iq}c_{jq}U_{ijq} \;\; = \;\; \mathrm{A}_{11q} + \mathrm{A}_{12q},$$

with

$$\mathrm{A}_{11q} = \sum_{i=1}^{g} c_{iq}^2 Q_{iq}, \quad \mathrm{A}_{12q} = \sum_{i=1}^{g} c_{iq}^2 U_{iq} + 2\sum_{\substack{i=1 \\ i<j}}^{g}\sum_{j=1}^{g} c_{iq}c_{jq}U_{ijq},$$

where $Q_{iq} = (E_{iq} - U_{iq})/n_i$, E_{iq}, U_{iq}, U_{ijq} are defined as for single contrast, except for each q now. The rest of the theory proceeds likewise, so that Theorem 2 and Corollary 3.1 stand

Table 1: Estimated size of $\widetilde{\mathrm{T}}$ for three distributions with three covariance triplets

		Normal			Uniform			Exponential		
n_1, n_2, n_3	p	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3
10, 20, 30	50	0.061	0.044	0.060	0.048	0.060	0.052	0.062	0.058	0.065
	100	0.051	0.051	0.057	0.052	0.054	0.056	0.055	0.055	0.058
	300	0.048	0.055	0.054	0.055	0.055	0.053	0.057	0.052	0.054
	500	0.050	0.056	0.049	0.051	0.055	0.053	0.055	0.057	0.057
	1000	0.044	0.052	0.051	0.046	0.048	0.051	0.052	0.054	0.055
20, 30, 50	50	0.050	0.048	0.045	0.049	0.052	0.045	0.049	0.057	0.054
	100	0.052	0.057	0.046	0.056	0.055	0.047	0.047	0.055	0.052
	300	0.054	0.053	0.054	0.052	0.048	0.051	0.055	0.054	0.053
	500	0.047	0.050	0.048	0.055	0.052	0.052	0.058	0.056	0.055
	1000	0.051	0.053	0.055	0.053	0.053	0.048	0.051	0.048	0.051
30, 50, 100	50	0.054	0.053	0.048	0.051	0.054	0.052	0.056	0.049	0.048
	100	0.057	0.049	0.055	0.050	0.051	0.055	0.055	0.054	0.052
	300	0.055	0.048	0.052	0.055	0.054	0.050	0.053	0.052	0.052
	500	0.055	0.047	0.050	0.054	0.052	0.051	0.053	0.050	0.047
	1000	0.049	0.051	0.053	0.049	0.051	0.052	0.049	0.053	0.051

valid for any T_q defined for qth contrast, using corresponding A_{11q} and A_{12q}. We therefore leave the unnecessarily repetitive details, and rather focus on the following important remarks which highlights the essential differences with the single contrast case.

First, the emphasis on making an orthogonal set of contrasts is due to the fact that such a set picks all information from the data without retaining much redundancies. It is further substantiated by the orthogonality condition, $\sum_{i=1}^{g} c_{iq}c_{iq'} = 0$, which, because of $\sum_{i=1}^{g} c_{iq} = 0$, mimics the numerator of a covariance.

Second, the theory of set of orthogonal contrasts pertains to the case of planned comparisons within the ambit of multiple testing. It differs from, *e.g*, Scheffé's method of all possible contrasts (Scheffé, 1959, Ch. 3), originally devised as a post-hoc strategy after global univariate ANOVA hypothesis is rejected. Scheffé's method allows infinitely many contrasts, although practically only a finite set is recommended and practically used in order to keep better error control.

Third, since many hypotheses are tested simultaneously, an error control mechanism is called for. With g relatively small or moderate in practice, a simple Bonferroni adjustment would suffice, which controls the family wise error rate in the strong sense. Otherwise, some researchers recommend a comparison-wise error control. For comprehensive theoretical results on multiple testing and error control procedures, see Dickhaus (2014). For a high-dimensional multiple testing framework, see Ahmad (2019a) and the references therein.

4. Simulations

We assess the accuracy of the proposed test statistics, particularly focusing on its robustness to normality assumption and validity under high-dimensional settings. For simplicity, we consider $\widetilde{\mathrm{T}}$ in Theorem 2 for $g = 3$. We generate p-dimensional random vectors

of sizes (n_1, n_2, n_3) from normal, exponential and uniform distributions, where

$$n_1 \in \{10, 20, 30\},\ n_2 = \{20, 30, 50\},\ n_3 = \{30, 50, 100\},\ p = \{50, 100, 300, 500, 1000\}.$$

All (n_1, n_2, n_3, p) quadruplets are used under three triplets of covariance structures, denoted, S_1: (CS, AR, UN), S_2: (CS1, CS2, AR) and S_3: (CS, AR1, AR2), which are composed of compound symmetric (CS), Autoregressive of order 1, AR(1), and unstructured (UN) matrices. They are defined as CS: $\mathbf{\Sigma} = (1-\rho)\mathbf{I}_p + \rho\mathbf{J}_p$, AR: $\text{Cov}(X_k, X_l) = \kappa\rho^{|k-l|}$, $\forall\ k, l$, and UN: $\mathbf{\Sigma} = (\sigma_{ij})_{i,j=1}^{p}$ with $\sigma_{ij} = 1(1)p$ $(i = j)$ and $\rho_{ij} = (i-1)/p$ $(i > j)$, respectively. Here, $\mathbf{I}$ denotes identity matrix and $\mathbf{J}$ is a matrix of 1s. We use $\kappa = 1$ for all three triplets, where CS1, CS2, and AR1, AR2, refer to CS and AR with $\rho = 0.3$ and $\rho = 0.7$, respectively.

The nominal level is set at $\alpha = 0.05$, for both size and power. Moreover, for power, we set H_1 by letting $\boldsymbol{\mu}_1$ deviate from H_0 in a monotonically increasing fashion, *i.e.*, $\boldsymbol{\mu}_1 = \delta_r \mathbf{p}_1$, $\mathbf{p}_1 = (1/p, \ldots, p/p)$, $\delta_r = 0.2(0.2)1$. Finally, we used two contrast vectors, $(c_1, c_2, c_3) = (1, 0, -1)$ and $(1, -2, 1)$, and due to similarity of results, only the first case is discussed here.

The estimated size and power are obtained by averaging over 2000 simulations. Table 1 reports the size for all distributions and Tables 2-3 report the power, respectively, for normal and uniform distributions. The power results for exponential distributions were very similar to those for uniform, and are therefore not reported.

All simulations are carried out in SAS/IML where data for the three multivariate distributions are generated by using appropriate arguments for corresponding built-in SAS/IML functions. For this, first, by using the square root of the assumed covariance matrix (*e.g*, compound symmetry), a sequence of multivariate normal vectors is generated which follows this structure. Then, by probability integral transform, it is converted into a uniform sequence, say U, which is adjusted for its mean and variance by subtracting a mean vector of length p (with each entry 1/2) and multiplying by the square root of a $p \times p$ diagonal

Table 2: Estimated power of $\widetilde{\text{T}}$ for normal distribution with three covariance triplets

		S_1			S_2			S_3		
n_1, n_2, n_3	p/δ	0.2	0.6	1.0	0.2	0.6	1.0	0.2	0.6	1.0
10, 20, 30	50	0.151	0.963	1.000	0.146	0.962	1.000	0.131	0.923	1.000
	100	0.203	0.998	1.000	0.195	0.985	1.000	0.187	0.998	1.000
	300	0.351	1.000	1.000	0.329	1.000	1.000	0.345	1.000	1.000
	500	0.441	1.000	1.000	0.461	1.000	1.000	0.438	1.000	1.000
	1000	0.528	1.000	1.000	0.545	1.000	1.000	0.582	1.000	1.000
20, 30, 50	50	0.263	1.000	1.000	0.269	1.000	1.000	0.252	1.000	1.000
	100	0.385	1.000	1.000	0.392	1.000	1.000	0.368	1.000	1.000
	300	0.700	1.000	1.000	0.699	1.000	1.000	0.714	1.000	1.000
	500	0.865	1.000	1.000	0.866	1.000	1.000	0.831	1.000	1.000
	1000	0.942	1.000	1.000	0.998	1.000	1.000	0.917	1.000	1.000
30, 50, 100	50	0.426	1.000	1.000	0.435	1.000	1.000	0.423	1.000	1.000
	100	0.637	1.000	1.000	0.647	1.000	1.000	0.649	1.000	1.000
	300	0.956	1.000	1.000	0.948	1.000	1.000	0.947	1.000	1.000
	500	1.000	1.000	1.000	0.994	1.000	1.000	0.999	1.000	1.000
	1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

covariance matrix (with each diagonal entry 1/12). The exponential distribution follows by an additional log-transformation of U followed by its corresponding adjustment.

We observe accurate size control under all parameters. In particular, the performance for small or moderate sample sizes and for increasing dimension, for all covariance triplets, is noteworthy. A slight fluctuation of size can be seen for the exponential distribution but it stabilizes itself for even moderate sample sizes. Of particular mention is the power which is not only reasonably high, but also increases for increasing p as well as for increasing n_i.

The performance of the statistic for non-normal cases further implies its robustness under the general model. The overall performance of the statistic supports its use in practice for high-dimensional data with moderate sample sizes and departures from normality.

Table 3: Estimated power of $\widetilde{\mathrm{T}}$ for uniform distribution with three covariance triplets

		S_1			S_2			S_3		
n_1, n_2, n_3	p/δ	0.2	0.6	1.0	0.2	0.6	1.0	0.2	0.6	1.0
10, 20, 30	50	0.143	0.961	1.000	0.134	0.965	1.000	0.142	0.956	1.000
	100	0.198	0.995	1.000	0.190	0.998	1.000	0.181	0.999	1.000
	300	0.315	1.000	1.000	0.329	1.000	1.000	0.351	1.000	1.000
	500	0.463	1.000	1.000	0.472	1.000	1.000	0.446	1.000	1.000
	1000	0.586	1.000	1.000	0.619	1.000	1.000	0.530	1.000	1.000
20, 30, 50	50	0.255	1.000	1.000	0.240	0.998	1.000	0.243	1.000	1.000
	100	0.349	1.000	1.000	0.368	1.000	1.000	0.368	1.000	1.000
	300	0.697	1.000	1.000	0.686	1.000	1.000	0.705	1.000	1.000
	500	0.884	1.000	1.000	0.805	1.000	1.000	0.811	1.000	1.000
	1000	0.883	1.000	1.000	0.936	1.000	1.000	0.960	1.000	1.000
30, 50, 100	50	0.412	1.000	1.000	0.415	1.000	1.000	0.447	1.000	1.000
	100	0.660	1.000	1.000	0.618	1.000	1.000	0.628	1.000	1.000
	300	0.957	1.000	1.000	0.958	1.000	1.000	0.954	1.000	1.000
	500	0.998	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000
	1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

5. Discussion and remarks

A test statistic for contrast comparison of mean vectors is introduced when the dimension of the vectors is large, even exceeding the number of vectors. Relaxing normality assumptions, properties of the test statistic, including its limit under high-dimensional set up, is provided for a general multivariate model and a few mild assumptions. The statistic is simple and composed of computationally efficient estimators. Simulations are used to demonstrate the theoretical properties of the test statistic. An extension to a set of orthogonal contrasts is also given.

Acknowledgements

The author is thankful to the editors, and particularly to a reviewer whose comments helped improve the section on simulations.

References

Ahmad, R. (2014). A U-statistic approach for a high-dimensional two-sample mean testing problem under non-normality and Behrens-Fisher setting. *Annals of the Institute of Statistical Mathematics*, **66**, 33-61.

Ahmad, R. (2017). Location-invariant multi-sample U-tests for covariance matrices with large dimension. *Scandinavian Journal of Statistics*, **44**, 500-523.

Ahmad, R. (2019a). Multiple comparisons of high-dimensional mean vectors under general conditions. *Journal of Statistical Computations and Simulation*, **89**, 1044-1059.

Ahmad, R. (2019b). A unified approach to testing mean vectors with large dimensions. *AStA: Advances in Statistical Analysis*, **103**, 593-618.

Ahmad, R. (2022). Tests for proportionality of matrices with large dimension. *Journal of Multivariate Analysis*, **189**, 104865.

Anderson, T. W. (2003). *Introduction to Multivariate Statistical Analysis*, 3rd edition, Wiley New York.

Dickhaus T. (2014). *Simultaneous Statistical Inference.* Springer, New York.

Hayter, A. J. (2014). Inferences on linear combinations of normal means with unknown and unequal variances. *Sankhya*, **76A**, 257-279.

Jiang, J. (2010). *Large Sample Techniques for Statistics.* Springer, New York.

Koroljuk, V. S. and Borovskich, Y.V. (1994). *Theory of U-statistics.* Kluwer Press, Dordrecht.

Lee, A. J. (1990). *U-Statistics: Theory & Practice*, Marcel Dekker, New York.

Mardia, K. V, J. T. Kent, and C. C. Taylor (2024). *Multivariate Analysis*, 2nd edition, Wiley, New York.

Scheffé, H. (1959). *The Analysis of Variance*, Wiley, New York.

Searle, S. R. (1971). *Linear Models.* Wiley, New York.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press, Cambridge.

APPENDIX

A. Some basic results

Lemma 1: For $\mathbf{Z}_{ik} \in \mathbb{R}^p$, $k = 1, \ldots, n_i$, defined in Model (11), let $\mathbf{Z}_{ik}^{\mathrm{T}}\mathbf{Z}_{ik}$ and $\mathbf{Z}_{ik}^{\mathrm{T}}\mathbf{Z}_{ir}$, $k \neq r$, be quadratic and bilinear form of independent components from sample i, and $\mathbf{Z}_{ik}^{\mathrm{T}}\mathbf{Z}_{jl}$, $k \neq l$, $i \neq j$, be the bilinear form composed of vectors from two independent samples. Also let γ be as defined in Assumption 1 and $\odot$ denotes the Hadamard product. Then $\mathrm{E}(\mathbf{Z}_{ik}^{\mathrm{T}}\mathbf{Z}_{ir}) = 0$, $\mathrm{E}(\mathbf{Z}_{ik}^{\mathrm{T}}\mathbf{Z}_{jl}) = 0$, $\mathrm{E}(\mathbf{Z}_{ik}^{\mathrm{T}}\mathbf{Z}_{ik}) = \|\mathbf{\Gamma}_i\|^2$, $\mathrm{E}(\mathbf{Z}_{ik}^{\mathrm{T}}\mathbf{Z}_{ir})^2 = \|\mathbf{\Sigma}_i\|^2$, $\mathrm{E}(\mathbf{Z}_{ik}^{\mathrm{T}}\mathbf{Z}_{jl})^2 = \|\mathbf{\Gamma}_i\mathbf{\Gamma}_j\|^2$. Further,

$$\begin{aligned}
\mathrm{E}(\mathbf{Z}_{ik}^{\mathrm{T}}\mathbf{Z}_{ik})^2 &= 2\|\mathbf{\Sigma}_i\|^2 + [\|\mathbf{\Gamma}_i\|^2]^2 + M_1 \\
\mathrm{E}(\mathbf{Z}_{ik}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{Z}_{ik})^2 &= 2\|\mathbf{\Sigma}_i^2\|^2 + [\|\mathbf{\Sigma}_i\|^2]^2 + M_2 \\
\mathrm{E}(\mathbf{Z}_{ik}^{\mathrm{T}}\mathbf{Z}_{ir})^4 &= 6\|\mathbf{\Sigma}_i^2\|^2 + 3[\|\mathbf{\Sigma}_i\|^2]^2 + M_3 \\
\mathrm{E}(\mathbf{Z}_{ik}^{\mathrm{T}}\mathbf{Z}_{jl})^4 &= 6\|\mathbf{\Sigma}_i\mathbf{\Sigma}_j\|^2 + 3[\|\mathbf{\Gamma}_i\mathbf{\Gamma}_j\|^2]^2 + M_4,
\end{aligned}$$

with $M_1 = (\gamma - 3)\operatorname{tr}(\mathbf{\Sigma}_i \odot \mathbf{\Sigma}_i)$, $M_2 = (\gamma_i - 3)\operatorname{tr}(\mathbf{\Sigma}_i^2 \odot \mathbf{\Sigma}_i^2)$, $M_3 = 6(\gamma - 3)\operatorname{tr}(\mathbf{\Sigma}_i^2 \odot \mathbf{\Sigma}_i^2) + (\gamma_i - 3)^2\operatorname{tr}(\mathbf{\Sigma}_i \odot \mathbf{\Sigma}_i)^2$, and $M_4 = 6(\gamma - 3)\operatorname{tr}(\mathbf{\Sigma}_i^2 \odot \mathbf{\Sigma}_j^2) + (\gamma - 3)^2\operatorname{tr}(\mathbf{\Sigma}_i \odot \mathbf{\Sigma}_i)\operatorname{tr}(\mathbf{\Sigma}_j \odot \mathbf{\Sigma}_j)$.

All moments in Lemma 1 reduce to those under normality for $\gamma = 3$; see Searle (1971).

Lemma 2: (Jiang, 2010, Page 183) Let $Y_1, Y_2, \ldots$ be iid r.vs. with $\mathrm{E}(Y_i) = 0$, $\mathrm{Var}(Y_i) = 1$, and b_{ni} be constants, $1 \leq i \leq n$. Then $\sum_{i=1}^n b_{ni}Y_i \xrightarrow{\mathcal{D}} N(0,1)$ as $n \to \infty$, if $\max_i b_{ni}^2 \to 0$.

B. Main proofs

B.1. Proof of theorem 1

With $\mathrm{E}(Q_i) = \|\mathbf{\Gamma}_i\|^2/n_i$, $\mathrm{E}(U_i) = \|\boldsymbol{\mu}_i\|^2$, $\mathrm{E}(U_{ij}) = \langle\boldsymbol{\mu}_i, \boldsymbol{\mu}_j\rangle$, we get, by independence,

$$\mathrm{E}(\mathrm{A}_{11}) = \sum_{i=1}^{g} c_i^2\|\mathbf{\Gamma}_i\|^2/n_i = \operatorname{tr}(\mathbf{\Sigma}_0)$$

$$\mathrm{E}(\mathrm{A}_{12}) = \sum_{i=1}^{g} c_i^2\|\boldsymbol{\mu}_i\|^2 + 2\sum_{\substack{i=1 \\ i<j}}^{g}\sum_{j=1}^{g} c_ic_j\langle\boldsymbol{\mu}_i, \boldsymbol{\mu}_j\rangle = \left(\sum_{i=1}^{g} c_i\boldsymbol{\mu}_i\right)^{\mathrm{T}}\left(\sum_{i=1}^{g} c_i\boldsymbol{\mu}_i\right) = \|\boldsymbol{\mu}_0\|^2.$$

$$\begin{aligned}
\mathrm{Var}(\mathrm{A}_{12}) &= \mathrm{Var}\left(\sum_{i=1}^{g} c_i^2U_i + 2\sum_{i<j} c_ic_jU_{ij}\right) \\
&= \mathrm{Var}\left(\sum_{i=1}^{g} c_i^2U_i\right) + 4\,\mathrm{Var}\left(\sum_{\substack{i=1 \\ i<j}}^{g}\sum_{j=1}^{g} c_ic_jU_{ij}\right) + 4\,\mathrm{Cov}\left(\sum_{i=1}^{g} c_i^2U_i, \sum_{\substack{i=1 \\ i<j}}^{g}\sum_{j=1}^{g} c_ic_jU_{ij}\right) \\
&= \sum_{i=1}^{g} c_i^4\,\mathrm{Var}(U)_i + 4\sum_{\substack{i=1 \\ i<j}}^{g}\sum_{j=1}^{g} c_i^2c_j^2\,\mathrm{Var}(U_{ij}) + 8\sum_{\substack{i=1 \\ i<j<j'}}^{g}\sum_{j=1}^{g}\sum_{j'=1}^{g} c_i^2c_jc_{j'}\,\mathrm{Cov}(U_{ij}, U_{ij'}) \\
&\quad + 8\sum_{\substack{i=1 \\ i<i'<j}}^{g}\sum_{i'=1}^{g}\sum_{j=1}^{g} c_ic_{i'}c_j^2\,\mathrm{Cov}(U_{ij}, U_{i'j}) + 4\sum_{\substack{i=1 \\ i<j}}^{g}\sum_{j=1}^{g} c_i^3c_j\,\mathrm{Cov}(U_i, U_{ij}) + 4\sum_{\substack{i=1 \\ i<j}}^{g}\sum_{j=1}^{g} c_i^3c_j \\
&\quad \mathrm{Cov}(U_j, U_{ij})
\end{aligned}$$

where the remaining covariances vanish when all indices are unequal. Using the second order moments of one- and two-sample U-statistics (see *e.g* Koroljuk and Borovskich, 1994), *i.e.*,

$$\begin{aligned}\mathrm{Var}(U_{n_i}) = &\quad 2\left[2(n_i-1)\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{\mu}_i + \|\boldsymbol{\Sigma}_i\|^2\right]/n_i(n_i-1)\\ \mathrm{Var}(U_{n_in_j}) = &\quad [n_i\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_j\boldsymbol{\mu}_i + n_j\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{\mu}_j + \|\boldsymbol{\Gamma}_i\boldsymbol{\Gamma}_j\|^2]/n_in_j\end{aligned}$$

with (see also Ahmad, 2019b) $\mathrm{Cov}(U_{n_i}, U_{n_in_j}) = 2\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{\mu}_i/n_i$, $\mathrm{Cov}(U_{n_j}, U_{n_in_j}) = 2\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_j\boldsymbol{\mu}_j/n_j$, $\mathrm{Cov}(U_{n_in_j}, U_{n_in_{j'}}) = \boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{\mu}_{j'}/n_i$, and $\mathrm{Cov}(U_{n_in_j}, U_{n_{i'}n_j}) = \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_j\boldsymbol{\mu}_{i'}/n_j$, we get

$$\begin{aligned}\mathrm{Var}(\mathrm{A}_{12}) &= 2\sum_{i=1}^{g}\frac{c_i^4\|\boldsymbol{\Sigma}_i\|^2}{n_i(n_i-1)} + 4\sum_{\substack{i=1\\i<j}}^{g}\sum_{j=1}^{g}\frac{c_i^2c_j^2\|\boldsymbol{\Gamma}_i\boldsymbol{\Gamma}_j\|^2}{n_in_j} + 4\sum_{i=1}^{g}\frac{c_i^4\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{\mu}_i}{n_i}\\ &\quad + 4\left(\sum_{\substack{i=1\\i<j}}^{g}\sum_{j=1}^{g}\frac{c_i^2c_j^2\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_j\boldsymbol{\mu}_i}{n_j} + \sum_{\substack{i=1\\i<j}}^{g}\sum_{j=1}^{g}\frac{c_j^2c_i^2\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{\mu}_j}{n_i}\right)\\ &\quad + 8\left(\sum_{\substack{i=1\\i<j}}^{g}\sum_{j=1}^{g}\frac{c_ic_j^3\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_j\boldsymbol{\mu}_j}{n_j} + \sum_{\substack{i=1\\i<j}}^{g}\sum_{j=1}^{g}\frac{c_i^3c_j\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{\mu}_i}{n_i}\right)\\ &\quad + 8\left(\sum_{\substack{i=1\\i<j<j'}}^{g}\sum_{j=1}^{g}\sum_{j'=1}^{g}\frac{c_i^2c_jc_{j'}\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{\mu}_{j'}}{n_i} + \sum_{\substack{i=1\\i<i'<j}}^{g}\sum_{i'=1}^{g}\sum_{j=1}^{g}\frac{c_ic_{i'}c_j^2\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_j\boldsymbol{\mu}_{i'}}{n_j}\right).\end{aligned}$$

Slightly re-arranging the terms, we get the required expression as

$$\begin{aligned}\mathrm{Var}(\mathrm{A}_{12}) &= 2\sum_{i=1}^{g}c_i^4\frac{\|\boldsymbol{\Sigma}_i\|^2}{n_i(n_i-1)} + 4\sum_{\substack{i=1\\i<j}}^{g}\sum_{j=1}^{g}c_i^2c_j^2\frac{\|\boldsymbol{\Gamma}_i\boldsymbol{\Gamma}_j\|^2}{n_in_j} + 4\sum_{i=1}^{g}(c_i\boldsymbol{\mu}_i)^{\mathrm{T}}\frac{c_i^2\boldsymbol{\Sigma}_i}{n_i}(c_i\boldsymbol{\mu}_i)\\ &\quad + 4\left(\sum_{\substack{i=1\\i<j}}^{g}\sum_{j=1}^{g}(c_i\boldsymbol{\mu}_i)^{\mathrm{T}}\frac{c_j^2\boldsymbol{\Sigma}_j}{n_j}(c_i\boldsymbol{\mu}_i) + \sum_{\substack{i=1\\i<j}}^{g}\sum_{j=1}^{g}(c_j\boldsymbol{\mu}_j)^{\mathrm{T}}\frac{c_i^2\boldsymbol{\Sigma}_i}{n_i}(c_j\boldsymbol{\mu}_j)\right)\\ &\quad + 8\left(\sum_{\substack{i=1\\i<j}}^{g}\sum_{j=1}^{g}(c_i\boldsymbol{\mu}_i)^{\mathrm{T}}\frac{c_j^2\boldsymbol{\Sigma}_j}{n_j}(c_j\boldsymbol{\mu}_j) + \sum_{\substack{i=1\\i<j}}^{g}\sum_{j=1}^{g}(c_j\boldsymbol{\mu}_j)^{\mathrm{T}}\frac{c_i^2\boldsymbol{\Sigma}_i}{n_i}(c_i\boldsymbol{\mu}_i)\right)\\ &\quad + 8\left(\sum_{\substack{i=1\\i<i'<j}}^{g}\sum_{j=1}^{g}\sum_{j'=1}^{g}(c_i\boldsymbol{\mu}_i)^{\mathrm{T}}\frac{c_j^2\boldsymbol{\Sigma}_j}{n_j}(c_{i'}\boldsymbol{\mu}_{i'}) + \sum_{\substack{i=1\\i<j<j'}}^{g}\sum_{i'=1}^{g}\sum_{j=1}^{g}(c_j\boldsymbol{\mu}_j)^{\mathrm{T}}\frac{c_i^2\boldsymbol{\Sigma}_i}{n_i}(c_{j'}\boldsymbol{\mu}_{j'})\right)\\ &= 2\sum_{i=1}^{g}c_i^4\frac{\|\boldsymbol{\Sigma}_i\|^2}{n_i(n_i-1)} + 4\sum_{\substack{i=1\\i<j}}^{g}\sum_{j=1}^{g}c_i^2c_j^2\frac{\|\boldsymbol{\Gamma}_i\boldsymbol{\Gamma}_j\|^2}{n_in_j}. \qquad (19)\end{aligned}$$

B.2. Proof of theorem 2

The strategy, as explained around Theorem 2, is to combine the consistency of A_{11} and weak limit of A_{12}. First, $\mathrm{E}(\mathrm{E}_i) = \mathrm{tr}(\boldsymbol{\Sigma}_i)/n_i + \|\boldsymbol{\mu}_i\|^2$, $\mathrm{E}(U_i) = \|\boldsymbol{\mu}_i\|^2$, give $\mathrm{E}(Q_i) = \mathrm{tr}(\boldsymbol{\Sigma}_i)/n_i$, independent of $\boldsymbol{\mu}_i$. As c_i are known constants, we get, for $\mathrm{A}_{11} = \sum_{i=1}^{g} c_i^2 Q_i$, by independence,

$$\mathrm{Var}(\mathrm{A}_{11}) = \sum_{i=1}^{g} c_i^4 \,\mathrm{Var}(Q_i),$$

where $Q_i = (\mathrm{E}_i - U_i)/n_i$. It thus suffices to focus on Q_i. From Lemma 1 and Sec. B.1,

$$\begin{aligned}
\mathrm{Var}(Q_i) &\leq \frac{1}{n_i^2}\left\{\mathrm{Var}(\mathrm{E}_i) + \mathrm{Var}(U_i)\right\} = \frac{1}{n_i^2}\left\{\frac{1}{n_i}\,\mathrm{Var}(\|\mathbf{X}_{ik}\|^2) + \frac{2\|\boldsymbol{\Sigma}_i\|^2}{n_i(n_i-1)}\right\} \\
&\leq \frac{1}{n_i^2}\left\{\frac{(\gamma_i-1)\|\boldsymbol{\Sigma}_i\|^2}{n_i} + \frac{2\|\boldsymbol{\Sigma}_i\|^2}{n_i(n_i-1)}\right\} = \frac{\gamma_i+1}{n_i^3}\|\boldsymbol{\Sigma}_i\|^2 \\
&\leq (\gamma_i+1)c_i^2 O\left(\frac{1}{n_i}\right),
\end{aligned}$$

under the assumptions. It proves the consistency of Q_i, hence of A_{11}, as $n_i, p \to \infty$. Now consider T in (10) which, using the consistency of A_{11}, can be written as

$$\mathrm{T} - 1 = \frac{\mathrm{A}_{12}}{\mathrm{tr}(\boldsymbol{\Sigma}_0)} \cdot \frac{\mathrm{tr}(\boldsymbol{\Sigma}_0)}{\mathrm{A}_{11}} = \frac{\mathrm{A}_{12}}{\mathrm{tr}(\boldsymbol{\Sigma}_0)}[1 + o_P(1)].$$

Using moments in Sec. B.1 and, for convenience, ignoring the $o_P(1)$ factor, we have

$$\mathrm{E}(\mathrm{T} - 1) = \frac{\|\boldsymbol{\mu}_0\|^2}{\mathrm{tr}(\boldsymbol{\Sigma}_0)}, \quad \sigma_1^2 = \frac{2\|\boldsymbol{\Sigma}_0\|^2 + \mathrm{R}}{[\mathrm{tr}(\boldsymbol{\Sigma}_0)]^2},$$

which, under H_0, reduce, respectively, to $\mathrm{E}(\mathrm{T} - 1) = 0$, $\sigma_0^2 = 2\|\boldsymbol{\Sigma}_0\|^2/[\mathrm{tr}(\boldsymbol{\Sigma}_0)]^2$, where R is given in Theorem 1. Denote $\mathbf{U} = (\mathbf{U}_1^{\mathrm{T}},\ \mathbf{U}_2^{\mathrm{T}})^{\mathrm{T}}$, where the sub-vectors,

$$\mathbf{U}_1 = (c_1^2 U_1, \ldots, c_g^2 U_g)^{\mathrm{T}}, \quad \mathbf{U}_2 = (c_1c_2U_{12}, \ldots, c_1c_gU_{1g}, c_2c_1U_{21}, c_2c_3U_{23}, \ldots, c_{g-1}c_gU_{g-1,g})^{\mathrm{T}}$$

are composed of one- and two-sample U-statistics of all distinct pairs, respectively. We can write $\mathrm{A}_{12} = \mathbf{1}_G^{\mathrm{T}}\mathbf{U}$, with $\mathbf{1}_G$ a vector of all 1s of dimension $G = g + g(g-1) = g^2$. Note that, elements in $\mathbf{U}_2$ such as U_{12} and U_{21} are same, by symmetry of the kernel, but are repeated to count all possible cases, so that A_{12} can be represented as a linear combination of the entire vector $\mathbf{U}$. We note that $\mathrm{E}(\mathrm{A}_{12}) = \mathbf{1}^{\mathrm{T}}\,\mathrm{E}(\mathbf{U}) = \|\boldsymbol{\mu}_0\|^2$ and $\mathrm{Var}(\mathrm{A}_{12}) = \mathbf{1}^{\mathrm{T}}\,\mathrm{Cov}(\mathbf{U})\mathbf{1} = 2\|\boldsymbol{\Sigma}_0\|^2 + \mathrm{R}$, as in Theorem 1, where

$$\mathrm{Cov}(\mathbf{U}) = \begin{pmatrix} \mathrm{Cov}(\mathbf{U}_1) & \mathrm{Cov}(\mathbf{U}_1, \mathbf{U}_2) \\ \mathrm{Cov}(\mathbf{U}_2, \mathbf{U}_1) & \mathrm{Cov}(\mathbf{U}_2) \end{pmatrix}.$$

It follows that $\mathrm{Cov}(\mathbf{U}_1)$ and $\mathrm{Cov}(\mathbf{U}_2)$, on the diagonal of $\mathrm{Cov}(\mathbf{U})$, lead to $2\|\boldsymbol{\Sigma}_0\|^2$ in $\mathrm{Var}(\mathrm{A}_{12})$, where $\mathrm{Cov}(\mathbf{U}_1, \mathbf{U}_2)$ leads to R. Further, under independence, $\mathrm{Cov}(\mathbf{U}_1)$ is a diagonal matrix, and off-diagonal elements of $\mathrm{Cov}(\mathbf{U}_2)$, *i.e.*, $\mathrm{Cov}(U_{ij}, U_{i'j'})$, are also zero when $i \neq i'$, $j \neq j'$. The rest of the terms in $\mathrm{Cov}(\mathbf{U})$ are of the form, *e.g*, $\mathrm{Cov}(U_{ij}, U_{i'j}) = \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_j\boldsymbol{\mu}_{i'}/n_i$, which

constitute R and, under the assumptions, are uniformly bounded in the limit, and the same holds for the elements of off-diagonal blocks, $\mathrm{Cov}(\mathbf{U}_1, \mathbf{U}_2)$. However, $\mathrm{R} = 0$ under H_0, and also $\mathrm{R}/[\mathrm{tr}(\boldsymbol{\Sigma}_0)]^2 \to 0$ asymptotically under H_1, so that $\sigma_1^2/\sigma_0^2 \to 1$ in the limit. Hence, $\mathrm{Cov}(\mathbf{U})/[\mathrm{tr}(\boldsymbol{\Sigma}_0)]^2$ can be considered as a diagonal matrix for the limit.

Further, $\mathrm{E}(\mathrm{T}-1)$ is uniformly bounded, and so is $2\|\boldsymbol{\Sigma}_0\|^2/[\mathrm{tr}(\boldsymbol{\Sigma}_0)]^2 \leq 2$, under the assumptions, where these bounds remain intact for any p, so that we can use a sequential limit. Writing $\mathbf{1}^{\mathrm{T}}(\mathbf{U} - \mathrm{E}(\mathbf{U})) = \mathrm{A}_{12} - \mathrm{E}(\mathrm{A}_{12})$, with corresponding elements $U_i - \mathrm{E}(U_i)$ and $U_{ij} - \mathrm{E}(U_{ij})$, and associated kernels, $\langle \overline{\mathbf{X}}_{ik}, \overline{\mathbf{X}}_{ir} \rangle - \|\boldsymbol{\mu}_i\|^2$, and $\langle \overline{\mathbf{X}}_{ik}, \overline{\mathbf{X}}_{jl} \rangle - \langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j \rangle$, it follows, from the asymptotic theory of U-statistics (Koroljuk and Borovskich, 1994), that, for any p,

$$n_i c_i^2 U_i \xrightarrow{\mathcal{D}} \sum_{s=1}^{p} \lambda_{is}(z_{is}^2 - 1) \quad \text{and} \quad \sqrt{n_i n_j}\, U_{n_i n_j} \xrightarrow{\mathcal{D}} \sum_{s=1}^{p} \lambda_{is}\lambda_{js} z_{is} z_{js},$$

as $n_i \to \infty$, where z_{is}, z_{js} are iid $N(0,1)$ variables, and independent of each other, and λ_{is} are the eigenvalues of $\boldsymbol{\Sigma}_i$. Now, taking p and the denominator into account, and applying Lemma 2 for $p \to \infty$, the required limit follows by a simple application of the Cramér-Wold device and Slutsky's theorem (van der Vaart, 1998), as was similarly done in Ahmad (2019b).

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 611- 637
http://www.ssca.org.in/journal

Distribution of the Hölder Mean of P-Values with Applications to Multiple Testing

Jiangtao Gou[1] **and Ajit C. Tamhane**[2]
[1]*Department of Mathematics and Statistics,*
Villanova University, Villanova, PA 19085, USA
[2]*Department of Industrial Engineering and Management Sciences,*
Northwestern University, Evanston, IL 60208, USA

Received: 29 April 2024; Revised: 10 October 2024; Accepted: 16 October 2024

Abstract

We study the null distribution of the Hölder mean with a scalar parameter $m \in (-\infty, +\infty)$ of i.i.d. P-values for testing $n \geq 2$ null hypotheses subject to the familywise error rate (FWER) control. We find the exact critical values for $n = 2, 3$ and the asymptotic critical values for $n > 3$ for selected values of m. We use them in a closed multiple testing procedure (MTP) which we illustrate by a numerical example. We compare the powers of the tests of the intersection hypothesis $H_0 = \cap_{i=1}^{n} H_i$ for $n = 2$ and 3 using the Hölder means with different values of m to find the best choice. Asymptotic critical values are not very accurate (are generally too conservative) and so power comparisons are not performed for larger n.

Key words: Arithmetic mean; Closed procedure; Distribution theory; Familywise error rate; Geometric mean; Harmonic mean; Hölder mean; Power comparison.

AMS Subject Classifications: 62E99

1. Introduction

In Gou and Tamhane (2024) we studied the null distribution of the harmonic mean of the P-values with application to multiple testing. We compared the resulting multiple testing procedure (MTP) with the commonly used P-value based MTPs of Holm (1979), Hochberg (1988) and Hommel (1988) and found it to be generally more powerful.

The arithmetic, geometric and harmonic means are special cases of Hölder mean, so it is natural to ask whether in the class of all the Hölder mean based MTPs, if there is some subclass that is more powerful under certain non-null configurations of interest. However, we must first derive the null distribution of the Hölder mean and obtain its critical values. This is the main focus of the present paper. In Section 6 we give a closed MTP (Marcus *et al.*, 1976) that uses the Hölder means for testing multiple hypotheses.

Corresponding Author: Ajit C. Tamhane
Email: atamhane@northwestern.edu

Consider testing $n \geq 2$ hypotheses, $H_1, \ldots, H_n$, subject to the strong familywise error rate (FWER) control requirement (Hochberg and Tamhane, 1987):

$$\text{FWER} = \Pr\{\text{Reject at least one true } H_i\} \leq \alpha \tag{1}$$

where $\alpha \in (0, 1)$ is prespecified. Let $P_1, \ldots, P_n$ denote the P-values associated with the hypotheses $H_1, \ldots, H_n$. The overall null hypothesis is denoted by $H_0 = \cap_{i=1}^n H_i$. We will assume that under H_0, the P_i's are independent and identically distributed (i.i.d.) uniform random variables over $[0, 1]$. This assumption is relaxed to allow for de[endent P-values in simulations reported in Section 8.

Consider a given real-valued parameter $m \in (-\infty, \infty)$ and weights

$$w_i > 0 \ (i = 1, \ldots, n) \ \text{ such that } \ \sum_{i=1}^{n} w_i = 1. \tag{2}$$

Then the weighted Hölder mean of the P-values $P_1, \ldots, P_n$ with parameter m is defined as

$$\bar{P}_n(m, w) = \left(\frac{1}{n} \sum_{i=1}^{n} w_i P_i^m \right)^{1/m}. \tag{3}$$

The unweighted Hölder mean corresponds to $w_1 = \cdots = w_n = 1/n$ and is denoted simply by $\bar{P}_n(m)$, dropping w in the notation. The arithmetic, geometric and harmonic means are special cases of the Hölder mean for $m = 1, 0$ and -1, respectively. The Hölder means for selected values of m have been previously considered by Vovk and Wang (2020) and by Tian *et al.* (2023). Here we study them in more detail with focus on their exact null distributions for $n = 2$ and their asymptomatic null distributions for $n > 2$.

The outline of the paper is as follows. Section 2 gives expressions for the c.d.f. of the unweighted Hölder mean for general m and $n = 2$. Section 3 gives expressions for the cumulative distribution function (c.d.f.) of the weighted Hölder mean for selected values of m and the expressions for their lower α critical values for $n = 2$. Section 4 gives the critical values for $n = 3$. Section 5 derives the asymptotic null distributions of the unweighted Hölder mean. Section 6 gives the closed MTP based on the Hölder means. Section 7 gives a numerical example to illustrate this MTP for harmonic, geometric and arithmetic means. Section 8 gives a numerical type I error and power comparisons for testing $H_0 = \cap_{i=1}^n$ for $n = 2$ and as well as for type I error for selected $n \geq 10$. Finally Section 9 gives concluding remarks. Derivations of all analytical results and proofs of theorems are presented in the Appendix.

2. Null distribution of unweighted Hölder mean for general m and $n = 2$

Before we state the main theorem of this section about the null distribution of $\bar{P}_2(m)$, we show in Figure 1 how the rejection boundaries in the (P_1, P_2) space change with m for selected values of $m = -\infty, -1, 0, 1, \infty$ for fixed $\alpha = 0.25$. (A large value of α is chosen so that the plotted rejection boundaries are distinguishable from each other.) The rejection boundaries also change with α but their relative behavior with respect to m remains the same.

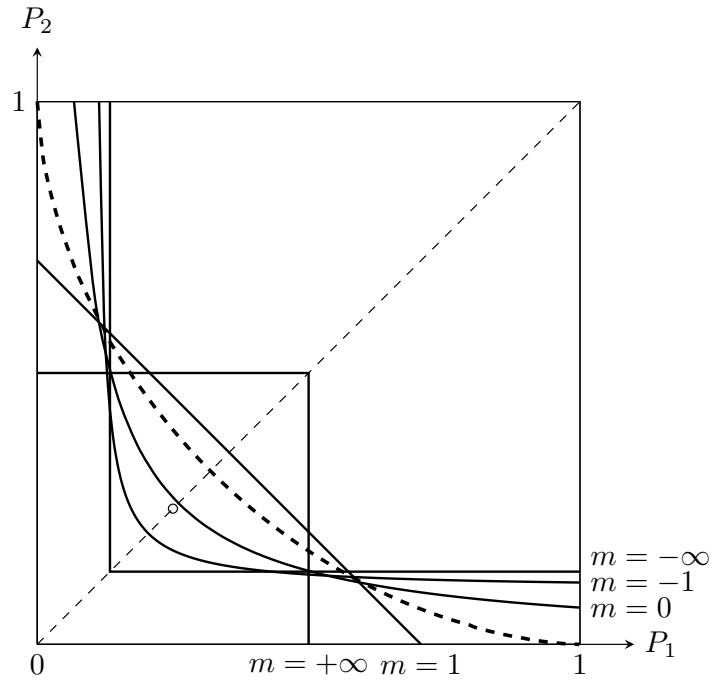


Figure 1: Rejection boundaries for selected values of m

Notice from this figure that the rejection boundaries for $m = 0, -1$ and $-\infty$ go from the top edge of the square to the right edge, while for $m = 1$ and ∞ they go from the bottom edge to the left edge. By the continuity in m and symmetry in P_1 and P_2, it follows that there exists an $m = m^* \in (-\infty, \infty)$ and an associated critical value $c = c^* \in (0, 1)$, for which the rejection boundary connects the top left corner $(P_1, P_2) = (0, 1)$ to the bottom right corner $(P_1, P_2) = (1, 0)$. This rejection boundary is shown by a dotted line in the figure and we refer to it as the *critical boundary*.

Given that the rejection boundary is defined by $P_1^m + P_2^m = 2c^m$ and the critical boundary passes through the points $(0, 1)$ and $(1, 0)$, it follows that for the critical boundary we have $2(c^*)^{m^*} = 1$ or $c^* = (1/2)^{1/m^*}$.

The following numerical example illustrates the calculation of m^* and c^* for $\alpha = 0.05$. First note that

$$\alpha = \Pr\{\bar{P}_2(m) \le c\} = \Pr\{P_1 \le (2c^m - P_2^m)^{1/m}\} = \int_0^1 (2c^m - x^m)^{1/m} dx.$$

Substitute $m^* = 1/3$ and $2(c^*)^{m^*} = 1$ in the above integral, which then becomes

$$\alpha = \int_0^1 (1 - x^{1/3})^3 dx.$$

Now put $1 - x^{1/3} = y$. Then $dx = 3(1 - y)^2 dy$. So we get

$$\alpha = 3 \int_0^1 y^3 (1 - y)^2 dy = \frac{3}{60} = 0.05.$$

Thus $m^* = 1/3$ and $c^* = (1/2)^3 = 0.125$ gives $\alpha = 0.05$. This pair of (m^*, c^*) values is shown in Table 1 along with other pairs of values for selected α values computed using MATLAB function `fsolve()`.

Table 1: m^* and c^* values for selected α for $n = 2$

α	m^*	c^*
0.010	0.2336	0.0515
0.025	0.2812	0.0850
0.050	0.3333	0.1250
0.100	0.4113	0.1854

In the following theorem we give expressions for the c.d.f. of $\bar{P}_2(m)$ for general m. First let $F_2(x; m) = \Pr\{\bar{P}_2(m) \leq x\}$ denote the c.d.f. of $\bar{P}_2(m)$. Also let $B_p(a, b)$ denote the incomplete beta function defined as

$$B_p(a, b) = \int_0^p x^{a-1}(1-x)^{b-1}dx,$$

where $p \leq 1$. When $p = 1$ we have the complete beta function denoted by $B_1(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$.

Theorem 1: The c.d.f. of $\bar{P}_2(m)$ is given by

$$F_2(x, m) = \begin{cases} (2x^m - 1)^{1/m} + \frac{4^{1/m}x^2}{m}\left[B_{1/2x^m}\left(\frac{1}{m}, \frac{1}{m} + 1\right)\right. \\ \quad \left. -B_{1-1/2x^m}\left(\frac{1}{m}, \frac{1}{m} + 1\right)\right], \quad \text{when } 0 \leq x \leq 1, \quad m \leq m^*, m \neq 0 \\ \frac{4^{1/m}x^2}{m}B_1\left(\frac{1}{m}, \frac{1}{m} + 1\right), \quad \text{when } 0 \leq x \leq 2^{-1/m}, \quad m > m^*. \end{cases} \qquad \square$$

The case $m = 0$ is covered in Part 5 of Theorem 2 and hence is not included here. We don't need to compute the c.d.f. for $x > 2^{-1/m}$ when $m > m^*$ because the corresponding α values are too large to be practically useful.

3. Exact null distribution of weighted Hölder mean for selected values of m and $n = 2$

In this section we obtain the c.d.f. of the weighted Hölder mean, denoted by $F_2(x; m, w)$ for selected m values. These results are given in the following theorem. The lower α critical values in each case can be found by solving the equation $F_2(x; m, w) = \alpha$ for x. Explicit expressions for the critical values are given where available. We denote these critical values by $c_2(m, \alpha)$.

Theorem 2: This theorem has nine parts corresponding to the nine selected m values, $m = -\infty, -2, -1, -0.5, 0, 0.5, 1, 2$ and $+\infty$.

Part 1 ($m = -\infty$):
As $m \to -\infty$, $\bar{P}_n(m, w) \to P_{\min}$ for any choice of weights. Assuming $P_{\min}$ is unique, its c.d.f. and lower α critical value are given by

$$F_n(x; -\infty) = 1 - (1 - x)^n \quad \text{and} \quad c_n(-\infty, \alpha) = 1 - (1 - \alpha)^{1/n}.$$

Part 2 ($m = -2$):
For $m = -2$ the c.d.f. of

$$\bar{P}_2(-2, w) = \left(\frac{w_1}{P_1^2} + \frac{w_2}{P_2^2}\right)^{-1/2}$$

is given by

$$F_2(x; -2, w) = x\left[\sqrt{w_1(1 - w_2x^2)} + \sqrt{w_2(1 - w_1x^2)}\right]. \tag{4}$$

For equal weights this simplifies to $F_2(x; -2) = x\sqrt{2 - x^2}$. The lower α critical value for equal weights is $c_2(-2, \alpha) = \sqrt{1 - \sqrt{1 - \alpha}}$.

Part 3 ($m = -1$):
For $m = -1$ the c.d.f. of the weighted harmonic mean,

$$\bar{P}_2(-1, w) = \left(\frac{w_1}{P_1} + \frac{w_2}{P_2}\right)^{-1},$$

is given by

$$F_2(x; -1, w) = x + w_1w_2x^2 \ln\left[1 + \frac{1 - x}{w_1w_2x^2}\right].$$

For equal weights this simplifies to

$$F_2(x; -1) = x + \frac{x^2}{4} \ln\left[1 + \frac{4(1 - x)}{x^2}\right]. \tag{5}$$

There is no closed form solution to the equation $F_2(x; -1) = \alpha$.

Part 4 ($m = -0.5$):
For $m = -0.5$, for equal weights the c.d.f. of

$$\bar{P}_2(-0.5) = \left[\frac{1}{2}\left(\frac{1}{\sqrt{P_1}} + \frac{1}{\sqrt{P_2}}\right)\right]^{-2}$$

is given by

$$F_2(x, -0.5) = \frac{x}{(2 - \sqrt{x})^2} + \frac{x}{8}\left(6\sqrt{x} - x - \frac{x(4 + \sqrt{x})}{2 - \sqrt{x}} + 3x \ln\left(\frac{(2 - \sqrt{x})^2}{x}\right) + 2 - \frac{2x}{(2 - \sqrt{x})^2}\right). \tag{6}$$

There is no closed form solution to the equation $F_2(x; -0.5) = \alpha$.

Part 5 ($m = 0$):
For $m = 0$ the c.d.f. of the weighted geometric mean

$$\bar{P}_2(0, w) = P_1^{w_1}P_2^{w_2}$$

is given by

$$F_2(x; 0, w) = \left(1 - \frac{w_2}{w_1}\right)x^{1/w_1} + \left(1 - \frac{w_1}{w_2}\right)x^{1/w_2} \quad w_1 \neq w_2.$$

For equal weights the c.d.f. is given by

$$F_2(x;0) = \Pr\left\{\chi_4^2 > -4\ln x\right\} = x^2(1-2\ln x) \tag{7}$$

and its lower α critical value equals

$$c_2(0,\alpha) = \exp\left(-\frac{1}{4}\chi_{4,\alpha}^2\right), \tag{8}$$

where $\chi_{4,\alpha}^2$ is the upper α critical point of the χ_4^2 distribution.

Part 6 ($m = 0.5$):
For $m = -0.5$, the c.d.f. of

$$\bar{P}_2(0.5,w) = \left(w_1\sqrt{P_1} + w_2\sqrt{P_2}\right)^2.$$

for equal weights is given by

$$F_2(x,-0.5) = \frac{8x^2}{3}, \tag{9}$$

The lower α critical value equals

$$c_2(0.5,\alpha) = \sqrt{\frac{3\alpha}{8}} \quad \text{for} \quad \alpha \le \frac{1}{6}.$$

Part 7 ($m = 1$):
For $m = 1$ the c.d.f. of the weighted arithmetic mean, assuming $w_1 \le w_2$, is given by

$$F_2(x;1,w) = \begin{cases} \frac{x^2}{2w_1w_2}, & 0 \le x \le w_1, \\ \frac{2x-w_1}{2w_2}, & w_1 \le x \le w_2, \\ 1 - \frac{(1-x)^2}{2w_1w_2}, & w_2 < x \le 1. \end{cases}$$

For equal weights this simplifies to

$$F_2(x,1) = \begin{cases} 2x^2, & 0 \le x \le 1/2, \\ 1-2(1-x)^2, & 1/2 < x \le 1. \end{cases} \tag{10}$$

The lower α critical value for equal weights is given by $c_2(1,\alpha) = \sqrt{\alpha/2}$ if $\alpha \le 1/2$.

Part 8 ($m = 2$):
For $m = 2$, assuming that $w_1 \le w_2$, the c.d.f. of

$$\bar{P}_2(2,w) = (w_1P_1^2 + w_2P_2^2)^{1/2}$$

is given by

$$F_2(x;2,w) = \begin{cases} \frac{\pi x^2}{4\sqrt{w_1w_2}}, & x \le \sqrt{w_1}, \\ \frac{\sqrt{w_1(x^2-w_1)} + x^2\tan^{-1}\left(\sqrt{\frac{w_1}{x^2-w_1}}\right)}{2\sqrt{w_1w_2}}, & \sqrt{w_1} < x \le \sqrt{w_2}, \\ \frac{1}{2}\left(\sqrt{\frac{x^2-w_2}{w_1}} + \sqrt{\frac{x^2-w_1}{w_2}}\right) & \\ \quad + \frac{x^2}{2\sqrt{w_1w_2}}\left(\tan^{-1}\left(\sqrt{\frac{w_1}{x^2-w_2}}\right) - \tan^{-1}\left(\sqrt{\frac{x^2-w_1}{w_1}}\right)\right), & x > \sqrt{w_2}. \end{cases}$$

For equal weights this simplifies to

$$F_2(x;2) = \begin{cases} \frac{\pi x^2}{2} & x \le \sqrt{1/2}, \\ \sqrt{2x^2-1} + x^2 \tan^{-1}\left(\frac{1-x^2}{\sqrt{2x^2-1}}\right) & x > \sqrt{1/2}. \end{cases} \tag{11}$$

For $\alpha > \pi/4$, there is no closed form solution to the equation $F_2(x;2) = \alpha$. For $\alpha \le \pi/4$, we have

$$c_2(2,\alpha) = \sqrt{\frac{2\alpha}{\pi}}.$$

Part 9 ($m = \infty$):
As $m \to +\infty$, $\bar{P}_n(m,w) \to P_{\max}$ for any choice of weights. Assuming $P_{\max}$ is unique, its c.d.f. and lower α critical value are given by

$$F_n(x;\infty) = x^n \quad \text{and} \quad c_n(\infty,\alpha) = \alpha^{1/n}. \qquad \square$$

Table 2 summarizes the formulae for finding the critical values $c_2(m,\alpha)$ for the nine selected values of m and $\alpha = 0.05$. From this table we see that the critical value increases with m. This is true in general for any $n \ge 2$ as stated in Theorem 3.

Table 2: Critical values $c_2(m,\alpha)$ for selected $m, n = 2$ and $\alpha = 0.05$

m	Formula for $c_2(m,\alpha)$	$c_2(m,\alpha)$	m	Formula for $c_2(m,\alpha)$	$c_2(m,\alpha)$
$-\infty$	$c_2(-\infty,\alpha) = 1-\sqrt{1-\alpha}$	0.0253	0.5	$c_2(0.5,\alpha) = \sqrt{\frac{3\alpha}{8}}$ if $\alpha \le 1/6$	0.1369
-2	$c_2(-2,\alpha) = \sqrt{1-\sqrt{1-\alpha}}$	0.0354	1	$c_2(1,\alpha) = \sqrt{\frac{\alpha}{2}}$ if $\alpha \le 1/2$	0.1581
-1	Solve $x + \frac{x^2}{2}\ln\left[1+\frac{4(1-x)}{x^2}\right] = \alpha$	0.0460	2	$c_2(2,\alpha) = \sqrt{\frac{2\alpha}{\pi}}$ if $\alpha \le \pi/4$	0.1784
-0.5	Solve Eqn. (6) $= \alpha$	0.0616	∞	$c_2(\infty,\alpha) = \sqrt{\alpha}$	0.2236
0	Solve $x^2(1-2\ln x) = \alpha$	0.0933			

Theorem 3: For any fixed $\alpha \in (0,1)$ and $n \ge 2$ the critical value $c_n(m,\alpha)$ is an increasing function of m. $\square$

4. Exact critical values for $n = 3$

The rejection region for $n = 3$ is defined by

$$\left(\frac{P_1^m + P_2^m + P_3^m}{3}\right)^{1/m} \le c,$$

where $c \in (0,1)$ is a critical constant depending on α and m. Just as the critical bound for $n = 2$ passes through the points $(1,0)$ and $(0,1)$ in the (P_1,P_2) space, the critical surface for $n = 3$ passes through the points $(1,0,0),(0,1,0)$ and $(0,0,1)$. The corresponding critical (m^*,c^*) thus satisfy $c^* = (1/3)^{1/m^*}$. The type I error probability α for testing $H_0 = H_1 \cap H_2 \cap H_3$ is given by

$$\int_0^{3^{1/m}c} \int_0^{(3c^m - p_3^m)^{1/m}} \int_0^{(3c^m - p_2^m - p_3^m)^{1/m}} dp_1 \; dp_2 \, dp_3 = \alpha.$$

Table 3: Critical values $c_3(m, .05)$ for selected m values and $\alpha = 0.05$

m	$c_3(m, 0.05)$
$-\infty$	0.0170
-2	0.0289
-1	0.0443
-0.5	0.0691
0	0.1226
0.5	0.1839
0.6848	0.2010
1	0.2231
2	0.2639
$+\infty$	0.3684

Setting $3c^m = 1$ for the critical surface, the above equation reduces to

$$\int_0^1 \int_0^{(1-p_3^m)^{1/m}} (1 - p_2^m - p_3^m)\, dp_2\, dp_3 = \alpha.$$

For $\alpha = 0.05$ the above equation can be solved using the MATLAB function `fsolve()` for m resulting in $m^* = 0.6848$ and $c^* = (1/3)^{1/m^*} = 0.2010$. Analogous to the $n = 2$ case, different rejection regions and hence different integral expressions must be evaluated for $m > m^*$ and $m < m^*$. Before we do that for $m = 0$ (geometric mean) we have $-2n \ln(\bar{P}_3(0)) \sim \chi^2_{2n}$ and hence $c_n(\alpha) = \exp(-(1/2n)\chi^2_{2n,\alpha})$. Therefore

$$c_3(0, 0.05) = \exp(-(1/6)\chi^2_{6,.05}) = 0.1226.$$

Omitting the analytical details we give in Table 3 the critical values $c_3(m, 0.05)$ for selected m values. These are used in the type I error rate and power simulations in Section 8.

5. Asymptotic null distribution of the unweighted Hölder mean

The exact null distribution of the Hölder mean is difficult to derive in general for $n > 2$. Hence we resort to asymptotics. The P_i^m are i.i.d. with a beta distribution with parameters $a = 1/m$ and $b = 1$. The mean and variance of this distribution are

$$E(P_i^m) = \frac{1}{m+1} \quad \text{and} \quad \text{Var}(P_i^m) = \frac{m^2}{(m+1)(2m+1)}. \tag{12}$$

Note that $\text{Var}(P_i^m)$ exists (is finite) for $m > -1/2$ and does not exist (is either infinite or negative) for $m \leq -1/2$. So the standard Lindeberg-Lévy central limit theorem (CLT) applies in the former case, but not in the latter in which case $\bar{P}_n(m)$ is not asymptotically normal. Hence we treat the two cases separately.

5.1. The case $m > -1/2$

The case $m = 0$ is covered in Part 5 of Theorem 2 since it does not require asymptotics.

By the CLT,

$$\frac{\left(\frac{1}{n}\sum_{=1}^{n} P_i^m - \frac{1}{m+1}\right)}{\sqrt{\frac{m^2}{n(m+1)(2m+1)}}} \longrightarrow N(0,1)$$

as $n \to \infty$. The lower α critical value for $\frac{1}{n}\sum_{i=1}^{n} P_i^m$ is then given by

$$\frac{1}{m+1} - z_\alpha \sqrt{\frac{m^2}{n(m+1)(2m+1)}}, \tag{13}$$

where z_α is the $100(1-\alpha)$ percentile of the $N(0,1)$ distribution. However, we require the asymptotic critical values of $\bar{P}_n(m) = \left(\frac{1}{n}\sum_{i=1}^{n} P_i^m\right)^{1/m}$. One method (Method 1) is to take the $(1/m)$th power of (13). Another method (Method 2) is to use the delta method to find the mean and variance of $\bar{P}_n(m)$ and apply the CLT approximation to it.

The delta method gives

$$E(\bar{P}_n(m)) = E\left(\frac{1}{n}\sum_{i=1}^{n} P_i^m\right)^{1/m} \approx \left(\frac{1}{m+1}\right)^{1/m} \tag{14}$$

and

$$\mathrm{Var}(\bar{P}_n(m)) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} P_i^m\right)^{1/m} \approx \frac{(m+1)^{1-2/m}}{n(2m+1)}. \tag{15}$$

The derivation of these two formulae is given in the Appendix. The lower α critical value for $\bar{P}_n(m)$ using Method 2 is given by

$$\left(\frac{1}{m+1}\right)^{1/m} - z_\alpha \sqrt{\frac{(m+1)^{1-2/m}}{n(2m+1)}}. \tag{16}$$

The critical values obtained by both these methods are given in Table 4 for selected values of m and n. Which method gives more accurate results depends on m and n.

Table 4: Asymptotic lower $\alpha = 0.05$ critical values of $\bar{P}_n(m)$ for $m > -1/2$

m	Method	n				
		10	20	50	100	1000
−0.25	Method 1	0.1752	0.2065	9.2403	0.2600	0.2970
	Method 2	0.1148	0.1739	0.2263	0.2527	0.2963
0	Exact	0.2079	0.2481	0.2884	0.3104	0.3490
0.5	Method 1	0.2668	0.3142	0.3594	0.3834	0.4247
	Method 2	0.2442	0.3029	0.3549	0.3811	0.4244
1.0	Method 1	0.2877	0.3499	0.4050	0.4329	0.4788
	Method 2	0.2877	0.3499	0.4050	0.4329	0.4788
2.0	Method 1	0.2544	0.3787	0.4618	0.4984	0.5536
	Method 2	0.3447	0.4129	0.4733	0.5038	0.5541

The accuracy of these critical values is evaluated by simulating their associated type I errors in Section 8.3.

5.2. The case $-1 < m \leq -1/2$

Here $X_i = P_i^m$ follows a beta distribution with $a = -1/m$ and $b = 1$. Its p.d.f. and c.d.f. are given by

$$f_{X_i}(x) = \left(-\frac{1}{m}\right) x^{\frac{1}{m}-1} \quad \text{and} \quad F_{X_i}(x) = 1 - x^{\frac{1}{m}} \quad \text{for } x \geq 1. \tag{17}$$

The variance of this distribution is ∞, so the standard Lindeberg-Lévy CLT does not apply and $(1/n)\sum_{i=1}^n P_i^m$ is not asymptotically normal. So we apply the generalized CLT (Gnedenko and Kolmogorov, 1954; Ibragimov and Linnik, 1971; Petrov, 1975) stated below.

Theorem 4: Let $X_1, \ldots, X_n$ be i.i.d. random variables with the distribution function $F_X(x)$ satisfying the conditions

$$F_X(x) \sim k_1|x|^{-a^*} \quad \text{as} \quad x \longrightarrow -\infty$$

and

$$1 - F_X(x) \sim k_2|x|^{-a^*} \quad \text{as} \quad x \longrightarrow +\infty$$

with $a^* > 0$. Then there exist sequences $\{\mu_n\}$ and $\{\sigma_n\}$ where $\sigma_n > 0$ such that the distribution of the centered and normalized sum

$$Z_n = \frac{\sum_{i=1}^n X_i - \mu_n}{\sigma_n}$$

weakly converges to a stable distribution (denoted by $S(a, b)$) with parameters $a = \min\{a^*, 2\}$ and $b = (k_2 - k_1)/(k_2 + k_1)$ as $n \to +\infty$. The centering and normalizing values μ_n and σ_n depend on the parameters a and b. □

Let $c^*(\alpha)$ denote the upper α critical value of the stable distribution $S(1, 1)$. Then the critical value of the $S(a, b)$ distribution is

$$c(\alpha) = a + bc^*(\alpha).$$

A discussion of the stable distribution and methods of approximating the critical value $c^*(\alpha)$ is given in the Appendix.

The asymptotic critical value of $\sum_{i=1}^n X_i = \sum_{i=1}^n P_i^m$ is $\mu_n + c(\alpha)\sigma_n$. Then the critical value of $\bar{P}_n(m) = [(1/n)\sum_{i=1}^n P_i^m]^{1/m}$ can be approximated by making the corresponding transformation as

$$c_n(m, \alpha) = \left[(1/n)(\mu_n + c(\alpha)\sigma_n)\right]^{1/m}. \tag{18}$$

Table 6 gives the critical values computed using this method, which we refer to as Method 0. The results regarding the values of μ_n and σ_n used in the three cases discussed below are due to Mijnheer (1975), Samorodnitsky and Taqqu (1994) and Uchaikin and Zolotarev (1999). Some selected values of $c^*(\alpha)$ are given in Table 5.

We now apply Theorem 4 to different cases for values of $m \leq -1/2$.

Table 5: Selected values of $c^*(\alpha)$

α	0.01	0.025	0.05	0.10	0.20
$c^*(\alpha)$	65.9760	27.1899	14.0048	7.1287	3.3843

Case 1 ($m = -1/2$)

From (17) we obtain the p.d.f. and c.d.f. of $X_i = P_i^{-1/2}$ as

$$f_{X_i}(x) = 2x^{-3} \quad \text{and} \quad F_{X_i}(x) = 1 - x^{-2} \quad \text{for } x \geq 1.$$

Therefore $k_1 = 0, k_2 = 1$ and $a^* = 2$. So $a = 2$ and $b = 1$. For these values of a and b it has been shown that (see the previously mentioned references)

$$\mu_n = nE\left(X_i\right) = nE\left(P_i^{-1/2}\right) = 2n \quad \text{and} \quad \sigma_n = \sqrt{n \ln n}.$$

Furthermore, the stable law $S(2,1)$ is simply the $N(0,(\sqrt{2})^2)$ distribution, so

$$\frac{\sum_{i=1}^n P_i^{-1/2} - 2n}{\sqrt{2n \ln n}} \quad \longrightarrow N(0,1^2).$$

Thus the lower α critical value of $\sum_{i=1}^n P_i^{-1/2}$ is $2n - z_\alpha \sqrt{2n \ln n}$ from which the lower α critical value of $\bar{P}_n(-1/2)$ can be approximated as

$$c_n\left(-1/2, \alpha\right) = \left[\frac{1}{n}\left\{2n - z_\alpha \sqrt{2n \ln n}\right\}\right]^{-2}.$$

Case 2 ($-1 < m < -1/2$)

From (17) we get $k_1 = 0, k_2 = 1$ and $a^* = 1/m$. Thus we have $a = -1/m$ $(1 < a < 2)$ and $b = 1$. For these values of a and b it has been shown that (see the previously mentioned references)

$$\mu_n = nE(X_i) = nE\left(P_i^m\right) = \frac{na}{a-1} \quad \text{and} \quad \sigma_n = \left(\frac{n\pi}{2\Gamma(a)\sin(a\pi/2)}\right)^{1/a}.$$

Therefore

$$\frac{\sum_{i=1}^n P_i^m - \frac{na}{a-1}}{\left(\frac{n\pi}{2\Gamma(a)\sin(a\pi/2)}\right)^{1/a}} \quad \longrightarrow S(a,b).$$

The asymptotic lower α critical value of $S(a,b)$ is

$$c(\alpha) = -\frac{1}{m} + c^*(\alpha).$$

Hence the approximate lower α critical value of $\bar{P}_n(m)$ is

$$c_n(m,\alpha) = \left[\frac{1}{n}\left\{\frac{na}{a-1} - c(\alpha)\left(\frac{n\pi}{2\Gamma(a)\sin(a\pi/2)}\right)^{1/a}\right\}\right]^{1/m}.$$

Case 3 ($m \leq -1$) The case $m = -1$ corresponds to the harmonic mean and is discussed in detail in Gou and Tamhane (2024). So we consider only the case $m < -1$. From (17) we get $k_1 = 0, k_2 = 1$ and $a^* = 1/m$. Thus we have $a = -\frac{1}{m}$ and $b = 1$ where $0 < a < 1$. For these values of a and b it has been shown that (see the previously mentioned references)

$$\mu_n = 0 \quad \text{and} \quad \sigma_n = \left(\frac{n\pi}{2\Gamma(a)\sin(a\pi/2)}\right)^{1/a}.$$

Therefore

$$\frac{\sum_{i=1}^{n} P_i^m}{\left(\frac{n\pi}{2\Gamma(a)\sin(a\pi/2)}\right)^{1/a}} \longrightarrow S(a,b).$$

Since a and b are the same as in Case 2, $c(\alpha)$ is also the same. Hence the approximate lower α critical value of $\bar{P}_n(m)$ is

$$\left[\frac{1}{n}\left\{-c(\alpha)\left(\frac{n\pi}{2\Gamma(a)\sin(a\pi/2)}\right)^{1/a}\right\}\right]^{1/m}.$$

Table 6: Asymptomatic lower $\alpha = 0.05$ critical values of $\bar{P}_n(m)$ for $m \leq -1/2$ using Method 0

m	n				
	10	20	50	100	1000
-0.5	0.1030	0.1189	0.1423	0.1601	0.2079
-1.0	0.0412	0.0400	0.0386	0.0376	0.0346
-2.0	0.0159	0.0112	0.0071	0.0050	0.0016
-3.0	0.0110	0.0069	0.0037	0.0024	0.0005

The accuracy of these critical values is evaluated by simulating their associated type I errors in Section 8.3.

6. A closed multiple testing procedure (MTP)

Our testing strategy will be to use the closure method (Marcus *et al.*, 1976) based on $\bar{P}_n(m)$ with a preselected m as the test statistic. The closure method begins by testing the overall null hypothesis $H_0 = \cap_{i=1}^{n} H_i$ at level α. If H_0 is rejected then it tests all subset null hypotheses of size $n-1$ each at level α. If any subset null hypothesis is not rejected then all its subsets are accepted by implication. This ensures coherence (Gabriel, 1969). On the other hand, if any subset null hypothesis of size $n' \leq n$ is rejected then all its subsets of size $n'-1$ that are not already accepted by implication are tested each at level α.

This procedure does not have a simple stepwise shortcut like the Holm and the Hochberg procedures have. However, these computations can be substantially reduced as follows. When testing all subsets of size $n' \leq n$, first test the subset with the largest P-values. If it is significant then all other subsets of size n' will also be significant and need not be tested. Otherwise test the subset with the smallest P-values. If it is nonsignificant then

all other subsets of size n' will also be nonsignificant and need not be tested. This method is illustrated in the numerical example in Section 7. A simple `R` code can be used to compute the Hölder means.

Dobriban (2020) has given an alternative shortcut which he called fast closed testing (FACT) algorithm. It is particularly efficient when n is large. He showed that when the hypotheses are exchangeable, we don't need to test all $2^n - 1$ intersection hypotheses, but only $n(n+1)/2$ of them. For example, if $n = 5$ then instead of testing all $2^5 - 1 = 31$ subsets, we only need to test $5(5+1)/2 = 15$ of them, a saving of 50%. As n grows larger, obviously saving increases. Here we don't use this algorithm as it would require much explanation.

7. Numerical example

Consider a dose response study in which $n = 5$ doses are tested for efficacy, labeled from the highest to the lowest as 1 through 5. Suppose that the P-values for the comparisons with placebo (zero dose) are as follows:

$$P_1 = 0.01,\ P_2 = 0.02,\ P_3 = 0.03,\ P_4 = 0.04,\ P_5 = 0.30.$$

Denote the corresponding hypotheses by $H_1, \ldots, H_5$. Because of space constraints we will only briefly illustrate three MTPs: harmonic mean MTP (denoted by HMP), geometric mean MTP (denoted by GMP) and arithmetic mean MTP (denoted by AMP). We will use $\alpha = 0.05$.

Harmonic Mean Procedure (HMP): The critical values for HMP are

$$c_1 = 0.0500,\ c_2 = 0.0460,\ c_3 = 0.0443,\ c_4 = 0.0433,\ c_5 = 0.0425.$$

Step 1: Test the whole set $\{1, 2, 3, 4, 5\}$. The harmonic mean for this set is $0.0236 < c_5 = 0.0425$, so we reject it.

Step 2: Test the subset $\{2, 3, 4, 5\}$ of size 4 with the largest P-values. The harmonic mean for this subset is $0.0358 < c_4 = 0.0433$, so we reject it.

Step 3: Test the subset $\{3, 4, 5\}$ of size 3 with the largest P-values. The harmonic mean for this subset is $0.0486 > c_3 = 0.0443$. Therefore we accept intersection hypotheses associated with all subsets of $\{3, 4, 5\}$. The next largest harmonic mean is associated with the subset $\{2, 4, 5\}$ and is $0.0383 < c_3 = 0.0443$, which is thus rejected and hence all other subsets of size 3 are rejected.

Step 4: Test only those subsets of size 2 that include 1 or 2 or both. The subset $\{2, 5\}$ has the largest harmonic mean $0.0375 < c_2 = 0.0460$ and hence the subsets $\{1, 5\}$ is also rejected.

Step 5: Test only $\{1\}$ and $\{2\}$. Since P_1 and P_2 are $< c_1 = 0.05$, both H_1 and H_2 are rejected.

Thus HMP rejects two hypotheses, H_1 and H_2.

Having explained how HMP operates, we will present the application of GMP and AMP rather briefly, since they operate similarly.

Geometric Mean Procedure (GMP): The critical values for GMP are

$$c_1 = 0.0500,\ c_2 = 0.0933,\ c_3 = 0.1226,\ c_4 = 0.1439,\ c_5 = 0.1603.$$

At the first step we get $\bar{P}_5(0, \{1,2,3,4,5\}) = 0.0373 < c_5 = 0.1603$, so reject $\{1,2,3,4,5\}$. Next $\bar{P}_4(0, \{2,3,4,5\}) = 0.0518 < c_4 = 0.1439$, so reject $\{2,3,4,5\}$. Next $\bar{P}_3(0, \{3,4,5\}) = 0.0711 < c_3 = 0.1226$, so reject $\{3,4,5\}$. Next $\bar{P}_2(0, \{4,5\}) = 0.1095 > c_2 = 0.0933$ and $\bar{P}_2(0, \{3,5\}) = 0.0949 > c_2 = 0.0933$, so these subsets are accepted while the subset $\{1,2\}$ is rejected since $\bar{P}_2(0, \{1,2\}) = 0.0141 < c_2 = 0.0933$. Finally since P_1 and P_2 are $< c_1 = 0.05$, both H_1 and H_2 are rejected.

Arithmetic Mean Procedure (AMP): The critical values for AMP are

$$c_1 = 0.0500,\ c_2 = 0.1581,\ c_3 = 0.2231,\ c_4 = 0.2617,\ c_5 = 0.2869.$$

At the first step we get $\bar{P}_5(1, \{1,2,3,4,5\}) = 0.0800 < c_5 = 0.2869$, so reject $\{1,2,3,4,5\}$. Next $\bar{P}_4(1, \{2,3,4,5\}) = 0.0975 < c_4 = 0.2617$, so reject $\{2,3,4,5\}$. Next $\bar{P}_3(1, \{3,4,5\}) = 0.1233 < c_3 = 0.2231$, so reject $\{3,4,5\}$. Next $\bar{P}_2(1, \{2,5\}) = 0.1600, \bar{P}_2(1, \{3,5\}) = 0.1605, \bar{P}_2(1, \{4,5\}) = 0.1700$ are all $> c_2 = 0.1581$ and so are not rejected while all other pairs of hypotheses are rejected including $\{1,5\}$ for which $\bar{P}_2(1, \{1,5\}) = 0.1550$. So only H_1 remains to be tested and since $P_1 = 0.01 < c_1 = 0.05$, it is rejected. Thus AMP only rejects H_1.

8. Type I error and power simulations

8.1. Power simulations for $n = 2$

To save space, we report only the power of the test of $H_0 = H_1 \cap H_2$ for $n = 2$. Note that if the closed test procedure is consonant (Gabriel, 1969)), *i.e.*, if it rejects H_0 then it also rejects at least one of H_1 or H_2. Therefore the power of the test of H_0 is also the power of the corresponding closed MTP. It is easy to show that the closed MTP given above is not consonant for $n = 2$ if $c_2(m, \alpha) > \alpha$. In that case it is possible to have $P_1, P_2 > \alpha$ but $\bar{P}_2(m) < c_2(m, \alpha)$. So H_0 is rejected but neither H_1 nor H_2. For example, consider $m = 1$ (arithmetic mean). Let $P_1 = P_2 = 0.15$. Then $\bar{P}_2(1) = 0.15 < c_2(1, 0.05) = 0.1581$, but $P_1 = P_2 = 0.15 > c_1(1, 0.05) = 0.05$. From Table 2 we see that MTPs are consonant if $m \leq -1$ for $\alpha = 0.05$.

The power comparison setup is as follows. Let $X_1 \sim N(\mu_1, 1)$ and $X_2 \sim N(\mu_2, 1)$ with $\text{Corr}(X_1, X_2) = \rho \geq 0$. Further let $P_1 = 1 - \Phi(X_1)$ and $P_2 = 1 - \Phi(X_2)$. Under the alternative hypothesis ($\mu_1 \neq 0$ or $\mu_2 \neq 0$) the power can be expressed as a bivariate normal integral for all m. So it can be evaluated using numerical integration and does not need to be simulated. The integral expressions for power are omitted for brevity. The power is evaluated for $m = -\infty, -2, -1, -1/2, 0, 1/2, 1, 2, +\infty$ and for six configurations of (μ_1, μ_2) either $\mu_1 = 0$ and $\mu_2 = 1, 2, 3$ or $\mu_1 = \mu_2 = 1, 2, 3$. The power results for $\rho = 0$ are given in Table 7.

We also conducted power comparisons for $\rho = -0.5$ and $\rho = +0.5$, but we don't show them in Table 7. Furthermore, we also evaluated the Pr(Type I Error) under the overall null

hypothesis $\mu_1 = \mu_2 = 0$ for $\rho = 0, -0.5$ and $+0.5$. This probability is 0.05 under $\rho = 0$ by design and is confirmed by simulation and hence is not shown in Table 7. We see that for $\rho = -0.5$ the Pr(Type I Error) is slightly liberal for $m = -2$ and $-\infty$ while it is conservative for other values pf m. On the other hand, for $\rho = +0.5$ the Pr(Type I Error) is slightly conservative for $m = -2$ and $-\infty$ while it is quite liberal for other values of m.

Table 7: Power for rejecting $H_0 = H_1 \cap H_2$ for selected values of m, (μ_1, μ_2) and $\alpha = 0.05$

m	P(Type I Error)		Power					
			(μ_1, μ_2)					
	$(0,0)$	$(0,0)$	$(0,1)$	$(0,2)$	$(0,3)$	$(1,1)$	$(2,2)$	$(3,3)$
	$\rho = -0.5$	$\rho = +0.5$	$\rho = 0$					
$-\infty$	0.0506	0.0459	0.1909	0.5303	0.8559	0.3110	0.7678	0.9781
-2	0.0508	0.0478	0.1915	0.5311	0.8562	0.3159	0.7783	0.9809
-1	0.0478	0.0507	0.1928	0.5320	0.8561	0.3283	0.7982	0.9849
-0.5	0.0424	0.0572	0.1945	0.5309	0.8538	0.3487	0.8245	0.9890
0	0.0260	0.0736	0.1925	0.5086	0.8307	0.3886	0.8646	0.9937
0.5	0.0094	0.0957	0.1601	0.3182	0.4502	0.4024	0.8729	0.9939
1	0.0102	0.1005	0.1464	0.2481	0.3000	0.3838	0.8425	0.9867
2	0.0111	0.1021	0.1389	0.2175	0.2471	0.3679	0.8167	0.9800
∞	0.0124	0.1024	0.1996	0.2208	0.8559	0.3538	0.7966	0.9751

Figure 2 shows the plots of power with left panel showing the plots when H_1 is true and H_2 is false ($\mu_1 = 0, \mu_2 = 1$, 2 or 3) and right panel showing the plots when both H_1 and H_2 are equally false ($\mu_1 = \mu_2 = 1$, 2 or 3).

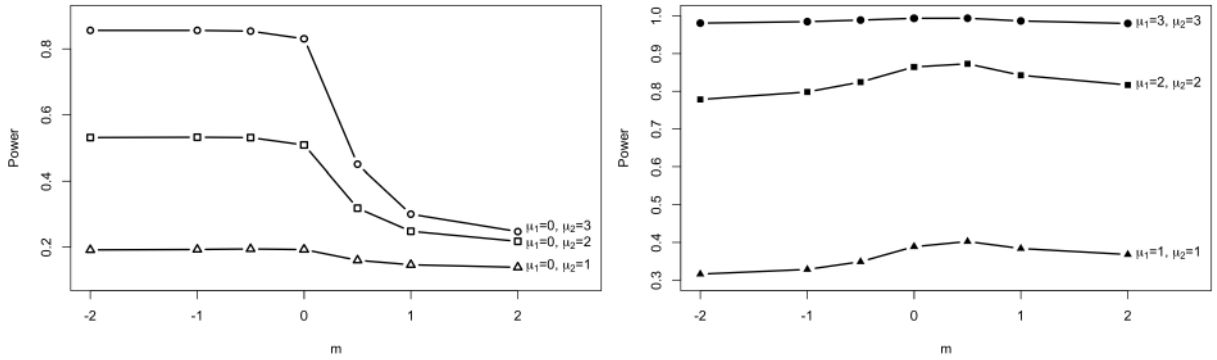


Figure 2: Plots of power for rejecting $H_0 = H_1 \cap H_2$ using different m (left panel: $\mu_1 = 0, \mu_2 = 1, 2, 3$, right panel: $\mu_1 = \mu_2 = 1, 2, 3$)

The Pr(Type I Error) is fairly well controlled when $\rho = -0.5$, but not when $\rho = +0.5$. The power results show that the maximum (or nearly maximum) power is achieved close to $m = -1$ (harmonic mean) when one hypothesis is true and the other is false. When both hypotheses are equally false, $m = 0.5$ yields the maximum power. The plots are fairly flat in the vicinity of the maximum, so any value of m close to the true optimum would work nearly equally well.

8.2. Type I error and power simulations for $n = 3$

Using the critical values obtained in Section 4 we evaluated the type I error of rejecting $H_0 = H_1 \cap H_2 \cap H_3$ (which is also the FWER of any consonant closed MTP) under independence and positive and negative dependence. The P-values are generated in the same manner as for $n = 2$ by simulating equicorrelated trivariate normal random variables with zero means and common correlation $\rho = 0$ for independence, $\rho = 0.5$ for positive dependence and $\rho = -0.25$ for negative dependence and transforming them to P-values. The number of replications were 10^6. The simulation results for type I error are presented in Table 8. Notice that the type I error rate is controlled accurately under independence for all m and conservatively for $\rho = -0.25$ when $m \geq -1.0$; however, it is not controlled for $\rho = 0.5$ when $m \geq -1.0$.

Table 8: Simulated type I error for rejecting $H_0 = H_1 \cap H_2 \cap H_3$ under independence ($\rho = 0$), positive dependence ($\rho = 0.5$) and negative dependence ($\rho = -0.25$)

m	$\rho = 0$	$\rho = 0.5$	$\rho = -0.25$
$-\infty$	0.0501	0.0435	0.0506
-2.0	0.0501	0.0454	0.0500
-1.0	0.0500	0.0512	0.0477
-0.5	0.0499	0.0636	0.0416
0.0	0.0495	0.0941	0.0231
0.5	0.0497	0.1270	0.0102
1.0	0.0496	0.1401	0.0106
2.0	0.0497	0.1447	0.0126
$+\infty$	0.0496	0.1438	0.0152

Next we consider power for rejecting $H_0 = H_1 \cap H_2 \cap H_3$. We considered three different configurations: $(\mu_1, \mu_2, \mu_3) = (0, 0, \delta), (0, \delta, \delta)$ and (δ, δ, δ) where $\delta = 2$. The simulated powers for different m are summarized in Table 9.

Table 9: Simulated powers for rejecting $H_0 = H_1 \cap H_2 \cap H_3$ under independence for different m ($\rho = -0.25$)

m	(μ_1, μ_2, μ_3)		
	(0,0,2)	(0,2,2)	(2,2,2)
$-\infty$	0.4709	0.7043	0.8354
-2.0	0.4720	0.7146	0.8507
-1.0	0.4734	0.7362	0.8798
-0.5	0.4717	0.7656	0.9147
0.0	0.4324	0.8008	0.9537
0.5	0.2529	0.7307	0.9626
1.0	0.1829	0.5161	0.9414
2.0	0.1496	0.3994	0.9023
$+\infty$.	0.1291	0.3342	0.8633

First we note that as in the case of $n = 2$, maximum power is achieved at $m = -1$ (harmonic mean) when only one H_i is false, with optimum m increasing as more hypotheses

Table 10: Simulated type I error for rejecting $H_0 = \cap_{i=1}^n H_i$ for $n \geq 10$ using asymptotic approximations to critical values Using Method 1 and Method 2 When $\alpha = 0.05$

m	Method*	n				
		10	20	50	100	1000
-2	Method 0	0.0498	0.0500	0.0502	0.0501	0.0499
-0.5	Method 0	0.0543	0.0499	0.0455	0.0430	0.0362
-0.25	Method 1	0.0784	0.0803	0.0819	0.0818	0.0797
	Method 2	0.0208	0.0365	0.0515	0.0591	0.0719
0.25	Method 1	0.0414	0.0389	0.0370	0.0354	0.0334
	Method 2	0.0168	0.0219	0.0260	0.0280	0.0311
0.5	Method 1	0.0261	0.0248	0.0239	0.0233	0.0222
	Method 2	0.0132	0.0161	0.0184	0.0194	0.0210
1	Method 1	0.0092	0.0095	0.0098	0.0098	0.0100
	Method 2	0.0092	0.0095	0.0098	0.0098	0.0100
2	Method 1	0.0003	0.0009	0.0016	0.0016	0.0021
	Method 2	0.0059	0.0043	0.0034	0.0030	0.0025

* Method 0 uses the generalized central limit theorem (GCLT); see Section 5.2. Methods 1 and 2 use the central limit theorem (CLT); see Section 5.1.

become false: optimum $m = -0.5$ when two hypotheses are false and optimum $m = 0$ (geometric mean) when all three hypotheses are false. The power first increases with m and then decreases rapidly as m approaches $+\infty$.

8.3. Type I error simulations for $n \geq 10$

To check the accuracy of the asymptotic approximations to the critical values computed using Method 1 and Method 2 in Tables 4 and 6 we performed simulations of type I error for rejecting the overall null hypothesis $H_0 = \cap_{i=1}^n H_i$. The results are reported in Table 10. These results show that the asymptotic approximations are not very accurate and better approximations need to be found. Method 2 gives generally conservative approximations (estimated type I error rate is $< \alpha = 0.05$) except for $m = -0.25$ and $n \geq 50$, while Method 1 gives anti-conservative approximations for all values of n when $m = -0.25$; otherwise it is conservative. Generally, Method 2 is more conservative than Method 1.

9. Concluding remarks and practical recommendations

In this paper we have exhaustively studied the null distribution of the Hölder mean with the exact distribution for $n = 2$ and the asymptotic distribution for large n. We have also obtained the exact critical values for $n = 3$. The exact null distribution in closed form is also available for all $n > 2$ in special cases, *e.g.*, minimum, maximum and geometric mean and can be obtained using the convolution method in other cases, in particular, the harmonic mean and arithmetic mean. The asymptotic approximations to critical values are generally too conservative and better approximations need to be found.

These null distributions and their critical values are employed in a closed MTP. The power of the test of $H_0 = \cap_{i=1}^n$ for $n = 2$ and 3 for different values of m is evaluated for

six different configurations of (μ_1, μ_2) and three different configurations of (μ_1, μ_2, μ_3) and optimum chpoces of m are found. The power comparisons show that for $n = 2$ if only one null hypothesis is false ($\mu_1 = 0, \mu_2 > 0$) then the test based on the harmonic mean ($m = -1$) gives the maximum power and if both null hypotheses are equally false ($\mu_1 = \mu_2 > 0$) then the test based on the Hölder mean with $m = -0.5$ gives the maximum power. Similarly, for $n = 3$, if only one null hypothesis is false. Similarly, maximum power is achieved at $m = -1$ (harmonic mean) when only one H_i is false, with optimum m increasing as more hypotheses become false: optimum $m = -0.5$ when two hypotheses are false and optimum $m = 0$ (geometric mean) when all three hypotheses are false. Since the power plots in the vicinity of the maximum power are fairly flat, our practical recommendation is to use $m = -0.5$ or$m = -1$.

In this paper the power comparisons are limited to the test of H_0 for $n = 2$ and 3. Power comparisons are not made for $n \geq 10$ because the asymptotic critical values are too conservative.

Acknowledgements

We are grateful to Professor Nairanjana Dasgupta, a guest editor of this issue, for inviting us to submit an article. It is indeed a great honor to publish our work in honor and memory of Professor C. R. Rao, one of the greatest statisticians. Professor Rao made fundamental contributions to distribution theory and this is our small contribution to the area. We dedicate this article to his memory as a token of our appreciation of him.

References

Dobriban, E. (2020). Fast closed testing for exchangeable local tests. *Biometrika*, **107**, 761–768.

Gabriel, K. R. (1969). Simultaneous test procedures–some theory of multiple comparisons. *The Annals of Mathematical Statistics*, **40**, 224–250.

Gnedenko, B. V. and Kolmogorov, A. N. (1954). *Limit Distributions for Sum of Independent Random Variables*. Addison-Wesley, Cambridge, Massachusetts.

Gou, J. and Tamhane, A. C. (2024). A closed multiple test procedure based on the harmonic means of p-values. *submitted*.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**, 800–802.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. John Wiley and Sons, New York, New York.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, **75**, 383–386.

Ibragimov, I. A. and Linnik, Y. V. (1971). *Independent and Stationarily Sequences of Random Variables*. Wolters-Noordnoff, Groningen.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**, 655–660.

Mijnheer, J. (1975). *Sample Path Properties of Stable Processes*. Mathematical Centre tracts. Mathematisch Centrum.

Petrov, V. V. (1975). *Sums of Independent Random Variables*. Springer-Verlag, Berlin, Heidelberg.

Rektorys, K. (1969). *Survey of Applicable Mathematics*. The M.I.T. Press, Cambridge, Massachusetts.

Samorodnitsky, G. and Taqqu, M. S. (1994). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman and Hall, New York, New York.

Tian, J., Chen, X., Katsevich, E., Goeman, J., and Ramdas, A. (2023). Large-scale simultaneous inference under dependence. *Scandinavian Journal of Statistics*, **50**, 750–796.

Uchaikin, V. V. and Zolotarev, V. M. (1999). *Chance and Stability: Stable Distributions and their Applications*. VSP International Science Publishers, Utercht, Netherlands.

Vovk, V. and Wang, R. (2020). Combining p-values via averaging. *Biometrika*, **107**, 791–808.

ANNEXURE

Appendix: Proofs and Derivations

Proof of Theorem 1 As seen from Figure 3, for $m < m^*$, the rejection boundary is convex while for $m > m^*$, the rejection boundary is concave. The corresponding rejection regions are below the rejection boundaries. Therefore the regions of integration are different for evaluating the integral below:

$$F_2(x;m) = \Pr\left\{\frac{P_1^m + P_2^m}{2} \le x^m\right\} = \int (2x^m - y^m)dy. \tag{19}$$

The two regions are

$$R_1 = \{0 \le P_1 \le (2x^m - P_2^m)^{1/m}, 0 \le P_2 \le 1\} \quad (m \le m^*).$$

and

$$R_2 = \{0 \le P_1 \le 2^{1/m}x, 0 \le P_2 \le (2x^m - P_1^m)^{1/m} \quad (m > m^*)$$

The Case ($m \le m^*$): By integrating (19) over the region R_1, we get

$$\begin{aligned}
&F_2(x;m)\\
&= \int_0^{(2x^m-1)^{1/m}} (2x^m - y^m)^{1/m}dy\\
&= \int_0^{(2x^m-1)^{1/m}} dy + \int_{(2x^m-1)^{1/m}}^{1} (2x^m - y^m)^{1/m}dy\\
&= (2x^m-1)^{1/m} + (2^{1/m}x)^2 \int_{(1-1/2x^m)^{1/m}}^{1/2^{1/m}x} (1-u^m)^{1/m}du \ \text{ (by putting } \ u = y/(2^{1/m}x)).
\end{aligned}$$

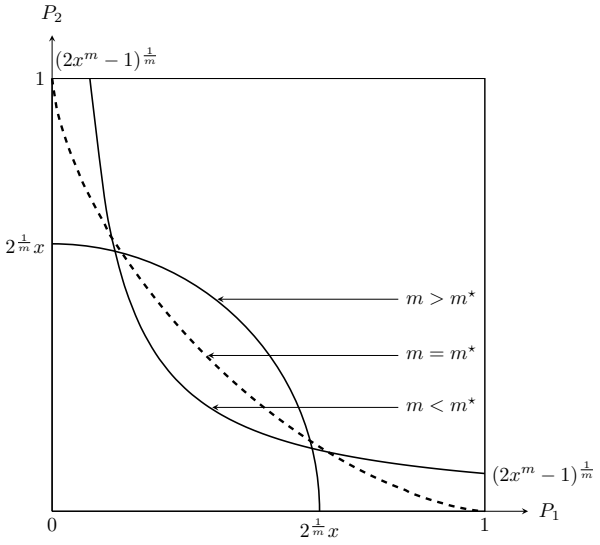


Figure 3: Rejection regions R_1 for $m \leq m^*$ and R_2 for $m > m^*$ which are below the respective boundaries

Hence

$$F_2(x;m) = (2x^m-1)^{1/m} + 4^{1/m}x^2\left[\int_0^{1/2^{1/m}x}(1-u^m)^{1/m}du - \int_0^{1-1/2^{1/m}x}(1-u^m)^{1/m}du\right].$$

Now put $u^m = v$. Hence $du = (1/m)v^{1/m-1}dv$. Thus we get

$$\begin{aligned}F_2(x;m) &= (2x^m-1)^{\frac{1}{m}} + \frac{4^{\frac{1}{m}}x^2}{m}\left[\int_0^{1/2x^m} v^{1/m-1}(1-v)^{1/m}dv - \int_0^{1-1/2x^m} v^{1/m-1}(1-v)^{1/m}dv\right] \\ &= (2x^m-1)^{1/m} + \frac{4^{1/m}x^2}{m}\left[B_{1/2x^m}\left(\frac{1}{m},\frac{1}{m}+1\right) - B_{1-1/2x^m}\left(\frac{1}{m},\frac{1}{m}+1\right)\right].\end{aligned}$$

The Case $m > m^*$: By integrating (19) over the region R_2, we get

$$\begin{aligned}F_2(x;m) &= \int_0^{2^{1/m}x}(2x^m-y^m)^{1/m}dy \\ &= 4^{1/m}x^2\int_0^1(1-u^m)^{1/m}du \quad (\text{by putting } u = \tfrac{y}{2^{1/m}x}).\end{aligned}$$

Now put $u^m = v$. Hence $du = (1/m)v^{1/m-1}dv$. Thus we get

$$\begin{aligned}F_2(x;m) &= \frac{4^{1/m}x^2}{m}\int_0^1 v^{1/m-1}(1-v)^{1/m}dv \\ &= \frac{4^{1/m}x^2}{m}B_1\left(\frac{1}{m},\frac{1}{m}+1\right) \\ &= \frac{4^{1/m}x^2}{m}\frac{\Gamma\left(\frac{1}{m}\right)\Gamma\left(\frac{1}{m}+1\right)}{\Gamma\left(\frac{2}{m}+1\right)}.\end{aligned}$$

assuming $2^{\frac{1}{m}}x \leq 1$ that is equivalent to $x \leq 2^{-1/m}$.

Proof of Theorem 2:

Part 1 ($m = -\infty$):

Assume that $P_1 = P_{\min}$ is unique. Then

$$\left(\sum_{i=1}^{n} w_i P_i^m\right)^{1/m} = P_1 \left(w_1 + \sum_{i=2}^{n} w_i \left(\frac{P_i}{P_1}\right)^m\right)^{1/m}$$
$$\to P_1 \text{ as } m \to -\infty \text{ since } w_1^{1/m} \to 1 \text{ and } \left(\frac{P_i}{P_1}\right)^m \to 0 \ \ \forall i > 2.$$

The c.d.f. of $P_1 = P_{\min}$ is

$$F_n(x; -\infty, w) = \Pr\{P_{\min} \leq x\} = 1 - \prod_{i=1}^{n} \Pr\{P_i > x\} = 1 - (1-x)^n.$$

Equating this to α and solving for x, we get

$$c_n(-\infty, \alpha) = 1 - (1-\alpha)^{1/n}$$

Part 2 ($m = -2$):

We have

$$F_2(x; -2, w) = \Pr\left\{\left(\frac{w_1}{P_1^2} + \frac{w_2}{P_2^2}\right)^{-1/2} \leq x\right\}$$
$$= \Pr\left\{P_2 \leq \sqrt{\frac{w_2 x^2 P_1^2}{P_1^2 - w_1 x^2}}\right\}.$$

Now note that if $P_1 \leq \sqrt{\frac{w_1 x^2}{1 - w_2 x^2}}$ then $P_2 \leq 1$. Hence the above probability equals

$$F_2(x; -2, w) = \Pr\left\{P_1 \leq \sqrt{\frac{w_1 x^2}{1 - w_2 x^2}}, P_2 \leq 1\right\} + \Pr\left\{P_1 > \sqrt{\frac{w_1 x^2}{1 - w_2 x^2}}, P_2 \leq \sqrt{\frac{w_2 x^2 P_1^2}{P_1^2 - w_1 x^2}}\right\}$$
$$= \sqrt{\frac{w_1 x^2}{1 - w_2 x^2}} + \int_{\sqrt{\frac{w_1 x^2}{1 - w_2 x^2}}}^{1} \sqrt{\frac{w_2 x^2 y^2}{y^2 - w_1 x^2}} dy$$
$$= \sqrt{\frac{w_1 x^2}{1 - w_2 x^2}} + \left.\frac{\sqrt{w_2 x^2 y^2 (y^2 - w_1 x^2)}}{y}\right|_{\sqrt{\frac{w_1 x^2}{1 - w_2 x^2}}}^{1}$$
$$= \sqrt{\frac{w_1 x^2}{1 - w_2 x^2}} + x\sqrt{w_2(1 - w_1 x^2)} - \sqrt{\frac{w_1 x^2}{1 - w_2 x^2}} + x\sqrt{w_1(1 - w_2 x^2)}$$
$$= x\left(\sqrt{w_1(1 - w_2 x^2)} + \sqrt{w_2(1 - w_1 x^2)}\right).$$

Part 3 ($m = -1$):

This case corresponds to *weighted harmonic mean*. Its c.d.f. is derived in the following.

$$\begin{aligned}
F_2(x;-1,w) &= \Pr\left(\left(\frac{w_1}{P_1}+\frac{w_2}{P_2}\right)^{-1} \le x\right) \\
&= \Pr\left\{P_2 \le \frac{w_2 x P_1}{P_1 - w_1 x}\right\} \\
&= \Pr\left\{P_1 \le \frac{w_1 x}{1-w_2 x}, P_2 \le 1\right\} + \Pr\left\{P_1 > \frac{w_1 x}{1-w_2 x}, P_2 \le \frac{w_2 x P_1}{P_1 - w_1 x}\right\} \\
&= \frac{w_1 x}{1-w_2 x} + \int_{\frac{w_1 x}{1-w_2 x}}^{1} \frac{w_2 x y}{y - w_1 x} dy \\
&= \frac{w_1 x}{1-w_2 x} + \left\{w_2 x y + w_1 w_2 x^2 \ln\left|\frac{y}{w_2 x} - \frac{w_1}{w_2}\right|\right\}\Bigg|_{\frac{w_1 x}{1-w_2 x}}^{1}
\end{aligned}$$

$$\begin{aligned}
&= \frac{w_1 x}{1-w_2 x} + \frac{w_2 x(1-x)}{1-w_2 x} + w_1 w_2 x^2 \ln\left[\frac{(1-w_1 x)(1-w_2 x)}{w_1 w_2 x^2}\right] \\
&= x + w_1 w_2 x^2 \ln\left[\frac{(1-w_1 x)(1-w_2 x)}{w_1 w_2 x^2}\right] \\
&= x + w_1 w_2 x^2 \ln\left[1 + \frac{1-x}{w_1 w_2 x^2}\right].
\end{aligned}$$

In Step 5 above we have used the standard formula from Rektorys (1969): For $a \neq b \neq 0$,

$$\int \frac{y dy}{ay+b} = \frac{y}{a} - \frac{b}{a^2}\ln|ay+b|.$$

Part 4 ($m = -0.5$):

$$\begin{aligned}
& F_2(x;-0.5) \\
= \;& \frac{1}{\left(\frac{2}{\sqrt{x}}-1\right)^2} + \int_{\frac{1}{\left(\frac{2}{\sqrt{x}}-1\right)^2}}^{1} \frac{1}{\left(\frac{2}{\sqrt{x}}-\frac{1}{\sqrt{y}}\right)^2} dy \\
= \;& \frac{1}{\left(\frac{2}{\sqrt{x}}-1\right)^2} + \frac{x}{8}\cdot\left(x^{3/2}/(\sqrt{x}-2\sqrt{y}) + 4\sqrt{xy} + 3x\ln(2\sqrt{y}-\sqrt{x}) + 2y\right)\Big|_{\frac{1}{\left(\frac{2}{\sqrt{x}}-1\right)^2}}^{1} \\
= \;& \frac{x}{(2-\sqrt{x})^2} + \frac{x}{8}\left(-\frac{x^{3/2}}{2-\sqrt{x}} + 4\sqrt{x} + 3x\ln(2-\sqrt{x}) + 2\right. \\
& \left. -x + 2\sqrt{x} - \frac{4x}{2-\sqrt{x}} - 3x\ln\left(\frac{x}{2-\sqrt{x}}\right) - \frac{2x}{(2-\sqrt{x})^2}\right) \\
= \;& \frac{x}{(2-\sqrt{x})^2} + \frac{x}{8}\left(6\sqrt{x} - x - \frac{x(4+\sqrt{x})}{2-\sqrt{x}} + 3x\ln\left(\frac{(2-\sqrt{x})^2}{x}\right) + 2 - \frac{2x}{(2-\sqrt{x})^2}\right).
\end{aligned}$$

The lower α critical value is obtained by solving the equation obtained by setting the above expression equal to α.

Part 5 ($m = 0$):

We have

$$F_2(x;0,w) = \Pr\left(P_1^{w_1} P_2^{w_2} \le x\right).$$

Now note that if $P_1 \le x^{1/w_1}$ then $P_2 \le 1$. Therefore

$$\begin{aligned}
F_2(x;0,w) &= \Pr\{P_1 \le x^{1/w_1}, P_2 \le 1\} + \Pr\left\{P_1 > x^{1/w_1}, P_2 \le \frac{x^{1/w_2}}{P_1^{w_1/w_2}}.\right\} \\
&= x^{1/w_1} + \int_{x^{1/w_1}}^{1} \frac{x^{1/w_2}}{y^{w_1/w_2}} dy \\
&= x^{1/w_1} + x^{1/w_2} \frac{w_1}{w_2 - w_1} \left[y^{\frac{w_2 - w_1}{w_2}} \right]_{x^{1/w_1}}^{1} \\
&= x^{1/w_1} + x^{1/w_2} \frac{w_1}{w_2 - w_1} \left[1 - x^{\frac{w_2 - w_1}{w_1 w_2}} \right] \\
&= \begin{cases} x^2(1 - 2\ln x), & w_1 = w_2 = 1/2 \\ \frac{1}{1 - \frac{w_2}{w_1}} x^{\frac{1}{w_1}} + \frac{1}{1 - \frac{w_1}{w_2}} x^{\frac{1}{w_2}} & w_1 \ne w_2. \end{cases}
\end{aligned}$$

For an alternative proof of the case $w_1 = w_2 = 1/2$, note that for any $n \ge 2$,

$$\begin{aligned}
F_2(x;0,w) &= \Pr\left\{ \left(\prod_{i=1}^{n} P_i \right)^{1/n} \le x \right\} \\
&= \Pr\left\{ -\frac{2}{n} \sum_{i=1}^{n} \ln P_i > -2\ln x \right\} \\
&= \Pr\left\{ -2 \sum_{i=1}^{n} \ln P_i > -2n \ln x \right\} \\
&= \Pr\left\{ \chi^2_{2n} > -2n \ln x \right\}.
\end{aligned}$$

Now consider $n = 2$. Then by putting $u = t/2$ in the integral below we get

$$F_n(x;0,w) = \int_{-4\ln x}^{\infty} \frac{1}{2^2 \Gamma(2)} t e^{-t/2} dt = \int_{-2\ln x}^{\infty} u e^{-u} du = x^2(1 - 2\ln x).$$

Part 6 ($m = 0.5$):

We have

$$\begin{aligned}
F_2(x, 0.5) &= \int_0^{2^2 x} (2\sqrt{x} - \sqrt{y})^2 dy \\
&= -\frac{8}{3} \sqrt{x y^3} + 4xy + y^2/2 \Big|_0^{4x} \\
&= \frac{8x^2}{3}.
\end{aligned}$$

Equating $8x^2/3 = \alpha$ we get the lower α critical value as $c_2(0.5, \alpha) = \sqrt{3\alpha/8}$. These expressions for the c.d.f. and the α critical value are valid for all α less than or equal to

$$\int_0^1 (1 - x^{1/2})^2 dx = \frac{x^2}{2} - \frac{4x^{3/2}}{3} + x\Big|_0^1 = \frac{1}{6}.$$

Part 7 ($m = 1$):

Assuming $w_1 \leq w_2$, the c.d.f. $F_2(x; 1, w)$ is given by the areas of the regions in the (P_1, P_2) space as follows.

1. $(0 \leq x \leq w_1)$: In this case the region of interest is the triangle shown in Figure 4 (a). Its area equals
$$F_2(x; 1, w) = \frac{1}{2}\left(\frac{x}{w_1} \times \frac{x}{w_2}\right) = \frac{x^2}{2w_1w_2}.$$
2. $(w_1 < x \leq w_2)$: In this case the region of interest is the quadrilateral shown in Figure 4 (b). Its area equals
$$F_2(x; 1, w) = \frac{1}{2}\left(\frac{x}{w_2} + \frac{x - w_1}{w_2}\right) = \frac{2x - w_1}{2w_2}.$$
3. $(w_2 < x \leq 1)$: In this case the region of interest is the trapezoid shown in Figure 4 (c). Its area equals
$$\begin{aligned} F_2(x; 1, w) &= 1 - \frac{1}{2}\left(1 - \frac{x - w_1}{w_2}\right)\left(1 - \frac{x - w_2}{w_1}\right) \\ &= 1 - \frac{1}{2}\left(\frac{w_2 - x + w_1}{w_2}\right)\left(\frac{w_1 - x + w_2}{w_1}\right) \\ &= 1 - \frac{(1 - x)^2}{2w_1w_2}. \end{aligned}$$

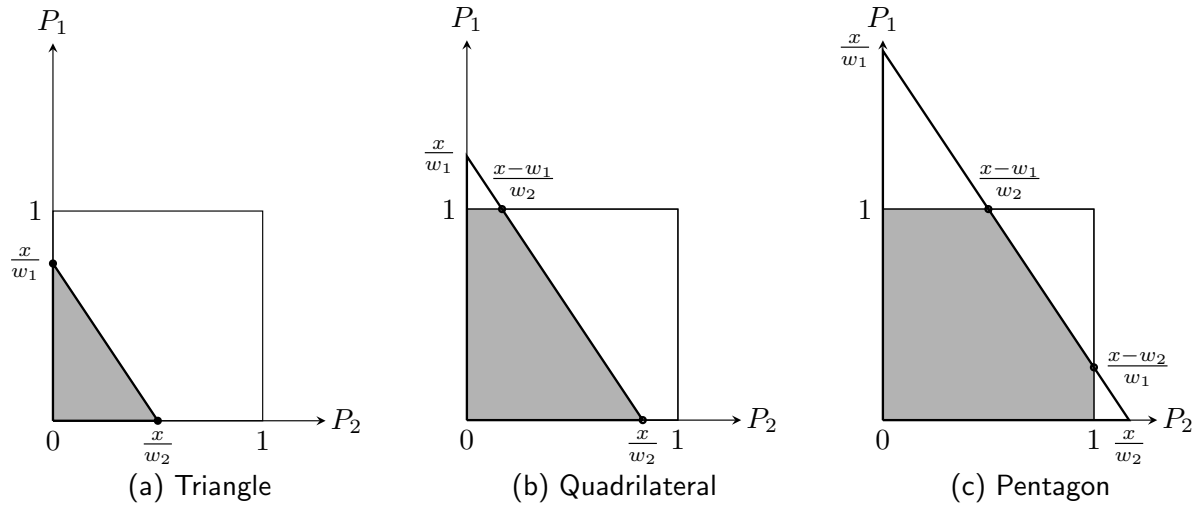
Part 8 ($m = 2$):

Assume $w_1 \leq w_2$. We consider three cases.

Case 1 ($x \leq \sqrt{w_1}$) Then

$$\begin{aligned} F_2(x; 2, w) &= \Pr\{\sqrt{w_1P_1^2 + w_2P_2^2} \leq x\} \\ &= \Pr\{w_1P_1^2 + w_2P_2^2 \leq x^2\} \\ &= \Pr\left\{\frac{P_1^2}{(x^2/w_1)} + \frac{P_2^2}{(x^2/w_2)} \leq 1\right\} \\ &= \frac{\pi x^2}{4\sqrt{w_1w_2}}. \end{aligned}$$

using the formula for the area of an ellipse.



(a) Triangle (b) Quadrilateral (c) Pentagon

Figure 4: Rejection regions for weighted arithmetic mean

Case 2 ($\sqrt{w_1} < x \leq \sqrt{w_2}$)

$$\begin{aligned} F_2(x;2,w) &= \Pr\{\sqrt{w_1P_1^2 + w_2P_2^2} \leq x\} \\ &= \int_0^1 \sqrt{\frac{x^2 - w_1y^2}{w_2}}dy \\ &= \left.\frac{\left(y\sqrt{w_1(x^2 - w_1y^2)} + x^2\tan^{-1}\left(\frac{\sqrt{w_1}y}{\sqrt{x^2-w_1y^2}}\right)\right)}{2\sqrt{w_1w_2}}\right|_0^1 \\ &= \frac{\sqrt{w_1(x^2 - w_1)} + x^2\tan^{-1}\left(\frac{\sqrt{w_1}y}{\sqrt{x^2-w_1y^2}}\right)}{2\sqrt{w_1w_2}}. \end{aligned}$$

Case 3 ($x > \sqrt{w_2}$)

$$\begin{aligned} F_2(x;2,w) &= \Pr\left(\sqrt{w_1P_1^2 + w_2P_2^2} \leq x\right) \\ &= \sqrt{\frac{x^2 - w_2}{w_1}} + \int_{\sqrt{\frac{x^2-w_2}{w_1}}}^1 \sqrt{\frac{x^2 - w_1y^2}{w_2}}dy \\ &= \sqrt{\frac{x^2 - w_2}{w_1}} + \left.\frac{\left(\sqrt{w_1}y\sqrt{x^2 - w_1y^2} + x^2\tan^{-1}\left(\frac{\sqrt{w_1}y}{\sqrt{x^2-w_1y^2}}\right)\right)}{2\sqrt{w_1w_2}}\right|_{\sqrt{\frac{p^2-w_2}{w_1}}}^1 \\ &= \sqrt{\frac{x^2 - w_2}{w_1}} + \frac{\sqrt{w_1(x^2 - w_1)} + x^2\tan^{-1}\left(\frac{\sqrt{w_1}}{\sqrt{x^2-w_1}}\right)}{2\sqrt{w_1w_2}} \end{aligned}$$

$$
\begin{aligned}
& -\frac{\sqrt{w_1 w_2}\sqrt{\frac{x^2-w_2}{w_1}} + x^2 \tan^{-1}\left(\sqrt{\frac{x^2-w_2}{w_2}}\right)}{2\sqrt{w_1 w_2}} \\
= \; & \frac{1}{2}\left(\sqrt{\frac{x^2-w_2}{w_1}} + \sqrt{\frac{x^2-w_1}{w_2}}\right) + \frac{x^2}{2\sqrt{w_1 w_2}}\left(\tan^{-1}\left(\sqrt{\frac{w_1}{x^2-w_2}}\right) - \tan^{-1}\left(\sqrt{\frac{x^2-w_2}{w_2}}\right)\right)
\end{aligned}
$$

Part 9 ($m = \infty$)

Assume that $P_n = P_{\max}$ is unique. Then

$$
\begin{aligned}
\left(\sum_{i=1}^{n} w_i P_i^m\right)^{1/m} &= P_n \left(\sum_{i=1}^{n} w_i \left(\frac{P_i}{P_n}\right)^m\right)^{1/m} \\
&= P_n \left(w_n + \sum_{i=1}^{n-1} w_i \left(\frac{P_i}{P_n}\right)^m\right)^{1/m} \\
&\to P_n \quad \text{as } m \to \infty \text{ since } w_n^{1/m} \to 1 \text{ and } \left(\frac{P_i}{P_n}\right)^m \to 0
\end{aligned}
$$

The c.d.f. of $P_n = P_{\max}$ is

$$
F_n(x; \infty, w) = \Pr\{P_{\max} \le x\} = \prod_{i=1}^{n} \Pr\{P_i \le x\} = x^n.
$$

Equating this to α and solving for x, we get $c_n(\infty, \alpha) = \alpha)^{1/n}$.

Proof of Theorem 3:
Consider two values of m, $m' < m''$.Then we have

$$
\Pr\{\bar{P}_n(m') \le c_n(m', \alpha)\} = \Pr\{\bar{P}_n(m'') \le c_n(m'', \alpha)\} = \alpha.
$$

From the power mean inequality we have $\bar{P}_n(m') \le \bar{P}_n(m'')$. Therefore

$$
\begin{aligned}
\Pr\{\bar{P}_n(m') \le c_n(m'', \alpha)\} &= \Pr\{\bar{P}_n(m') \le \bar{P}_n(m'') \le c_n(m'', \alpha)\} \\
&\quad + \Pr\{\bar{P}_n(m') \le c_n(m'', \alpha) \le \bar{P}_n(m'')\} \\
&\ge \Pr\{\bar{P}_n(m') \le \bar{P}_n(m'') \le c_n(m'', \alpha)\} \\
&= \Pr\{\bar{P}_n(m'') \le c_n(m'', \alpha)\} \\
&= \alpha.
\end{aligned}
$$

Since $\Pr\{\bar{P}_n(m') \le c_n(m'', \alpha)\} \ge \alpha$ and $\Pr\{\bar{P}_n(m') \le c_n(m', \alpha)\} = \alpha$, it follows that $c_n(m', \alpha) \le c_n(m'', \alpha)$.

Derivation of the Mean and Variance of $\frac{1}{n}\left(\sum_{i=1}^{n} P_i^m\right)^{1/m}$ Using the Delta Method
Denote $\frac{1}{n}\left(\sum_{i=1}^{n} P_i^m\right) = X$. From (12) it follows that

$$
E(X) = \frac{1}{m+1} \quad \text{and} \quad \mathrm{Var}(X) = \frac{m^2}{n(m+1)(2m+1)}.
$$

Now let $g(X) = X^{1/m}$. By the delta method we have $E[g(X)] \approx \left(\frac{1}{m+1}\right)^{1/m}$ and

$$\begin{aligned}\mathrm{Var}[g(X)] &\approx \mathrm{Var}(X)[g'(m)]^2 \\ &= \frac{m^2}{n(m+1)(2m+1)}\left[-\frac{(m+1)^{-(1+1/m)}}{m}\right]^2 \\ &= \frac{m^2}{n(m+1)(2m+1)}\frac{(m+1)^{2(1-1/m)}}{m^2} = \frac{(m+1)^{1-2/m)}}{n(2m+1)}.\end{aligned}$$

Distribution of $S(1,1)$

There is no explicit closed formula for the distribution of $S(1,1)$. However, we can calculate it numerically from its characteristic function given by

$$\varphi(t \mid a, b, \mu, \sigma) = \exp\left\{it\mu - |t\sigma|^a(1 - ib \cdot \mathrm{sgn}(t) \cdot \Psi)\right\} \tag{20}$$

where $i = \sqrt{-1}$, $a \in (0, 2]$ is a stability parameter, $b \in [-1, 1]$ is a skewness parameter, $\mu \in (-\infty, \infty)$ is a shift parameter, $\sigma > 0$ is a scale parameter and

$$\Psi = \begin{cases} \tan\left(\frac{\pi a}{2}\right) & (a \neq 1), \\ -\frac{2}{\pi}\ln|t| & (a = 1). \end{cases}$$

The p.d.f. of $S(1,1)$ can be found by applying the inverse Fourier transform to its characteristic function:

$$f^*(x) = \frac{1}{2n}\int_{-\infty}^{\infty} \varphi(t)e^{-ixt}dt. \tag{21}$$

The c.d.f. can be found from $F^*(x) = \int_{-\infty}^{x} f^*(t)dt$. These operations can be done numerically using the `MATLAB` function `makedist()`. Figure 5 shows the plots of the p.d.f. of $S(0.5,1), S(1,1)$ and $S(1.5,1)$computed using the above numerical method.

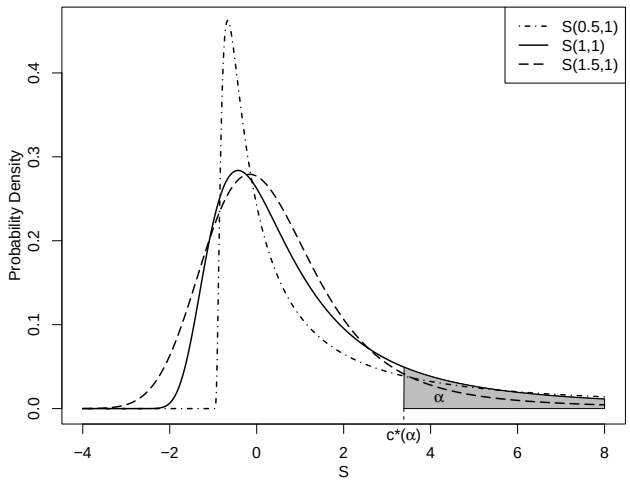


Figure 5: The p.d.f.s of the stable distributions $S(0.5,1)$, $S(1,1)$ and $S(1.5,1)$.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 639–666
http://www.ssca.org.in/journal

Identification of Changes in Temperature and Precipitation in Cities Across the Contiguous United States through High Dimensional Change Point Analysis

Abhishek Kaul[1], Alexandros Paparas[2], Venkata K. Jandhyala[1] and Stergios B. Fotopoulos[3]

[1]*Department of Mathematics and Statistics, Washington State University, USA*
[2]*Department of Information Systems and Business Analytics, Eastern Washington University, USA*
[3]*Department of Finance and Management Science, Washington State University, USA*

Received: 06 June 2024; Revised: 16 October 2024; Accepted: 19 October 2024

Abstract

In this article, we carry out a simultaneous study of changes in temperature and precipitation variables in several cities across the contiguous United States for the time period 1948-2023. Included among the seven climatic variables that we consider are both extremes as well as averages. The data on all the climatic variables in this article is sourced from Global Summary of the Year (GSOY), provided by the National Center for Environmental Information (NCEI) under the National Oceanic and Atmospheric Administration (NOAA). The main goal of the article is to simultaneously detect abrupt changes in the averages of the seven climatic variables by implementing recently developed high dimensional change point methodology. The methodology identifies three years, namely, 1957, 1989, and 2010 as the years of changes taking place in the climatic variables. Extensive follow up analysis is carried out to determine clusters of cities such that the nature of changes are similar within each cluster and differ significantly between the clusters. The clustering is done based upon the magnitudes of change computed for each change year at each city and for each climatic variable. The clusters enabled us to identify regions within US in which increases/decreases have occurred at any given change year. For example, temperatures were found to decrease in the change year 1957 and this decrease occurred predominantly in the northeastern, southeastern and southern regions of the United States. More comprehensive summary of our findings can be found in the discussion section of the article. Some plausible reasons for changes such as solar dimming and solar brightening were discussed in the concluding remarks section.

Key words: Temperature and precipitation; High-dimensional; Change points; Climatic extremes.

AMS Subject Classifications: 62F12; 62P12

Corresponding Author: Venkata K Jandhyala
Email: jandhyala@wsu.edu

1. Introduction

The study of climatic changes through important climatic variables is fundamental to a proper understanding of the prevailing climatic conditions as well as the conditions we can expect in the years ahead. Such an understanding will enable humans to adapt to changing conditions and plan for taking timely actions to prevent the onslaught of extreme climatic conditions. Among all climatic variables, temperature and precipitation are by far the most important variables for purposes of observation, analysis and understanding.

In this paper, our main goal is to study changes in temperature and precipitation at 91 stations spread throughout the contiguous United States. Changes in temperature and precipitation can be studied separately for the two variables or could also be studied together in a combined way. Also, changes could be studied in averages, extremes, thresholds and each of such studies brings its own understanding of the variables under study. In this article we consider seven variables, all representing one of precipitation or temperature from each of the cities considered in this analysis. Specifically, we have considered two precipitation and five temperature variables:

PRCP1 Frequency of days with precipitation exceeding one inch
PRCP Total annual precipitation
TMAX32 Instances of maximum temperature dropping below 32°F
TMAX90 Occurrences of maximum temperature surpassing 90°F
TAVG Average annual temperature
TMAX Average annual maximum temperature
TMIN Average annual minimum temperature

There is a large body of literature on climatic studies including those on temperature and precipitation. Among them, most of the existing studies on temperature focus on extremes only. For example, heat waves in 1995 and 1999 resulted in 739 and 110 excess deaths, respectively, in the city of Chicago alone. Based on regional climate model simulations (RCMs), Kunkel *et al.* (2010) predicted that there is a high probability of heat waves of unprecedented severity by the end of twenty first century if the high emissions path is followed. Oswald (2018) studied spatially continuous homogenized climate data to examine changes in regularity of heat waves including nighttime and daytime temperatures across the continental United States. The analysis showed prevalence of heatwaves between mid-70s and 2015. This was preceded by a decrease since 1948, the beginning of the dataset. Earlier Oswald and Rood (2014) studied extreme heat event days (EHEs) in the continental US based on daily maximum, minimum temperatures. The study period was 1930-2010 and results showed negative trends in the interior while positive trends showed in coastal and southern areas. While decreases occurred between 1930-1970, these decreases were offset by increases between 1970-2010. Gaffen and Ross (1998) examined trends in the frequency of days with anomalously high apparent temperatures (ATs) across the United States from 1949 to 1995 and observed that the annual frequency of extreme minimum ATs increased at the greatest number of stations, particularly in the eastern and western United States. Extending the data for the years 1949-2010, Grundstein and Dowd (2011) found that an increase in occurrence of 1-day extreme minimum ATs was particularly notable, especially in the eastern and western United States. Lee *et al.* (2014) examined monthly maximum and minimum temperatures from 932 stations located across the contigouous US for the years

1897-2010 and found estimated trend for monthly maximum had a mean of 0.47°C/Century while the estimated trend for monthly minimum had a mean of 1.65° C/Century.

Studies that focus on precipitation changes are equally important. Events of extreme precipitation are among the costliest of natural disasters. They are associated with flooding, damage to infrastructure, and loss of life. In the United States alone, extreme precipitation events have caused more than $200 billion damages during 1988-2017, with an increasing trend in costs as these events have become more frequent. In a recent study, Martinez-Villalobos and Neelin (2023) used a probability distribution for precipitation and predicted that about 13% of the globe and 25% of the tropics have displayed increases in extreme precipitation. While studying changes in extreme precipitation in the northeastern united states, Nazarian *et al.* (2022) found that extreme precipitation increased throughout the region with the largest changes seen in the summer. Implementing dynamically downscaled simulation, most recently Nazarian *et al.* (2024) predicted that both mean and extreme precipitation will increase to the east of the Sierra Madre highland and that extreme precipitation events can be expected to double throughout the region. Earlier, under a predicted 2°C of global warming, Rupp *et al.* (2022) found large variability in the magnitude of extreme precipitation across the western United States. Specifically, they found that majority of the region showed heavier tails for extreme precipitation under warming, while plateaus of eastern Oregon and Washington, and the crest of the Sierra Nevada, showed a lightening of tails. Armal *et al.* (2018) developed a Bayesian multilevel model using data from 1244 rainfall stations throughout the contiguous united states and found statistically significant trends in extreme rainfall frequency in 742 of the 1244 stations. These stations were predominantly in US Southeast and Northeast regions. Also, the trends in 274 out of the 742 stations can be attributed to El Nino Southern Oscillation, the North Atlantic Oscillation, the Pacific decadal oscillation, and the Atlantic multidecadal oscillation along with changes in global surface temperature anomalies. These 274 stations are mainly found in the U.S. northwest, west, and southwest climate regions.

There are several articles in the literature that study changes in both temperature and precipitation. It is important to review some of such studies as well to obtain a better understanding of changes in these two climatic factors that are interdependent. Robinson (2021) reviews the observational evidence for climate-driven increases in extremes most relevant to the continental United States. Wang *et al.* (2015) applied dynamical-statistical downscaling approach for studying climate change impacts at local scales. They applied the methodology for projecting future climate over the province of Ontario, Canada and found that there would be a significant warming trend throughout this century for the entire province while less precipitation is projected for most of the selected weather stations. Later, Zhou *et al.* (2018) predicted that there will be an increasing pattern of temperature and precipitation extremes over Canada over two time- slices (*i.e.*, 2046-2065 and 2076-2095). The effects of climate change and global warming on Alaska are unequivocal. From 1949 to 2012, the annual mean temperature increased 1.78°C and annual precipitation increased 3.1mm; winter changes were most dramatic, with temperatures climbing 3.78°C and precipitation increasing by 7.2mm (Bieniek *et al.* (2014)). Isaac and Van Wijngaarden (2012) analyzed hourly values of temperatures and relative humidity observed at 309 stations located across North America for the period 1948-2010. Trends were determined based on straight line fits and results showed significant warming trends in the mid western US, Canadian prairies, and western arctic. Lai and Dzombak (2019) analyzed time series of historical annual aver-

age temperature, total precipitation, and extreme weather indices for 103 (for temperature indices) and 115 (for precipitation indices) U.S. cities with climate records starting from as early as 1870. Applying linear regression modelling, Lai and Dzombak (2019) constructed 95% confidence intervals for the mean rate of change. The results showed increases in annual average temperature and precipitation although there were spatial and temporal variations. Cities in the Northeast and Midwest showed significant increases in precipitation while no increases in temperature in Southeast regions were found. Earlier, Griffiths and Bradley (2007) examined changes in five temperature and five precipitation extreme indicators from the northeast US for the period 1926-2000. Their empirical orthogonal function (EOF) analysis showed increases in both temperature and precipitation extremes. High correlation was found between number of frost days and warm nights and Atlantic Oscillation (AO).

The above short review of studies regarding temperature and precipitation changes within the contiguous United States makes it clear that such climatic studies must continue. Then only the scientific community will be able to have a proper understanding of the dynamic behavior of various climatic variables including temperature and precipitation. It is also clear from the above review that change point methodology ((Zhao and Chu, 2010; Wang *et al.*, 2010; Villarini *et al.*, 2013; Lee *et al.*, 2014)) is a powerful way for modeling changes in climatic variables. In this article, we shall adapt this frequently implemented method for capturing changes in temperature and precipitation variables within the contiguous United States over the period 1948-2023.

The change point methodology has long been a tool for climatologists for estimation of unknown time points at which abrupt changes might have occurred in one or more climatic variables. See, *e.g.*, Jandhyala *et al.* (2013); Beaulieu *et al.* (2012); Reeves *et al.* (2007), and the many references therein. Change point models for climatic data can be implemented individually in a univariate way for each city, or can also be implemented simultaneously for all cities in a multivariate way. Clearly, simultaneous modelling accounts for dependencies among cities that would otherwise have been ignored. Moreover, changes detected from a univariate analysis are with respect to the corresponding variance in the one the dimensional series. Whereas, a change recovered from a multivariate series measures the total change (in ℓ_2 magnitude) across all components with respect to the total variance across all components. This distinction highlights the main advantage of multivariate change point estimation, *i.e.*, it brings out systemic macro-level (in this case country-wide) temporal changes providing a more robust perspective on large scale climatic changes. In contrast changes that are recovered componentwise may be localized at a city or other regional level that may instead be indicative of localized weather variations instead of the large scale climate.

In recent years, change point methods have been developed for modeling and analyzing high dimensional data where parameter size is much larger than the sample size. These high dimensional change point methods enable the implementation of models that were previously considered intractable. While high dimensional change point methods have been applied for the analysis of financial data (Cai and Wang (2023)), socio-economic data (Kaul *et al.* (2019)), and mortality data (Chen *et al.* (2023)), there has not yet been application of this methodology for modeling and analyzing climatological data. In this paper, our goal is to carry out a comprehensive high dimensional modeling and analysis of temperature and precipitation data from cities across the contiguous United States. We shall first present a brief review of recent advances in high dimensional change point methods.

Fixed dimensional mean shift models and other variants have existed for several decades with well-known monographs being available, *e.g.*,Csorgo and Horváth (1997). The multivariate and high dimensional versions of these non-stationary models have seen significant recent research with an overwhelming proportion devoted to estimation methodologies for change points. A common thread of available methods is the use of general purpose algorithm's such as *binary segmentation* Venkatraman (1992), *wild binary segmentation* Fryzlewicz (2014) and *minimal partitioning via dynamic programming* Jackson *et al.* (2005). The first two work as extensions of single change point methods to multiple changes. From a methods perspective, the literature on estimation of change points under high dimensions can be forked into two general approaches, **(a).** Regularized cumulative sum (CUSUM) based recovery that is typically built on ℓ_1, ℓ_2 or ℓ_∞ aggregations of a cumulative sum metric. **(b).** Regularized M-estimation type recovery that is typically built upon a squared loss or a likelihood function. The former considered in (Enikeeva and Harchaoui, 2019; Jirak, 2015) which are based on an ℓ_2, ℓ_∞ aggregation, respectively. Other CUSUM based estimators include (Cho and Fryzlewicz, 2015; Cho, 2016; Wang and Samworth, 2018) amongst others, with the last allowing for sparsity of parameters and thus allowing for high dimensional means. Approach (b) is taken in (Wang *et al.*, 2020; Kaul *et al.*, 2021). Algorithmic advancements pertaining to minimal partitioning that is particularly critical for M-estimation type change point recovery is developed in Killick *et al.* (2012). Several other types of high dimensional change point models have also been studied in the recent literature, *e.g.*, linear regression, Bernoulli networks, graphical models, see, *e.g.*, (Kaul *et al.*, 2019, 2023; Lee *et al.*, 2016; Bhattacharjee *et al.*, 2020; Wang *et al.*, 2021) and several others. The problem of post-estimation inference on change points is a much lesser studied aspect in comparison to estimation alone, however some recent works have developed significant results under large data designs. Fundamental results under univariate $p = 1$ designs are available in *e.g.*, (Bai, 1994; Eichinger and Kirch, 2018; Cho and Kirch, 2022; Fotopoulos *et al.*, 2010). The case of diverging p is considered in Bhattacharjee *et al.* (2017). The article Kaul *et al.* (2021) which considers the high dimensional case, in a single change point setting ($N = 1$).

The article is organized as follows. Section 2 describes data analyzed in the article. Section 3 discusses published results on high dimensional change point methods that are utilized for the analysis in the paper, and Section 4 presents the implementation of high dimensional change point methods and their results. Section 5 is dedicated to a comprehensive discussion of the results and Section 6 ends the paper with some concluding remarks.

2. Temperature and precipitation data from contiguous United States

The data on temperature and precipitation variables is spread across the contiguous United States. It originates from the Global Summary of the Year (GSOY), provided by the National Center for Environmental Information (NCEI) under the National Oceanic and Atmospheric Administration (NOAA). It is available publicly and can be accessed from the NOAA GSOY Database. While the complete NCEI dataset is more comprehensive, we have meticulously collected data only on temperature and precipitation variables from 91 cities for the period 1948-2023 spread across the 48 contiguous states of the US. The dataset collected and analyzed in this article includes two precipitation and five temperature variables. Amongst these variables, three are discrete and the remaining are continuous. These include: PRCP1: # of days in a year with precipitation exceeding one inch, and PRCP: total annual

precipitation measured in mm; and five temperature variables - TMAX32: # of instances in days of maximum temperature dropping below 32°F, TMAX90: # of occurrences in days of maximum temperature surpassing 90°F, TAVG: average annual temperature in °C computed by adding the unrounded monthly average temperatures and dividing by 12, TMAX: average annual maximum temperature in °C obtained as average of the mean monthly maximum temperatures, and TMIN: average annual minimum temperature in °C obtained as average of the mean monthly minimum temperatures. It may be noted that among the seven climatic variables, the variables PRCP1, TMAX32, TMAX90, TMAX, and TMIN represent extremes with PRCP1 being the only extreme variable for precipitation.

The collected data spanning years 1948-2023 demonstrates substantial diversity in temporal scope and geography. Thus, the selected cities ensure comprehensive coverage of various geographical regions and climatic conditions of the US. The dataset includes not only big metropolitan cities, but also rural areas surrounding these urban centers, offering a comprehensive representation of climatic conditions beyond the city limits. Along the East Coast, cities such as New York, Boston, and Philadelphia offer insights into the climatic nuances of the Northeast. Moving southward, vibrant urban centers like Atlanta, Miami, and New Orleans provide a glimpse into the subtropical climates of the Southeast. Across the Midwest, cities like Chicago, Minneapolis, and Kansas City showcase the variability of continental climates. In the Great Lakes region, cities such as Buffalo, Cleveland, and Milwaukee experience the moderating effects of the large bodies of water, influencing their climate patterns. On the West Coast, cities such as Los Angeles, San Francisco, and Seattle offer perspectives on the mild coastal climates of the Pacific. In the Southwest, cities like Phoenix, Las Vegas, and Albuquerque experience arid desert climates, while Denver and Salt Lake City experience the high-altitude conditions of the Rocky Mountains. Our dataset also includes cities in the Mountain West, Great Plains, and Pacific Northwest, providing a comprehensive understanding of climatic variations across the United States.

3. High dimensional methods for identifying change points in time series

We adopt a high dimensional multiple mean shift framework to model the considered climate data, specifically,

$$y_t = \sum_{j=1}^{N+1} \theta_{(j)}^0 \mathbf{1}[\tau_{j-1}^0 < t \leq \tau_j^0] + \varepsilon_t, \quad \text{for } t = 1, ..., T, \tag{1}$$

wherein $y_t = (y_{t1}, y_{t2}, ..., y_{tp})^T \in \mathbb{R}^p$ denotes the underlying temperature (5 variables for each city) and precipitation (2 variables for each city) variable across all considered cities.

There are 91 cities in the data set, resulting in $p = 628$ variables. The Model 1 assumes there are an unknown number $N \in \mathbb{N}^+ = \{1, 2, ...\}$ of change points in the underlying mean vectors $\theta_{(j)}^0 \in \mathbb{R}^p$, $j = 1, ..., (N+1)$. , where their locations in the sampling period are denoted by $\tau^0 = (\tau_1^0, \tau_2^0, ..., \tau_N^0)^T \subseteq \{1, ..., T\}^N$. Our analysis to follow allows for spatial dependence across variables y_{tj}, $j = 1, ..., p$, *i.e.*, it allows for a dependence between temperature and precipitation variables as well as across cities. However, we assume temporal independence.

Remark 1: While the modelling structure adopted induces a large number of parameters, our chosen methodology is capable of allowing such high dimensionality as explained below.

A potential alternative for reduced modelling dimensions is to perform a coarser aggregation of cities into regional blocks (*e.g.*, North, Northeast, East, Midwest, South, Southeast, Southwest, West, and Northwest), however such an approach may lead to compromise on the post-hoc identification of the natural homogeneity of climatic changes amongst considered cities and instead force these to be on the chosen coarser grid.

We utilize the method and results of Kaul *et al.* (2021) for inference on change points. They proposed an iterative estimation procedure between squared loss based change point recovery and a ℓ_1-regularized squared loss recovery of mean estimates. While this article is developed under the assumption of a single change point $N = 1$, we utilize its natural extensions to multiple change points via the extensively chosen principle of binary segmentation.

The following provides a brief description of the methods and main results of Kaul *et al.* (2021) utilized here. Let $\bar{y} = \left(\sum_{t=1}^{T} y_t/T\right)$ and $x_t = (y_t - \bar{y})$, be the globally centered observations. Then under a single change point ($N = 1$), define a squared loss,

$$Q\Big(\tau, \theta\Big) = \sum_{t=1}^{\tau} \|x_t - \theta_{(1)}\|_2^2 + \sum_{t=\tau+1}^{T} \|x_t - \theta_{(2)}\|_2^2. \tag{2}$$

Additionally define ℓ_1 regularized mean estimated at any given τ as,

$$\hat{\theta}_{(j)}(\tau) = k_{\lambda_j}\Big(x_{(j)}(\tau)\Big), \quad j = 1, 2 \tag{3}$$

with $k_\lambda(x) = \text{sign}(x)(|x| - \lambda)_+$, $\lambda > 0$, $x \in \mathbb{R}^p$, is the *soft-thresholding* operator, where $\text{sign}(\cdot)$, $|\cdot|$, and $(\cdot)_+$[1] are applied component-wise. Then Algorithm 1 provides a twice-iterative method for recovery of the change point, where $\gamma > 0$ is a tuning parameter.

Next we briefly discuss properties of the estimator $\hat{\tau}$ that are relevant for our analysis. These properties assume suitable regularity conditions. Among the two most relevant ones, first we allow for spatial dependence with the underlying distribution being of a sub-exponential type (see *e.g.*, Vershynin (2019)). Next, given the high dimensional nature of the considered problem, an underlying sparsity of the mean parameters is also assumed. Further details are omitted here in view of simplicity of exposition.

The change point estimate from Algorithm 1 possesses desirable statistical properties in context of both estimation and inference, despite the underlying high dimensionality of mean parameters. To characterize the limiting distribution of the estimate $\tilde{\tau}$, we require the following negative drift two-sided random walk initializing at the origin,

$$\mathcal{C}_\infty(\zeta) = \begin{cases} \sum_{t=1}^{\zeta} z_t, & \zeta \in \mathbb{N}^+ \\ 0, & \zeta = 0 \\ \sum_{t=1}^{-\zeta} z_t^*, & \zeta \in \mathbb{N}^-, \end{cases} \tag{4}$$

Here z_t, z_t^* are independent copies of a normal distribution $\mathcal{N}(-\xi_\infty^2, 4\xi_\infty^2\sigma_\infty^2)$, which are also independent over all t. Here the parameters $\xi_\infty = \lim_{T\to\infty} \xi > 0$ and $\sigma_\infty^2 = \lim_{T\to\infty} \sigma^2$, where both ξ and σ^2 are as defined as follows,

$$\eta^0 = \Big(\theta^0_{(1)} - \theta^0_{(2)}\Big), \quad \xi = \|\eta^0\|_2, \quad \text{and} \quad \sigma^2 = \eta^{0T}\Sigma\eta^0/\xi^2$$

[1]For $x \in \mathbb{R}$, $(x)_+ = x$, if $x \geq 0$, and $x = 0$ if $x < 0$.

Algorithm 1 (KFJS 2021): Estimation of τ^0 with boundary selection (under $N = 1$)

(Initialize): Select a preliminary evenly spaced coarse grid $\mathcal{D} \subset \{1, ..., T\}$ of cardinality $\log T$. Select an initializer $\check{\tau} \in \mathcal{D}$ as the best fitting value to the data $\left\{x_t\right\}_{t=1}^{T}$.

Step 1: Obtain estimates $\check{\theta}_{(j)} = \hat{\theta}_{(j)}(\check{\tau})$, $j = 1, 2$, and update change point estimates as

$$\hat{\tau} = \underset{\tau \in \{1,...,(T-1)\}}{\arg\min} \; Q(\tau, \check{\theta}),$$

and perform an ℓ_0 regularization as

$$\hat{\tau}^* = \begin{cases} T(\text{no change}) & \text{if } \{Q(T, \check{\theta}) - Q(\hat{\tau}, \check{\theta})\} < \gamma \\ \hat{\tau} & \text{else.} \end{cases}$$

Step 2: If $\hat{\tau}^* = T$ the set $\tilde{\tau} = T$, else if $\hat{\tau}^* > 0$, obtain estimates $\hat{\theta}_{(j)} = \hat{\theta}_{(j)}(\hat{\tau})$, $j = 1, 2$, and refit change point as,

$$\tilde{\tau} = \underset{\tau \in \{1,...,(T-1)\}}{\arg\min} \; Q(\tau, \hat{\theta}),$$

(Output): $\tilde{\tau}$

where, $\Sigma = E(\varepsilon_t \varepsilon_t^T)$ is the underlying covariance structure of Model 1. Finally, normality of the increments z_t in (4) is also a consequence of the normality assumption on the distribution underlying Model 1. Then, we have,

$$(\tilde{\tau} - \tau^0) \Rightarrow \underset{\zeta \in \mathbb{Z}}{\arg\max}\, \mathcal{C}_\infty(\zeta), \tag{5}$$

We utilize (5) to construct asymptotically valid confidence intervals for the change point parameters. Specifically, these are obtained as $[\tilde{\tau} - q_{(1-\alpha/2)}, \tilde{\tau} + q_{(1-\alpha/2)}]$ where $q_{(1-\alpha/2)}$ is the $(1 - \alpha/2)^{th}$ quantile of the considered arg max of a two sided random walk with a negative drift. Since no analytical form of this distribution is available, we obtain these quantiles via a monte-carlo simulation, *i.e.*, simulating the two-sided random walk process and in turn obtaining realizations from the distribution under consideration.

The above results are under a single change point assumption, whereas the model and data under consideration have multiple change points. For this extension, we adopt the fairly standard practice of implementing binary segmentation, *i.e.*, recursively split data into binary partitions until no further change points are observed. This process utilizes Algorithm 1 in each recursive step, however, this algorithm is implemented only upto the ℓ_0 regularization of Step 1 (stated as Algorithm 2). The entire process of estimating multiple change points is then stated as Algorithm 3 (KFJS+BS) below.

As suggested in Kaul *et al.* (2021) a further local refitting is carried out of the change point estimates (analog of Step 2 of Algorithm 1). Specifically, Let $\hat{\tau}$ and $\hat{\mathbb{N}}$ represent the location and number of change point estimates obtained from Algorithm 2 and $\hat{\theta}(\hat{\tau})$ represent

Algorithm 2 (KFJS 2021): Estimation of τ^0 with boundary selection (under $N = 1$)

(Initialize): Select a preliminary evenly spaced coarse grid $\mathcal{D} \subset \{1, ..., T\}$ of cardinality $\log T$. Select an initializer $\check{\tau} \in \mathcal{D}$ as the best fitting value to the data $\left\{x_t\right\}_{t=1}^{T}$.

Step 1: Obtain estimates $\check{\theta}_{(j)} = \hat{\theta}_{(j)}(\check{\tau})$, $j = 1, 2$, and update change point estimates as

$$\hat{\tau} = \underset{\tau \in \{1,...,(T-1)\}}{\arg\min} \; Q(\tau, \check{\theta}),$$

and perform an ℓ_0 regularization as

$$\hat{\tau}^* = \begin{cases} T(\text{no change}) & \text{if } \{Q(T, \check{\theta}) - Q(\hat{\tau}, \check{\theta})\} < \gamma \\ \hat{\tau} & \text{else.} \end{cases}$$

(Output): $\hat{\tau}^*$

Algorithm 3 (KJFS+BS): Extension of KJFS to multiple changes via binary segmentation

(Initialize): $\hat{\tau}_{\text{st}} = \phi$ collecting all change points to be estimated.
Implement $\hat{\tau}$= Algorithm 2 $\Big(\{1, ..., T\}\Big)$.
If $\hat{\tau} = T$ (no change) then **Stop**
Else $\hat{\tau}_{\text{up}} = (\tau_{\text{st}}, \hat{\tau})$ (updated vector of estimated change points)
While length$(\hat{\tau}_{up})$ > length$(\hat{\tau}_{st})$ **do**
$\hat{\tau}_{st} = \hat{\tau}_{up}$
for $m \in 1 : (\text{length}(\tau_{st}) + 1)$ **do**
partition$_m = \{\tau_{st(m-1)}, ..., \tau_{st(m)}\}$
$\hat{\tau} = \text{Algorithm 2}(\text{partition}_{\text{m}})$
If $\hat{\tau}$ is away from boundary of sampling period of partition **then**
$\hat{\tau}_{up} = (\hat{\tau}_{st}, \hat{\tau})$
(Output): all estimated change points of vector $\hat{\tau}_{up}$ sorted in ascending order.

the mean estimates obtained from the associated partitioning via 3. Further, let,

$$Q_j\Big(\tau_j, \tau_{-j}, \theta\Big) = \sum_{t=\tau_{j-1}+1}^{\tau_j} \|x_t - \theta_{(j)}\|_2^2 + \sum_{t=\tau_j+1}^{\tau_{j+1}} \|x_t - \theta_{(j+1)}\|_2^2. \tag{6}$$

Then define the locally refitted estimates as,

$$\tilde{\tau}_j := \tilde{\tau}_j\Big(\hat{\tau}_{-j}, \hat{\theta}\Big) = \underset{\hat{\tau}_{j-1} < \tau_j < \hat{\tau}_{j+1}}{\arg\min} \; Q_j\Big(\tau_j, \hat{\tau}_{-j}, \hat{\theta}\Big), \quad j = 1, ..., \hat{N} \tag{7}$$

Then confidence intervals for the parameters τ_j^0's are obtained by utilizing the change point estimates $\tilde{\tau}_j$ and by a piecewise application of (5).

4. Results of the high dimensional change point analysis

As described in Section 2, the data consists of two precipitation and five temperature variables collected from stations at 91 cities spread across the contiguous United States over the period 1948-2023. We implement the analysis with all the seven climatic variables in the model ($p = 628$). Here, it should be noted that the value of p is lower than what one would expect. This happened because some of the variables had no variability in their values throughout the sampling period, and hence such variables were removed from the analysis. All computations are carried out in the statistical software *R*.

Before we begin the presentation of results, we would like to bring to the attention of the reader the inadequacy of identifying merely the point estimates of unknown change points. The first step of the high dimensional change point analysis is the implementation of Algorithm 3, mainly for point estimation of change points in the climatic data. For each change point identified in the mean vector, one may only conclude that there is a change in at least one of the component climatic variables.

This leaves still the question of identifying further the actual climatic variables in which changes have occurred in their means. For this purpose we perform component-wise t-tests for a comparison of pre and post means across estimated change points. A further Bonferroni correction is made based on the number of tests performed in order to control the family wise error rate of the procedure. Based on this inferential procedure, we draw conclusions about changes in the climatic variables comprehensively.

Table 1: Estimated change points in years via the implementation of algorithm-3

Climatic variables (#)	**Number of parameters (p)**	**Estimated change points**
Temperature and precipitation (7)	628	1957, 1989, 2010

Table 2: Confidence interval for each of the three change points detected via Algorithm 3 together with estimated jump sizes (ξ) and estimated variances(σ^2_∞)

Estimated change point	**95% confidence interval**	**Estimated jump size (ξ)**	**Estimated variance (σ^2_∞)**
1957	(1955, 1959)	16.02	36.82
1989	(1986, 1992)	14.2	44.59
2010	(2009, 2011)	18.09	43.34

As for presentation of results, we begin with presenting in Table 1 the change points identified by Algorithm 3. The fitted model with all the seven temperature and precipitation variables consisted of three change points estimated in years as 1957, 1989, 2010. Confidence intervals for the true change points along with estimated jump size, and estimated variance are presented in Table 2. As described above, confidence intervals for change point in each of the component climatic variables enabled us to determine whether a change has occurred in that component variable or not. Upon applying this method at each of the three change

points, we were able to determine the number of changes identified at each city and for each climatic variable. Results from this analysis are presented in Tables 3-5, and Tables 6-7. Specifically, Table 3 consists of list of cities that have undergone a change in their mean at the change year 1957 and the listing is made for each of the seven climatic variables PRCP1, PRCP, TMAX32, TMAX90, TAVG, TMAX, TMIN. Table 4 and Table 5 consist of similar listings of cities for change years 1989 and 2010, respectively. Tables 6 and 7 consist of the listing of all 91 cities together with the number of change points in each climatic variable for any given city.

Upon identifying changes, it is important to compute the magnitudes of change in each climatic variable at each city. The magnitudes of change enable us to clearly understand the nature and severity of changes in the climatic variables under consideration. Moving further, based upon the magnitudes of change, we can also identify clusters of cities so that different clusters may identify groups of cities with different magnitudes while maintaining similarity in changes within each cluster. Often such clusters of cities can be associated with a particular region, and such information is extremely important for interpreting changes in climatic studies. Among the plethora of clustering methods, K-means clustering stands out as a widely adopted technique for segmenting datasets into a predefined number of groups, denoted as 'k clusters'. Its primary objective is to categorize objects into clusters, maximizing intra-class similarity while minimizing inter-class dissimilarity. In the K-means approach, each cluster is characterized by a centroid, computed as the mean of points within the cluster. The process begins with specifying the desired number of clusters (k), followed by the random selection of k objects from the dataset to serve as initial centroids. Subsequently, each remaining object is assigned to the nearest centroid based on Euclidean distance, a step known as the 'cluster assignment' step. The algorithm then updates the mean value of each cluster, termed the 'centroid update' step, iteratively repeating these steps until convergence is attained. Convergence indicates stability, signifying that cluster assignments remain unchanged between successive iterations.

In this study, the K-means clustering method was implemented using the 'kmeans' function from the 'cluster' (Maechler *et al.* (2013)) and 'factoextra' (Kassambara and Mundt (2021)) packages in R. The clusters resulting from the K-means cluster analysis for each of the three change points together with a comment on the nature of each cluster are presented in Tables 8-10 The actual magnitudes of change in each cluster for each climatic variable are presented in Table 11.

5. Discussion of results

We shall begin our discussion with Tables 1-2 that identify the change points in years through the application of Algorithm 3 to data on temperature and precipitation variables. The change years for the model with all seven temperature and precipitation variables are 1957, 1989, and 2010. The 95% confidence intervals presented in Table 2 for each of the three true change years are very tight (at most +/-3 years), thus indicating the high precision with which the change years have been estimated. Further, we look at Tables 3-5 lists cities that have undergone a change, respectively, in the years 1957, 1989, and 2010, for each of the seven climatic variables. Focusing on the two precipitation variables PRCP1 and PRCP we notice that changes in PRCP1 occurred at 9 cities in 1957, at 6 cities in 1989, and at only 3 cities in 2010, whereas similar numbers for PRCP are 5, 8, and 1, respectively. Similar city

Table 3: List of cities that have undergone a change in the year 1957 for each of the seven climatic variables, namely, PRCP1, PRCP, TMAX32, TMAX90, TAVG, TMAX, TMIN

Variable	Cities
PRCP1	Baton Rouge, Brownsville, Columbia, Eugene, Milwaukee, New York City, Oklahoma City, Tallahassee, Wichita
PRCP	Albuquerque, Baton Rouge, New York City, Tallahassee, Wichita
TMAX32	Atlanta, Birmingham, Buffalo, Charleston, Charlotte, Chattanooga, Cincinnati, Cleveland, Columbus, Knoxville, Lexington, Louisville, Nashville, New Orleans, Philadelphia, Pittsburgh, Richmond, Saint Louis, Wichita
TMAX90	Atlanta, Austin, Boise, Charlotte, Chattanooga, Cleveland, Columbia, Columbus, Jacksonville, Knoxville, Lexington, Macon, Montgomery, Portland OR, Raleigh, Sacramento, San Francisco, Seattle, Tucson, Wichita
TAVG	Albuquerque, Atlanta, Augusta, Bakersfield, Baton Rouge, Birmingham, Brownsville, Buffalo, Burlington, Charlotte, Chattanooga, Cincinnati, Cleveland, Columbia, Columbus, Detroit, El Paso, Fresno, Greensboro, Jacksonville, Knoxville, Lexington, Little Rock, Los Angeles, Louisville, Macon, Montgomery, Nashville, New Orleans, Philadelphia, Phoenix, Pittsburgh, Portland OR, Raleigh, Reno, Richmond, Sacramento, San Antonio, San Diego, San Francisco, Seattle, Tallahassee, Virginia Beach, Wichita
TMAX	Atlanta, Augusta, Austin, Baton Rouge, Birmingham, Buffalo, Burlington, Charlotte, Chattanooga, Cleveland, Columbia, Columbus, Detroit, Greensboro, Houston, Jacksonville, Knoxville, Lexington, Los Angeles, Louisville, Macon, Miami, Mobile, Montgomery, Nashville, New Orleans, Philadelphia, Pittsburgh, Portland OR, Raleigh, Sacramento, San Antonio, San Francisco, Seattle, Virginia Beach, Wichita
TMIN	Albuquerque, Atlanta, Augusta, Bakersfield, Baton Rouge, Birmingham, Brownsville, Burlington, Charlotte, Chattanooga, Cincinnati, Cleveland, Columbia, Dayton, Detroit, El Paso, Eugene, Fresno, Jacksonville, Knoxville, Las Vegas, Little Rock, Los Angeles, Macon, Madison, Miami, Nashville, New Orleans, Philadelphia, Phoenix, Pittsburgh, Portland, Raleigh, Reno, Richmond, Sacramento, San Diego, San Francisco, Seattle, Tallahassee, Virginia Beach, Wichita

Table 4: List of cities that have undergone a change in the year 1989 for each of the seven climatic variables, namely, PRCP1, PRCP, TMAX32, TMAX90, TAVG, TMAX, TMIN

Variable	Cities
PRCP1	Albany, Dayton, Fort Wayne, Knoxville, Macon, Tallahassee
PRCP	Albany, Concord, Dayton, Fargo, Fort Wayne, Madison, Rochester, Tallahassee
TMAX32	Albany, Albuquerque, Allentown, Amarillo, Atlanta, Augusta, Birmingham, Burlington, Charleston, Charlotte, Chattanooga, Cleveland, Colorado Springs, Columbia, Columbus, Dallas, Greensboro, Harrisburg, Indianapolis, Little Rock, Louisville, Macon, Memphis, Milwaukee, Mobile, Nashville, New York City, New Orleans, Oklahoma, Philadelphia, Pittsburgh, Providence, Raleigh, Richmond, Saint Louis, Tulsa, Virginia Beach, Wichita
TMAX90	Austin, Boise, Brownsville, Fargo, Miami, New Orleans, Raleigh, San Diego, Sioux Falls, Tallahassee, Tucson
TAVG	Albany, Albuquerque, Allentown, Amarillo, Atlanta, Augusta, Austin, Baton Rouge, Birmingham, Boise, Boston, Brownsville, Buffalo, Burlington, Charleston, Charlotte, Chattanooga, Cheyenne, Chicago, Cincinnati, Cleveland, Columbia, Columbus, Concord, Dallas, Denver, Des Moines, Detroit, El Paso, Fargo, Fort Wayne, Fresno, Greensboro, Harrisburg, Hartford, Houston, Indianapolis, Knoxville, Las Vegas, Lexington, Little Rock, Louisville, Madison, Memphis, Miami, Milwaukee, Minneapolis, Montgomery, Nashville, New York City, New Orleans, Oklahoma City, Orlando, Philadelphia, Phoenix, Pittsburgh, Portland OR, Portland ME, Providence, Raleigh, Reno, Richmond, Saint Louis, Salt Lake, San Antonio, Springfield, Tallahassee, Tampa, Tucson, Tulsa, Virginia Beach, Washington DC, Wichita
TMAX	Allentown, Amarillo, Atlanta, Augusta, Austin, Baton Rouge, Birmingham, Boise, Brownsville, Buffalo, Burlington, Charleston, Charlotte, Chattanooga, Cleveland, Columbus, Concord, Dallas, Denver, Detroit, El Paso, Fort Wayne, Greensboro, Harrisburg, Indianapolis, Little Rock, Louisville, Memphis, Miami, Milwaukee, Montgomery, New Orleans, Oklahoma, Orlando, Philadelphia, Phoenix, Pittsburgh, Portland ME, Providence, Raleigh, Saint Louis, San Antonio, San Diego, San Francisco, Tallahassee, Tucson, Virginia Beach
TMIN	Albany, Albuquerque, Atlanta, Austin, Billings, Birmingham, Boise, Boston, Brownsville, Buffalo, Burlington, Charleston, Charlotte, Chattanooga, Cheyenne, Chicago, Cincinnati, Cleveland, Columbia, Columbus, Concord, Dallas, Dayton, Des Moines, Detroit, El Paso, Fargo, Fort Wayne, Fresno, Greensboro, Harrisburg, Hartford, Houston, Indianapolis, Knoxville, Las Vegas, Lexington, Little Rock, Louisville, Madison, Memphis, Miami, Milwaukee, Minneapolis, Nashville, New York City, New Orleans, Oklahoma, Omaha, Philadelphia, Phoenix, Pittsburgh, Portland OR, Portland ME, Providence, Raleigh, Reno, Richmond, Rochester, Saint Louis, Salt Lake, San Antonio, Seattle, Sioux Falls, Springfield, Tallahassee, Tampa, Tucson, Tulsa, Virginia Beach, Washington DC, Wichita

Table 5: List of cities that have undergone a change in the year 2010 for each of the seven climatic variables, namely, PRCP1, PRCP, TMAX32, TMAX90, TAVG, TMAX, TMIN

Variable	Cities
PRCP1	Baton Rouge, Cincinnati, Raleigh
PRCP	Eugene
TMAX32	None
TMAX90	Albuquerque, Amarillo, Austin, Bakersfield, Boise, Brownsville, Colorado Springs, Dallas, Denver, Des Moines, El Paso, Eugene, Houston, Miami, Nashville, Orlando, Reno, Saint Louis, Seattle, Wichita
TAVG	Albany, Albuquerque, Allentown, Amarillo, Atlanta, Augusta, Austin, Bakersfield, Baton Rouge, Birmingham, Boise, Boston, Brownsville, Burlington, Charleston, Charlotte, Chattanooga, Cincinnati, Cleveland, Colorado Springs, Columbia, Concord, Dallas, Dayton, El Paso, Fresno, Greensboro, Harrisburg, Hartford, Houston, Jacksonville, Knoxville, Las Vegas, Lexington, Louisville, Miami, Montgomery, Nashville, New York City, New Orleans, Omaha, Orlando, Philadelphia, Phoenix, Portland ME, Providence, Raleigh, Reno, Richmond, Rochester, Sacramento, Salt Lake, San Antonio, San Diego, Seattle, Spokane, Tallahassee, Tampa, Tucson, Virginia Beach, Washington DC, Wichita
TMAX	Albany, Albuquerque, Allentown, Amarillo, Atlanta, Augusta, Austin, Bakersfield, Boston, Brownsville, Burlington, Charleston, Charlotte, Cheyenne, Cleveland, Colorado Springs, Columbia, Dallas, Dayton, El Paso, Eugene, Fresno, Hartford, Houston, Jacksonville, Las Vegas, Lexington, Louisville, Macon, Miami, Montgomery, Nashville, New Orleans, Orlando, Phoenix, Portland ME, Reno, Rochester, Sacramento, Saint Louis, Salt Lake, San Diego, Seattle, Tallahassee, Tampa, Tucson, Virginia Beach, Washington DC, Wichita
TMIN	Albany, Albuquerque, Allentown, Amarillo, Atlanta, Augusta, Austin, Bakersfield, Baton Rouge, Birmingham, Boise, Boston, Brownsville, Buffalo, Burlington, Charleston, Chattanooga, Cincinnati, Cleveland, Colorado Springs, Columbia, Columbus, Concord, Dallas, Dayton, El Paso, Fresno, Greensboro, Harrisburg, Hartford, Houston, Jacksonville, Knoxville, Las Vegas, Louisville, Miami, Montgomery, Nashville, New York City, New Orleans, Omaha, Orlando, Philadelphia, Phoenix, Pittsburgh, Portland OR, Portland ME, Raleigh, Reno, Richmond, Rochester, Salt Lake, San Antonio, Seattle, Spokane, Tallahassee, Tampa, Tucson, Virginia Beach, Washington DC

count for temperature variables are: TMAX 32 – 19, 38, 0; TMAX90 – 20, 11, 20; TAVG – 44, 79, 63; TMAX – 36, 48, 49; and TMIN – 42, 72, 60, respectively. It is also informative to see the same numbers for each of the three change years. Thus the number of cities in which changes have occurred at each of the change years in climatic variables PRCP1, PRCP, TMAX32, TMAX90, TAVG, TMAX, and TMIN, respectively are: in 1957 – 9, 5, 19, 20, 44, 36, 42; in 1989 – 6, 8, 38, 11, 79, 48, 72; and in 2010 – 3, 1, 0, 20, 63, 49, 60.

Clearly, changes in temperature variables dominate the changes in precipitation variables. Also, changes in continuous variables (PRCP, TAVG, TMAX, TMIN) are significantly higher compared to changes in the three discrete variables (PRCP1, TMAX32, TMAX90). Perhaps this can be anticipated ahead because the information content in continuous variables is much more than that available in discrete variables and hence changes in continuous variables can be detected with higher precision. Among continuous temperature variables, changes in TAVG (44, 79, 63) and TMIN (42, 72, 60) are significantly higher compared to changes in TMAX (36, 48, 49). While the three variables had similar number of changes in

Table 6: List of cities along with corresponding number of change points for data on PRCP1, PRCP, TMAX32, TMAX90, TAVG, TMAX, and TMIN during the years 1948-2023

City	State	PRCP1	PRCP	TMAX32	TMAX90	TAVG	TMAX	TMIN
Albany	New York	1	1	1	0	2	1	2
Albuquerque	New Mexico	0	1	1	1	3	1	3
Allentown	Pennsylvania	0	0	1	0	2	2	1
Amarillo	Texas	0	0	1	1	2	2	1
Atlanta	Georgia	0	0	2	1	3	3	3
Augusta	Georgia	0	0	1	0	3	3	2
Austin	Texas	0	0	0	3	2	3	2
Bakersfield	California	0	0	0	1	2	1	2
Baton Rouge	Louisiana	2	1	0	0	3	2	2
Billings	Montana	0	0	0	0	0	0	1
Birmingham	Alabama	0	0	2	0	3	2	3
Boise	Idaho	0	0	0	3	2	1	2
Boston	Massachusetts	0	0	0	0	2	1	2
Brownsville	Texas	1	0	0	2	3	2	3
Buffalo	New York	0	0	1	0	2	2	2
Burlington	Vermont	0	0	1	0	3	3	3
Charleston	South Carolina	0	0	2	0	2	2	2
Charlotte	North Carolina	0	0	2	1	3	3	2
Chattanooga	Tennessee	0	0	2	1	3	2	3
Cheyenne	Wyoming	0	0	0	0	1	1	1
Chicago	Illinois	0	0	0	0	1	0	1
Cincinnati	Ohio	1	0	1	0	3	0	3
Cleveland	Ohio	0	0	2	1	3	3	3
Color. Spring	Colorado	0	0	1	1	1	1	1
Columbia	South Carolina	1	0	1	1	3	2	3
Columbus	Ohio	0	0	2	1	2	2	2
Concord	New Hampshire	0	1	0	0	2	1	2
Dallas	Texas	0	0	1	1	2	2	2
Dayton	Ohio	1	1	0	0	1	1	3
Denver	Colorado	0	0	0	1	1	1	0
Des Moines	Iowa	0	0	0	1	1	0	1
Detroit	Michigan	0	0	0	0	2	2	2
El Paso	Texas	0	0	0	1	3	2	3
Eugene	Oregon	1	1	0	1	0	1	1
Fargo	North Dakota	0	1	0	1	1	0	1
Fort Wayne	Indiana	1	1	0	0	1	1	1
Fresno	California	0	0	0	0	3	1	3
Greensboro	North Carolina	0	0	2	0	3	2	2
Harrisburg	Pennsylvania	0	0	2	0	2	1	2
Hartford	Connecticut	0	0	0	0	2	1	2
Houston	Texas	0	0	0	1	2	2	2
Indianapolis	Indiana	0	0	1	0	1	1	1
Jacksonville	Florida	0	0	0	1	2	2	2
Kansas City	Kansas	0	0	0	0	0	0	0
Knoxville	Tennessee	1	0	1	1	3	1	3

Table 7: List of cities along with corresponding number of change points for data on PRCP1, PRCP, TMAX32, TMAX90, TAVG, TMAX, and TMIN during the years 1948-2023. (Continued).

City	State	PRCP1	PRCP	TMAX32	TMAX90	TAVG	TMAX	TMIN
Las Vegas	Nevada	0	0	0	0	2	1	3
Lexington	Kentucky	0	0	1	1	3	2	1
Little Rock	Arkansas	0	0	1	0	2	1	2
Los Angeles	California	0	0	0	0	1	1	1
Louisville	Kentucky	0	0	2	0	3	3	2
Macon	Georgia	1	0	1	1	1	2	1
Madison	Wisconsin	0	1	0	0	1	0	2
Memphis	Tennessee	0	0	1	0	1	1	1
Miami	Florida	0	0	0	2	2	3	3
Milwaukee	Wisconsin	1	0	1	0	1	1	1
Minneapolis	Minnesota	0	0	0	0	1	0	1
Mobile	Alabama	0	0	1	0	0	1	0
Montgomery	Alabama	0	0	0	1	3	3	1
Nashville	Tennessee	0	0	2	1	3	2	3
New York City	New York	1	1	1	0	2	0	2
New Orleans	Louisiana	0	0	2	2	3	3	3
Oklahoma City	Oklahoma	1	0	1	0	1	1	1
Omaha	Nebraska	0	0	0	0	1	0	2
Orlando	Florida	0	0	0	1	2	2	1
Philadelphia	Pennsylvania	0	0	2	0	3	2	3
Phoenix	Arizona	0	0	0	0	3	2	3
Pittsburgh	Pennsylvania	0	0	2	0	2	2	3
Portland	Oregon	0	0	0	1	2	1	3
Portland	Maine	0	0	0	0	2	2	2
Providence	Rhode Island	0	0	1	0	2	1	1
Raleigh	North Carolina	1	0	1	2	3	2	3
Reno	Nevada	0	0	0	1	3	1	3
Richmond	Virginia	0	0	2	0	3	0	3
Rochester	New York	0	1	0	0	1	1	2
Sacramento	California	0	0	0	1	2	2	1
Saint Louis	Missouri	0	0	2	1	1	2	1
Salt Lake City	Utah	0	0	0	0	2	1	2
San Antonio	Texas	0	0	0	0	3	2	2
San Diego	California	0	0	0	1	2	2	1
San Francisco	California	0	0	0	1	1	2	1
Seattle	Washington	0	0	0	2	2	2	3
Sioux Falls	South Dakota	0	0	0	1	0	0	1
Spokane	Washington	0	0	0	0	1	0	1
Springfield	Missouri	0	0	0	0	1	0	1
Tallahassee	Florida	2	2	0	1	3	2	3
Tampa	Florida	0	0	0	0	2	1	2
Tucson	Arizona	0	0	0	2	2	2	2
Tulsa	Oklahoma	0	0	1	0	1	0	1
Virginia Beach	Virginia	0	0	1	0	3	3	3
Washington	DC	0	0	0	0	2	1	2
Wichita	Kansas	1	1	2	2	3	2	2

Table 8: Clusters based on KMEANS clustering algorithm implemented upon actual differences in averages of the climatic variables before and after the change point in the year 1957.

Cluster	Cities	Cluster characteristic
1	Albany, Allentown, Amarillo, Billings, Boise, Boston, Charleston, Cheyenne, Chicago, Colorado Springs, Concord, Dallas, Dayton, Denver, Des Moines, Eugene, Fargo, Fort Wayne, Hartford, Houston, Indianapolis, Kansas City, Las Vegas, Madison, Memphis, Miami, Milwaukee, Minneapolis, Mobile, Oklahoma City, Omaha, Orlando, Portland ME, Providence, Rochester, Salt Lake City, Sioux Falls, Spokane, Springfield, Tampa, Tucson, Tulsa, Washington DC	No big changes in climatic variables
2	Bakersfield, Fresno, Los Angeles, Phoenix, Portland OR, Reno, Sacramento, San Diego, San Francisco, Seattle	High increases in TMAX90, TAVG, TMAX, and TMIN
3	Atlanta, Austin, Charlotte, Chattanooga, Columbia, Jacksonville, Knoxville, Lexington, Macon, Montgomery, Raleigh	Decrease in TMAX90, TAVG, TMAX, TMIN and increase in TMAX32, PRCP1
4	Albuquerque, Augusta, Birmingham, Brownsville, Burlington, Detroit, El Paso, Greensboro, Little Rock, Nashville, New Orleans, Richmond, San Antonio, Virginia Beach	Small increases in TMAX32, PRCP and small decreases in TAVG, TMAX, TMIN
5	Baton Rouge, New York City, Tallahassee, Wichita	High increases in PRCP1 and PRCP
6	Buffalo, Cincinnati, Cleveland, Columbus, Harrisburg, Louisville, Philadelphia, Pittsburgh, Saint Louis	Big increases in TMAX32, and Cecreases in TMAX90, TAVG, TMAX, TMIN

Table 9: Clusters based on KMEANS clustering algorithm implemented upon actual differences in averages of the climatic variables before and after the change point in the year 1989.

Cluster	Cities	Cluster characteristic
1	Tallahassee	High decrease in PRCP1 and PRCP
2	Albuquerque, Atlanta, Birmingham, Boise, Buffalo, Charleston, Charlotte, Chicago, Columbia, Dallas, Detroit, El Paso, Fresno, Greensboro, Hartford, Houston, Las Vegas, Little Rock, Memphis, Minneapolis, Nashville, Oklahoma City, Phoenix, Portland OR, Portland ME, Reno, Richmond, Salt Lake City, San Antonio, Tampa, Virginia Beach, Wichita	High increase in TMIN
3	Albany, Concord, Dayton, Fargo, Fort Wayne, Knoxville, Madison	High increases in PRCP1 and PRCP
4	Amarillo, Augusta, Bakersfield, Baton Rouge, Billings, Boston, Cheyenne, Cincinnati, Colorado Springs, Denver, Des Moines, Eugene, Jacksonville, Kansas City, Lexington, Los Angeles, Macon, Mobile, Montgomery, New York City, Omaha, Orlando, Rochester, Sacramento, San Diego, San Francisco, Seattle, Sioux Falls, Spokane, Springfield, Tulsa, Washington DC	No big changes in any of the variables
5	Allentown, Burlington, Chattanooga, Cleveland, Columbus, Harrisburg, Indianapolis, Louisville, Milwaukee, Philadelphia, Pittsburgh, Providence, Saint Louis	Decrease in TMAX32, and an increase in TAVG and TMIN
6	Austin, Brownsville, Miami, New Orleans, Raleigh, Tucson	Increase in TMAX90, TAVG, TMAX, TMIN

Table 10: Clusters based on KMEANS clustering algorithm implemented upon actual differences in averages of the climatic variables before and after the change point in the year 2010.

Cluster	Cities	Cluster characteristic
1	Albany, Allentown, Atlanta, Augusta, Boston, Burlington, Charleston, Charlotte, Cleveland, Columbia, Dayton, Fresno, Hartford, Jacksonville, Las Vegas, Lexington, Louisville, Montgomery, New Orleans, Phoenix, Portland ME, Rochester, Sacramento, Salt Lake City, San Diego, Seattle, Tallahassee, Tampa, Tucson, Virginia Beach, Washington DC	Increase in TAVG, TMAX, TMIN
2	Billings, Buffalo, Cheyenne, Chicago, Columbus, Denver, Des Moines, Detroit, Fargo, Fort Wayne, Indianapolis, Kansas City, Little Rock, Los Angeles, Macon, Madison, Memphis, Milwaukee, Minneapolis, Mobile, Oklahoma City, Pittsburgh, Portland OR, Providence, San Francisco, Sioux Falls, Springfield, Tulsa	No significant changes
3	Albuquerque, Amarillo, Austin, Bakersfield, Brownsville, Colorado Springs, Dallas, El Paso, Houston, Miami, Nashville, Orlando, Reno, Saint Louis, Wichita	High increases in TMAX90 and TMAX
4	Birmingham, Boise, Chattanooga, Concord, Greensboro, Harrisburg, Knoxville, New York City, Omaha, Philadelphia, Richmond, San Antonio, Spokane	Increases in TMAX90, TAVG, TMIN
5	Baton Rouge, Cincinnati, Raleigh	Increase in PRCP1
6	Eugene	High increase in TMAX90 and high decrease in PRCP

Table 11: Magnitudes of change for clusters in each change year and for each of the climatic variables representing temperature and precipitation.

Change Year	Cluster	PRCP1	PRCP	TMAX32	TMAX90	TAVG	TMAX	TMIN
		(days)	(mm)	(days)	(days)	°C	°C	°C
1957	1	0.210	0.000	0.011	-0.135	0.000	-0.034	0.013
	2	0	0	0	2.323	0.849	0.403	1.132
	3	0.406	0	1.758	-18.384	-0.749	-0.954	-0.568
	4	0.18	4.653	0.991	0	-0.669	-0.487	-0.692
	5	4.639	254.368	1.366	-5.02	-0.556	-0.459	-0.613
	6	0	0	11.221	-3.03	-0.646	-0.676	-0.402
1989	1	-4.403	-198.544	0.000	14.092	0.597	0.652	0.542
	2	0	0	-1.319	0.226	0.839	0.324	1.177
	3	1.579	111.116	-1.084	-0.654	0.622	0.173	0.912
	4	0.066	2.569	-0.705	-0.398	0.219	0.119	0.266
	5	0	0	-8.491	0	0.967	0.799	1.117
	6	0	0	-0.438	20.025	0.894	0.866	0.921
2010	1	0	0	0	0.108	0.87	0.817	0.86
	2	0	0	0	0.879	0.019	0.054	0.095
	3	0	0	0	16.226	0.824	0.954	0.735
	4	0	0	0	0.656	0.728	0	0.861
	5	3.117	0	0	0	0.716	0	0.82
	6	0	-226.546	0	8.399	0	0.728	0

Table 12: Summary of observations made about changes in climatic variables that occurred in 1957.

Climatic Variable	Cluster	Region	Increase/Decrease
PRCP1	3	Southeastern (3)	Increase (3)
	5	Eastern half (5)	High increase (5)
PRCP	4	South-southeastern (4)	Small increase (4)
	5	Eastern half (5)	High increase (5)
TMAX32	3	Southeastern (3)	Increase (3)
	4	South-southeastern (4)	Small increase (4)
	6	Northeastern (6)	Big increase
TMAX90	2	West coast (2)	High increase (2)
	3	Southeastern (3)	Decrease (3)
	6	Northeastern (6)	Decrease (6)
TAVG	2	West coast (2)	High increase (2)
	3	Southeastern (3)	Decrease (3)
	4	South-southeastern (4)	small decrease (4)
	6	Northeastern (6)	Decrease (6)
TMAX	2	West coast (2)	High increase (2)
	3	Southeastern (3)	Decrease (3)
	4	South-southeastern (4)	Small decrease (4)
	6	Northeastern (6)	Decrease (6)
TMIN	2	West coast (2)	High increase (2)
	3	Southeastern (3)	Decrease (3)
	6	Northeastern (6)	Decrease (6)

Table 13: Summary of observations made about changes in climatic variables that occurred in 1989.

Climatic Variable	Cluster	Region	Increase/Decrease
PRCP1	1	Tallahassee (1)	High decrease (1)
	3	Eastern (3)	High increase (3)
PRCP	1	Tallahassee (1)	High decrease (1)
	3	Eastern (3)	High increase (3)
TMAX32	5	Northeastern (5)	Decrease (5)
TMAX90	6	Southern (6)	Increase (6)
TAVG	5	Northeastern (5)	Increase (5)
	6	Southern (6)	Increase (6)
TMAX	6	Southern (6)	Increase (6)
TMIN	2	Throughout (2)	High increase (2)
	5	Northeastern (5)	Increase (5)
	6	Southern (6)	Increase (6)

Table 14: Summary of observations made about changes in climatic variables that occurred in 2010.

Climatic Variable	Cluster	Region	Increase/Decrease
PRCP1	5	Baton Rouge, Cincinnati, Raleigh (5)	Increase (5)
PRCP	6	Eugene (6)	High decrease (6)
TMAX32	–	——	——
TMAX90	3 4 6	Central (3) Eastern (4) Eugene (6)	High increase (3) Increase (4) High increase (6)
TAVG	1 4	Eastern or Western (1) Eastern (4)	Increase (1) Increase (4)
TMAX	1 3	Eastern or Western (1) Central (3)	Increase (1) High increase (3)
TMIN	1 4	Eastern or Western (1) Eastern (4)	Increase (1) Increase (4)

Table 15: Region wise representation of changes in temperature and precipitation variables

Region	Temperature variables	Year of change	Increase/decrease
Northeastern	TMAX32	1957	High increase
	TMAX, TAVG, TMAX, TMIN, TMAX32	1957, 1989	Decrease
	TAVG, TMIN, TAVG	1989, 2010	Increase
Eastern	PRCP1, PRCP	1957, 1989	High increase
	PRCP	1957	Small increase
	TMAX90, TAVG, TMAX, TMIN	2010	Increase
Southeastern	PRCP1, PRCP, TMAX32	1957	Increase
	TMAX90, TAVG, TMAX, TMIN	1957	Decrease
Southern	PRCP	1957	Small increase
	TAVG, TMAX	1957	Small decrease
	TMAX90, TAVG, TMAX, TMIN	1989	Increase
Central	TMAX90	2010	High increase
	TMAX	2010	Increase
West Coast	TMAX90, TAVG, TMAX, TMIN	1957	High increase
	TAVG, TMAX, TMIN	2010	Increase
Throughout	TMIN	1989	High increase

1957, TAVG and TMIN had much higher number of cities that changed in 1989 and 2010 compared to number of cities that TMAX has changed in the same two change years. The same can be observed from the number of change points in each of the climatic variables at each of the 91 cities. This phenomenon should be understood with a deeper understanding of how higher extreme temperatures change compared to average and lower extreme temperature changes.

We shall now discuss results from cluster analysis based on magnitudes of change presented in Tables 8-10 and Table 11. There are six clusters in the change year 1957 (Table 8), and the magnitudes of change for these six clusters are presented in Table 11. Clearly, there are identifiable differences between the clusters. Cities in Cluster 2 (Bakersfield, Fresno, Los Angeles, Phoenix, Portland OR, Reno, Sacramento, San Diego, San Francisco, Seattle; darker orange) belonging to the west coastal region of the US have shown high increases in TMAX90 (2.323 days), TAVG (0.849°C), TMAX (0.403°C), and TMIN (1.132°C). All cities in Cluster 3 (Atlanta, Austin, Charlotte, Chattanooga, Columbia, Jacksonville, Knoxville, Lexington, Macon, Montgomery, Raleigh) belong to the southeastern region, and these cities have shown increased average change in PRCP1 (0.406 mm), TMAX32 (1.758 days), and significantly decreased changes in TMAX90 (-18.384 days), TAVG (0.749°C), TMAX (0.954°C), and TMIN (0.568°C). Cities in Cluster 6 (Buffalo, Cincinnati, Cleveland, Columbus, Harrisburg, Louisville, Philadelphia, Pittsburgh, Saint Louis) are all in the northeastern region, and cities in this cluster have very high average increase in TMAX32 (11.221 days) and decreases in TAVG (0.646°C), TMAX (0.676°C), and TMIN (0.402°C). Cities in Cluster 4 (Albuquerque, Augusta, Birmingham, Brownsville, Burlington, Detroit, El Paso, Greensboro, Little Rock, Nashville, New Orleans, Richmond, San Antonio, Virginia Beach) are mostly seen in south-southeastern parts of the US and these cities have experienced small increases in TMAX32 (0.991 days), PRCP (4.653 mm), and small decreases in TAVG (0.669°C), TMAX (0.487°C), and TMIN (0.692°C). Cluster 5 (Baton Rouge, New York City, Tallahassee, Wichita) has only four cities in it and these cities are located only in the eastern half of the US map and these cluster of cities may be characterized to show high increases in PRCP1 (4.639 days) and PRCP (254.368 mm). Finally, Cluster 1 (rest of the cities), which as most number of cities these cities have no significant changes, and are all spread evenly throughout the US. Overall, it is clear from Table 8 that there have been more decreasing tends in the temperature variables, and thus the change year 1957 can be viewed as indicative of the beginning of a cooling period. It is also worth noting the very large increase of 254.368 mm of precipitation in PRCP at cluster 5 cities.

Moving on to change year 1989, there are again six clusters in this change year as well (Table 9, Table 11). Among these, cities in Cluster 3 (Albany, Concord, Dayton, Fargo, Fort Wayne, Knoxville, Madison) are spread in the eastern part of the US, cities in Cluster 5 (Allentown, Burlington, Chattanooga, Cleveland, Columbus, Harrisburg, Indianapolis, Louisville, Milwaukee, Philadelphia, Pittsburgh, Providence, Saint Louis) are all clustered in northeastern part of the US, and cities in Cluster 6 (Austin, Brownsville, Miami, New Orleans, Raleigh, Tucson) are all lined up in the southern part of the US. Among the remaining two clusters, Cluster 1 has only one city (Tallahassee) with high decrease in PRCP1(-4.403 days) and PRCP (-198.544 mm), and cluster 2 consisting of large number of cities can be characterized as having large increase in TMIN (1.177°C). Cluster 4 has the largest number of cities and the cities in this cluster show no significant change in their averages. Cities in Cluster 3 have high increases in PRCP1 (1.579 days) and PRCP

(111.116 days); Custer 5 cities show a decrease in TMAX32 (-8.491°C), and an increase in TAVG (0.967) and TMIN (1.117), and cluster 6 cities showed increase in TMAX90 (20.025 days), TAVG (0.894°C), TMAX (0.866°C), and TMIN (0.921°C). Overall, Table 11 makes it clear that the magnitudes of change in this cluster are mostly positive, particularly for temperature variables and thus the change year 1989 can be seen as ending the cooling period that began in 1957 and that there is a transition into the beginning of warmer periods.

Among clusters in change year 2010 (Table 10), Eugene , OR identifies itself as Cluster 6. This city on the west coast can be identified with large drop in PRCP (-226.546 mm) and a large increase in TMAX90 (8.399 days). A large drop in average precipitation together with a large increase in the number of extremely hot days implies that Eugene might have begun undergoing impactful climatic change in 1989, moving towards drought like conditions. Next Cluster 5 (Baton Rouge, Cincinnati, Raleigh) stands out as a cluster with strong increase in PRCP1 (3.117 days). Cities in Cluster 3 (Albuquerque, Amarillo, Austin, Bakersfield, Brownsville, Colorado Springs, Dallas, El Paso, Houston, Miami, Nashville, Orlando, Reno, Saint Louis, Wichita), located mostly in the central region of the US have undergone large increases in TMAX90 (16.226 days) and TMAX (0.954°C), essentially showing increases in extremely hot conditions, both in duration and intensity. Cluster 4 (Birmingham, Boise, Chattanooga, Concord, Greensboro, Harrisburg, Knoxville, New York City, Omaha, Philadelphia, Richmond, San Antonio, Spokane) with cities located mostly on the eastern region began undergoing moderately large increases in temperature variables TMAX90, TAV, and TMIN. Cluster 1 (Albany, Allentown, Atlanta, Augusta, Boston, Burlington, Charleston, Charlotte, Cleveland, Columbia, Dayton, Fresno, Hartford, Jacksonville, Las Vegas, Lexington, Louisville, Montgomery, New Orleans, Phoenix, Portland ME, Rochester, Sacramento, Salt Lake City, San Diego, Seattle, Tallahassee, Tampa, Tucson, Virginia Beach, Washington DC) with cities located mostly in either east coast or west coast has also undergone increasing trends in TAVG, TMAX and TMIN variables. Cluster 2 with large number of cities located throughout US showed no significant changes in any of the variables. With the exception of Eugene that showed large drop in PRCP, a striking feature of this change year is that there are no negative changes in any of the averages across all clusters and all variables. The increases are all in temperature variables only, and thus, the change year 2010 can be seen as a shift towards even warmer conditions that began in 1989.

The discussion of results will be enriched much more through a proper compilation of various observations made about changes that occurred in the years 1957, 1989, and 2010. We have done such a compilation of observations for each of 1957, 1989, and 2010, and these compilations are presented in Tables 12-14, respectively. There is much to learn from a proper understanding of the information contained in each of these tables. We begin with a careful look at Table 12 where observations are summarized about changes that occurred in 1957. The Precipitation variables PRCP1 and PRCP had moderate increases in the southeastern region, and high increases in the eastern half of the US. As for temperature variables, there is much similarity in the changes that occurred in TMAX90, TAVG, TMAX and TMIN variables. All of these four temperature variables have undergone high increase in the west coast and a decrease in southeastern as well as northeastern parts of the US. Only TMAX32 variable has undergone an increase in southeastern and northeastern regions. The summary from Table 13 for the change year 1989 reveals that there was a high decrease in the two precipitation variables PRCP1 and PRCP at Tallahassee, and high increase in PRCP1 and

PRCP in the eastern region. Among temperature variables even though TMAX32 decreased in the northeastern region, other variables TMAX90, TAVG, TMAX and TMIN have all increased in the southern region, and the temperature has increased in the northeastern for TAVG and TMIN also. As for changes in the year 2010, Table 14 shows the precipitation variable PRCP1 increased at Baton Rouge, Cincinnati, and Raleigh whereas there was sharp decrease PRCP at Eugene. Among temperature variables no changes were observed in TMAX32. Similar changes occurred in the three variables TAVG, TMAX, and TMIN with increasing temperatures seen in western and eastern regions. The TMAX90 temperature variable has undergone high increases in central region while significant increases occurred in the eastern region.

Region wise representation of changes presented in Table 15 also allow us to further understand the nature of the changes that occurred in both temperature and precipitation variables. Changes in Precipitation variables, PRCP1 and PRCP, occurred in eastern, southeastern, and southern areas of the US. All the changes in both the precipitation variables occurred in either 1957 or 1989, and moreover, all changes have led to varying levels (small to high) of increases only. In particular, increases in PRCP occurred in all three regions, whereas increases in PRCP1 occurred only in eastern and southeastern regions, that too in 1957. As for temperature changes, there were both decreases and increases in the temperature variables. The decreases were limited to northeastern, southeastern, and southern regions and the decreases in temperature variables occurred mostly in 1957 only. All changes that occurred in temperature variables in 2010 have been increases only, and these increases have occurred in northeastern, eastern and central regions. The year 1989 saw decreases in the northeastern region and otherwise increases in southern region while high increases occurred throughout in TMIN only.

Finally, we have computed overall magnitudes of change for each climatic variable over the 75-year long sampling period 1948-2023. The computed overall average changes are: PRCP1: 0.193 days; PRCP: 5.559 mm; TMAX32: -0.166 days; TMAX90: 0.660 days; TAVG: 0.333°C; TMAX: 0.186°C; TMIN: 0.429°C. Clearly, at an overall level, there were no significant changes in the averages of the two precipitation variables PRCP1 and PRCP as well as the two discrete temperature variables TMAX32 and TMAX90. The overall changes in TAVG, TMAX and TMIN are of much interest. These changes observed over the 75-year period can be better compared with previous works in the literature if we convert these average changes into °C/100 years. Upon doing so we find the changes in averages as – TAVG: 0.444°C /100 yr; TMAX: 0.248°C /100 yr; TMIN: 0.572°C /100 yr.

The above average changes per century are highly influenced by the cooling period that began in 1957 and continued till 1989. Hence, in order to understand more recent trends in temperature changes, it is better to compute the overall magnitudes of changes in TAVG, TMAX and TMIN for the period 1990-2023, a 33 year period. We found these 33-year period changes in averages as – TAVG: 0.595°C; TMAX: 0.404°C; TMIN: 0.699°C. Assuming present temperature trends would continue till the end of the century, the same changes when projected as °C/100 yr are – TAVG: 1.803°C /100 yr; TMAX: 1.224°C /100 yr; TMIN: 2.118°C /100 yr. Of course, the assumption that current temperature trends would continue till the end of the century can be seen to be unrealistic and in this sense the above °C/100 yr increases should be viewed as being conservative.

Comparing the changes in temperature variables with existing literature, even if global in scope, Hawkins and Jones (2013) remarked that more recent analyses support average temperature increases at the rate of 0.500°C/100 yr, first projected by Callendar (1938). In comparison, our current study projects change in average temperature for the US as 0.444°C /100 yr. Based upon a change point modeling, Lee *et al.* (2014) concluded that monthly maximum had a mean change of 0.47°C/Century while the mean change for the monthly minimum was 1.65°C/Century. Results for our monthly maximum TMAX and monthly minimum TMIN showed increases in both extremes. For the whole data period 1948-2023, the increase in TMAX is 0.248°C /100 yr and the same for TMIN is 0.572°C /100 yr. However, if we consider increases for the data period 1990-2023, then the increases in the two extremes are much higher with the increase in TMAX at 1.224°C /100 yr and the increase in TMIN at 2.118°C /100 yr.

6. Concluding remarks

In this study, we have applied recently developed method of high dimensional change point analysis for identifying changes in temperature and precipitation variables based upon data from 91 stations from contiguous United States for the period 1948-2023. A total of seven climatic variables have been considered for studying changes and among these, one precipitation variable and four temperature variables represent extremes. The analysis has identified changes occurring in the years 1957, 1989, and 2010. The magnitudes of changes in the variables and relevant areas where changes have occurred has all been discussed in sufficient detail in the previous section. Here, we shall focus briefly on reasons behind the changes identified by the methodology. First, it is important to note that the change point methodology applied in this study only enables to identify changes but doesn't dwell into reasons behind any of the changes identified by the method. Thus, we need to collect such information from published literature. Changes in climatic variables can occur due to anthropogenic factors or due to various natural phenomena including volcanic eruptions, solar radiation fluctuations, ocean fluctuations such as Pacific Decadal Oscillations (PDO) *etc.* Abrupt changes in climatic variables can also occur due to undocumented causes such as changes in measuring instrumentations that do not get recorded, unrecorded shifts in station locations, *etc.* Anthropogenic causes are those human activities such as industrialization pollution, deforestation, urbanization, *etc.*, that lead to emitting harmful greenhouse gases into the atmosphere.

Wild *et al.* (2005) discuss about evidence of solar dimming caused by air pollution between the period 1958-1985 and the reversal of solar dimming to solar brightening subsequent to 1985. It is possible that the solar dimming between 1958-1985 may have induced the temperature declines that our analysis has identified between the years 1959-1989, a time period that closely matches with solar dimming period. Since solar dimming is a global phenomenon, it is possible that the temperature declines during the identified period may not be limited to the United States alone. Also, the solar dimming apparently does not impact uniformly throughout the United States since the temperature declines have been noticed predominantly in the northeastern, southeastern and southern regions of the US. Conversely, the solar brightening that began after 1985 might explain the observed increases uniformly in all temperature variables subsequent to the year 1989. Greater increases in temperature variables observed since 2010 require further investigation.

Acknowledgements:

The authors thank the reviewer(s) and the Editors for constructive comments and suggestions that have led to improved discussions in the paper.

References

Armal, S., Devineni, N., and Khanbilvardi, R. (2018). Trends in extreme rainfall frequency in the contiguous united states: Attribution to climate change and climate variability modes. *Journal of Climate*, **31**, 369–385.

Bai, J. (1994). Least squares estimation of a shift in linear processes. *Journal of Time Series Analysis*, **15**, 453–472.

Beaulieu, C., Chen, J., and Sarmiento, J. L. (2012). Change-point analysis as a tool to detect abrupt climate variations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **370**, 1228–1249.

Bhattacharjee, M., Banerjee, M., and Michailidis, G. (2017). Common change point estimation in panel data from the least squares and maximum likelihood viewpoints. *arXiv preprint arXiv:1708.05836*.

Bhattacharjee, M., Banerjee, M., and Michailidis, G. (2020). Change point estimation in a dynamic stochastic block model. *The Journal of Machine Learning Research*, **21**, 4330–4388.

Bieniek, P. A., Walsh, J. E., Thoman, R. L., and Bhatt, U. S. (2014). Using climate divisions to analyze variations and trends in alaska temperature and precipitation. *Journal of Climate*, **27**, 2800–2818.

Cai, H. and Wang, T. (2023). Estimation of high-dimensional change-points under a group sparsity structure. *Electronic Journal of Statistics*, **17**, 858–894.

Callendar, G. S. (1938). The artificial production of carbon dioxide and its influence on temperature. *Quarterly Journal of the Royal Meteorological Society*, **64**, 223–240.

Chen, Y., Wang, T., and Samworth, R. J. (2023). Inference in high-dimensional online changepoint detection. *Journal of the American Statistical Association*, **119**, 1461–1472.

Cho, H. (2016). Change-point detection in panel data via double cusum statistic. *Electronic Journal of Statistics*, **10**, 2000–2038.

Cho, H. and Fryzlewicz, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **77**, 475–507.

Cho, H. and Kirch, C. (2022). Bootstrap confidence intervals for multiple change points based on moving sum procedures. *Computational Statistics & Data Analysis*, **175**, 107552.

Csorgo, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. John Wiley & Sons Chichester.

Eichinger, B. and Kirch, C. (2018). A MOSUM procedure for the estimation of multiple random change points. *Bernoulli*, **24**, 526 – 564.

Enikeeva, F. and Harchaoui, Z. (2019). High-dimensional change-point detection under sparse alternatives. *The Annals of Statistics*, **47**, 2051 – 2079.

Fotopoulos, S. B., Jandhyala, V. K., and Khapalova, E. (2010). Exact asymptotic distribution of change-point mle for change in the mean of gaussian sequences. *The Annals of Applied Statistics*, **4**, 1081–1104.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, **42**, 2243–2281.

Gaffen, D. J. and Ross, R. J. (1998). Increased summertime heat stress in the us. *Nature*, **396**, 529–530.

Griffiths, M. L. and Bradley, R. S. (2007). Variations of twentieth-century temperature and precipitation extreme indicators in the northeast united states. *Journal of Climate*, **20**, 5401–5417.

Grundstein, A. and Dowd, J. (2011). Trends in extreme apparent temperatures over the united states, 1949–2010. *Journal of Applied Meteorology and Climatology*, **50**, 1650–1653.

Hawkins, E. and Jones, P. D. (2013). On increasing global temperatures: 75 years after callendar. *Quarterly Journal of the Royal Meteorological Society*, **139**, 1961–1963.

Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tsai, T. T. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, **12**, 105–108.

Jandhyala, V., Fotopoulos, S., MacNeill, I., and Liu, P. (2013). Inference for single and multiple change-points in time series. *Journal of Time Series Analysis*, **34**, 423–446.

Jirak, M. (2015). Uniform change point tests in high dimension. *The Annals of Statistics*, **43**, 2451–2483.

Kassambara, A. and Mundt, F. (2021). Factoextra: extract and visualize the results of multivariate data analyses, R package version 1.0. 7. 2020.

Kaul, A., Fotopoulos, S. B., Jandhyala, V. K., and Safikhani, A. (2021). Inference on the change point under a high dimensional sparse mean shift. *Electronic Journal of Statistics*, **15**, 71–134.

Kaul, A., Jandhyala, V. K., and Fotopoulos, S. B. (2019). An efficient two step algorithm for high dimensional change point regression models without grid search. *Journal of Machine Learning Research*, **20**, 1–40.

Kaul, A., Zhang, H., Tsampourakis, K., Michailidis, G., Kaul, A., Zhang, H., Tsampourakis, K., and Michailidis, G. (2023). Inference on the change point under a high dimensional covariance shift. *Journal of Machine Learning Research*, **24**, 1–68.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, **107**, 1590–1598.

Kunkel, K. E., Liang, X.-Z., and Zhu, J. (2010). Regional climate model projections and uncertainties of us summer heat waves. *Journal of Climate*, **23**, 4447–4458.

Lai, Y. and Dzombak, D. A. (2019). Use of historical data to assess regional climate change. *Journal of Climate*, **32**, 4299–4320.

Lee, J., Li, S., and Lund, R. (2014). Trends in extreme us temperatures. *Journal of Climate*, **27**, 4209–4225.

Lee, S., Seo, M. H., and Shin, Y. (2016). The lasso for high dimensional regression with a possible change point. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **78**, 193–210.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., Roudier, P., and Gonzalez, J. (2013). Package 'cluster'. *Dosegljivo na*, **980**.

Martinez-Villalobos, C. and Neelin, J. D. (2023). Regionally high risk increase for precipitation extreme events under global warming. *Scientific Reports*, **13**, 5579.

Nazarian, R. H., Brizuela, N. G., Matijevic, B. J., Vizzard, J. V., Agostino, C. P., and Lutsko, N. J. (2024). Projected changes in mean and extreme precipitation over northern mexico. *Journal of Climate*, **24**.

Nazarian, R. H., Vizzard, J. V., Agostino, C. P., and Lutsko, N. J. (2022). Projected changes in future extreme precipitation over the northeast united states in the nacordex ensemble. *Journal of Applied Meteorology and Climatology*, **61**, 1649–1668.

Oswald, E. M. (2018). An analysis of the prevalence of heat waves in the united states between 1948 and 2015. *Journal of Applied Meteorology and Climatology*, **57**, 1535–1549.

Oswald, E. M. and Rood, R. B. (2014). A trend analysis of the 1930–2010 extreme heat events in the continental united states. *Journal of Applied Meteorology and Climatology*, **53**, 565–582.

Reeves, J., Chen, J., Wang, X. L., Lund, R., and Lu, Q. Q. (2007). A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, **46**, 900–915.

Robinson, W. A. (2021). Climate change and extreme weather: A review focusing on the continental united states. *Journal of the Air & Waste Management Association*, **71**, 1186–1209.

Rupp, D. E., Hawkins, L. R., Li, S., Koszuta, M., and Siler, N. (2022). Spatial patterns of extreme precipitation and their changes under˜ 2° c global warming: a large-ensemble study of the western usa. *Climate Dynamics*, **59**, 2363–2379.

Venkatraman, E. S. (1992). *Consistency Results in Multiple Change-Point Problems*. Stanford University.

Vershynin, R. (2019). *High-Dimensional Probability*, volume NA. Cambridge, UK: Cambridge University Press.

Villarini, G., Smith, J. A., and Vecchi, G. A. (2013). Changing frequency of heavy rainfall over the central united states. *Journal of Climate*, **26**, 351–357.

Wang, D., Yu, Y., and Rinaldo, A. (2020). Univariate mean change point detection: Penalization, cusum and optimality. *Electronic Journal of Statistics*, **14**, 1917–1961.

Wang, D., Yu, Y., and Rinaldo, A. (2021). Optimal change point detection and localization in sparse dynamic networks. *The Annals of Statistics*, **49**, 203 – 232.

Wang, T. and Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80**, 57–83.

Wang, X., Huang, G., Lin, Q., Nie, X., and Liu, J. (2015). High-resolution temperature and precipitation projections over ontario, canada: a coupled dynamical-statistical approach. *Quarterly Journal of the Royal Meteorological Society*, **141**, 1137–1146.

Wang, X. L., Chen, H., Wu, Y., Feng, Y., and Pu, Q. (2010). New techniques for the detection and adjustment of shifts in daily precipitation data series. *Journal of Applied Meteorology and Climatology*, **49**, 2416–2436.

Wild, M., Gilgen, H., Roesch, A., Ohmura, A., Long, C. N., Dutton, E. G., Forgan, B., Kallis, A., Russak, V., and Tsvetkov, A. (2005). From dimming to brightening: Decadal changes in solar radiation at earth's surface. *Science*, **308**, 847–850.

Zhao, X. and Chu, P.-S. (2010). Bayesian changepoint analysis for extreme events (typhoons, heavy rainfall, and heat waves): An rjmcmc approach. *Journal of Climate*, **23**, 1034–1046.

Zhou, X., Huang, G., Baetz, B. W., Wang, X., and Cheng, G. (2018). Precis-projected increases in temperature and precipitation over canada. *Quarterly Journal of the Royal Meteorological Society*, **144**, 588–603.

Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 667–693
http://www.ssca.org.in/journal

A Heisenberg-esque Uncertainty Principle for Simultaneous (Machine) Learning and Error Assessment?

Xiao-Li Meng
Department of Statistics, Harvard University

Received: 28 September 2024; Revised: 31 November 2024; Accepted: 02 December 2024

Abstract

A highly cited and inspiring article by Bates *et al.* (2024) demonstrates that the prediction errors estimated through cross-validation, Bootstrap or Mallow's C_P can all be independent of the actual prediction errors. This essay hypothesizes that these occurrences signify a broader, Heisenberg-like uncertainty principle for learning: optimizing learning and assessing actual errors using the same data are fundamentally at odds. Only suboptimal learning preserves untapped information for actual error assessments, and vice versa, reinforcing the 'no free lunch' principle. To substantiate this intuition, a Cramér-Rao-style lower bound is established under the squared loss, which shows that the relative regret in learning is bounded below by the square of the correlation between any unbiased error assessor and the actual learning error. Readers are invited to explore generalizations, develop variations, or even uncover genuine 'free lunches.' The connection with the Heisenberg uncertainty principle is more than metaphorical, because both share an essence of the Cramér-Rao inequality: marginal variations cannot manifest individually to arbitrary degrees when their underlying co-variation is constrained, whether the co-variation is about individual states or their generating mechanisms, as in the quantum realm. A practical takeaway of such a learning principle is that it may be prudent to reserve some information specifically for error assessment rather than pursue full optimization in learning, particularly when intentional randomness is introduced to mitigate overfitting.

Key words: C. R. Rao; Cramér-Rao bound; Cross validation; Epistemology; Heisenberg uncertainty principle; Machine learning; Quantum mechanics; Uniformly minimum variance unbiased estimator.

AMS Subject Classifications: 62K05, 05B05

1. A Rao-esque apology and a quantum-leap excuse

Many of the advances in statistics and machine learning are about using data as efficiently and reliably as possible to achieve a host of learning objectives, such as inference, prediction, classification, *etc.* Being statistically efficient typically means to optimize over some criterion that amounts to minimizing learning errors based on the data at hand, whether

Corresponding Author: Xiao-Li Meng
Email: meng@stat.harvard.edu

in a brute-force fashion, such as minimizing a χ^2 distance or adopting the L^2-loss directly on the target of learning, or through deeper principles, *e.g.*, by maximizing a likelihood function or a posterior density. Since the actual learning errors themselves cannot be known without an external benchmark, we seek clever and reliable ways to assess them, whether for training machine learning algorithms, constructing confidence intervals, or checking Bayesian models.

Naturally, we wish to be able to optimally use our data for both purposes: to most efficiently learn whatever we can learn, and to most reliably assess the errors in whatever we cannot learn. However, since any information on the actual learning error can be used to improve the learning itself, we should be mindful that optimizing one endeavor comes at the expense of the other. To emphasize this no-free lunch principle, this essay first revisits seemingly quaint examples and classical results to remind ourselves that this principle has been in action for as long as statistical inference exists. However, such an issue has not received much emphasis apparently because principled statistical methods, such as likelihood or Bayesian methods, automatically prioritize optimal learning over error assessment.

Yet time has changed. Machine learning and other pattern-seeking methods require much intuition and judgment to tune well, when their theoretical guiding principles are not well developed or digested. Substituting—not merely supplementing—virtual trials and errors for sapient contemplation and introspection is becoming increasingly habitual, making us more vulnerable to wishful thinking, misinformed intuitions, and misguided common sense. To better prepare students and newcomers to our progressively empiricism-slanted culture of learning, this essay then recasts a classical result regarding UMVUE to the broader class of problems of unbiased learning, and establishes a mathematical inequality that captures the aforementioned Heisenberg-esque uncertainty principle for simultaneous learning and error assessment under the squared loss.

This inequality is a low-hanging fruit in establishing a general theory for understanding the competing nature between optimal learning and actual error assessing. Nevertheless, it can help us anticipate and better appreciate further results such as those obtained in Bates *et al.* (2024), which show that the error estimates from cross validation and other popular methods can be independent of actual learning error. The uncertainty principle tells us that this should not come as a surprise. Rather, the independence is an indication that the corresponding learning is optimal in some sense.

Since this essay was prepared for this special issue in memory of Professor C. R. Rao, it seems fitting to quote Rao (1962), a discussion article presented[1] to the Royal Statistical Society in England (RSS):

> "While thanking the Royal Statistical Society for giving me an opportunity to read a paper at one of its meetings, I must apologize for choosing a subject which may appear somewhat classical. But I hope this small attempt intended to state in precise terms what can be claimed about m.l. estimates, in large samples, will at least throw some light on current controversies."

Rao (1962) was a paper on "Efficient estimates and optimum inference procedures in

[1]As a reminder of C. R. Rao's remarkable personal and professional longevity, this presentation took place before my parents had decided to conceive me.

large samples" (and his "m.l." referred to maximum likelihood, not machine learning), one of a series of fundamental articles that he authored during what is now considered an era of classical mathematical statistics. Therefore, initially I was somewhat surprised by Rao's apologetic sentiment—one that I ought to adopt myself for bringing up UMVUE in an era where few statistics students would recognize the acronym without Googling it. However, upon reflection, and considering his training under R. A. Fisher and the characteristically wry culture of RSS discussion at that time, I suspect Rao's apology was more of a gentle reminder to not ignore established literature or wisdom when facing new problems. I am therefore grateful to the editors of this special issue, especially Bhramar Mukherjee, for the opportunity to honor Professor C. R. Rao with one more example of the value of such a reminder: how classical statistical results can offer insights and contextualization for modern work in data science like Bates *et al.* (2024).

I am also deeply grateful to Bhramar for her extraordinary patience in allowing me two extra months to complete this essay, without which I would have embarrassed myself significantly more by writing about the Heisenberg Uncertainty Principle (HUP) while knowing almost surely nothing even about classical mechanics[2]. The connection between Cramér-Rao inequality and HUP has long been suspected, but I was unaware of any statistical literature on the connection between the two (however, during this work, I was made aware of such results in information theory—see Section 7).

Unfortunately, I had found neither the time nor the courage to explore quantum physics. Bhramar's invitation gave me a great excuse to delve into it, though clearly it has been a quantum leap (or dive). I am therefore deeply grateful to the physicists, philosophers, and statisticians (see acknowledgment) who generously took the time to educate and inspire me, introducing me to numerous articles that, no doubt, will require another quantum-leap excuse to digest fully. These include physics literature on quantum Cramér-Rao bounds and quantum Fisher information (*e.g.*, Tóth and Petz, 2013; Tóth and Fröwis, 2022), as well as statistical writings on the relevance of quantum uncertainty to statistics (*e.g.*, Gelman and Betancourt, 2013), to name just a few.

Nevertheless, to set readers' expectations realistically, this essay offers nothing about HUP that isn't already in Wikipedia. I wrote much of it as reading notes to educate myself, so, paraphrasing a most memorable chiasmus from an RSS discussion: "The parts of the paper that are true are not new, and parts that are new are not true" (McCullagh, 1999). My hope, however, is that these notes may still be of use to those who share my curiosity (and innocence). I also hope that my attempt to extend the notion of covariance to quantum operators might encourage us to step out of our comfort zones without stepping out of our minds.

Intellectually, quantum indeterminacy is a captivating and challenging topic, especially for those of us who have been probability-law abiding citizens. To my knowledge, currently only a few statisticians—most notably Richard Gill[3]—have studied it systematically. Therefore, even if everything "new" in this essay ends up merely demonstrating that humans can out-hallucinate ChatGPT, I'd still be content dedicating it to the legendary C.

[2]Majoring in pure math in 1980s China means that I had taken no courses outside of mathematics, with the exception of mandatory ones for regulating students' bodies or minds.

[3]See https://www.math.leidenuniv.nl/ gillrd/

R. Rao. Throughout his extraordinary career, Professor Rao applied his statistical insight and mathematical skills to establish and solidify the foundations of statistics. As quantum computing looms on the horizon, some statisticians should be leading the way in building the foundations of quantum data science, as articulated in the discussion article "When Quantum Computation Meets Data Science: Making Data Science Quantum" by Wang (2022), a prominent statistician exploring quantum computing's role in data science. Thus, even if this essay inspires only one future C. R. Rao of quantum data science, it won't take a quantum leap to believe that Professor Rao would embrace my dedication.

More broadly, I would find great professional satisfaction (and justification for my insomnia) if this essay serves as a reminder that time-honored statistical theory and wisdom have much to offer as we statisticians are increasingly called to step outside our comfort zones—from embracing machine learning to anticipating quantum computing. By learning from and contributing to other fields, especially time-tested ones such as philosophy and physics, we can enhance the intellectual impact of our discipline.

2. A paradox of error assessment?

Let us start with an excursion to the classical statistical sanctuary most frequently adopted in statistical research and pedagogy: we have an independently and identically distributed (i.i.d) normal sample, $X_1, \cdots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, and we are interested in making inferences about μ. It is well-known that the maximum likelihood estimator (MLE) for μ is the sample mean $\bar{X}_n$. The actual error of the MLE then is $\delta = \bar{X}_n - \mu$. It is textbook knowledge that the sample mean $\bar{X}_n$ and the sample variance S_n^2 are independent under the normal model $N(\mu, \sigma^2)$. This fact is critical for establishing perhaps the most celebrated pivotal quantity in statistics, $t = \sqrt{n}(\bar{X} - \mu)/S_n$, *i.e.*, the t statistic, because of the existence of the parameter-free distribution of t for any $n \geq 2$, thanks to the aforementioned independence.

But this independence also implies a seemingly paradoxical fact that has received no mention in any textbook (that I am aware of): that $\hat{\delta}^2 \equiv S_n^2/n$ apparently is the worst estimate of the square of the actual error $\delta^2 = (\bar{X}_n - \mu)^2$, because $\hat{\delta}^2$ and δ^2 are independent of each other for any choice of $\theta = \{\mu, \sigma^2\}$. In what other context would a statistician (knowingly) suggest estimating an unknown with an independent quantity?

The article by Bates *et al.* (2024) reminds us that this seemingly paradoxical phenomenon is far more prevalent than we may have realized. To recast their findings in a broader setting but with a scalar estimand for notational simplicity, consider the possibly heteroscedastic linear regression setting,

$$Y_i = \theta X_i + \epsilon_i, \quad \text{where} \quad \mathrm{E}[\epsilon_i|\mathbf{X}] = 0, \mathrm{V}(\epsilon_i|\mathbf{X}) = \sigma_i^2, \quad i = 1, \ldots, n. \tag{1}$$

and conditioning on $\mathbf{X} = \{X_1, \ldots, X_n\}$, $\{\epsilon_1, \ldots, \epsilon_n\}$ are mutually independent. As Bates *et al.* (2024) reminds us, when $\{\epsilon_1, \ldots, \epsilon_n\}$ are i.i.d $N(0, \sigma^2)$, the least-squares estimator for θ, $\hat{\theta}_{\text{LS}} = \sum_{i=1}^n Y_i X_i / \sum_{i=1}^n X_i^2$ is independent of the residual $R = \{\hat{r}_i = Y_i - \hat{\theta} X_i, i = 1, \ldots, n\}$, for any given $\{\theta, \sigma^2\}$. Consequently, since the true predictive error depends on the data only through $\hat{\theta}_{\text{LS}}$, and cross-validation error estimators are functions only of the residuals, the true and estimated errors are independent of each other. The results obviously apply to any

error estimates that depend on data only through R, which is the case virtually for all the common estimators in practice, as demonstrated in Bates *et al.* (2024).

It is well-known (*e.g.*, Casella and Berger, 2024) that under the i.i.d normal setting, $\hat{\theta}_{\text{LS}}$ is the MLE and indeed UMVUE (uniformly minimum variance unbiased estimator) because its variance reaches the Cramér-Rao bound. Even without the normality, we know that $\hat{\theta}_{\text{LS}}$ is BLUE (best linear unbiased estimator) and it is linearly uncorrelated with the residual R under the squared loss, because it is the orthogonal projection of Y onto the space expanded by $\mathbf{X}$ when σ_i is invariant of i.

Although rarely mentioned in textbooks, this optimality-orthogonality duality appears in essentially all inferential paradigms. Geometrically speaking, the equivalence is due to the fact that the linear correlation between two variables is the cosine of the angle between them in the L^2 space, and optimal projection is the orthogonal projection. Probabilistically, the ubiquity of this duality is manifested by the so-called "Eve's law" (Blitzstein and Hwang, 2014), an instance of the Pythagorean theorem in the L^2 space.

That is, under any joint distribution, $p(H, G)$, as long as it generates finite second moments, $\text{Cov}[H - \text{E}(H|G), \text{E}(H|G)] = 0$, because $\text{E}(H|G)$ is the orthogonal projection of H to the space of L^2 functions that are measurable with respect to the σ-field generated by G. Consequently, the Pythagorean theorem is in force:

$$\begin{aligned}\text{V}(H) &= \text{E}\left[H - \text{E}(H)\right]^2 = \text{E}[H - \text{E}(H|G)]^2 + \text{E}[\text{E}(H|G) - \text{E}(H)]^2 \\ &= \text{E}[\text{V}(H|G)] + \text{V}[\text{E}(H|G)], \end{aligned} \tag{2}$$

which is Eve's law. The ubiquity of the duality is due to the fact that the expectation operator in (2) can be taken with any kind of distribution: posterior (predictive) distributions for Bayesian inferences, super-population distributions as typical for likelihood inference (as in the $N(\mu, \sigma^2)$ example), or randomization distributions as in finite-population calculations (as adopted in Meng, 2018).

Nevertheless, this duality is a qualitative statement, as it does not quantify what happens for non-optimal estimation or learning. As demonstrated below, this duality can be extended quantitatively by tethering the deficiency in learning with the relevancy in assessing the actual learning errors. This quantification crystallizes the reason for the apparent paradox, and it can help reduce wasted efforts in pursuit of the impossible. It also makes it clearer that there is no real paradox, much like how Simpson's paradox is not a paradox once its workings are revealed and understood (*e.g.* Liu and Meng, 2014; Gong and Meng, 2021).

The title of the next section says it all: there is no free lunch. If there is any data information left—after learning—for assessing the actual error, then we can reduce the actual error by removing the part that can be predicted by the untapped data information. This implies our learning is not optimal, and vice versa. Section 3 illustrates this fact in the context of heteroscedastic regression, followed by a broad reflection in Section 4 on its implications in the context of error assessment without external benchmarks, a statistical magic. Sections 5 and 6 then establish respectively the exact and asymptotic inequalities that capture the learning uncertainty principle under the squared loss.

To facilitate a formal comparison with HUP using the notion of co-variation, Sec-

tion 7 discusses the generalization of the measure of co-variance from real-valued variables to complex-valued variables and functions. Section 8 then applies the generalization to the case of HUP by defining co-variances between mechanisms (*e.g.*, the position and momentum *operators*) rather than between the states they generate (*e.g.*, the actual position and momentum states). With these preparations, Section 9 compares the learning-error inequality, Cramér-Rao inequality, and HUP inequality, highlighting their shared essence from a statistical perspective.

Section 10 reflects on various philosophical issues surrounding uncertainty principles in general, and HUP in particular, with insights from the encyclopedic essay by Hilgevoord and Uffink (2024). Section 11 briefly touches on the trade-off between quantitative and qualitative studies, prompted by a discussion in Hilgevoord and Uffink (2024), and how intellectual inquires can benefit from their happy marriage. This leads to a piece of advice from Professor Rao on living a happy life, which serves as a fitting conclusion to this essay in his memory. However, to encourage students to engage with this essay to the fullest extent of their attention spans, Section 12 provides a prologue, especially for those who may not enjoy technical appendices but wish the essay were even longer.

3. Once again, there is no free lunch

Consider the heteroscedastic setting (1), where we know that BLUE is given by the weighted LS, in the form of

$$\hat{\theta}_w = \frac{\sum_{i=1}^n w_i Y_i X_i}{\sum_{i=1}^n w_i X_i^2}, \tag{3}$$

when the weights $w_i \propto \sigma_i^{-2}, i = 1, \ldots, n$. Now consider an arbitrarily weighted $\hat{\theta}_w$, and its correlation—denoted by ρ—with the corresponding residual $R_w = \{\hat{r}_{w,i} = Y_i - \hat{\theta}_w X_i; i = 1, \ldots, n\}$. For conveying the main idea, the case of $n = 2$ is sufficient. As a special case of the general expression given in Appendix A, we have, conditioning on $\mathbf{X}$ (but we suppress this conditioning notation-wise unless necessary),

$$\rho^2(\hat{\theta}_w, \hat{r}_{w,i}) = \frac{X_1^2 X_2^2 (w_1\sigma_1\sigma_2^{-1} - w_2\sigma_2\sigma_1^{-1})^2}{(w_1^2 X_1^2\sigma_1^2 + w_2^2 X_2^2\sigma_2^2)(X_1^2\sigma_1^{-2} + X_2^2\sigma_2^{-2})}, \quad i = 1, 2, \tag{4}$$

which is zero if and only if $w_i \propto \sigma_i^{-2}, i = 1, 2$ (as long as $X_i \neq 0, i = 1, 2$). That is, $\hat{\theta}_w$ is BLUE (or the MLE if we assume normality) if and only if $\hat{\theta}_w$ is uncorrelated with $\hat{r}_{w,i}$. More importantly, expression (4) tells us exactly how the statistical efficiency of $\hat{\theta}_w$ is directly linked to this correlation.

Specifically, let $\hat{\theta}_{\text{BLUE}}$ be the optimally weighted LS estimator with weight $w_i \propto \sigma_i^{-2}, i = 1, 2$, and RR_w be the relative regret of an arbitrarily weighted $\hat{\theta}_w$ under the squared loss, that is,

$$RR_w = \frac{\text{V}(\hat{\theta}_w) - \text{V}(\hat{\theta}_{\text{BLUE}})}{\text{V}(\hat{\theta}_w)} = 1 - \frac{(w_1 X_1^2 + w_2 X_2^2)^2}{(w_1^2 X_1^2\sigma_1^2 + w_2^2 X_2^2\sigma_2^2)(X_1^2\sigma_1^{-2} + X_2^2\sigma_2^{-2})}. \tag{5}$$

Whereas it may not be immediate from (4) and (5), one can verify directly that

$$\rho^2(\hat{\theta}_w, \hat{r}_{w,i}) = RR_w, \quad i = 1, 2, \tag{6}$$

for any choice of weights w or values of $\{\sigma_i^2, i = 1, 2\}$. This means that if we want to increase the magnitude of the correlation between $\hat{\theta}_w$ and $\hat{r}_{w,i}$, we must sacrifice the efficiency of $\hat{\theta}_w$, and vice versa.

However, why might we want to increase $|\rho(\hat{\theta}_w, \hat{r}_{w,i})|$? Consider the case where our learning target is $c\theta$, with c being a constant. For example, we take $c = 1$ when the regression coefficient θ is the target, or $c = X^*$ when the learning target is the mean of Y when $X = X^*$. In such cases, the actual error is given by $\delta_w = c(\hat{\theta}_w - \theta)$. We can assess δ_w via $\hat{\delta}_w = \tilde{c}\hat{r}_{w,1}$ for some choice of $\tilde{c}$ (recall $\hat{r}_{w,1} + \hat{r}_{w,2} = 0$ and hence a single residual suffices). Because

$$\rho^2(\delta_w, \hat{\delta}_w) = \rho^2(c\hat{\theta}_w, \tilde{c}\hat{r}_{w,1}) = \rho^2(\hat{\theta}_w, \hat{r}_{w,1}), \tag{7}$$

we see that by moving $\rho^2(\hat{\theta}_w, \hat{r}_{w,1})$ away from zero, we will have an assessment $\hat{\delta}_w$ of the actual error δ_w that has a degree of conditional relevancy, that is, $\hat{\delta}_w$ is at least correlated with δ_w conditioning on the setting (1). But this gain of relevancy is achieved necessarily by increasing the relative regret (recall the relative regret for $c\hat{\theta}_w$ is invariant to the value of c), that is, by sacrificing the efficiency of $\hat{\theta}_w$, because

$$\rho^2(\delta_w, \hat{\delta}_w) = RR_w, \tag{8}$$

thanks to (6)-(7).

If our learning target is to predict (a new) Y^* when $X = X^*$, then the actual prediction error is $\delta_w^* = Y^* - \hat{\theta}_w X^*$. In such cases, the prediction risk under the squared loss is

$$\mathrm{E}(Y^* - \hat{\theta}_w X^*)^2 = \mathrm{V}(Y^*) + (X^*)^2 \mathrm{V}(\hat{\theta}_w).$$

Because $\mathrm{V}(Y^*)$ and $(X^*)^2$ are invariant to the weights, we obtain the relative regret for prediction $RR_w^* = \gamma RR_w$, where RR_w is from (5) and the adjustment factor γ is given by

$$\gamma = \frac{(X^*)^2 \mathrm{V}(\hat{\theta}_w)}{\mathrm{V}(Y^*) + (X^*)^2 \mathrm{V}(\hat{\theta}_w)}. \tag{9}$$

Furthermore, because $\hat{\delta}_w = \tilde{c}\hat{r}_{w,1}$ is independent of Y^*, $\mathrm{Cov}(\delta_w^*, \hat{\delta}_w) = -X^* \mathrm{Cov}(\hat{\theta}_w, \hat{\delta}_w)$. Hence,

$$\rho^2(\delta_w^*, \hat{\delta}_w) = \frac{(X^*)^2 \mathrm{Cov}^2(\hat{\theta}_w, \hat{\delta}_w)}{\left[\mathrm{V}(Y^*) + (X^*)^2 \mathrm{V}(\hat{\theta}_w)\right] \mathrm{V}(\hat{\delta}_w)} = \gamma \rho^2(\hat{\theta}_w, \hat{\delta}_w). \tag{10}$$

Consequently, the identity (8) holds for both estimation and prediction, implying the same trade-off between optimal learning and relevant error assessment.

Section 5 below will provide a general inequality that captures this trade-off under squared loss, for which identity (8) is a special case. But before presenting that result, we must ask: if we cannot relevantly assess the actual error δ, then what kind of errors have we been assessing? And that is exactly one of the two questions raised in the title of Bates *et al.* (2024): Cross-validation: what does it estimate and how well does it do it? The following section supplements Bates *et al.* (2024) to answer this question more broadly and more pedagogically.

4. Jay Leno's irony and a statistical magic

During one of the years the United States census took place (likely 2000-2001), co-median Jay Leno brought up the issue of under-counting on his *Tonight Show*. He began by informing the audience that the U.S. Census Bureau had just reported that approximately p percentage of the population had not been counted. With an arch smile, he then quipped, "But I don't understand—if they knew they missed p percentage of people, why didn't they just add it back?" (The actual value p he used now lies deep in my memory.)

The audience was amused, as was I, though perhaps for different reasons—what amused me was the very appearance of such a nerdy joke on a mainstream comedy show. Humor is often rooted in life's ironies, and whoever crafted this joke clearly understood the irony in announcing both an estimate and its error. In the case of the U.S. Census, the irony—or more accurately, the *magic*—is not as profound as it may seem. The estimation of undercount relies on external data, such as demographic analysis, post-enumeration surveys, administrative records, and other sources. The term *magic* is used here because statistical inference can appear magical to uninitiated yet inquisitive minds. How can one estimate an unknown quantity, and then estimate the error of that estimation, without any external knowledge of the true value?

The magic begins with a sleight of hand—in this case, the word *error* does not refer to the actual error, as a layperson might assume. Instead, we aim to understand the statistical properties of the actual error by imagining its variations across hypothetical replications. The construction of these replications depends on the philosophical framework one subscribes to, with the two main schools being frequentist and Bayesian (but see Lin (2024b) for a spectrum between them). Perhaps surprisingly, the key to resolving the apparent paradox in Section 2 lies in adopting insights from both perspectives.

To see this, consider again the normal example where the true error is $\delta = \bar{X}_n - \mu$. In the frequentist framework, the hypothetical replications consist of all possible copies of $D = \mathbf{X} = \{X_1, \ldots, X_n\}$ generated from $N(\mu, \sigma^2)$ with the *same* but unknown parameter values $\theta = \{\mu, \sigma^2\}$. In this replication setting, the expected value of δ^2, which is the sampling variance of $\bar{X}_n$, equals σ^2/n. It is well-known that under the same replication framework, the expectation of $\hat{\delta}^2 = S_n^2/n$ is also σ^2/n.

Thus, while δ^2 and $\hat{\delta}^2$ are independent of each other for any given $\theta = \{\mu, \sigma^2\}$, they share the same expectation within the frequentist framework. By invoking the same leap of faith that underpins the frequentist approach—trusting and transferring average behaviors to assess individual cases—we justify $\hat{\delta}^2$ as an estimate of δ^2. Such a leap of faith exists regardless of the goal of our data exercise, be it prediction, estimation, or attribution (significance testing), albeit with increased levels of intolerance to the inaccuracy in error assessing, as revealed by the insightful article of Efron (2020).

For Bayesians, such a leap of faith is unconvincing or even "irrelevant" in the sense of Dempster (1963), as the actual error can differ significantly from its expectation. The independence between $\hat{\delta}^2$ and δ^2 suggests that accepting this leap would require a religious level of faith. In the Bayesian framework, the relevant hypothetical replications include all possible values of $\theta = \{\mu, \sigma^2\}$ (and their associated probabilities) that could have generated the same data set D, and therefore the same $\{\bar{X}_n, S_n^2\}$.

However, for such a replication setting to be realized—for instance, via a simulation—a prior distribution for $\theta = \{\mu, \sigma^2\}$ must be assumed. This postulation represents the Bayesian leap of faith in actual implementations, since it is virtually certain that a part of the assumption is faith-based instead of knowledge-driven; for a broader discussion on the necessity of such leaps across all major schools of statistical inference—Bayesian, Fiducial, and Frequentist (BFF)—see Craiu *et al.* (2023) and more comprehensively the *Handbook on BFF Inference* edited by Berger *et al.* (2024).

Although we shall not take a Bayesian excursion here, we can borrow the Bayesian concept of allowing $\theta = \{\mu, \sigma^2\}$ to have a distribution in order to establish a joint replication setting, where both D and $\theta = \{\mu, \sigma^2\}$ vary. This framework is relevant (for frequentists) when recommending the same statistical procedure across multiple studies with normal data, where both D and $\theta = \{\mu, \sigma^2\}$ may differ from study to study. In the machine learning world—or any domain reliant on training data—such a joint replication setting can be visualized as potential training datasets drawn from related populations, which makes transfer learning a meaningful endeavor (*e.g.*, Abba *et al.*, 2024).

For our normal example, given any proper prior on θ, it can be shown (see Appendix B) that over any proper joint replication of $\{D, \theta\}$,

$$\rho(\hat{\delta}^2, \delta^2) = \frac{\gamma_{\sigma^2}^2}{\sqrt{\frac{n+1}{n-1}\gamma_{\sigma^2}^2 + \frac{2}{n-1}}\sqrt{3\gamma_{\sigma^2}^2 + 2}}, \tag{11}$$

where γ_{σ^2} is the coefficient of variation of σ^2 with respect to the (proper) prior distribution of σ^2. This correlation is non-negative, providing a plausible measure of how relevant $\hat{\delta}^2$ is for assessing δ^2. It is zero if and only if $\mathrm{V}(\sigma^2) = 0$, meaning that we revert to the situation of conditioning on a fixed σ^2: since S_n^2 is invariant to μ, $\hat{\delta}^2$ and δ^2 remain independent when conditioned on σ^2 alone. The fact that (11) is a monotonic increasing function of γ_{σ^2} implies that the relevance of $\hat{\delta}^2$ for assessing δ^2 increases as the heterogeneity among the studies—in terms of the within-study variation indexed by σ^2—grows. This monotonicity is intuitive, given that S_n^2 is an unbiased and asymptotically efficient estimator of σ^2, and $\hat{\delta}^2$ is useful for comparing the magnitudes of δ^2 across studies with different σ^2 values. However, the fact that this correlation can never exceed $1/\sqrt{3} \approx 0.577$ is unexpected. For those of us who believe that mathematical results are never coincidental, contemplating the intricacies of this bound might induce insomnia (while serving as a cure for many others).

This joint replication framework clarifies the role of $\hat{\delta}^2$ as an *adaptive benchmark* for assessing the statistical properties of δ^2 over the hypothetical replications. *That* is statistical magic—the ability to establish cross-study comparisons based on a single study. More broadly, the magic lies in creating hypothetical "control" replications $\{\tilde{D}, \tilde{\theta}\}$ from the actual "treatment" $\{D, \theta\}$ at hand, as elaborated in Liu and Meng (2016), borrowing the metaphor of individualized treatment.

Generally speaking, the magic relies on two tricks: (I) creating replications within D, and (II) linking those replications to the imagined variations of D through the within-D replications from (I). The first trick is applicable when the mechanism generating the data D inherently includes (higher resolution) replications, either by design (*e.g.*, simple random sampling) or by declaration (*e.g.*, imposing an i.i.d. structure as a working assumption).

The second trick is enabled by theoretical understanding (*e.g.*, the relationship between the distribution of the sample mean and the distribution of the individual samples) or by simulations and approximations that are enabled by (I), such as the Bootstrap (see Craiu *et al.*, 2023, for a discussion).

The magic metaphor also serves as a reminder that magic relies on illusions, and interpreting average errors as actual ones is one such illusion. With that understanding, we might wonder if it's possible to assess the actual error with greater relevance. For example, in the normal case, one might ask whether a different error estimate $\check{\delta}$ could be more relevant for $\delta = \bar{X}_n - \mu$, in the sense that $\rho(\check{\delta}, \delta) > 0$ given any value of $\theta = \{\mu, \sigma^2\}$. The classical statistical literature offers a fairly clear answer to this question, as discussed below.

5. From UMVUE to an uncertainty principle for unbiased learning

The celebrated Cramér–Rao bound, more broadly known as the information inequality (see Lehmann and Casella, 2006, Ch. 2), tells us that if $\hat{\theta}$ is an unbiased estimator for θ under a parametric model $f(D|\theta)$, then under mild conditions, $\mathrm{V}(\hat{\theta}) \geq I^{-1}(\theta)$, where $I(\theta)$ is the expected Fisher information. For the normal example, when we take $\theta = \mu$ (temporarily assuming σ^2 is known), we have $\mathrm{V}(\bar{X}_n) = \sigma^2/n = I^{-1}(\mu)$, where $I(\mu)$ is the expected Fisher information from $f(X_1, \ldots, X_n|\mu)$. Thus, we know $\bar{X}_n$ is UMVUE for μ.

It is well-known that an estimator $\hat{\theta}$ is UMVUE if and only if it is uncorrelated with any unbiased estimator U for zero for any θ (see Lehmann and Casella, 2006, Ch. 2), that is, $\mathrm{E}_\theta[(\hat{\theta} - \theta)U] = 0$, whenever $\mathrm{E}_\theta(U) = 0$. Since $\hat{\theta} - \theta$ is simply the actual error δ, this result implies that conditioning on θ, it is impossible to have an error assessment $\hat{\delta}$ for δ that is both unbiased and relevant at the same time, *i.e.*, $\mathrm{E}_\theta(\hat{\delta}) = 0$ and $\rho^2_\theta(\hat{\delta}, \delta) > 0$ cannot hold simultaneously for any θ, where we inject the subscript θ in ρ_θ to explicate that the correlation is with respect to $f(D|\theta)$ for fixed θ.

Intuitively, if any unbiased error assessment $\hat{\delta}$ is correlated with δ, then some part of the actual error δ is predictable by $\hat{\delta}$. This means that we could improve $\hat{\theta}$ without losing its unbiasedness, which contradicts the fact that $\hat{\theta}$ is already an UMVUE. An astute reader may quickly recognize that this insight has much broader implications than merely for UMVUEs. The following result is a proof of this realization, using the same proof strategy as for UMVUE, but establishes a broader quantitative result than the aforementioned qualitative "if and only if" result for UMVUE. The result is presented in the scalar case for simplicity, but its multivariate counterpart can be derived easily using corresponding matrix notation.

Specifically, let $Q \in \mathbb{R}$ be our target of learning, which could represent a future outcome, a model parameter, a latent trait, *etc.* Suppose the state space of our data D is Ω and $\hat{Q} : \Omega \to \mathbb{R}$ is our learning algorithm, or a *learner* for Q. For any learner $\hat{Q}$, let $\hat{\delta}_{\hat{Q}} : \Omega \to \mathbb{R}$ be an assessment (*e.g.*, an estimator) of the exact (additive) error of $\hat{Q}$, namely, $\delta_{\hat{Q}} = \hat{Q} - Q$. Let $L(\hat{Q}, Q)$ be the loss function, and $\mathcal{P} = \{P_s(D; Q), s \in S\}$ be the family of distributions under which we calculate the learning risk: $R_s(\hat{Q}) = \mathrm{E}_s[L(\hat{Q}, Q)]$. Note that Q may be a function of s (*e.g.*, when estimating the model parameter s) or it may be a random variable itself (*e.g.*, a future realization), in which case the notation $P_s(D; Q)$ represents the joint distribution over D and Q.

Theorem 1: Let $L(\hat{Q}, Q) = (\hat{Q} - Q)^2$ be the squared loss, and let $L^2_{\mathcal{P}}$ denote the collection of all square-integrable functions with respect to $\mathcal{P}$. Define

$$\mathcal{Q} = \{\hat{Q}(D) \in L^2_{\mathcal{P}} : \mathrm{E}_s(\hat{Q} - Q) = 0, \forall s \in S\} \tag{12}$$

as the collection of unbiased learners of Q with respect to $\mathcal{P}$. For any $\hat{Q} \in \mathcal{Q}$, define

$$\mathcal{E}(\hat{Q}) = \{\hat{\delta}_{\hat{Q}}(D) \in L^2_{\mathcal{P}} : \mathrm{E}_s(\hat{\delta}_{\hat{Q}}) = 0, \forall s \in S\} \tag{13}$$

as the collection of corresponding unbiased error assessors for $\delta_{\hat{Q}}$. Suppose there exists an optimal learner $\hat{Q}^{\mathrm{opt}} \in \mathcal{Q}$, with risk $R_s^{\mathrm{opt}} < \infty$ under $f_s, s \in S$. Then:

(I) For any $\hat{Q} \in \mathcal{Q}$ and any corresponding $\hat{\delta}_{\hat{Q}} \in \mathcal{E}(\hat{Q})$, we have

$$\rho_s^2(\delta_{\hat{Q}}, \hat{\delta}_{\hat{Q}}) \leq \frac{R_s(\hat{Q}) - R_s^{\mathrm{opt}}}{R_s(\hat{Q})} \equiv RR_s(\hat{Q}), \quad \forall s \in S, \tag{14}$$

where $RR_s(\hat{Q})$ is the relative regret of $\hat{Q}$ under distribution P_s, and it is set to zero if $R_s(\hat{Q}) = 0$.

(II) Equality $\rho_s^2(\delta_{\hat{Q}}, \hat{\delta}_{\hat{Q}}) = RR_s(\hat{Q})$ holds for any particular s if and only if R_s^{opt} is attainable in the sub-class $\mathcal{Q}(\hat{Q}, \hat{\delta}_{\hat{Q}}) = \{\hat{Q} - \lambda\hat{\delta}_{\hat{Q}} : \forall \lambda \in \mathbb{R}\} \subset \mathcal{Q}$.

Proof: For any given $\hat{Q} \in \mathcal{Q}$ (which is non-empty since $\hat{Q}^{\mathrm{opt}} \in \mathcal{Q}$) and any $\hat{\delta}_{\hat{Q}} \in \mathcal{E}(\hat{Q})$ (which is non-empty since $\hat{\delta}_{\hat{Q}} \equiv 0$ is always included), we define $\hat{Q}_\lambda = \hat{Q} - \lambda\hat{\delta}_{\hat{Q}}$ for any constant $\lambda \in \mathbb{R}$. Under our assumptions, $\mathrm{E}_s(\hat{Q}_\lambda - Q) = 0$, and $\hat{Q}_\lambda \in L^2_{\mathcal{P}}$, implying $\hat{Q}_\lambda \in \mathcal{Q}$. Since $\hat{Q}_\lambda - Q = \delta_{\hat{Q}} - \lambda\hat{\delta}_{\hat{Q}}$ and it has mean zero under $f_s(D; Q)$, we have

$$R_s^{\mathrm{opt}} \leq R_s(\hat{Q}_\lambda) = \mathrm{V}_s(\delta_{\hat{Q}} - \lambda\hat{\delta}_{\hat{Q}}) = \mathrm{V}_s(\delta_{\hat{Q}}) + \lambda^2\mathrm{V}_s(\hat{\delta}_{\hat{Q}}) - 2\lambda\mathrm{Cov}_s(\delta_{\hat{Q}}, \hat{\delta}_{\hat{Q}}), \quad \forall s \in S. \tag{15}$$

Since the left-hand side of this inequality is free of λ, the inequality holds when we minimize the right-hand side over $\lambda \in \mathbb{R}$, which is achieved at $\lambda = \lambda^* = \mathrm{Cov}_s(\delta_{\hat{Q}}, \hat{\delta}_{\hat{Q}})/\mathrm{V}_s(\hat{\delta}_{\hat{Q}})$, assuming $\mathrm{V}_s(\hat{\delta}_{\hat{Q}}) > 0$. (When $\mathrm{V}_s(\hat{\delta}_{\hat{Q}}) = 0$, $\rho_s(\delta_{\hat{Q}}, \hat{\delta}_{\hat{Q}}) = 0$; hence (14) holds trivially, and we can set $\lambda^* = 0$.) Thus, we obtain

$$R_s^{\mathrm{opt}} \leq \mathrm{V}_s(\delta_{\hat{Q}}) \left[1 - \rho_s^2(\delta_{\hat{Q}}, \hat{\delta}_{\hat{Q}})\right], \quad \forall s \in S,$$

which yields (14) since $R_s(\hat{Q}) = \mathrm{V}_s(\delta_{\hat{Q}})$ when $\mathrm{E}_s(\delta_{\hat{Q}}) = 0$. This proves part (I).

Part (II) follows from (15) as well, because the equality holds there if and only if R_s^{opt} is attainable by $\hat{Q}_{\lambda^*} \in \mathcal{Q}(\hat{Q}, \hat{\delta}_{\hat{Q}})$. This includes the case with $\mathrm{V}_s(\hat{\delta}_{\hat{Q}}) = 0$, where the result holds trivially, because then $\rho_s(\delta_{\hat{Q}}, \hat{\delta}_{\hat{Q}}) = 0$ and $R_s(\hat{Q}) = R_s^{\mathrm{opt}}$, *i.e.*, $\hat{Q}$ itself is optimal. □

The immediate implication of inequality (14) is that there is no free lunch. If we want to increase the relevance of our assessment $\hat{\delta}_{\hat{Q}}$ for the actual error $\delta_{\hat{Q}}$ by increasing their

correlation, we must also increase the relative regret for $\hat{Q}$, effectively sacrificing degrees of freedom of learning for the error assessment. Conversely, the less regret in $\hat{Q}$, the less relevant its error assessment will be to the actual error. In the extreme case, when $\hat{Q} = \hat{Q}^{\text{opt}}$, we arrive at the following result, where by a *relevant error assessor* we mean it is linearly correlated with the actual error of the learner.

Corollary 1: Under the same setup as in Theorem 1, the following two assertions cannot hold simultaneously:

(A) $\hat{Q} \in \mathcal{Q}$ is an *optimal and unbiased learner* for Q under P_s; and

(B) $\hat{Q}$ has an *unbiased and relevant error assessor* $\hat{\delta}_{\hat{Q}} \in \mathcal{E}(\hat{Q})$.

6. Beyond unbiased learning and error assessing

A key limitation of Theorem 1 is the requirement that both the learner and error assessor must be unbiased. An immediate generalization is to consider cases where both are asymptotically unbiased, under an asymptotic regime with respect to some information index ι, such as the size of data. Mathematically, given a sequence of error order e_ι such that $\limsup_{\iota\to\infty} |e_\iota| = 0$, we can modify the classes of the learners and error assessors in (12) and (13) respectively by

$$\mathcal{Q}_\iota = \{\hat{Q}(D) \in L^2_{\mathcal{P}} : \mathrm{E}_s[\hat{Q}(D) - Q] = O(e_\iota), \forall s \in S\}, \tag{16}$$

$$\mathcal{E}_\iota(\hat{Q}) = \{\hat{\delta}_{\hat{Q}}(D) \in L^2_{\mathcal{P}} : \mathrm{E}_s(\hat{\delta}_{\hat{Q}}) = O(e_\iota), \forall s \in S\}, \tag{17}$$

where $O(e_\iota)$ is the standard notation for being of the same order as e_ι. That the error assessor $\hat{\delta}_{\hat{Q}}$ must share the same order of expectation as the actual error $\delta_{\hat{Q}}$ is a necessary requirement to render the term 'error assessor' meaningful, as otherwise anything could be regarded as $\hat{\delta}_{\hat{Q}}$. With these modifications, we have the following asymptotic counterpart of Theorem 1.

Theorem 2: Assume the same setup as Theorem 1, but with $\mathcal{Q}$ and $\mathcal{E}(\hat{Q})$ extended respectively to $\mathcal{Q}_\iota$ and $\mathcal{E}_\iota(\hat{Q})$. We then have

$$\rho_s^2(\delta_{\hat{Q}}, \hat{\delta}_{\hat{Q}}) \le R R_s(\hat{Q}) + O(e_\iota^2), \quad \forall s \in S, \tag{18}$$

where e_ι is a sequence of vanishing error rates that determines the asymptotic regime.

Proof: For $\hat{Q} \in \mathcal{Q}_\iota$ and $\hat{\delta}_{\hat{Q}} \in \mathcal{E}_\iota(\hat{Q}_\iota)$, we can write $\mathrm{E}_s(\delta_{\hat{Q}}) = a_\iota$ and $\mathrm{E}_s(\hat{\delta}_{\hat{Q}}) = b_\iota$ where $a_\iota = O(e_\iota)$ and $b_\iota = O(e_\iota)$ by our assumption. Hence for $\hat{Q}_\lambda = \hat{Q} - \lambda\hat{\delta}_{\hat{Q}}$, $\mathrm{E}_s(\hat{Q}_\lambda - Q) = a_\iota - \lambda b_\iota = Q(e_\iota)$ for any λ, implying that $\hat{Q}_\lambda \in \mathcal{Q}_\iota$. Let λ^* be the minimizer of $\mathrm{V}_s\left[\delta_{\hat{Q}} - \lambda\hat{\delta}_{\hat{Q}}\right]$, as defined in the proof of Theorem 1. The optimality of R_s^{opt} then implies that

$$\begin{aligned} R_s^{\text{opt}} \le R_s(\hat{Q}_{\lambda^*}) &= \mathrm{V}_s\left[\delta_{\hat{Q}} - \lambda^*\hat{\delta}_{\hat{Q}}\right] + \left[\mathrm{E}_s(\delta_{\hat{Q}} - \lambda^*\hat{\delta}_{\hat{Q}})\right]^2 \\ &= \mathrm{V}_s(\delta_{\hat{Q}})\left[1 - \rho_s^2(\delta_{\hat{Q}}, \hat{\delta}_{\hat{Q}})\right] + (a_\iota - \lambda^* b_\iota)^2. \\ &\le R_s(\delta_{\hat{Q}})\left[1 - \rho_s^2(\delta_{\hat{Q}}, \hat{\delta}_{\hat{Q}})\right] + (a_\iota - \lambda^* b_\iota)^2. \end{aligned}$$

But this proves the inequality (18) because $(a_\iota - \lambda^* b_\iota)^2 = O^2(e_\iota) = O(e_\iota^2)$. □

A major application of Theorem 2 is for the maximum likelihood estimator $\hat{Q}_{\text{MLE}}$, which under regularity conditions is efficient and asymptotically normal (*e.g.*, Lehmann and Casella, 2006) and hence it is asymptotically optimal under the squared loss. Theorem 2 says that asymptotically, there cannot be any relevant error assessor $\hat{\delta}_{\text{MLE}} \in \mathcal{E}_\iota(\hat{Q}_{\text{MLE}})$ that is asymptotically correlated with the actual error $\delta_{\text{MLE}} = \hat{Q}_{\text{MLE}} - Q$. When $\{\hat{\delta}_{\text{MLE}}, \delta_{\text{MLE}}\}$ are jointly asymptotically normal, then Theorem 2 would imply that any such $\hat{\delta}_{\text{MLE}}$ will be asymptotically independent of δ_{MLE}. It is worthy noting that the same would hold for any estimator that is asymptotically normal and optimal (under quadratic loss), such as those studied in the classic work by Wald (1943) and Le Cam (1956).

Because the asymptotic variance of the MLE can be well approximated by the inverse of Fisher information, especially the observed Fisher information (Efron and Hinkley, 1978), the preceding result might lead some readers to wonder if the MLE and the observed Fisher information are asymptotically independent, or at least the MLE and the inverse of the observed Fisher information $I_{obs}^{-1}(\hat{Q})$ are asymptotically uncorrelated. The normal example given in Section 2 may be especially suggestive, since the MLE for μ, $\bar{X}_n$, is independent of $I_{obs}^{-1}(\hat{\mu}) = n/\hat{\sigma}^2_{\text{MLE}} = n^2/[(n-1)S_n^2]$. However it will be a mistake to generalize from this example.

Consider the same normal model $N(\mu, \sigma^2)$, but our goal now is to estimate the variance σ^2. The MLE for σ^2 is $\hat{\sigma}^2_{\text{MLE}} = (n-1)S_n^2/n$, and the corresponding observed Fisher information (pretending μ is known) is $I^{-1}(\hat{\sigma}^2_{\text{MLE}}) = 2\hat{\sigma}^4_{\text{MLE}}/n$; hence they have a deterministic relationship. However, this is not a contradiction to Theorem 2 because $I^{-1}(\hat{\sigma}^2_{\text{MLE}})$ is not an unbiased assessment of the actual error, but rather its variance. Since the variance is effectively an index of the *problem difficulty* for estimation (as termed in Meng, 2018), it is entirely natural to expect that the variance can vary closely with the value of the estimand. The normal mean problem is a special case because it is a location family, for which shifting the mean only changes the value of the estimand, but does not alter the difficulty of its estimation. This point is reinforced if we reparameterize σ^2 via $\eta = \log \sigma^2$, which yields $\hat{\eta}_{\text{MLE}} = \log \hat{\sigma}^2_{\text{MLE}}$ and $I^{-1}(\hat{\eta}^2_{\text{MLE}}) = 2/n$, and they are now trivially independent of each other, because $\hat{\eta}_{\text{MLE}} - \eta \sim \log \chi^2_{n-1} - \log(n-1)$ is a location family.

The consideration of the relationship between the MLE and the Fisher information provides a natural segue to the following discussion involving the relationship between inequity (14) and the Cramér-Rao low bound. As is well documented[4], the seminal work by Rao (1945) was prompted by a question raised during a lecture Rao gave in 1943 on whether there could be a small-sample counterpart of the asymptotic efficiency for MLE as captured by the Fisher information. However, the significance of this work goes beyond accenting the role of Fisher information, because the Cramér-Rao inequality can be viewed as a statistical counterpart of the fundamental Heisenberg Uncertainty Principle (Griffiths and Schroeter, 2018) via the notion of *co-variation*, as explored in the next three sections.

[4]See the video on C.R. Rao: A Life in Statistics II at `https://www.youtube.com/watch?v=eaxjUxoCx5w&t=324s`

7. Measuring co-variation without probabilistic joint-state specifications

In statistical and (ordinary) probabilistic literature, the most commonly adopted measure of the co-variation of two real-valued random variables G and H is their covariance $\mathrm{Cov}(G, H)$ (which includes correlation once G and H are standardized) defined via their joint probabilistic distribution $F_{G,H}(g, h)$:

$$\mathrm{Cov}(G, H) = \int\int (g - \mu_G)(h - \mu_h) F_{G,H}(dg, dh) = \langle (g - \mu_G), (h - \mu_h) \rangle_F, \tag{19}$$

where μ_G and μ_H are respectively the means of G and H, which, without loss of generality, we will assume to be zero for the subsequent discussions for notational simplicity. The subscript F in the inner product notation highlights the critical dependence of $\mathrm{Cov}(G, H)$ on their joint distribution $F(g, h)$. The elegant Hoeffding identity (Hoeffding, 1940)

$$\mathrm{Cov}(G, H) = \int\int [F_{G,H}(g, h) - F_G(g) F_H(h)]\, dg dh, \tag{20}$$

where F_G and F_H are the marginal (cumulative) distributions, further highlights how the covariance measures the co-variation in G and H as captured by their joint distribution, with respect to their benchmark distribution under the assumption of independence.

For HUP, it seems natural to take $G = x$, the position of a particle, and $H = p$, its momentum, to follow the standard notation in quantum mechanics. It is textbook knowledge (*e.g.* Landau and Lifshitz, 2013; Griffiths and Schroeter, 2018) that densities of the position x and momentum p are given by $|\psi(x)|^2$ and $|\varphi(p)|^2$ respectively, where $\psi(x)$ is a complex-valued position wave function, and the momentum wave function $\varphi(p)$ is a scaled Fourier transform of $\psi(x)$ in the form of

$$\varphi(p) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \psi(x)\, e^{-ipx/\hbar} dx, \tag{21}$$

where the scale factor $\hbar = h/(2\pi)$, with $h = 6.6260701 \times 10^{-34}$, the Planck's constant. Clearly, $\psi(x)$ is the inverse Fourier transform of $\varphi(p)$, and together x and p form a pair of the so-called conjugate variables (Stam, 1959).

As a statistician, once I understood how the marginal distributions for x and p were constructed, I naturally asked for their joint distribution. This is where things become intriguing or puzzling to those of us who are trained to model non-deterministic relationships via probability, because (quantum) physicists' answer would be that there is no joint probability distribution for x and p—not that they are unknown, but that there cannot be one. Unlike the mystery of deep learning to statisticians—and its winning of the Nobel Prize in physics only makes it more intriguing or puzzling—I found good clues to the inadequacy of ordinary probability for dealing the quantum world by the very fact that its mathematical modeling involves non-commutative relationships, such as between operators or matrices.

Perhaps the easiest way to see potential complications with non-commutative relationship is to consider the problem of generalizing the notion of variance to co-variance with complex-valued variables. With real-valued random variables G and H having a joint distribution F, we know variance is the co-variance of a variable with itself, that

is, $\mathrm{V}(G) = \mathrm{Cov}(G, G)$. In other words, when we link variance with an inner product, *i.e.*, $\mathrm{V}(G) = \langle G, G\rangle_F$, there is a natural extension for covariance by defining $\mathrm{Cov}(G, H) = \langle G, H\rangle_F$. However, with the ordinary definition of the co-variance, this extension works only if the inner product is symmetric, that is, $\langle G, H\rangle_F = \langle H, G\rangle_F$, since $\mathrm{Cov}(G, H) = \mathrm{Cov}(H, G)$ in the *real* world.

This is where the complex world is, literally, more complex than the real world. For two complex-valued L^2 functions $u(y)$ and $v(y)$ on $y \in \Omega$, the inner product is not symmetric, because it is defined by

$$\langle u|v\rangle_\mu \equiv \int_\Omega \bar{u}(y)v(y)\mu(dy) \neq \langle v|u\rangle_\mu \equiv \int_\Omega \bar{v}(y)u(y)\mu(dy), \tag{22}$$

where $\bar{u}$ is the complex conjugate of u, and μ is a baseline measure, which does not need to be a probabilistic measure. This non-commutative property is at the heart of quantum mechanics, as reviewed in the next Section. It can also been seen with matrix mechanics, since for any two matrices A and B or more broadly operators, in general $AB \neq BA$. The very fact that a regular joint probability specificity must render $\mathrm{Cov}(u, v) = \mathrm{Cov}(v, u)$ should remind us that whatever 'joint specification' of u and v we come up with, it will be more nuanced than a direct probabilistic distribution for $\{u, v\}$ whenever (22) rears its head. This phenomenon is not unique to the quantum world, since a similar situation happens with the notion of quasi-score functions, which can violate a symmetry requirement for genuine score functions, as reviewed in Appendix C.

However, this complication does not imply that probabilistic thinking is out the window. Because $\overline{\langle v|u\rangle_\mu} = \langle u|v\rangle_\mu$, we see that if we define $\mathrm{Cov}(u, v) = \langle u|v\rangle_\mu$, then its magnitude, $|\mathrm{Cov}(u, v)| = |\mathrm{Cov}(v, u)|$ is symmetric. Therefore, as long as $|\mathrm{Cov}(u, v)|$ is used as a measure of the magnitude of the co-variation between u and v, we can treat it as if it were the magnitude of a standard probabilistic co-variance. In other words, the concept, or at least the essence of *co-variance*, can be extended to non-probabilistic settings, and this extension perhaps can help our appreciation of HUP from a statistical perspective, as detailed in the next Section.

8. A lower resolution co-variation: co-variance of generating mechanisms

In the quantum world, we have seen that a particle's position and momentum have their respectively well-defined probability distribution, and we can express $\mathrm{V}(x) = \langle f|f\rangle_\mu$ and $\mathrm{V}(p) = \langle g|g\rangle_\mu$, where $f(x) = x\psi(x)$ and $g(p) = p\phi(p)$. It is then mathematically tempting to define $\mathrm{Cov}(x, p) = \langle f|g\rangle_\mu$ and $\mathrm{Cov}(p, x) = \langle g|f\rangle_\mu$, using the notation of the previous section. This construction is problematic starting from the very notation $\mathrm{Cov}(x, p)$, since it may suggest that we are measuring the co-variance between the position and momentum as *states*, which creates an epistemic disconnect with the understanding that *a joint statehood* of x and p does not exist or cannot be constructed in the quantum world.

However, x and p clearly have physical relationships. Indeed the so-called Stam's uncertainty principle (Stam, 1959) establishes that

$$C^2\mathrm{V}(x) - J(p) \geq 0 \quad \text{and} \quad C^2\mathrm{V}(p) - J(x) \geq 0, \tag{23}$$

where $C = 4\pi$ for standard Fourier transform, and $C = 2/\hbar$ when we use the $\hbar$-scaled Fourier

transform (21). Here $J(p)$ is the Fisher information for the density of p, $f(p)$, that is,

$$J(p) = \int_{-\infty}^{\infty} \left[\frac{d \log f(p)}{dp}\right]^2 f(p) dp, \tag{24}$$

and similarly for $J(x)$. For readers who are unfamiliar with defining Fisher information for a density itself instead of its parameter, $J(p)$ is the same as the Fisher information for the location family $f(p-\theta)$, where θ shares the same state space as p (in the current case, the real line). In the same vein, the Cramér-Rao inequality can be applied to the density itself, which leads to $\mathrm{V}(x) \geq J^{-1}(x)$ and $\mathrm{V}(p) \geq J^{-1}(p)$. Consequently, as shown in Dembo (1990) and Dembo *et al.* (1991),

$$\mathrm{V}(x)\mathrm{V}(p) \geq C^{-2} = \frac{\hbar^2}{4}, \tag{25}$$

which is the same as the usual expression of HUP proved in Kennard (1927):

$$\Delta x \Delta p \geq \frac{\hbar}{2}, \tag{26}$$

where Δx and Δp denote respectively the standard deviation of x and p. Dembo (1990) and Dembo *et al.* (1991) also used (23) to prove that HUP implies the Cramér-Rao inequality.

The Stam's uncertainty principle is elegant, and it reveals a kind of relationship between two marginal distributions that is not commonly studied in statistical literature, because it bypasses the specification of a joint distribution between x and p. However, this does not rule out—and indeed it suggests—that we can consider quantifying the relationships between the *mechanisms* that generate x and p. A mechanism can generate a single state, many states, or no states at all—which is equivalent to presenting itself as a whole—at any given circumstance, such as temporal instance. Hence quantifying relationships among mechanisms is a broader construct than that for the states they generate.

For statistical readers, a reasonable analogy is to think about the notion of *likelihood*. When we employ a likelihood, we can consider a single likelihood value (*e.g.*, at the MLE), several likelihood values (*e.g.*, likelihood ratio tests), or not any particular value but the likelihood function as a whole (*e.g.*, for Bayesian inference). By considering co-variations at the (resolution) level of mechanisms instead of states, we may find it less foreign to contemplate indeterminacy of relationship, such as between two sets—including empty ones—of the states generated by related mechanisms.

Of course, one may wonder if any relationship between two mechanisms itself can be indeterminable. The logical answer is yes, but fortunately for quantum mechanics we do not need go that far. As any useful quantum mechanics textbook (Landau and Lifshitz, 2013; Griffiths and Schroeter, 2018) teaches us, the position mechanism and momentum mechanism can be represented mathematically via the so-called position operator $\hat{x}$ and momentum operator $\hat{p}$, to follow the notation in quantum mechanics, and they are tethered together when being applied to the same wave function $\psi(x)$ (in the position space[5]), that is

$$\hat{x} \circ \psi(x) = x\psi(x), \quad \text{and} \quad \hat{p} \circ \psi(x) = -i\hbar\psi'(x). \tag{27}$$

[5]One can define the operators equivalently in the conjugate momentum space via $\hat{p} \circ \varphi(p) = p\varphi(p)$ and $\hat{x} \circ \varphi(p) = i\hbar\varphi'(p)$, where the momentum wave function $\varphi(p)$ is the Fourier transform of $\psi(x)$ (Griffiths and Schroeter, 2018).

That is, the position operator acts on ψ by multiplying ψ with its argument, and the momentum operator acts on ψ by differentiating it, and multiplying it by $-i\hbar$, where $i = \sqrt{-1}$.

With these representations of the mechanisms, we can measure their co-variations induced by changing the state x in real line (as a univariate case) via the inner products, with respect to a common measure μ, typically Lebesgue measure. That is, we can define

$$\mathrm{Cov}(\hat{x}, \hat{p}) = \langle \hat{x} \circ \psi | \hat{p} \circ \psi \rangle_\mu = -i\hbar \int_{-\infty}^{\infty} x \bar{\psi}(x) \psi'(x)\, dx; \tag{28}$$

$$\mathrm{Cov}(\hat{p}, \hat{x}) = \overline{\mathrm{Cov}(\hat{x}, \hat{p})} = i\hbar \int_{-\infty}^{\infty} x \psi(x) \bar{\psi}'(x)\, dx = -i\hbar \left(1 + \int_{-\infty}^{\infty} x \bar{\psi}(x) \psi'(x)\, dx\right). \tag{29}$$

Here the last equality is obtained by integration by parts and by using the fact that $|\psi(x)|^2$ is a probability density and that $x|\psi(x)|^2$ vanishes at $x = \pm\infty$ (because physicists assume the mean position is finite). Together, expressions (28)-(29) imply that

$$\mathrm{Cov}(\hat{x}, \hat{p}) - \mathrm{Cov}(\hat{p}, \hat{x}) = i\hbar, \tag{30}$$

which is also the consequence of the so-called *canonical commutation relation* (Griffiths and Schroeter, 2018),

$$\hat{x} \circ \hat{p} - \hat{p} \circ \hat{x} = i\hbar, \tag{31}$$

which holds because $\hat{x} \circ (\hat{p} \circ f(x)) - \hat{p} \circ (\hat{x} \circ f(x)) = i\hbar f(x)$ for any differentiable function f.

An immediate consequence of (30) is that the magnitude of the covariances between $\hat{x}$ and $\hat{p}$ is bounded below regardless of the form of the wave function $\psi(x)$. This is because for any complex number z, $|z|^2 \geq |\mathrm{Im}(z)|^2 = |(z - \bar{z})/2i|^2$. Hence the identity (30) implies that

$$|\mathrm{Cov}(\hat{x}, \hat{p})|^2 \geq \left[\frac{\mathrm{Cov}(\hat{x}, \hat{p}) - \mathrm{Cov}(\hat{p}, \hat{x})}{2i}\right]^2 = \frac{\hbar^2}{4}. \tag{32}$$

As reviewed in the next section, inequality (32) implies HUP in the form of (26), just as Stam's uncertainty principle does. For that purpose, it is worth pointing out that marginally,

$$\mathrm{V}(\hat{x}) = \langle \hat{x} \circ \psi | \hat{x} \circ \psi \rangle_\mu = \int_{-\infty}^{\infty} x^2 \bar{\psi}(x) \psi(x) dx = \int_{-\infty}^{\infty} x^2 |\psi(x)|^2 dx; \tag{33}$$

$$\mathrm{V}(\hat{p}) = \langle \hat{p} \circ \psi | \hat{p} \circ \psi \rangle_\mu = \hbar^2 \int_{-\infty}^{\infty} \bar{\psi}'(x) \psi'(x)\, dx = \int_{-\infty}^{\infty} p^2 |\varphi(p)|^2 dp, \tag{34}$$

where the last equation in (34) is due to the fact that $\varphi(p)$ is the ($\hbar$-scaled) Fourier transformation of $\psi(x)$, as given in (21). These two equalities tell us that when we consider either the position or the momentum by itself, its mechanism-level variance, $\mathrm{V}(\hat{x})$ or $\mathrm{V}(\hat{p})$, and the state-level variance, $\mathrm{V}(x)$ or $\mathrm{V}(p)$, are the same. This renders the unity between the mechanism-level representation (as a distribution or operator) and the state-level representation (as a observable or latent variable), a distinction seldom made conceptually under the ordinary probability framework. However, this distinction can be crucial once we go outside the regular probability framework, as in the current context of measuring co-variations between the position and momentum.

9. Bounding co-variations: A commonality of uncertainty principles

With co-variances constructed broadly, we can study the similarities and differences between inequality (14) and the Cramér-Rao inequality, as well as their intrinsic connections with HUP. Specifically, both inequalities are based on bounding joint variations of two random objects, say, G and H, by their marginal variations. For (14), under the unbiasedness assumptions and using the notation given in Section 5, if we write $G = \delta_{\hat{Q}}$ and $H = \hat{\delta}_{\hat{Q}}$, then inequality (14) is the consequence of (omitting subscript s):

$$\mathrm{Cov}^2(G, H) \le \mathrm{V}(H)\left[\mathrm{V}(G) - R^{\mathrm{opt}}\right]. \tag{35}$$

For the Cramér-Rao inequality, we can take the same $G = \delta_{\hat{Q}} = \hat{Q} - Q$, where $\hat{Q}$ is an unbiased estimator for Q. We then let $H = S(\theta|D)$, the score function from a sampling model of our data D, $f(D|\theta)$, with $Q = Q(\theta)$. It is known that the Cramér-Rao inequality is the same as (*e.g.*, Lehmann and Casella, 2006)

$$[Q'(\theta)]^2 = \mathrm{Cov}^2(G, H) \le \mathrm{V}(H)\mathrm{V}(G), \tag{36}$$

where $Q'(\theta)$ is the derivative for $Q(\theta)$. (When $Q(\theta)$ is not differentiable, we can apply the bound given by Chapman and Robbins (1951)) in terms of likelihood ratio or elasticity.)

Evidently, inequality (36) is an application of the Cauchy-Schwartz inequality. In contrast, inequality (35) delivers a more precise bound because of the subtraction of the term R^{opt}. Indeed, inequality (35) is often an equality because the condition in (II) of Theorem 1 frequently holds in practice. Given the two inequalities share the same type of G, the difference must be attributable to something distinctive between the two H's. Whereas both H's have zero expectation, the first $H = \hat{\delta}_{\hat{Q}}$ is a *statistic*, required to be a function of data D only. In contrast, the second $H = S(\theta|D)$ is a *random function*, depending on both data D and the unknown θ. Since the actual error $\delta_{\hat{Q}} = \hat{Q} - Q(\theta)$ is also a random function, the second H can co-variate with G to a greater extent than the first H can. Consequently, $\mathrm{Cov}^2(G, H)$ can reach a looser upper bound in (36) than in (35). As an illustrative example, for estimating the normal mean under $N(\mu, \sigma^2)$, $Q = \bar{X}_n - \mu$ and $H = S(\mu|X) = n(\bar{X}_n - \mu)/\sigma^2 = nG/\sigma^2$, and hence (36) becomes equality, whereas such an H is clearly not permissible for (35).

Nevertheless, both inequalities reveal the tension between individual variations—features of their respective marginal distributions—and their co-variation, which reflects their relationships, probabilistic or not. For (36), in order to keep $\mathrm{Cov}^2(G, H)$ at the value of $[Q'(\theta)]^2 > 0$, the two variances $\mathrm{V}(H)$ and $\mathrm{V}(G)$ cannot be simultaneously small to an arbitrary degree, just as a rectangle cannot have arbitrarily small sides simultaneously when its area is bounded away from zero. This restriction leads to the Cramér-Rao lower bound. In (36), we purposefully write the Fisher information as the variance of the score function instead of the expectation of its negative derivative. The variance expression makes it clearer the co-variation essence of Cramér-Rao inequality, and draws a direct parallel with the inequality underlying HUP.

Specifically, using the notation and the inequality (32) of Section 8 and taking $G = \hat{x}$ and $H = \hat{p}$, we have

$$\frac{\hbar^2}{4} \le |\mathrm{Cov}(G, H)|^2 \le \mathrm{V}(H)\mathrm{V}(G), \tag{37}$$

Comparing (37) with (36), we see that the Cramér-Rao bound and the Heisenberg uncertainty principle are consequences of essentially the same statistical phenomenon, that is, two marginal variances necessarily compete with each for being arbitrarily small, when the corresponding covariance is constrained in magnitude from below.

In contrast, for (35), the trade-off is between the covariance and one of the marginal variances. To see this clearly, we can assume $\mathrm{V}(H) = 1$, which does not offend the assumption that $\mathrm{E}(H) = 0$. Inequality (35) then becomes

$$\mathrm{Cov}^2(G, H) \leq \mathrm{V}(G) - R^{\mathrm{opt}} = R_G, \tag{38}$$

where R_G is the regret of G. On the surface, the changes of covariance and $\mathrm{V}(G)$ appear to be coordinated instead of in competition, because the larger $\mathrm{Cov}^2(G, H)$, the larger $\mathrm{V}(G)$. The reverse holds when the inequality is equality (which often is the case), and more broadly larger $\mathrm{V}(G)$—and hence larger regret—at least allows more room for $\mathrm{Cov}^2(G, H)$ to grow. But this is exactly where the tension lies when we want to improve both the learning and error assessment; improving learning means to reduce R_G and hence have a *smaller* $\mathrm{V}(G)$, but improving error assessment requires a *larger* $\mathrm{Cov}^2(G, H)$.

10. Elementary mathematics, advanced statistics, and inspiring philosophy

Mathematically, the proof of either (36) or (37) is elementary, yet the implications of either inequality, as we know, are profound. Similarly, the inequality (35) is built upon equally elementary mathematics, and the work of Bates *et al.* (2024) has already suggested its potential impact. However, many more studies remain, particularly regarding alternative loss functions, where the relevance of error assessment may not align with covariance. From a probabilistic standpoint, a thorough theoretical exploration of the relevance of an error assessor, $\hat{\delta}$, for the true error δ should involve investigating the joint distribution of $\hat{\delta}$ and δ. In this context, irrelevance can be characterized by the independence between $\hat{\delta}$ and δ.

On a broader level, formulating a general trade-off between learning and error assessment remains a complex task. This challenge stems from the need to define and measure the actual information utilized during learning and to identify relevant replications when assessing errors. Both 'information' and 'learning' are elusive notions, having taken on numerous interpretations throughout history, many of which require a refined understanding. For instance, even in the case of classical likelihood inference within parametric models, the role of conditioning in error assessment continues to provoke theoretical and practical debates.

I was reminded of this reality by an astrostatistics project involving correcting conceptual and methodological errors in astrophysics for conducting model fitting and goodness-of-fit assessment via the popular C-statistics, which is the likelihood ratio statistic under a Poisson regression model (Cash, 1979). When the project started, I naively believed that it would be merely an exercise of applying classical likelihood theory and methods, perhaps with some clever computational tricks or approximations to render them practically efficient and hence appealing to astrophysicists.

As reported in Chen *et al.* (2024), however, the issue about whether one should condition on the MLE itself or not in the context of goodness-of-fit testing, is a rather nuanced one. The issue is closely related to the issue of conditioning on ancillary statistics,

since for testing distributional shape, the parametric parameters are *nuisance objects* (as termed in Meng, 2024) and their MLE can be intuitively perceived as locally ancillary (Cox, 1980; Severini, 1993) because the distribution shape of the MLE will be normal to the first order (under regularity conditions) despite the shape of the distribution being tested. However, it is not exactly ancillary, and to decide when conditioning is beneficial (*e.g.*, leading to a more powerful test) in any sample settling is not a straightforward matter. Higher order asymptotics can help provide insight, but communicating them intuitively is a tall order even for statisticians, let alone for astrophysicists or any scientists (including data scientists). However, regardless of whether low-level mathematics or high/tall order of statistics are involved, the ultimate challenge of contemplating and formulating uncertainty principles is epistemological, or even metaphysical. For readers interested in philosophical contemplation—and I'd expect that statisticians should be in that group because statistics is essentially *applied epistemology*[6], I highly recommend the over 50 pages entry titled "The Uncertainty Principle" by Hilgevoord and Uffink (2024) in *The Stanford Encyclopedia of Philosophy*.[7] It is an erudite and thought-provoking essay about the intellectual journey of Heisenberg's uncertainty principle. Even or perhaps especially the name "uncertainty principle" has an interesting story behind it, because initially the name did not contain either 'uncertainty' or 'principle'.

As Hilgevoord and Uffink (2024) discussed, the term *uncertainty* has multiple meanings, and it is not obvious in which sense the phenomena revealed by Heisenberg (1927) qualifies as 'uncertainty'; indeed, historically terms such as "inaccuracy, spread, imprecision, indefiniteness, indeterminateness, indeterminacy, latitude" were used by various writers for what is now known as HUP. More intriguingly, Heisenberg did not postulate the finding as any kind of *principle*, but rather as *relations*, such as "inaccuracy relations" or "indeterminacy relations". The discussions in Section 8 certainly reflect the relational nature of HUP, because it is fundamentally about the co-variation of position and momentum at the mechanism level.

The entry by Hilgevoord and Uffink (2024) invites readers to consider a fundamental question that underpins these onomasiological reflections: Is the HUP a mere epistemic constraint, or a metaphysical limitation in nature? Unsurprisingly, this question is a source of ongoing dispute among philosophers of physics and even among physicists themselves. The most well-known historical debates are Heisenberg and Bohr's Copenhagen interpretation emphasizing the metaphysical indeterminacy, and the contrasting deterministic interpretation developed by de Broglie and Bohm, known as Bohmian mechanics (Hilgevoord and Uffink, 2024).

Given I have already greatly exceeded the deadline to submit this essay, I will refrain from revealing any further thrills provided in Hilgevoord and Uffink (2024), such as more recent debates about HUP, leaving readers to enjoy their own treasure hunt. But I will

[6]This was a characterization given by philosopher Hanti Lin during the JSM 2024, where Hanti and I co-organized a session where each philosopher presented for 20 minutes followed by a 15-min discussion by a statistician, and there were three pairs in total. (I made a mistake that embodied the statisticians' modesty: the estimated room size I provided to the JSM meeting department had an unacceptably negative bias.)

[7]SEP is simply a fountain of afflatus and a *Who's Who* in philosophy. Indeed SEP was where I came across Hanti Lin's 115-page entry on "Bayesian Epistemology" (Lin, 2024a), and led to my invitation to Hanti to serve as a co-editor to establish the "Meta Data Science" column (`https://hdsr.mitpress.mit.edu/meta-data-science`) for *Harvard Data Science Review*.

mention that this question has prompted me to wonder whether inequality (14) also suggests that any effort to assess the actual error is antithetical to probabilistic learning.

This is because the crux of probabilistic learning—unlike deterministic approaches, such as solving algebraic equations—lies in using distributions as our fundamental mathematical vehicles for carrying our states of knowledge (or lack thereof) and for transporting data into information that furthers learning. From this distributional perspective, assessing the actual error means to assess the *distribution* of the actual error, which is all we need to, for example, provide the usual confidence regions. It does suffer from the leap of faith problem as discussed in Section 4, but then that is a universal predicament to any form of empirical learning, as far as I can imagine.

11. From uncertainty principles to happy marriages...

A further inspiration from Hilgevoord and Uffink (2024) is its discussion on the relationship between the original semi-quantitative argument made by Heisenberg (1927) and the mathematical formalism established by Kennard (1927). Kennard's inequality (26) is precise, but can be perceived as narrow, for instance, in its reliance on standard deviation to describe "uncertainty." A similar limitation applies to inequality (14), which assesses relevance through linear correlation, a measure that surely is not universally appropriate for capturing the notion of relevance.

More broadly, much remains to be examined regarding the trade-offs between the flexibility of qualitative frameworks, which embrace the nuances and ambiguities of natural language, and the rigor of quantitative formulations, which offer the precision of mathematical language but often at the risk of being overly restrictive or idealized. Reflecting on these trade-offs is essential to learning. Statisticians and data scientists, in particular, can draw from centuries of philosophical inquiry into epistemology, as exemplified by the discussions surrounding the HUP and the like. In truth, when thoughtfully practiced, data science embodies—or ought to embody—a harmonious blend of quantitative and qualitative thinking and reasoning. This was the central theme of my *Harvard Data Science Review* editorial, "Data Science: A Happy Marriage of Quantitative and Qualitative Thinking?" (Meng, 2021), inspired by Tanweer *et al.* (2021)'s compelling article, "Why the Data Revolution Needs Qualitative Thinking." Maintaining this harmony, akin to sustaining a functioning marriage, requires commitment from all parties and a willingness to compromise. Ultimately, it calls for the wisdom to recognize that individual fulfillment and happiness—whether in marriage, mentorship, or mind melding or mating—depends profoundly on collective well-being. Professor Rao certainly embodied this wisdom.

I vividly recall my first visit to Pennsylvania State University as a seminar speaker, shortly after Professor Rao's 72nd birthday on September 10, 1992. During the seminar lunch, Professor Rao graciously joined us. We—students and early-career researchers (myself included, back when my hair was dense almost surely everywhere)—felt honored by his presence. All questions naturally revolved around statistics, except for one that made us all chuckle: "Professor Rao, how does one live a long and happy life?"

Without missing a beat, and with his characteristic paced, confident cadence, Rao replied, "Keep your wife happy."

12. A prologue or an invitation

For those who would like this article to conclude with a statistical Q&A: During the elevator ride following my seminar, which carried the seemingly oxymoronic title "A Bayesian p-value" (a deliberate contrast to the title of Meng (1994)), Professor Rao turned to me and asked, "Do people still use p-values?" To which I responded...

Well, I'll leave that as a missing data point, inviting you to impute your own favorite answer. Alternatively, if you prefer, find a deliberately embedded mathematical (but petty) error in this article and exchange it for the answer by emailing `meng@stat.harvard.edu` (as long as God permits me to respond).

Acknowledgments

I am deeply grateful to physicists Aurore Courtoy, Louis Lyons, Thomas Junk, and Pavel Nadolsky, as well as statistician Yazhen Wang, for their careful and patient explanations regarding the non-existence of a joint probabilistic distribution of a particle's position and momentum. I am equally indebted to Hanti Lin for elucidating the philosophical debates surrounding the Heisenberg Uncertainty Principle.

My thanks also extend to editor Bhramar Mukherjee, to whom I owe a profound debt, and to Peter Bickel, Joe Blitzstein, Radu Craiu, Walter Dempsey, Benedikt Höltgen, Peter McCullagh, Pavlos Msaouel, Steve Stigler, Robert Tibshirani, Théo Voldoire, and Bob Williamson for collectively providing insightful comments and sharing relevant literature—some of which may inspire a sequel to this essay.

I also thank Julie Vu and Sicheng Zhou for their meticulous proofreading efforts; naturally, any remaining errors are entirely my own (though I wish they weren't!). Finally, I acknowledge partial financial support from the NSF during the period when this essay was conceived and completed.

References

Abba, M. A., Williams, J. P., and Reich, B. J. (2024). A Bayesian shrinkage estimator for transfer learning. *arXiv:2403.17321*.

Bates, S., Hastie, T., and Tibshirani, R. (2024). Cross-validation: what does it estimate and how well does it do it? *Journal of the American Statistical Association*, **119**, 1434–1445.

Berger, J., Meng, X. L., Reid, N., and Xie, M. (2024). *Handbook of Bayesian, Fiducial, and Frequentist Inference*. CRC Press.

Blitzstein, J. K. and Hwang, J. (2014). *Introduction to Probability*. CRC Press, Boca Raton, FL, 1st edition.

Casella, G. and Berger, R. (2024). *Statistical Inference*. CRC Press.

Cash, W. (1979). Parameter estimation in astronomy through application of the likelihood ratio. *The Astrophysical Journal*, **228**, 939.

Chapman, D. G. and Robbins, H. (1951). Minimum variance estimation without regularity assumptions. *The Annals of Mathematical Statistics*, **22**, 581–586.

Chen, Y., Li, X., Meng, X. L., van Dyk, D. A., Bonamente, M., and Kashyap, V. (2024). Boosting C-statistics in astronomy via conditioning: More power, less computation. Technical report, Department of Statistics, University of Michigan.

Cox, D. R. (1980). Local ancillarity. *Biometrika*, **67**, 279–286.

Craiu, R. V., Gong, R., and Meng, X. L. (2023). Six statistical senses. *Annual Review of Statistics and Its Application*, **10**, 699–725.

Dembo, A. (1990). Information inequalities and uncertainty principles. *Department of Statistics, Stanford University, Stanford, CA, Technical Report*, **75**.

Dembo, A., Cover, T. M., and Thomas, J. A. (1991). Information inequalities and uncertainty principles. *IEEE Transactions on Information Theory*, **37**, 1501–1518.

Dempster, A. P. (1963). Further examples of inconsistencies in the fiducial argument. *The Annals of Mathematical Statistics*, **34**, 884–891.

Efron, B. (2020). Prediction, estimation, and attribution. *International Statistical Review*, **88**, S28–S59.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, **65**, 457–483.

Gelman, A. and Betancourt, M. (2013). Does quantum uncertainty have a place in everyday applied statistics. *Behavioral and Brain Sciences*, **36**, 285.

Gong, R. and Meng, X. L. (2021). Judicious judgment meets unsettling updating: Dilation, sure loss, and simpson's paradox. *Statistical Science*, **36**, 169–214. Discussion article with rejoinder.

Griffiths, D. J. and Schroeter, D. F. (2018). *Introduction to Quantum Mechanics*. Cambridge University Press.

Heisenberg, W. (1927). Über den anschaulichen inhalt der quantentheoretischen kinematik und mechanik. *Zeitschrift für Physik*, **43**, 172–198.

Hilgevoord, J. and Uffink, J. (2024). The uncertainty principle. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Stanford University. Spring 2024 edition.

Hillery, M., O'Connell, R. F., Scully, M. O., and Wigner, E. P. (1984). Distribution functions in physics: Fundamentals. *Physics Reports*, **106**, 121–167.

Hoeffding, W. (1940). Maßtabinvariante Korrelatiostheorie. *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin*, **5**, 179–233.

Kennard, E. H. (1927). Zur quantenmechanik einfacher bewegungstypen. *Zeitschrift für Physik*, **44**, 326–352.

Landau, L. D. and Lifshitz, E. M. (2013). *Quantum Mechanics: Non-relativistic Theory*, volume 3. Elsevier.

Le Cam, L. (1956). On the asymptotic theory of estimation and testing hypotheses. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 3, pages 129–157. University of California Press.

Lehmann, E. L. and Casella, G. (2006). *Theory of Point Estimation*. Springer Science & Business Media.

Lin, H. (2024a). Bayesian epistemology. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. 2024 Edition, originally published 2022.

Lin, H. (2024b). To be a Frequentist or Bayesian? Five positions in a spectrum. *Harvard Data Science Review*, **6**. https://hdsr.mitpress.mit.edu/pub/axvcupj4.

Liu, K. and Meng, X. L. (2014). Comment: A fruitful resolution to simpson's paradox via multiresolution inference. *The American Statistician*, **68**, 17–29.

Liu, K. and Meng, X. L. (2016). There is individualized treatment. Why not individualized inference? *Annual Review of Statistics and Its Application*, **3**, 79–111.

Lorce, C. and Pasquini, B. (2011). Quark wigner distributions and orbital angular momentum. *Physical Review D—Particles, Fields, Gravitation, and Cosmology*, **84**, 014015.

McCullagh, P. (1999). Discussion on some statistical heresies. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **48**, 34–35.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, volume 37 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, London, 2nd edition.

Meng, X. L. (1994). Posterior predictive p-values. *The Annals of Statistics*, **22**, 1142–1160.

Meng, X. L. (2018). Statistical paradises and paradoxes in big data (I): law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics*, **12**, 685–726.

Meng, X. L. (2021). Data science: A happy marriage of quantitative and qualitative thinking? *Harvard Data Science Review*, **3**. https://hdsr.mitpress.mit.edu/pub/pger71uh.

Meng, X. L. (2024). A BFFer's exploration with nuisance constructs: Bayesian p-value, H-likelihood, and Cauchyanity. In *Handbook of Bayesian, Fiducial, and Frequentist Inference, Eds J. Berger, XL. Meng, N. Reid and M. Xie*, pages 161–187. Chapman and Hall/CRC.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, **37**, 81–91.

Rao, C. R. (1962). Efficient estimates and optimum inference procedures in large samples. *Journal of the Royal Statistical Society: Series B (Methodological)*, **24**, 46–63.

Severini, T. A. (1993). Local ancillarity in the presence of a nuisance parameter. *Biometrika*, **80**, 305–320.

Stam, A. J. (1959). Some inequalities satisfied by the quantities of information of fisher and shannon. *Information and Control*, **2**, 101–112.

Tanweer, A., Gade, E. K., Krafft, P., and Dreier, S. (2021). Why the data revolution needs qualitative thinking. *Harvard Data Science Review*, **3**. https://hdsr.mitpress.mit.edu/pub/u9s6f22y.

Tóth, G. and Fröwis, F. (2022). Uncertainty relations with the variance and the quantum fisher information based on convex decompositions of density matrices. *Physical Review Research*, **4**, 013075.

Tóth, G. and Petz, D. (2013). Extremal properties of the variance and the quantum fisher information. *Physical Review A—Atomic, Molecular, and Optical Physics*, **87**, 032324.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, **54**, 426–482.

Wang, Y. (2022). When quantum computation meets data science: Making data science quantum. *Harvard Data Science Review*, **4**. https://hdsr.mitpress.mit.edu/pub/kpn45eyx.

APPENDIX

Appendix A: Derivations for the regression example in section 3

In general, the weighted estimate of θ can be written as

$$\hat{\theta}_w = \frac{\sum_{i=1}^n w_i X_i Y_i}{\sum_{i=1}^n w_i X_i^2},$$

with OLS corresponding to choosing $w_i = 1$ and BLUE given by $w_i = \sigma_i^{-2}$, for all i. Conditioning on $\mathbf{X}$ but for notational simplicity we suppress the conditioning notation in all expectations below, we have

$$\mathrm{V}(\hat{\theta}_w) = \frac{\sum_{i=1}^n w_i^2 X_i^2 \sigma_i^2}{[\sum_{i=1}^n w_i X_i^2]^2} = \frac{T_{w,\sigma}}{T_w^2}.$$

Let $\hat{r}_{w,j} = Y_j - \hat{\theta}_w X_j$. Because $\mathrm{E}(\hat{r}_{w,j}) = 0$, to calculate ρ, we only need to calculate

$$\begin{aligned}
\mathrm{E}[\hat{\theta}_w(Y_j - \hat{\theta}_w X_j)] &= \frac{\sum_{i=1}^n w_i X_i \mathrm{E}[Y_i Y_j]}{T_w} - \frac{\mathrm{E}[\sum_{i=1}^n w_i X_i Y_i]^2 X_j}{T_w^2} \\
&= \frac{\sum_{i=1}^n w_i X_i [\mathrm{Cov}(Y_i, Y_j) + \theta^2 X_i X_j]}{T_w} - \frac{[\sum_{i=1}^n w_i^2 X_i^2 \sigma_i^2 + \theta^2 T_w^2]\, X_j}{T_w^2} \\
&= \frac{(\theta^2 T_w + w_j \sigma_j^2) X_j}{T_w} - \frac{[T_{w,\sigma} + \theta^2 T_w^2]\, X_j}{T_w^2} = \frac{X_j}{T_w}\left[w_j \sigma_j^2 - \frac{T_{w,\sigma}}{T_w}\right];
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{V}(\hat{r}_{w,j}) &= \mathrm{V}\left[\frac{\sum_{i=1}^n w_i X_i (X_i Y_j - X_j Y_i)}{T_w}\right] = T_w^{-2} \mathrm{V}\left[\sum_{i \neq j}^n w_i X_i (X_i Y_j - X_j Y_i)\right] \\
&= T_w^{-2} \mathrm{E}\left\{\mathrm{V}\left[\sum_{i \neq j}^n w_i X_i (X_i Y_j - X_j Y_i) | Y_j\right]\right\} + \mathrm{V}\left\{\mathrm{E}\left[\sum_{i \neq j}^n w_i X_i (X_i Y_j - X_j Y_i) | Y_j\right]\right\} \\
&= T_w^{-2}\left\{\left[X_j^2 \sum_{i \neq j}^n w_i^2 X_i^2 \sigma_i^2\right] + \mathrm{V}\left[\sum_{i \neq j}^n w_i X_i^2 Y_j\right]\right\} \\
&= T_w^{-2}\left\{\left[X_j^2 (T_{w,\sigma} - w_j^2 X_j^2 \sigma_j^2)\right] + [T_w - w_j X_j^2]^2 \sigma_j^2\right\} \\
&= T_w^{-2}\left\{X_j^2 T_{w,\sigma} + \sigma_j^2 [T_w^2 - 2 T_w w_j X_j^2]\right\}.
\end{aligned}$$

Putting all the pieces together, we have

$$\mathrm{Corr}(\hat{\theta}_w, \hat{r}_{w,j}) = \frac{X_j\left(w_j \sigma_j^2 T_w - T_{w,\sigma}\right)}{\sqrt{T_{w,\sigma}\left[X_j^2 T_{w,\sigma} + \sigma_j^2 (T_w^2 - 2 T_w w_j X_j^2)\right]}}, \quad j = 1, 2. \tag{39}$$

For $n = 2, j = 1$, expression (39) simplifies to the desired (4) because

$$\begin{aligned}\mathrm{Corr}(\hat{\theta}_w, r_{w,1}) &= \frac{X_1 X_2^2 w_2(w_1\sigma_1^2 - w_2\sigma_2^2)}{\sqrt{[X_1^2 w_2^2 X_2^2\sigma_2^2 + w_2^2 X_2^4\sigma_1^2][w_1^2 X_1^2\sigma_1^2 + w_2^2 X_2^2\sigma_2^2]}} \\ &= \frac{X_1|X_2|(w_1\frac{\sigma_1}{\sigma_2} - w_2\frac{\sigma_2}{\sigma_1})}{\sqrt{[X_1^2\sigma_1^{-2} + X_2^2\sigma_2^{-2}][w_1^2 X_1^2\sigma_1^2 + w_2^2 X_2^2\sigma_2^2]}}.\end{aligned}$$

To calculate the relative regret (RR), we have

$$\mathrm{V}(\hat{\theta}_w) = \mathrm{V}\left[\frac{\sum_{i=1}^n w_i X_i Y_j}{T_w}\right] = \frac{w_1^2 X_1^2\sigma_1^2 + w_2^2 X_2^2\sigma_2^2}{[w_1 X_1^2 + w_2 X_2^2]^2}, \tag{40}$$

which also implies, by taking $w_i \propto \sigma_i^{-2}$,

$$\mathrm{V}(\hat{\theta}_{\mathrm{BLUE}}) = \frac{1}{(X_1^2\sigma_1^{-2} + X_2^2\sigma_2^{-2})}. \tag{41}$$

Putting together (40) and (41) yields the desired (5).

Appendix B: Derivation of (11) in section 4

Because $\hat{\delta}^2$ and δ^2 are independent given $\theta = \{\mu, \sigma^2\}$ and hence $\mathrm{Cov}(\hat{\delta}^2, \delta^2|\mu, \sigma^2) = 0$, we see over the joint replication,

$$\mathrm{Cov}(\hat{\delta}^2, \delta^2) = \mathrm{E}\left[\mathrm{Cov}(\hat{\delta}^2, \delta^2|\mu, \sigma^2)\right] + \mathrm{Cov}\left[\mathrm{E}(\hat{\delta}^2|\mu, \sigma^2), \mathrm{E}(\delta^2|\mu, \sigma^2)\right] = \frac{1}{n^2}\mathrm{V}(\sigma^2),$$

as long as the prior distribution for $\theta = \{\mu, \sigma^2\}$ is proper. Furthermore, conditioning on $\theta = \{\mu, \sigma^2\}$, $\delta^2 \sim \sigma^2\chi_1^2/n$ and $\hat{\delta}^2 \sim \sigma^2\chi_{n-1}^2/[n(n-1)]$ (where the two chi-square variables are independent of each other), we have

$$\begin{aligned}\mathrm{V}(\hat{\delta}^2) =& \mathrm{E}\left[\mathrm{V}(\hat{\delta}^2|\mu, \sigma^2)\right] + \mathrm{V}\left[\mathrm{E}(\hat{\delta}^2|\mu, \sigma^2)\right] = \frac{2}{(n-1)n^2}\mathrm{E}\left(\sigma^4\right) + \frac{1}{n^2}\mathrm{V}(\sigma^2); \\ \mathrm{V}(\delta^2) =& \mathrm{E}\left[\mathrm{V}(\delta^2|\mu, \sigma^2)\right] + \mathrm{V}\left[\mathrm{E}(\delta^2|\mu, \sigma^2)\right] = \frac{2}{n^2}\mathrm{E}\left(\sigma^4\right) + \frac{1}{n^2}\mathrm{V}(\sigma^2).\end{aligned}$$

Consequently, we see over the joint replication,

$$\mathrm{Corr}(\hat{\delta}^2, \delta^2) = \frac{\mathrm{V}(\sigma^2)}{\sqrt{2(n-1)^{-1}\mathrm{E}(\sigma^4) + \mathrm{V}(\sigma^2)}\sqrt{2\mathrm{E}(\sigma^4) + \mathrm{V}(\sigma^2)}},$$

which yields (11) because $\mathrm{E}(\sigma^4) = \mathrm{V}(\sigma^2) + [\mathrm{E}(\sigma^2)]^2$.

Appendix C: A quasi-score analogy for understanding the lack of joint probability

For statistically oriented readers, an instructive–though far from being perfect–analogy to the issue of the non-existence of a probabilistic model due to violations of symmetry or commutativity is the generalization from likelihood inference via the score function to estimation based on quasi-score functions. The correct score function, when available, provides the most efficient inference asymptotically (under regularity conditions). However, specifying the correct data-generating model often requires more information and resources than we typically possess.

In contrast, a quasi-score function only requires the specification of the first two moments of the data-generating model. This makes it a more practical and robust alternative to exact model-based inference, particularly in the presence of model misspecification. However, this robustness comes at the cost of reduced efficiency, reflecting the trade-off inherent in this approach.

Broadly speaking there are three types of pseudo scores: (I) those that are equivalent to the actual score; (II) those that are not equivalent to the actual score, but are equivalent to the score from a misspecified data generating model, and (III) those that cannot be derived from any probabilistic model.

Type (III) exists because any (differentiable) authentic score vector $(S_1(\theta), \ldots, S_d(\theta))^\top$ for a d-dimension parameter $\theta = (\theta_1, \ldots, \theta_d)^\top$ must satisfy

$$\frac{\partial S_i(\theta)}{\partial \theta_j} = \frac{\partial S_j(\theta)}{\partial \theta_i}, \quad \forall\ i, j = 1, \ldots, d, \tag{42}$$

because the corresponding (observed) Fisher information matrix, $-\frac{\partial S(\theta)}{\partial \theta}$, is symmetric. However, even some most innocent looking quasi-scores, such as for certain 2×2 contingency tables, the symmetry requirement of (42) can be easily violated, as demonstrated in Chapter 9 of McCullagh and Nelder (1989), which is an excellent source for understanding quasi scores and estimation equations in general.

The fact that violating the symmetry condition (42) rules out the possibility of being an actual score may help some of us imagine how the lack of symmetry or commutativity might rule out the existence of a probability specification, at least from a mathematical perspective. Furthermore, just as one can generalize from likelihood to quasi-likelihood of many shapes and forms—again see McCullagh and Nelder (1989)—the non-existence of a probabilistic distribution does not prevent us from forming quasi-distributions for various purposes, such as the Wigner quasiprobability distribution, which permits negative values, for position and momentum (x, p) (Hillery *et al.*, 1984; Lorce and Pasquini, 2011). Whether the mechanism-level covariances as given in (28)-(29) have the same magnitude as that from the Wigner quasiprobability distribution will be left as a homework exercise.

Publisher
Society of Statistics, Computer and Applications
Registered Office: I-1703, Chittaranjan Park, New Delhi- 110019, INDIA
Mailing Address: B-133, Ground Floor, C.R. Park, New Delhi-110019, INDIA
Tele: 011-40517662
https://ssca.org.in/
statapp1999@gmail.com
2024

Printed by : Galaxy Studio & Graphics
Mob: +91 9818 35 2203, +91 9582 94 1203
Email: galaxystudio08@gmail.com