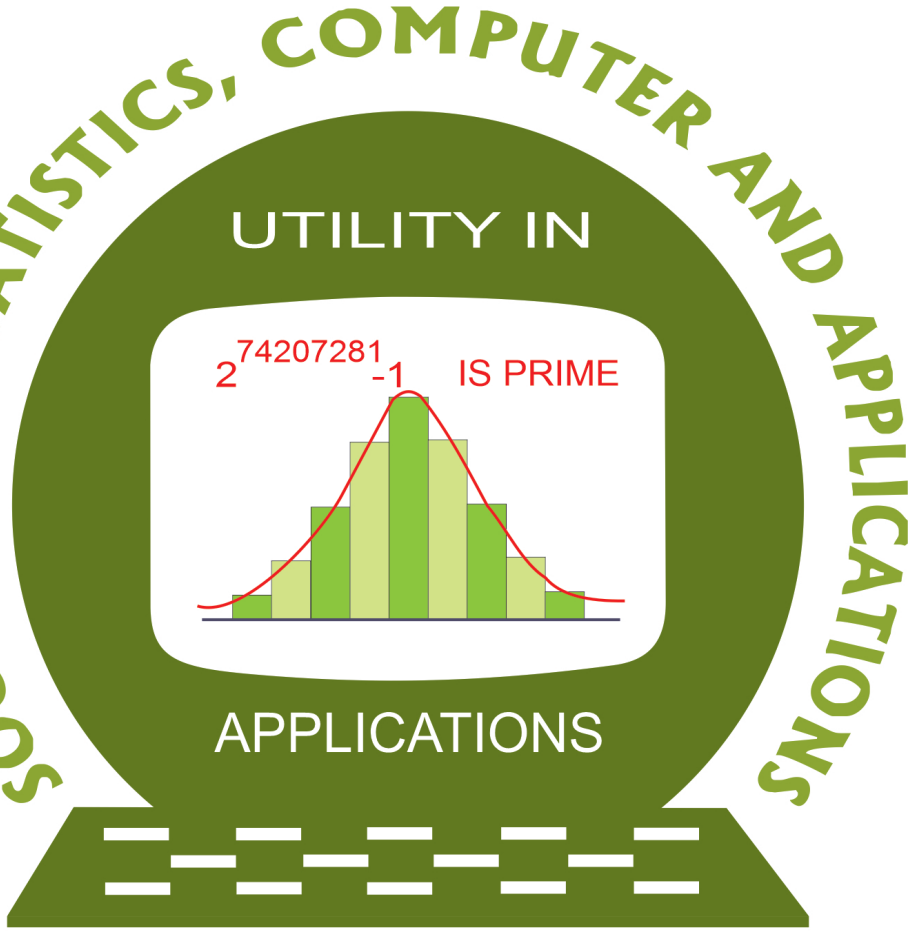
ISSN 2454-7395(online)

STATISTICS AND APPLICATIONS



FOUNDED 1998

Journal of the Society of Statistics, Computer and Applications
https://ssca.org.in/journal.html
Volume 24, No. 1, 2026 (New Series)

Society of Statistics, Computer and Applications

Council and Office Bearers

President-in-Chief

Late M.N. Das (Founder) Vinod Kumar Gupta

Presidents

Ashish Das Rajender Parsad

Executive Presidents

Jyotirmoy Sarkar Manish Sharma Manisha Pal

Patrons

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

Vice Presidents

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

Secretary

Vishal Deo

Foreign Secretary

Abhyuday Mandal

Treasurer

Ashish Das

Joint Secretaries

Bishal Gurung Raakhi Singh Shibani Roy Choudhury

Council Members

Aloke Lahiri, B. Re. Victor Babu, Banti Kumar, Dipak Roy Choudhury, Imran Khan, Mukesh Kumar, P.K. Das, Piyush Kant Rai, Rajni Jain, Ramasubramanian V., Raosaheb V. Latpate, Shalini Chandra, Sukanta Dash, V.M. Chako, Vishnu Vardhan R.

Ex-Officio Members (By Designation)

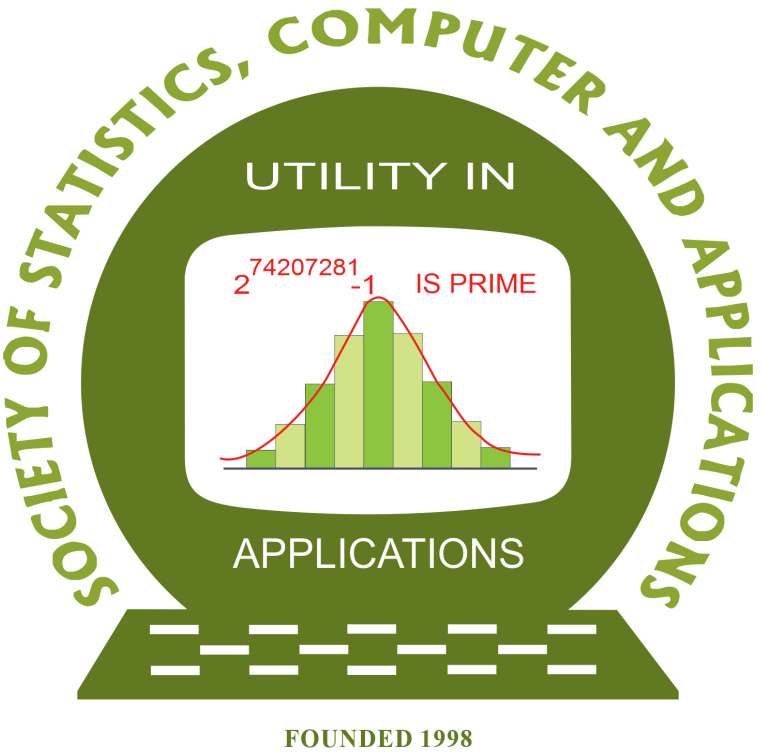
Director, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
Editor-in-Chief, Statistics and Applications
Chair Editors, Statistics and Applications
Executive Editors, Statistics and Applications

Society of Statistics, Computer and Applications

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 24, No. 1, 2026 (New Series)

We are pleased to announce the launch of the official **LinkedIn page** (https://www.linkedin.com/company/statistics-and-applications-journal-issn-2454-7395-online/)

and **YouTube channel** (https://www.youtube.com/@sscassca939)

of the journal Statistics and Applications.

These platforms are intended to facilitate dissemination and engagement with the academic community. The LinkedIn page will provide updates on journal activities, announcements, and access to published issues. The YouTube channel will host video recordings of invited talks, lectures, and other academic events organized under the auspices of the journal. In addition, video presentations corresponding to accepted papers, delivered by the respective corresponding authors, will be made available, and links to these recordings will be included in the abstracts of the published articles. Readers, authors, and researchers are encouraged to visit and follow these platforms.

Statistics and Applications

Volume 24, No. 1, 2026 (New Series)

Editorial Panel

Editor-in-Chief

V.K. Gupta, Former ICAR National Professor at IASRI, Library Avenue, Pusa, New Delhi -110012; vkgupta_1751@yahoo.co.in

Chair Editors

Durba Bhattacharya, Department of Statistics, St. Xavier's College (Autonomous), Kolkata - 700016; durba0904@gmail.com; durba@sxccal.edu

Sourish Das, Data Science Group, Chennai Mathematical Institute, H1 Sipcot IT Park, Siruseri, Chennai 603103, India; sourish.das@gmail.com; sourish@cmi.ac.in

Executive Editors

Snehanshu Saha, Computer Science and Information System, Head - APPCAIR (All Campuses), BITS Pillani K. K. Birla Goa Campus, India; snehanshus@goa.bits-pilani.ac.in

Sourabh Bhattacharya, Interdisciplinary Statistical Research Unit, Indian Statistical Unit, Kolkata, India; bhsourabh@gmail.com

Rajender Parsad (Senior), ICAR-IASRI, Library Avenue, Pusa, New Delhi - 110012; rajender1066@yahoo.co.in; rajender.parsad@icar.gov.in

Editors

Baidya Nath Mandal, Managing Editor, ICAR-Indian Agricultural Research Institute, Hazaribagh-825405, Jharkhand; mandal.stat@gmail.com

Suman Guha, Managing Editor, Department of Statistics, Presidency University, 86/1, College Street, Kolkata 700073, India; bst0404@gmail.com

Jyoti Gangwani, Production Executive, Formerly at ICAR-IASRI, Library Avenue, New Delhi 110012; jyoti0264@yahoo.co.in

Associate Editors

Abhyuday Mandal, Professor and Undergraduate Coordinator, Department of Statistics, University of Georgia, Athens, GA 30602; amandal@stat.uga.edu

Anirban Chakraborti, School of Computational and Integrative Sciences and School of Sanskrit and Indic Studies, Jawaharlal Nehru University, New Delhi 110067, India; anirban.chakraborti@gmail.com

Ashish Das, 210-C, Department of Mathematics, Indian Institute of Technology Bombay, Mumbai - 400076; ashish@math.iitb.ac.in; ashishdas.das@gmail.com

David Banks, Department of Statistical Science, Duke University, Durham, NC27708-0251 USA; david.banks@duke.edu

Deepayan Sarkar, Indian Statistical Institute, Delhi Centre, 7 SJS Sansanwal Marg, New Delhi - 110016; deepayan.sarkar@gmail.com; deepayan@isid.ac.in

Indranil Mukhopadhyay, Professor and Head, Human Genetics Unit, Indian Statistical Institute, Kolkata, India; indranilm100@gmail.com

Janet Godolphin, Department of Mathematics, University of Surrey, Guildford, GU2 7XH, UK; j.godolphin@surrey.ac.uk

Jiani Yin, Associate Director at Servier Biostatistics Project Lead Boston, MA, USA; jianiyin@gmail.com

Jyotirmoy Sarkar, Department of Mathematical Sciences, Indiana University Purdue University, Indianapolis, IN 46202-3216 USA; jsarkar@iupui.edu
K. Muralidharan, Professor, Department of Statistics, faculty of Science, Maharajah Sayajirao University of Baroda, Vadodara; lmv_murali@yahoo.com
K. Srinivasa Rao, Professor, Department of Statistics, Andhra University, Visakhapatnam, Andhra Pradesh; ksraoau@gmail.com
Lu Chen, NISS-NASS, USDA, USA, Research and Development Division, Sampling and Estimation Research Section; luchen459@gmail.com
M.R. Srinivasan, Visiting Professor, Chennai Mathematical Institute, Siruseri, Chennai-6031035, India; mrsrin8@gmail.com
Murari Singh, 65 Fulbert Crescent, Scarborough, ON M1S1C5, Canada; mandrsingh2010@gmail.com
Pranabendu Mishra, Computer Science Division, CMI, Chennai; pranabendu@cmi.ac.in
Pritam Ranjan, Indian Institute of Management, Indore - 453556; MP, India; pritam.ranjan@gmail.com
R. Vishnu Vardhan, Department of Statistics, Pondicherry University, Puducherry - 605014; vrstatsguru@gmail.com
Ramana V. Davuluri, Department of Biomedical Informatics, Stony Brook University School of Medicine, Health Science Center Level 3, Room 043 Stony Brook, NY 11794-8322, USA; ramana.davuluri@stonybrookmedicine.edu; ramana.davuluri@gmail.com
Rituparna Sen, Indian Statistical Institute Bengaluru, Karnataka 560059, India; ritupar.sen@gmail.com
S. Ejaz Ahmed, Faculty of Mathematics and Science, Mathematics and Statistics, Brock University, ON L2S 3A1, Canada; sahmed5@brocku.ca
Sat N. Gupta, Department of Mathematics and Statistics, 126 Petty Building, The University of North Carolina at Greensboro, Greensboro, NC -27412, USA; sngupta@uncg.edu
Satya Prakash Singh, Department of Mathematics and Statistics, IIT Kanpur, Kanpur 208016, UP, India; snghstyprksh@gmail.com
Satyaki Mazumdar, Indian Science Education and Research Kolkata, Mohanpur, Nadia-741246, West Bengal; satyaki@iiserkol.ac.in
Saumyadipta Pyne, Health Analytics Network, and Department of Statistics and Applied Probability, University of California Santa Barbara, USA; spyne@ucsb.edu, SPYNE@pitt.edu
Shuvo Bakar, Faculty of Medicine and Health, University of Sydney, Australia; shuvo.bakar@sydney.edu.au
Snigdhansu Chatterjee, School of Statistics, University of Minnesota, Minneapolis, MN -55455, USA; chatt019@umn.edu
Steffano Marchetti, Department of Economics and Management, University of Pisa, Via Ridolfi n, 10, 56124 Pisa, Italy; stefano.marchetti@unipi.it
T.V. Ramanathan; Department of Statistics; Savitribai Phule Pune University, Pune; madhavramanathan@gmail.com
Tapio Nummi, Faculty of Natural Sciences, Tampere University, Tampere Area, Finland; tapio.nummi@tuni.fi
Tirupati Rao Padi, Department of Statistics, Ramanujan School of Mathematical Sciences, Pondicherry University, Puducherry; drtrpadi@gmail.com
Utkarsh Tripathi, Solventum (3M Health Care), Pittsburgh Pennsylvania, USA; utkarshbitsp@gmail.com
V. Ramasubramanian, ICAR-NAARM, Rajendranagar, Hyderabad, Telangana – 500030; ram.vaidyanathan@gmail.com

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 24, No. 1, 2026 (New Series)
https://www.ssca.org.in/journal

CONTENTS

1.	Analysing the Power Gemeay Distribution: Properties and Diverse Applications *Lishamol Tomy, Anagha K. and Ahmed M. Gemeay*	1–26
2	An Inferential Study of Two Kumaraswamy Populations under Joint Ranked Set Sampling *Mahesh K. Bhingikar and D. P. Raykundaliya*	27–52
3	A Family of Additive-Multiplicative Frailty Models using the Inverse Gaussian as Frailty Distribution *Alok D. Dabade*	53–78
4	Competing Risks Analysis of Factors Influencing the Runs Scored by Top T20 Batsmen - A Survival Analysis Approach *M. Sathishkumar, M. Ramakrishnan and N. Viswanathan*	79–91
5	Understanding North Atlantic Climate Instabilities and Complex Interactions Using Data Science *Alka Yadav, Sourish Das, Anirban Chakraborti and Sudeep Shukla*	93–108
6	Stress-Strength Reliability Analysis of Power Function and Nakagami Distributions using Comparative Sampling *Surinder Kumar, Rahul Shukla and Bhupendra Meena*	109–125
7	Zero-One-Inflated Poisson-Garima Distribution and its Applications in Biomedical Studies *Divya A., Prasanth C. B. and Muhammed Anvar P.*	127–153
8	Semi-supervised Feature Selection using Maximum Mutual Information and Minimum Correlation through Augmented Learning *Arghya Kusum Das, Saptarsi Goswami, Amlan Chakrabarti and Basabi Chakraborty*	155–176
9	R-optimal Mixture Designs for Special Cubic Model *Mahesh Kumar Panda*	177–186
10	Poisson–Transmuted Geometric Convolution for Overdispersed Count Data *Anupama Nandi, Partha Jyoti Hazarika, Aniket Biswas, Morad Alizadeh, Hadi Saboori and Mohamed S. Eliwa*	187–220
11	Nonparametric Estimation and Analysis of Conditional Dynamic Failure Extropy in Bivariate Systems *Lekshmi Krishnan C. U. and E. I. Abdul Sathar*	221–236

12	ROC Curve for Binary Classification using X Lindley Distribution *Sandhya Singh and Saebugari Balaswamy*	237–248
13	Reliability Analysis of a Phased Mission System under Degradation using Wiener Process and Copulas *Satya Rani and Preeti Wanti Srivastava*	249–274
14	A Hyperspectral and Deep Learning Approach for Wheat Yield Prediction *Mohit Kumar, Alka Arora, Sudeep Marwaha, Viswanathan Chinnusamy, Sudhir Kumar, Soumen Pal, Mrinmoy Ray and Rajkumar Dhakar*	276–296
15	Bayesian Nonparametrics for Gene-Gene and Gene-Environment Interactions in Case-Control Studies: A Synthesis and Extension *Durba Bhattacharya and Sourabh Bhattacharya*	297–321
16	The Man Who Tamed Uncertainty: Andrei Kolmogorov and the Mathematics of Chance *Jyotishka Datta*	323–349
17	Corrigendum on "Prabhu-Ajgaonkar's 1967 Result Revisited" *Bikas Kumar Sinha and Manisha Pal*	351–352

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 24, No. 1, 2026 (New Series), pp 1–26
http://www.ssca.org.in/journal

Analysing the Power Gemeay Distribution: Properties and Diverse Applications

Lishamol Tomy[1], Anagha K.[2] and Ahmed M. Gemeay[3]

[1]*Department of Statistics, Deva Matha College, Kuravilangad, India*
[2]*Department of Statistics, St.Thomas College, Palai, Arunapuram, India*
[3]*Department of Mathematics, Faculty of Science, Tanta University, Tanta 31527, Egypt*

Received: 19 January 2025; Revised: 27 April 2025; Accepted: 30 April 2025

Abstract

This study proposes a novel statistical distribution called the power Gemeay distribution by modifying the Gemeay distribution. Various statistical properties like hazard rate function and its graphics, moments and related measures, order statistics, and incomplete moments are derived. Many estimation methods like the maximum likelihood method, Anderson-Darling estimation, right-tail Anderson-Darling estimation, and more are used to estimate the parameters. A simulation study is carried out to evaluate the efficiency of the estimation techniques. Based on the analysis of two real data sets, the novel distribution is more suitable than other existing models.

Key words: Gemeay distribution; Entropy; Order statistics; Anderson-Darling estimation.

AMS Subject Classifications: 60G25, 63M10, 63M20, 62P99

The video recording of the paper made under the SSCA's Online Lecture series is available at the Youtube channel URL `https://youtu.be/tKBrcQECaac`.

1. Introduction

Numerous statistical distributions have been established in recent years. These models can be used to describe data from the real world. Researchers are focusing on creating novel models to advance lifetime data investigations. Developing new statistical distributions is one aspect of such studies. To make the corresponding models more versatile and responsive to various types of data, researchers frequently modify the existing distributions. As a result, numerous distribution families have emerged and are still developing.

Distributions such as Exponential, Lindley, and XLindley possess a singular parameter, necessitating enhanced adaptability to real-time data. It is common to modify the distribution by adding parameters to the existing one to improve adaptability.

When the global pandemic of COVID-19 struck, experts put efforts into creating a

Corresponding Author: Anagha K
Email: anaghakkallely@gmail.com

new model that would be adaptable to COVID-19 data. As part of this, several models have been produced and a number more are in the works. Different models can be used to assess different aspects, such as the number of patients with COVID-19 or patients who died from the disease. Almongy *et al.* (2021), Abu El Azm *et al.* (2021), Muse *et al.* (2021) *etc.* are some of them. Afify *et al.* (2022b) proposed a novel model Marshall-Olkin reduced Kies (MORKi) flexible for COVID-19 recovery data.

Gemeay *et al.* (2023a) introduced a general two-parameter distribution (GTPD), a family of two-parameter continuous distributions, to extend the flexibility to the models. A specific case of the distribution, the Gemeay distribution (GD), was also proposed, which was more adaptable for the COVID-19 data than other models such as the exponential distribution, Frechet distribution, and Lindley distribution.

This study focuses on modifying GD by introducing an additional parameter through power transformation and analysing the properties of the distribution. The power transformation technique plays a vital role in improving the adaptability of models to accommodate diverse data structures. The transformation facilitates enhanced performance in practical applications when contrasted with standard-based models. We utilised $X = T^{1/\alpha}$ as it is essential to incorporate an additional shape parameter into our proposed model since employing $X = T^{\alpha}$ will designate α as a power scaling parameter instead of a tail shaping parameter. Various estimation approaches are employed, and the model's flexibility is evaluated. The motivation behind the new distribution is to enhance the model's adaptability for fitting actual data sets, unlike existing models in the literature. The model will augment prior efforts in this field and enable the modelling of survival statistics.

The power transformations have served as a technique for generating a new class of distributions. This has been extensively utilised across diverse domains owing to its adaptability and efficiency. Ghitany *et al.* (2013) proposed power Lindley distribution by employing power transformation to the Lindley distribution. Sarhan *et al.* (2014) developed a novel distribution and employed the same methodology to present the power-transformed version of this distribution. Al-Babtain *et al.* (2021) developed a novel flexible model named the Weibull Marshall-Olkin power Lindley (WMOPL) distribution to enhance the adaptability of the power Lindley distribution. The suggested WMOPL exhibits flexibility in modelling engineering data. Afify *et al.* (2022a), Meriem *et al.* (2022), Nagy *et al.* (2023), Yıldırım *et al.* (2023), Alsadat *et al.* (2023), Gemeay *et al.* (2023b), and others, have effectively utilised the technique.

The transformation enhances the distribution's adaptability, allowing it to effectively capture complex data patterns that traditional models might overlook. The introduced parameter by transformation is essential in defining the distribution's properties, enabling it to adjust to diverse data patterns more efficiently.

The structure of the article is as follows. The new distribution is formulated in Section 2, along with the potential forms of probability density function (PDF) and its statistical characteristics such as moments, incomplete moments, stochastic orders, *etc.* are discussed in Section 3. Additionally, Section 3 provides the potential forms of the hazard rate function (HRF). We review the strategies for estimating the parameters of our distribution in Section 4. The numerical simulation is examined in Section 5, and the adaptability of the distribution for COVID-19 data is analyzed in Section 6 along with one more real data. The topic is

concluded in Section 7.

2. Formulation of power Gemeay distribution

Gemeay *et al.* (2023a) presented a general two-parameter distribution, and they derived a new statistical model called Gemeay distribution (GD), its cumulative density function (CDF) defined as follows

$$G(x) = \frac{\theta^k \left(1 - e^{-\theta x}\right) + \theta^2 \Gamma(k+1, x\theta)}{\theta^2 k! + \theta^k}, \ x, \theta, k > 0, \tag{1}$$

where $\Gamma(a, z) = \int_0^z t^{a-1} e^{-t} dt$. The GD's PDF is

$$g(x) = \frac{\theta^{k+1} e^{-\theta x} \left(\frac{1}{\theta^2} + x^k\right)}{k! + \theta^{k-2}}. \tag{2}$$

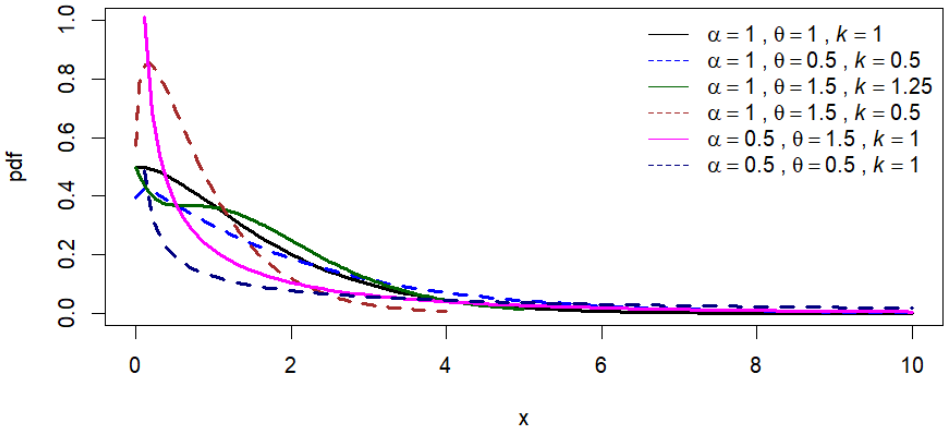
By using the following transformation $X = T^{\frac{1}{\alpha}}$ to (1), T follows GD, we obtain the CDF of the power Gemeay distribution (PGD) (for $x > 0$) as follows

$$F(x) = \frac{\theta^k \left(1 - e^{-\theta x^\alpha}\right) + \theta^2 \Gamma\left(k+1, x^\alpha \theta\right)}{\theta^2 k! + \theta^k}, \ x, \theta, k, \alpha > 0, \tag{3}$$

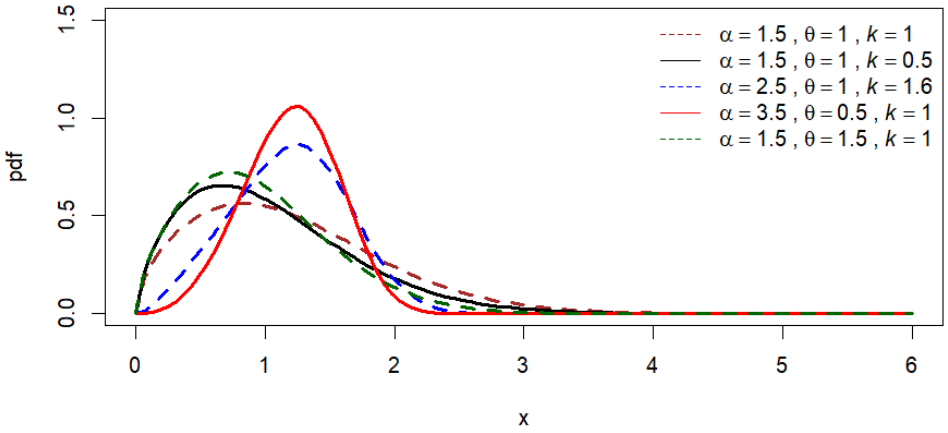
and its PDF is

$$f(x) = \frac{\alpha \theta^{k+1} x^{\alpha-1} e^{-\theta x^\alpha} \left(\theta^2 x^{\alpha k} + 1\right)}{\theta^2 k! + \theta^k}. \tag{4}$$

The PDFs of PGD for different parameter values are represented graphically in Figure 1.



(a)

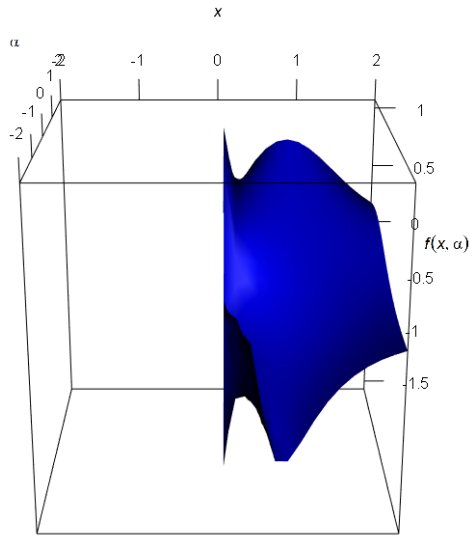


(b)

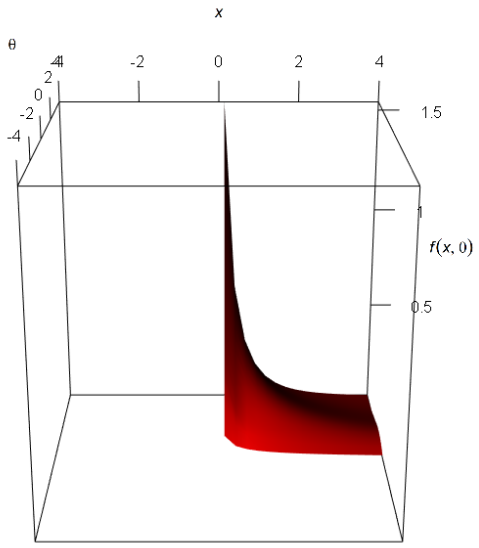
Figure 1: Plots of PDF with different parameter values

Figure 1a shows that when $\alpha \leq 1$, the PDF strictly declines for some choices of θ and k values while showing an increase initially and a decrease afterward for other choices of parameters. Figure 1b shows that the distribution is unimodal when $\alpha > 1$, regardless of the values of θ and k.

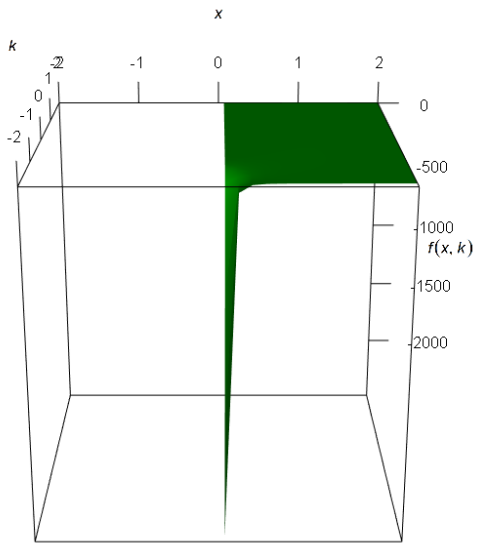
The three-dimensional (3-D) plot of the PDF of PGD with different choices of parameters are illustrated in Figure 2.



(a) PDF as a function of the variable x as well as the parameter α under the conditions of θ =0.75 and k=0.25



(b) PDF as a function of the variable x as well as the parameter θ under the conditions of α =0.5 and k=1.5



(c) PDF as a function of the variable x as well as the parameter k under the conditions of α =2.75 and θ=1.25

Figure 2: 3-D Plots of PDF with different choices of parameters

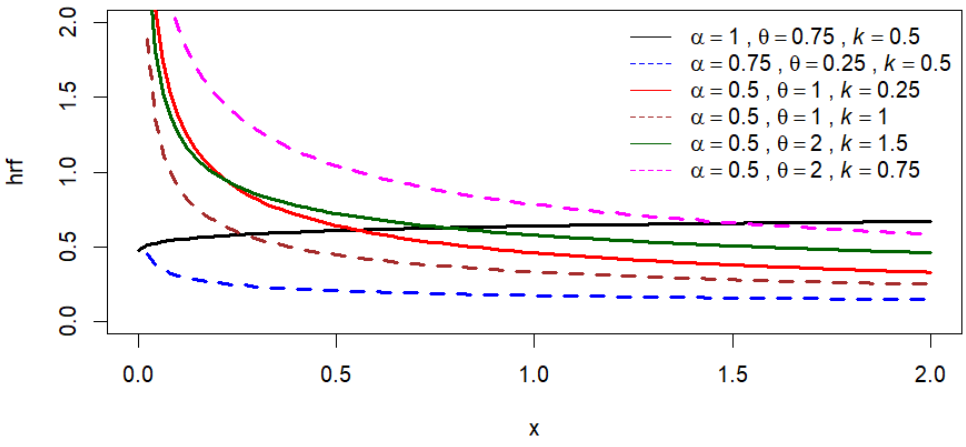
3. Statistical properties

3.1. Hazard rate function

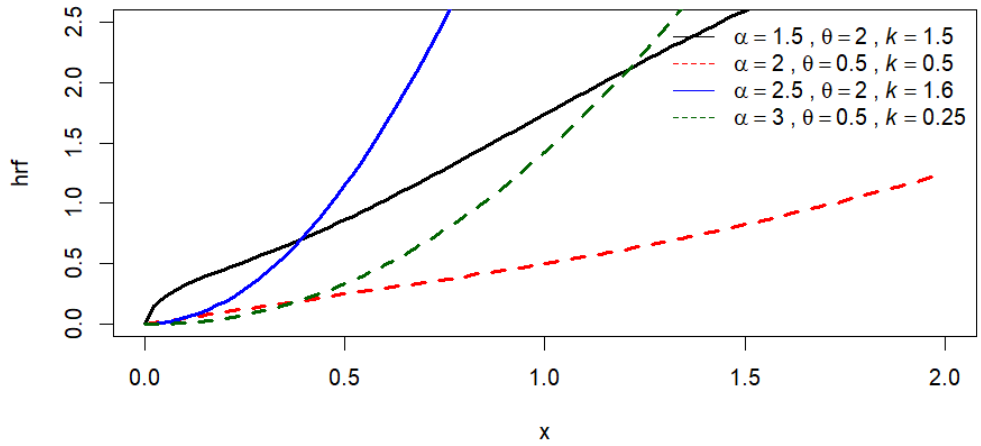
The HRF of the PGD is obtained as,

$$h(x) = \frac{\alpha\theta x^{\alpha-1}(\theta^2 x^{\alpha k}+1)}{\theta^{2-k}e^{\theta x^{\alpha}}\gamma(k+1, x^{\alpha}\theta)+1} \tag{5}$$

where $\gamma(a,z) = \int_z^{\infty} t^{a-1}e^{-t}dt$. The HRF of PGD is shown in Figure 3.



(a)



(b)

Figure 3: Plots of HRF with different parameter values

Figure 3a makes it abundantly evident that the HRF is decreasing in nature when $\alpha < 1$, regardless of the values of the other two parameters, θ and k. A constant line may be drawn on the graph when $\alpha = 1$. Figure 3b clearly shows that, regardless of θ and k values, the HRF increases in nature when $\alpha > 1$.

3.2. Moments and related measures

If $X \sim$ PGD,then the r^{th} moment of X can be obtained as follows;

$$E[X^r] = \int_0^\infty x^r f(x)dx = \frac{1}{\theta^{r/\alpha}}\left[\frac{(k+r/\alpha)!}{k! + \theta^{k-2}} + \frac{(r/\alpha)!}{\theta^{2-k}k! + 1}\right] \tag{6}$$

The first four moments can be obtained by putting $r = 1, 2, 3, 4$ respectively. Using this, we obtain the mean and variance of the distribution as follows.

$$Mean = \frac{1}{\theta^{1/\alpha}}\left[\frac{(k+1/\alpha)!}{k! + \theta^{k-2}} + \frac{(1/\alpha)!}{\theta^{2-k}k! + 1}\right]$$

$$Var(X) = \frac{1}{\theta^{2/\alpha}}\left[\frac{(k+2/\alpha)!}{k + \theta^{k-2}} + \frac{(2/\alpha)!}{\theta^{2-k}k! + 1} - \left(\frac{(k+1/\alpha)!}{k! + \theta^{k-2}} + \frac{(1/\alpha)!}{\theta^{2-k}k! + 1}\right)^2\right]$$

The skewness and kurtosis of PGD are derived as follows.

$$Skewness = \frac{E[X^3]}{(Var(X))^{3/2}} = \frac{\frac{1}{\theta^{3/\alpha}}\left[\frac{(k+3/\alpha)!}{k!+\theta^{k-2}} + \frac{(3/\alpha)!}{\theta^{2-k}k!+1}\right]}{\left(\frac{1}{\theta^{2/\alpha}}\left[\frac{(k+2/\alpha)!}{k+\theta^{k-2}} + \frac{(2/\alpha)!}{\theta^{2-k}k!+1} - \left(\frac{(k+1/\alpha)!}{k!+\theta^{k-2}} + \frac{(1/\alpha)!}{\theta^{2-k}k!+1}\right)^2\right]\right)^{3/2}}$$

$$Kurtosis = \frac{E[X^4]}{(Var(X))^2} = \frac{\frac{1}{\theta^{4/\alpha}}\left[\frac{(k+4/\alpha)!}{k!+\theta^{k-2}} + \frac{(4/\alpha)!}{\theta^{2-k}k!+1}\right]}{\left(\frac{1}{\theta^{2/\alpha}}\left[\frac{(k+2/\alpha)!}{k+\theta^{k-2}} + \frac{(2/\alpha)!}{\theta^{2-k}k!+1} - \left(\frac{(k+1/\alpha)!}{k!+\theta^{k-2}} + \frac{(1/\alpha)!}{\theta^{2-k}k!+1}\right)^2\right]\right)^2}$$

The coefficient of variation is given as,

$$C.V = \frac{\sqrt{Var(x)}}{E(X)} = \frac{\left(\frac{1}{\theta^{2/\alpha}}\left[\frac{(k+2/\alpha)!}{k+\theta^{k-2}} + \frac{(2/\alpha)!}{\theta^{2-k}k!+1} - \left(\frac{(k+1/\alpha)!}{k!+\theta^{k-2}} + \frac{(1/\alpha)!}{\theta^{2-k}k!+1}\right)^2\right]\right)^{1/2}}{\frac{1}{\theta^{1/\alpha}}\left[\frac{(k+1/\alpha)!}{k!+\theta^{k-2}} + \frac{(1/\alpha)!}{\theta^{2-k}k!+1}\right]}$$

The moment-generating function can be obtained as,

$$M_x(t) = \int_0^\infty e^{tx} f(x)dx = \frac{\alpha\theta^{k+1}}{\theta^2 k! + \theta^k}\left(\theta^2\Gamma(\alpha(k+1))(1-t)^{-\alpha(k+1)} + \Gamma\alpha(1-t)^{-\alpha}\right) \quad (7)$$

for $\alpha > 0$ and $Re(\alpha(k+1)) > 0$.
The characteristic function can be formulated by replacing t with it in Equation 7.

3.3. Entropy

The entropy can be defined as a measure of uncertainty or randomness. The entropy defined by Rényi (1961) is as follows,

$$H_z(X) = \frac{1}{1-z}log\left(\int_0^\infty f^z(x)dx\right) = \frac{1}{1-z}log\left[\left(\frac{\alpha\theta^{k+1}}{\theta^2 k! + \theta^k}\right)^z \sum_{t=0}^{z}\binom{z}{t}\frac{\theta^{2t}}{\alpha}\frac{\Gamma(kt + z/\alpha + 1)}{\theta^{kt+z/\alpha+1}}\right]$$

$H_z(x)$ tends to Shannon entropy as $z \to 1$.

3.4. Order statistics

Let $X_1, X_2, ...X_n$ be n random sample taken from a population. If the variables are arranged in ascending order of magnitude such that $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$, then $x_{(i)}$ is said to be the i^{th} order statistic, for $i = 1, 2, ..., n$ and specifically, $x_{(1)}$ is the first order statistic, and $x_{(n)}$ is the n^{th} order statistic.

The PDF of r^{th} order statistic $x_{(r)}$ can be obtained in general as

$$f_r(x) = \frac{n!}{(r-1)!(n-r)!}[F(x)]^{r-1}[1-F(x)]^{n-r}f(x) \quad (8)$$

For PGD, this can be derived as,

$$f_r(x) = \frac{n!}{(r-1)!(n-r)!}\left[\frac{\theta^k(1-e^{-\theta x^\alpha}) + \theta^2\Gamma(k+1, x^\alpha\theta)}{\theta^2 k! + \theta^k}\right]^{r-1} \times \left[\frac{\theta^2 k! + \theta^k - \theta^k(1-e^{-\theta x^\alpha}) - \theta^2\Gamma(k+1, x^\alpha\theta)}{\theta^2 k! + \theta^k}\right]^{n-r} \times \left[\frac{\alpha\theta^{k+1}x^{\alpha-1}e^{-\theta x^\alpha}(\theta^2 x^{\alpha k} + 1)}{\theta^2 k! + \theta^k}\right]$$

Putting $r = 1$, the PDF of first-order statistic is given as,

$$f_r(x) = n\left[\frac{\theta^2 k! + \theta^k - \theta^k(1 - e^{-\theta x^\alpha}) - \theta^2\Gamma(k+1, x^\alpha\theta)}{\theta^2 k! + \theta^k}\right]^{n-1} \times \left[\frac{\alpha\theta^{k+1}x^{\alpha-1}e^{-\theta x^\alpha}(\theta^2 x^{\alpha k} + 1)}{\theta^2 k! + \theta^k}\right]$$

The PDF of n^{th} order statistic is obtained as follows, by putting $r = n$.

$$f_r(x) = n\left[\frac{\theta^k(1 - e^{-\theta x^\alpha}) + \theta^2\Gamma(k+1, x^\alpha\theta)}{\theta^2 k! + \theta^k}\right]^{n-1} \times \left[\frac{\alpha\theta^{k+1}x^{\alpha-1}e^{-\theta x^\alpha}(\theta^2 x^{\alpha k} + 1)}{\theta^2 k! + \theta^k}\right]$$

3.5. Incomplete moments

According to Butler and McDonald (1989), the h^{th} incomplete moment for a density function can be defined as,

$$I(t; h) = \int_0^t x^h . f(x)dx \; ; \quad for \;\; x < t$$

For PGD, the h^{th} incomplete moment can be obtained as,

$$I(t; h) = \frac{\alpha\theta^{k-h/\alpha+1}}{\theta^2 k! + \theta^k}\left[\frac{(t^\alpha)^{2k}}{\alpha}\left(\Gamma(h/\alpha + k + 1) - \Gamma(h/\alpha + k + 1, t^\alpha\theta)\right) + \Gamma(h/\alpha) - \Gamma(h/\alpha, t^\alpha\theta)\right]$$

We have the first incomplete moment $I(t, 1)$ by putting $h = 1$ in the last equation. Moreover, the normalised incomplete moments are acquired as follows:

$$\phi(t; h) = \frac{I(t; h)}{E(y^h)} \tag{9}$$

where $E(y^h)$ is the h^{th} moment of Y where $y < t$. Putting $h = 0$ and $h = 1$ in Equation 9, we get the first two normalised incomplete moments, which are very useful in economics.

4. Estimation methods

In this section, we discussed the conventional estimation methods for determining the parameters of the PGD. These estimations are derived by optimising an objective function to maximise or minimise its value. For more information about these estimation methods, see Anderson and Darling (1952); Mukhtar *et al.* (2020); Choi and Bulgren (1968); Swain *et al.* (1988); Kao (1958); Torabi (2008).

In the context of estimating PGD parameters, the calculation of the PGD estimator involves utilising maximum likelihood estimation (MLE). This estimation technique revolves around the maximisation of the following equation.

$$\log L = n\log\left(\frac{\alpha\theta^{k+1}}{\theta^2 k! + \theta^k}\right) + \sum_{i=1}^{n}\log\left(\theta^2 x_i^{\alpha k} + 1\right) - \theta\sum_{i=1}^{n}x_i^\alpha + (\alpha - 1)\sum_{i=1}^{n}\log(x_i)$$

In the process of computing the PGD estimator, Anderson-Darling estimation (ADE) is employed, and this estimation technique revolves around the minimisation of the provided equation.

$$A(x_i) = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\log\frac{\theta^k\left(1-e^{-\theta x_{(i)}^{\alpha}}\right)+\theta^2\Gamma\left(k+1,x_{(i)}^{\alpha}\theta\right)}{\theta^2 k!+\theta^k} + \log\left(1-\frac{\theta^k\left(1-e^{-\theta x_{(n-i-1)}^{\alpha}}\right)+\theta^2\Gamma\left(k+1,x_{(n-i-1)}^{\alpha}\theta\right)}{\theta^2 k!+\theta^k}\right)\right]$$

In the process of obtaining the PGD estimator, we utilise right-tail Anderson-Darling estimation (RADE). This estimation procedure involves the minimisation of the equation provided, serving as a fundamental tool in precisely determining the PGD parameters.

$$R(x_i) = \frac{n}{2} - 2\sum_{i=1}^{n}\frac{\theta^k\left(1-e^{-\theta x_{(i)}^{\alpha}}\right)+\theta^2\Gamma\left(k+1,x_{(i)}^{\alpha}\theta\right)}{\theta^2 k!+\theta^k} - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\log\left(\frac{\theta^k\left(1-e^{-\theta x_{(i)}^{\alpha}}\right)+\theta^2\Gamma\left(k+1,x_{(i)}^{\alpha}\theta\right)}{\theta^2 k!+\theta^k}\right)$$

The PGD estimator is computed through the application of left-tailed Anderson-Darling estimation (LTADE). Its essence lies in the minimisation of the given equation.

$$L(x_i) = -\frac{3}{2}n + 2\sum_{i=1}^{n}\frac{\theta^k\left(1-e^{-\theta x_{(i)}^{\alpha}}\right)+\theta^2\Gamma\left(k+1,x_{(i)}^{\alpha}\theta\right)}{\theta^2 k!+\theta^k} - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\log\frac{\theta^k\left(1-e^{-\theta x_{(i)}^{\alpha}}\right)+\theta^2\Gamma\left(k+1,x_{(i)}^{\alpha}\theta\right)}{\theta^2 k!+\theta^k}$$

The determination of PGD parameters employs Cramér-von Mises estimation (CVME), a method that centers around the minimisation of the given equation.

$$C(x_i) = \frac{1}{12n} + \sum_{i=1}^{n}\left[\frac{\theta^k\left(1-e^{-\theta x_{(i)}^{\alpha}}\right)+\theta^2\Gamma\left(k+1,x_{(i)}^{\alpha}\theta\right)}{\theta^2 k!+\theta^k} - \frac{2i-1}{2n}\right]^2$$

In the process of computing the PGD estimator, we rely on the least-squares estimation (LSE) method, which involves minimizing the provided equation.

$$V(x_i) = \sum_{i=1}^{n}\left[\frac{\theta^k\left(1-e^{-\theta x_{(i)}^{\alpha}}\right)+\theta^2\Gamma\left(k+1,x_{(i)}^{\alpha}\theta\right)}{\theta^2 k!+\theta^k} - \frac{i}{n+1}\right]^2$$

In the computation of the PGD estimator, we employ the weighted least-squares estimation (WLSE) method, which centers on the minimisation of the equation provided.

$$W(x_i) = \sum_{i=1}^{n}\frac{(n+1)^2(n+2)}{i(n-i+1)}\left[\frac{\theta^k\left(1-e^{-\theta x_{(i)}^{\alpha}}\right)+\theta^2\Gamma\left(k+1,x_{(i)}^{\alpha}\theta\right)}{\theta^2 k!+\theta^k} - \frac{i}{n+1}\right]^2$$

To derive the PGD estimator, we employ the maximum product of spacing estimation (MPSE) method, which entails maximising the following equation.

$$\delta\left(x_i\right)=\frac{1}{n+1}\sum_{i=1}^{n+1}\log\nu_i(x_i),$$

$$\nu_i(x_i)=\frac{\theta^k\left(1-e^{-\theta x_{(i)}^{\alpha}}\right)+\theta^2\Gamma\left(k+1,x_{(i)}^{\alpha}\theta\right)}{\theta^2k!+\theta^k}-\frac{\theta^k\left(1-e^{-\theta x_{(i-1:n)}^{\alpha}}\right)+\theta^2\Gamma\left(k+1,x_{(i-1:n)}^{\alpha}\theta\right)}{\theta^2k!+\theta^k}$$

In determining the PGD estimator, the minimum spacing absolute distance estimation (MSADE) method is applied, involving minimising the provided equation.

$$\zeta\left(x_i\right)=\sum_{i=1}^{n+1}|\nu_i-\frac{1}{n+1}|$$

The calculation of the PGD estimator relies on the minimum spacing absolute-log distance estimation method (MSALDE). This method entails minimisation of the provided equation and is instrumental in accurately estimating PGD parameters.

$$\Upsilon_2\left(x_i\right)=\sum_{i=1}^{n+1}|\log\nu_i-\log\frac{1}{n+1}|$$

5. Numerical simulation

In this section, we delve into evaluating the effectiveness of various parameter estimation techniques for the PGD using an extensive dataset obtained through simulation. Our objective is to assess the performance of these estimation methods under different scenarios. Our simulations generated datasets with varying sample sizes, specifically, $n = 30$, 70, 100, 150, and 250. These datasets were generated by randomly sampling data points from the PGD's quantile function with different sets of parameter values. Our focus lies in understanding the behavior of the PGD estimators and evaluating the performance of these estimation techniques within this context. Additionally, we will employ a variety of metrics to evaluate the efficacy of different estimation methods, including an average of bias ($|Bias(\widehat{\boldsymbol{\iota}})|=\frac{1}{n}\sum_{i=1}^{n}|\widehat{\boldsymbol{\iota}}-\boldsymbol{\iota}|$), mean squared errors ($MSE=\frac{1}{n}\sum_{i=1}^{n}(\widehat{\boldsymbol{\iota}}-\boldsymbol{\iota})^2$), and mean relative errors ($MRE=\frac{1}{n}\sum_{i=1}^{n}|\widehat{\boldsymbol{\iota}}-\boldsymbol{\iota}|/\boldsymbol{\iota}$), $\boldsymbol{\iota}=(\theta,\ k,\ \alpha)$.

Tables from 5 to 12 (see Annexure) present the outcomes of simulating PGD parameters using ten distinct estimation techniques by providing the bias, MSE, and MRE for each sample size, for all the estimation techniques. Also ranks from 1 to 10 are assigned to the values starting from the lowest. It is noteworthy that all parameter estimates for the suggested distribution exhibit a high degree of reliability and proximity to their true values. The provided estimation techniques demonstrate precision, and in every considered scenario, all computed metrics show a decline as n increases. The performance of each estimation method is exceptional in identifying the proposed model parameters. Table 13 (see Annexure) provides an overview of the overall rankings for all estimation strategies and each rank is determined from the power presented in Tables 5-12. In our investigation, the MSADE method emerged as the most effective for assessing the parameter values in question, achieving a total score of 80.0, as depicted in Table 13.

6. Real data analysis

To assess the adaptability of the distribution, we apply PGD to real data sets. The data sets have been subjected to descriptive analysis, and several plots, including box plots and time to target (TTT) plots, have also been produced.

We contrast the model's adaptability with the GD and other well-known distributions like the Frechet distribution (FD), Frechet- Weibull distribution(FrWD), and Lomax distribution(LoD). To find the optimum model, metrics like the Akaike information criterion (AIC), Consistent Akaike information criterion (CAIC), Bayesian information criterion (BIC), the Kolmogrov-Smirnov (KS) statistic, and Anderson-Darling(A) statistic are utilized.

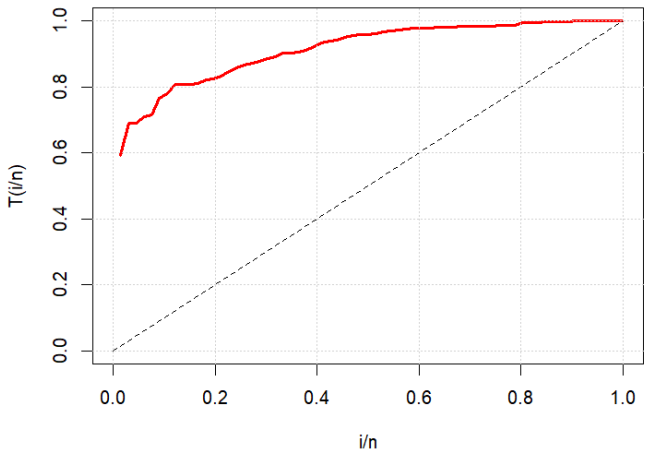
Dataset 1:

The data has been taken from Afify *et al.* (2022b) which indicates the recovery rate of COVID-19 infections in Spain from 3 March to 7 May 2020. The descriptive statistic of the data is given in Table 1.

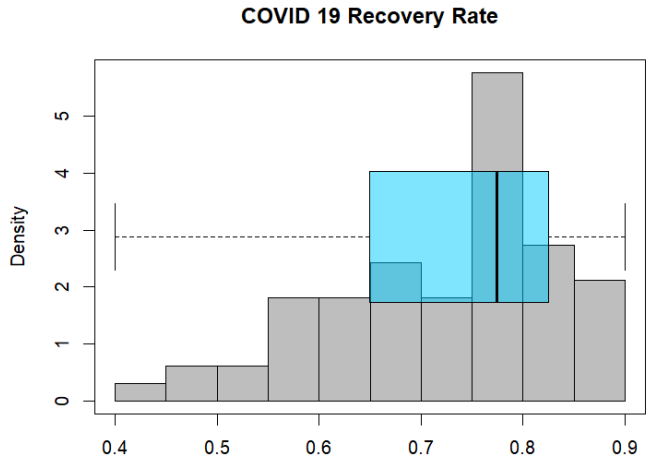
Table 1: Descriptive Statistics: COVID-19 recovery data

Minimum	1^{st} Quartile	Mean	Median	3^{rd} Quartile	Maximum	Skewness	Kurtosis
0.4286	0.6474	0.7240	0.7533	0.7975	0.8628	-0.688966	-0.4761233

Based on the data from Table 1 our data is left skewed and platykurtic in nature. It is evident from the histogram and box plot depicted in Figure 4b. Figure 4a shows the TTT plot, which is increasing.



(a) TTT plot of the COVID-19 data



(b) Histogram and box plot of COVID-19 data

Figure 4: TTT plot, Histogram, and box plot of the COVID-19 recovery data.

The goodness-of-fit statistics for the data set are shown in Table 2 together with the MLEs and SEs of the parameters of PGD and its competitors.

It is clear that PGD outperforms other distributions like GD, FrWD, FD, and LoD. The COVID data set analysis reveals that the model PGD meets all of the model selection criteria, making it the best model overall. It has a higher p-value of 0.9179 and a minimum value for statistics including AIC, CAIC, BIC and A as shown in Table 2. It makes the model adaptive for different lifetime data.

Table 2: Measures analysing goodness of fit of COVID-19 recovery data

Model	MLE (SE)	-L	AIC	CAIC	BIC	KS	A	P value (K-S test)
PGD	$\hat{\alpha}$ = 14.92(4.97) $\hat{\theta}$=14.02 (6.67) $\hat{k}$=-0.56 (0.21)	-59.76	-113.52	-113.13	-106.95	0.0682	0.6958	0.9179
GD	$\hat{\theta}$ = 3.709(0.448) $\hat{k}$=1.911 (0.333)	25.14	54.39	54.58	58.77	0.46	1.78	< 0.00001
LoD	$\hat{\alpha}$ = 17.912(9.68) $\hat{\lambda}$=12.795 (7.04)	46.45	96.913	97.104	101.29	0.481	1.942	< 0.000001
FD	$\hat{a}$=-1.47 (1.021) $\hat{m}$=2.13 (1.024) $\hat{s}$=17.31 (8.21)	-41.029	-76.059	-75.672	-69.49	0.1936	2.937	0.014
FrWD	$\hat{\alpha}$=4.47 (19.12) $\hat{\beta}$=0.976 (2.42) $\hat{\lambda}$=0.668 (1.35) $\hat{k}$=1.17 (5.04)	-35.19	-63.19	-62.53	-54.43	0.208	3.65	0.007

Figure 5 depicts the histogram of the data set along with the fitted PDF. The fitted PDF almost perfectly captures the shape of the histogram, as seen by an analysis of the diagram.

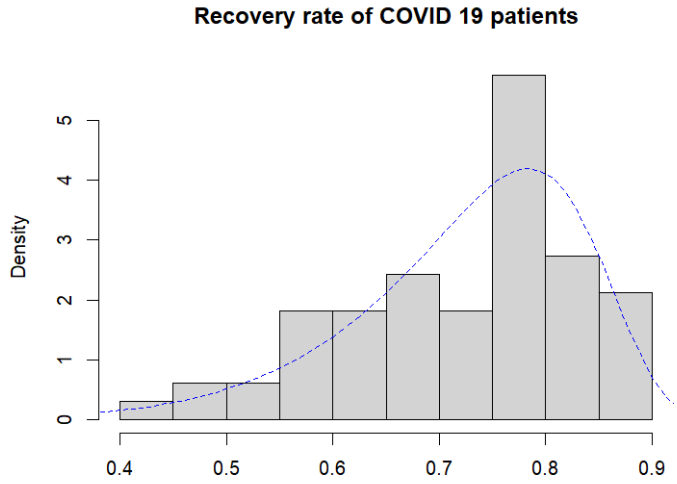


Figure 5: Histogram of COVID 19 recovery data with fitted pdf

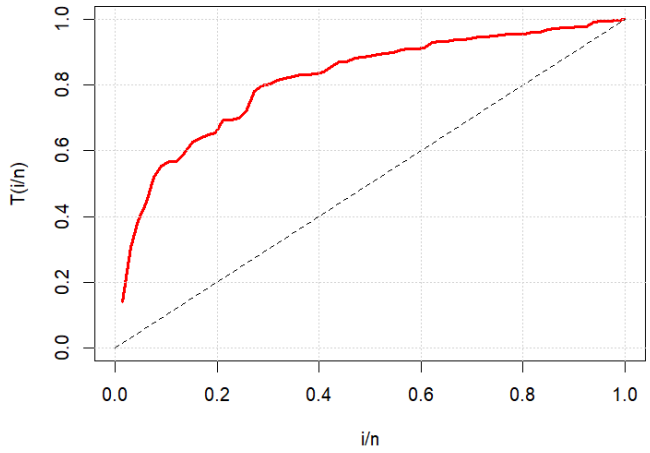
Data set 2:

The second data set, which provides the strengths of 1.5 cm glass fibers, was obtained from Nichols and Padgett (2006). Table 3 displays the descriptive statistics.

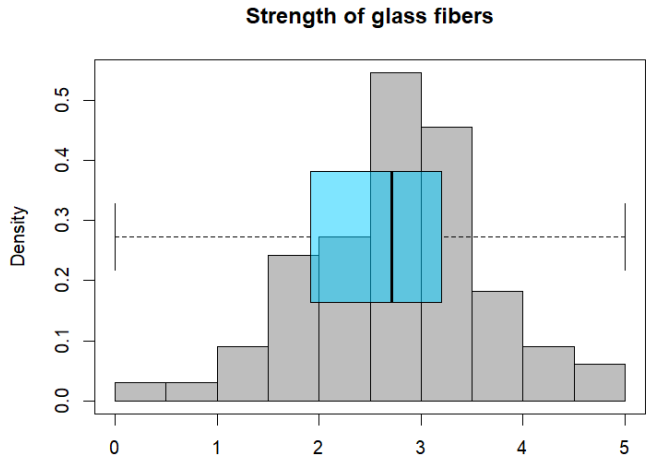
Table 3: Descriptive Statistics of strength of glass fibers data set

Minimum	1^{st} Quartile	Mean	Median	3^{rd} Quartile	Maximum	Skewness	Kurtosis
0.39	2.178	2.760	2.835	3.277	4.90	-0.1284	0.1267

It shows that our data is slightly negatively skewed and it is leptokurtic in nature and Figure 6b represents the histogram and box plot of the data set which justify the data in Table 3. Figure 6a shows the TTT plot which is increasing in nature.



(a) TTT plot of the strength of glass fibers



(b) Histogram and box plot of glass fibers data

Figure 6: TTT plot, Histogram, and box plot of the of glass fibers data.

The MLE, SE, and goodness of fit metrics for the glass fiber data for PGD and its competing models are presented in Table 4.

It is obvious that PGD performs better than other models. The model PGD is the best model overall, according to the analysis of the glass fibers data set, as it satisfies all model selection criteria. Table 4 shows the lowest values for the statistics AIC, CAIC, BIC, and A, as well as the higher p-value of 0.8427 for this study. The model becomes appropriate for various lifetime data as a result.

Table 4: Numerical values analyzing the goodness of fit for strength of glass fibers data

Model	MLE (SE)	-L	AIC	CAIC	BIC	KS	A	P value (K-S test)
PGD	$\hat{\alpha}$ = 2.163(0.376) $\hat{\theta}$=0.312 (0.196) $\hat{k}$=2.542 (0.511)	85.04	176.19	176.58	182.76	0.0758	0.3144	0.8427
GD	$\hat{\theta}$ = 3.49(0.564) $\hat{k}$=8.811 (1.48)	87.86	179.72	179.92	184.11	0.1005	0.684	0.517
LoD	$\hat{\alpha}$ = 14.38(7.31) $\hat{\lambda}$=38.51 (19.89)	137.43	274.06	274.25	278.44	0.361	1.41	<0.00001
FD	$\hat{a}$=-9.11 (4.1919) $\hat{m}$=11.38 (4.209) $\hat{s}$=12.105 (4.68)	95.48	195.88	196.275	202.457	0.1560	1.868	0.1212
FrWD	$\hat{\alpha}$=0.3375 (0.43) $\hat{\beta}$=2.0469 (9.83) $\hat{\lambda}$=1.7565 (1.82) $\hat{k}$=4.8753 (6.30)	121.19	250.39	251.045	259.148	0.23	5.38	0.0018

Figure 7 depicts the histogram of the data set along with the fitted PDF. This demonstrates that the distribution accurately depicts the form of the histogram.

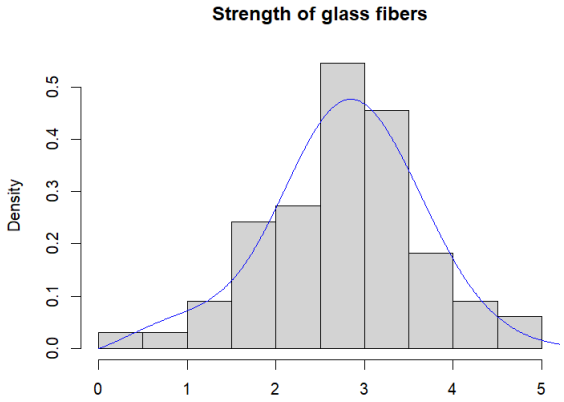


Figure 7: Histogram of glass fibers data with fitted pdf

7. Conclusion

In this paper, we proposed a new power Gemeay distribution with parameters α, θ and k. The PDF, CDF, and HRF are formulated, and the behavior of the PDF and HRF is assessed by evaluating the plots. Furthermore, some statistical properties are derived. The parameters are estimated using several methods like MLE, ADE, RADE, LTADE, CVME, LSE, WLSE, MPSE, MSADE, and MSALDE. A simulation study assesses the effectiveness of all ten estimation techniques. It should be emphasized that all of the distributional estimations show a high degree of accuracy and reliability concerning their true values. The best-fitted model among some existing distributions and the suggested distribution is determined by analysis of two real data sets, and it exceeds all other mentioned distributions in terms of the goodness of fit criterion.

Acknowledgements

We are sincerely grateful for the Editor's help and advice. We are extremely thankful of the reviewer's for insightful comments and suggestions.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Abu El Azm, W. S., Almetwally, E. M., Naji AL-Aziz, S., El-Bagoury, A. A. A. H., Alharbi, R., and Abo-Kasem, O. E. (2021). A new transmuted generalized Lomax distribution: Properties and applications to COVID-19 data. *Computational Intelligence and Neuroscience*, 1–14.

Afify, A. Z., Gemeay, A. M., Alfaer, N. M., Cordeiro, G. M., and Hafez, E. H. (2022a). Power-modified kies-exponential distribution: Properties, classical and Bayesian inference with an application to engineering data. *Entropy*, **24**, 883.

Afify, A. Z., Nassarand, M., Kumar, D., and Cordeiro, G. M. (2022b). A new unit distribution: Properties, inference, and applications. *Electronic Journal of Applied Statistical Analysis*, **15**, 438–462.

Al-Babtain, A. A., Kumar, D., Gemeay, A. M., Dey, S., and Afify, A. Z. (2021). Modeling engineering data using extended power-Lindley distribution: Properties and estimation methods. *Journal of King Saud University-Science*, **33**, 101582.

Almongy, H. M., Almetwally, E. M., Aljohani, H. M., Alghamdi, A. S., and Hafez, E. H. (2021). A new extended Rayleigh distribution with applications of COVID-19 data. *Results in Physics*, **23**, 104012.

Alsadat, N., Ahmad, A., Jallal, M., Gemeay, A. M., Meraou, M. A., Hussam, E., Elmetwally, E. M., and Hossain, M. M. (2023). The novel Kumaraswamy power Frechet distribution with data analysis related to diverse scientific areas. *Alexandria Engineering Journal*, **70**, 651–664.

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, **23**, 193–212.

Butler, R. J. and McDonald, J. B. (1989). Using incomplete moments to measure inequality. *Journal of Econometrics*, **42**, 109–119.

Choi, K. and Bulgren, W. G. (1968). An estimation procedure for mixtures of distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, **30**, 444–460.

Gemeay, A. M., Halim, Z., Abd El-Raouf, M. M., Hussam, E., Abdulrahman, A. T., Mashaqbah, N. K., Alshammari, N., and Makumi, N. (2023a). General two-parameter distribution: Statistical properties, estimation, and application on COVID-19. *Plos One*, **18**, e0281474.

Gemeay, A. M., Karakaya, K., Bakr, M., Balogun, O. S., Atchadé, M. N., and Hussam, E. (2023b). Power Lambert uniform distribution: Statistical properties, actuarial measures, regression analysis, and applications. *AIP Advances*, **13**.

Ghitany, M. E., Al-Mutairi, D. K., Balakrishnan, N., and Al-Enezi, L. (2013). Power Lindley distribution and associated inference. *Computational Statistics & Data Analysis*, **64**, 20–33.

Kao, J. H. K. (1958). Computer methods for estimating Weibull parameters in reliability studies. *IRE Transactions on Reliability and Quality Control*, 15–22.

Meriem, B., Gemeay, A. M., Almetwally, E. M., Halim, Z., Alshawarbeh, E., Abdulrahman, A. T., El-Raouf, M. A., and Hussam, E. (2022). [retracted] the power xLindley distribution: Statistical inference, fuzzy reliability, and COVID-19 application. *Journal of Function Spaces*, 9094078.

Mukhtar, M. S., El-Morshedy, M., Eliwa, M. S., and Yousof, H. M. (2020). Expanded Fréchet model: mathematical properties, copula, different estimation methods, applications and validation testing. *Mathematics*, **8**, 1949.

Muse, A. H., Tolba, A. H., Fayad, E., Abu Ali, O. A., Nagy, M., and Yusuf, M. (2021). Modelling the COVID-19 mortality rate with a new versatile modification of the log-logistic distribution. *Computational Intelligence and Neuroscience*, 8640794.

Nagy, M., Gemeay, A. M., Rajitha, C., Karakaya, K., Sağlam, Ş., Mansi, A., and Kilai, M. (2023). Power unit Gumbel type II distribution: Statistical properties, regression analysis, and applications. *AIP Advances*, **13**.

Nichols, M. D. and Padgett, W. J. (2006). A bootstrap control chart for Weibull percentiles. *Quality and Reliability Engineering International*, **22**, 141–151.

Rényi, A. (1961). On measures of entropy and information. *In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, **4**, 547–562.

Sarhan, A. M., Tadj, L., and Hamilton, D. C. (2014). A new lifetime distribution and its power transformation. *Journal of Probability and Statistics*, **2014**, 532024.

Swain, J. J., Venkatraman, S., and Wilson, J. R. (1988). Least-squares estimation of distribution functions in Johnson's translation system. *Journal of Statistical Computation and Simulation*, **29**, 271–297.

Torabi, H. (2008). A general method for estimating and hypotheses testing using spacings. *Journal of Statistical Theory and Applications*, **8**, 163–168.

Yıldırım, E., Ilıkkan, E. S., Gemeay, A. M., Makumi, N., Bakr, M., and Balogun, O. S. (2023). Power unit Burr-XII distribution: Statistical inference with applications. *AIP Advances*, **13**.

ANNEXURE

Table 5: Numerical values of the PGD simulation for $\theta = 0.25$, $k = 0.5$, and $\alpha = 0.75$.

n	Mea.	$\widehat{Est.}$	MLE	ADE	CVME	MPSE	OLSE	RTADE	WLSE	LTADE	MSADE	MSALDE
30	BIAS	$\hat{\theta}$	$0.25865^{\{5\}}$	$0.32403^{\{6\}}$	$0.32779^{\{7\}}$	$0.36521^{\{8\}}$	$0.43861^{\{10\}}$	$0.22234^{\{3\}}$	$0.39659^{\{9\}}$	$0.25818^{\{4\}}$	$0.1256^{\{1\}}$	$0.16549^{\{2\}}$
		$\hat{k}$	$0.79281^{\{5\}}$	$0.87253^{\{7\}}$	$0.83048^{\{6\}}$	$0.65183^{\{3\}}$	$0.97738^{\{9\}}$	$0.6618^{\{4\}}$	$0.98085^{\{10\}}$	$0.95445^{\{8\}}$	$0.14546^{\{1\}}$	$0.30555^{\{2\}}$
		$\hat{\alpha}$	$0.10985^{\{2\}}$	$0.12432^{\{5\}}$	$0.13937^{\{7\}}$	$0.1317^{\{6\}}$	$0.15504^{\{10\}}$	$0.11415^{\{4\}}$	$0.14265^{\{9\}}$	$0.14217^{\{8\}}$	$0.11054^{\{3\}}$	$0.10656^{\{1\}}$
	MSE	$\hat{\theta}$	$0.24582^{\{4\}}$	$0.39659^{\{6\}}$	$0.40477^{\{7\}}$	$0.54059^{\{8\}}$	$0.64788^{\{10\}}$	$0.20296^{\{3\}}$	$0.55552^{\{9\}}$	$0.28566^{\{5\}}$	$0.06825^{\{1\}}$	$0.15815^{\{2\}}$
		$\hat{k}$	$1.31754^{\{4\}}$	$1.77562^{\{7\}}$	$1.71171^{\{6\}}$	$1.87424^{\{8\}}$	$2.38251^{\{10\}}$	$0.99702^{\{3\}}$	$2.21934^{\{9\}}$	$1.7032^{\{5\}}$	$0.22187^{\{1\}}$	$0.58968^{\{2\}}$
		$\hat{\alpha}$	$0.02029^{\{3\}}$	$0.02469^{\{5\}}$	$0.03135^{\{7\}}$	$0.02804^{\{6\}}$	$0.03651^{\{10\}}$	$0.02135^{\{4\}}$	$0.03148^{\{8\}}$	$0.03256^{\{9\}}$	$0.0199^{\{2\}}$	$0.01853^{\{1\}}$
	MRE	$\hat{\theta}$	$1.03459^{\{5\}}$	$1.29613^{\{6\}}$	$1.31115^{\{7\}}$	$1.46085^{\{8\}}$	$1.75443^{\{10\}}$	$0.88934^{\{3\}}$	$1.58635^{\{9\}}$	$1.03274^{\{4\}}$	$0.5024^{\{1\}}$	$0.66195^{\{2\}}$
		$\hat{k}$	$1.58562^{\{5\}}$	$1.74507^{\{7\}}$	$1.66097^{\{6\}}$	$1.30366^{\{3\}}$	$1.95477^{\{9\}}$	$1.3236^{\{4\}}$	$1.9617^{\{10\}}$	$1.90891^{\{8\}}$	$0.29092^{\{1\}}$	$0.61111^{\{2\}}$
		$\hat{\alpha}$	$0.14647^{\{2\}}$	$0.16576^{\{5\}}$	$0.18583^{\{7\}}$	$0.1756^{\{6\}}$	$0.20673^{\{10\}}$	$0.1522^{\{4\}}$	$0.1902^{\{9\}}$	$0.18956^{\{8\}}$	$0.14739^{\{3\}}$	$0.14208^{\{1\}}$
	$\sum Ranks$		$35^{\{4\}}$	$54^{\{5\}}$	$60^{\{8\}}$	$56^{\{6\}}$	$88^{\{10\}}$	$32^{\{3\}}$	$82^{\{9\}}$	$59^{\{7\}}$	$14^{\{1\}}$	$15^{\{2\}}$
70	BIAS	$\hat{\theta}$	$0.15963^{\{4\}}$	$0.1918^{\{7\}}$	$0.21943^{\{8\}}$	$0.13679^{\{3\}}$	$0.24127^{\{10\}}$	$0.18063^{\{6\}}$	$0.22178^{\{9\}}$	$0.17231^{\{5\}}$	$0.07683^{\{1\}}$	$0.11258^{\{2\}}$
		$\hat{k}$	$0.63955^{\{4\}}$	$0.67896^{\{6\}}$	$0.71336^{\{7\}}$	$0.22915^{\{2\}}$	$0.73658^{\{8\}}$	$0.64138^{\{5\}}$	$0.74519^{\{9\}}$	$0.80685^{\{10\}}$	$0.09957^{\{1\}}$	$0.27747^{\{3\}}$
		$\hat{\alpha}$	$0.07085^{\{1\}}$	$0.08157^{\{5\}}$	$0.09285^{\{8\}}$	$0.07443^{\{3\}}$	$0.10021^{\{10\}}$	$0.08361^{\{6\}}$	$0.08964^{\{7\}}$	$0.09785^{\{9\}}$	$0.0746^{\{4\}}$	$0.07378^{\{2\}}$
	MSE	$\hat{\theta}$	$0.08664^{\{3\}}$	$0.12066^{\{7\}}$	$0.15022^{\{9\}}$	$0.10584^{\{5\}}$	$0.17945^{\{10\}}$	$0.1035^{\{4\}}$	$0.14769^{\{8\}}$	$0.10676^{\{6\}}$	$0.02032^{\{1\}}$	$0.05612^{\{2\}}$
		$\hat{k}$	$0.73117^{\{4\}}$	$0.83644^{\{6\}}$	$0.95092^{\{7\}}$	$0.43668^{\{3\}}$	$1.03546^{\{10\}}$	$0.73737^{\{5\}}$	$0.95386^{\{8\}}$	$0.99808^{\{9\}}$	$0.1092^{\{1\}}$	$0.33317^{\{2\}}$
		$\hat{\alpha}$	$0.00847^{\{1\}}$	$0.01108^{\{5\}}$	$0.01382^{\{8\}}$	$0.0101^{\{4\}}$	$0.01578^{\{10\}}$	$0.01138^{\{6\}}$	$0.01306^{\{7\}}$	$0.01498^{\{9\}}$	$0.00919^{\{3\}}$	$0.00911^{\{2\}}$
	MRE	$\hat{\theta}$	$0.63853^{\{4\}}$	$0.7672^{\{7\}}$	$0.87771^{\{8\}}$	$0.54717^{\{3\}}$	$0.96508^{\{10\}}$	$0.72254^{\{6\}}$	$0.88713^{\{9\}}$	$0.68924^{\{5\}}$	$0.30734^{\{1\}}$	$0.45034^{\{2\}}$
		$\hat{k}$	$1.2791^{\{4\}}$	$1.35791^{\{6\}}$	$1.42672^{\{7\}}$	$0.4583^{\{2\}}$	$1.47315^{\{8\}}$	$1.28276^{\{5\}}$	$1.49038^{\{9\}}$	$1.6137^{\{10\}}$	$0.19914^{\{1\}}$	$0.55494^{\{3\}}$
		$\hat{\alpha}$	$0.09446^{\{1\}}$	$0.10876^{\{5\}}$	$0.1238^{\{8\}}$	$0.09923^{\{3\}}$	$0.13362^{\{10\}}$	$0.11148^{\{6\}}$	$0.11952^{\{7\}}$	$0.13047^{\{9\}}$	$0.09946^{\{4\}}$	$0.09837^{\{2\}}$
	$\sum Ranks$		$26^{\{3\}}$	$54^{\{6\}}$	$70^{\{7\}}$	$28^{\{4\}}$	$86^{\{10\}}$	$49^{\{5\}}$	$73^{\{9\}}$	$72^{\{8\}}$	$17^{\{1\}}$	$20^{\{2\}}$
100	BIAS	$\hat{\theta}$	$0.13278^{\{4\}}$	$0.15621^{\{6\}}$	$0.18086^{\{8\}}$	$0.08393^{\{2\}}$	$0.20355^{\{10\}}$	$0.15778^{\{7\}}$	$0.18245^{\{9\}}$	$0.14482^{\{5\}}$	$0.06887^{\{1\}}$	$0.08869^{\{3\}}$
		$\hat{k}$	$0.59081^{\{4\}}$	$0.63705^{\{6\}}$	$0.6751^{\{7\}}$	$0.11845^{\{2\}}$	$0.68911^{\{8\}}$	$0.6111^{\{5\}}$	$0.69388^{\{9\}}$	$0.75662^{\{10\}}$	$0.09683^{\{1\}}$	$0.24613^{\{3\}}$
		$\hat{\alpha}$	$0.05959^{\{2\}}$	$0.06873^{\{5\}}$	$0.07822^{\{8\}}$	$0.05754^{\{1\}}$	$0.08547^{\{10\}}$	$0.07155^{\{6\}}$	$0.07664^{\{7\}}$	$0.0813^{\{9\}}$	$0.06404^{\{4\}}$	$0.06114^{\{3\}}$
	MSE	$\hat{\theta}$	$0.05403^{\{4\}}$	$0.07765^{\{7\}}$	$0.10351^{\{9\}}$	$0.05024^{\{3\}}$	$0.12863^{\{10\}}$	$0.0726^{\{6\}}$	$0.09407^{\{8\}}$	$0.06797^{\{5\}}$	$0.01697^{\{1\}}$	$0.03665^{\{2\}}$
		$\hat{k}$	$0.583^{\{4\}}$	$0.6769^{\{6\}}$	$0.79391^{\{8\}}$	$0.21273^{\{2\}}$	$0.85299^{\{10\}}$	$0.62988^{\{5\}}$	$0.76826^{\{7\}}$	$0.82993^{\{9\}}$	$0.1004^{\{1\}}$	$0.25214^{\{3\}}$
		$\hat{\alpha}$	$0.00608^{\{1\}}$	$0.008^{\{5\}}$	$0.00998^{\{8\}}$	$0.00617^{\{2\}}$	$0.01171^{\{10\}}$	$0.00829^{\{6\}}$	$0.00952^{\{7\}}$	$0.01036^{\{9\}}$	$0.00676^{\{4\}}$	$0.00647^{\{3\}}$
	MRE	$\hat{\theta}$	$0.53112^{\{4\}}$	$0.62486^{\{6\}}$	$0.72344^{\{8\}}$	$0.33572^{\{2\}}$	$0.81421^{\{10\}}$	$0.63112^{\{7\}}$	$0.72979^{\{9\}}$	$0.5793^{\{5\}}$	$0.27547^{\{1\}}$	$0.35477^{\{3\}}$
		$\hat{k}$	$1.18162^{\{4\}}$	$1.27409^{\{6\}}$	$1.3502^{\{7\}}$	$0.23689^{\{2\}}$	$1.37823^{\{8\}}$	$1.2222^{\{5\}}$	$1.38777^{\{9\}}$	$1.51324^{\{10\}}$	$0.19365^{\{1\}}$	$0.49226^{\{3\}}$
		$\hat{\alpha}$	$0.07945^{\{2\}}$	$0.09164^{\{5\}}$	$0.10429^{\{8\}}$	$0.07672^{\{1\}}$	$0.11396^{\{10\}}$	$0.0954^{\{6\}}$	$0.10219^{\{7\}}$	$0.1084^{\{9\}}$	$0.08539^{\{4\}}$	$0.08152^{\{3\}}$
	$\sum Ranks$		$29^{\{4\}}$	$52^{\{5\}}$	$71^{\{7.5\}}$	$17^{\{1\}}$	$86^{\{10\}}$	$53^{\{6\}}$	$72^{\{9\}}$	$71^{\{7.5\}}$	$18^{\{2\}}$	$26^{\{3\}}$
150	BIAS	$\hat{\theta}$	$0.10611^{\{4\}}$	$0.12412^{\{6\}}$	$0.15046^{\{9\}}$	$0.0439^{\{1\}}$	$0.16988^{\{10\}}$	$0.13703^{\{8\}}$	$0.13379^{\{7\}}$	$0.12264^{\{5\}}$	$0.05424^{\{2\}}$	$0.07316^{\{3\}}$
		$\hat{k}$	$0.53044^{\{4\}}$	$0.5773^{\{5\}}$	$0.63444^{\{8\}}$	$0.03948^{\{1\}}$	$0.66083^{\{9\}}$	$0.58086^{\{6\}}$	$0.62523^{\{7\}}$	$0.72204^{\{10\}}$	$0.07876^{\{2\}}$	$0.23075^{\{3\}}$
		$\hat{\alpha}$	$0.04914^{\{2\}}$	$0.05551^{\{5\}}$	$0.06505^{\{8\}}$	$0.0419^{\{1\}}$	$0.07134^{\{10\}}$	$0.06173^{\{7\}}$	$0.0587^{\{6\}}$	$0.06866^{\{9\}}$	$0.05276^{\{4\}}$	$0.05052^{\{3\}}$
	MSE	$\hat{\theta}$	$0.03585^{\{4\}}$	$0.04817^{\{6\}}$	$0.06652^{\{9\}}$	$0.01236^{\{2\}}$	$0.08195^{\{10\}}$	$0.05865^{\{8\}}$	$0.05101^{\{7\}}$	$0.04462^{\{5\}}$	$0.01011^{\{1\}}$	$0.02413^{\{3\}}$
		$\hat{k}$	$0.474^{\{4\}}$	$0.53568^{\{5\}}$	$0.64815^{\{8\}}$	$0.05504^{\{1\}}$	$0.70867^{\{9\}}$	$0.55634^{\{6\}}$	$0.58432^{\{7\}}$	$0.72035^{\{10\}}$	$0.07393^{\{2\}}$	$0.20945^{\{3\}}$
		$\hat{\alpha}$	$0.00428^{\{2\}}$	$0.0054^{\{5\}}$	$0.00693^{\{8\}}$	$0.00312^{\{1\}}$	$0.00831^{\{10\}}$	$0.00649^{\{7\}}$	$0.00576^{\{6\}}$	$0.00723^{\{9\}}$	$0.00454^{\{4\}}$	$0.00451^{\{3\}}$
	MRE	$\hat{\theta}$	$0.42442^{\{4\}}$	$0.49649^{\{6\}}$	$0.60186^{\{9\}}$	$0.1756^{\{1\}}$	$0.6795^{\{10\}}$	$0.54811^{\{8\}}$	$0.53515^{\{7\}}$	$0.49056^{\{5\}}$	$0.21698^{\{2\}}$	$0.29266^{\{3\}}$
		$\hat{k}$	$1.06088^{\{4\}}$	$1.15459^{\{5\}}$	$1.26888^{\{8\}}$	$0.07896^{\{1\}}$	$1.32165^{\{9\}}$	$1.16173^{\{6\}}$	$1.25046^{\{7\}}$	$1.44409^{\{10\}}$	$0.15752^{\{2\}}$	$0.4615^{\{3\}}$
		$\hat{\alpha}$	$0.06551^{\{2\}}$	$0.07401^{\{5\}}$	$0.08673^{\{8\}}$	$0.05586^{\{1\}}$	$0.09511^{\{10\}}$	$0.08231^{\{7\}}$	$0.07827^{\{6\}}$	$0.09155^{\{9\}}$	$0.07035^{\{4\}}$	$0.06736^{\{3\}}$
	$\sum Ranks$		$30^{\{4\}}$	$48^{\{5\}}$	$75^{\{9\}}$	$10^{\{1\}}$	$87^{\{10\}}$	$63^{\{7\}}$	$60^{\{6\}}$	$72^{\{8\}}$	$23^{\{2\}}$	$27^{\{3\}}$
250	BIAS	$\hat{\theta}$	$0.07797^{\{4\}}$	$0.0935^{\{5\}}$	$0.11638^{\{9\}}$	$0.02534^{\{1\}}$	$0.12017^{\{10\}}$	$0.10898^{\{8\}}$	$0.10273^{\{7\}}$	$0.09432^{\{6\}}$	$0.04487^{\{2\}}$	$0.05973^{\{3\}}$
		$\hat{k}$	$0.45976^{\{4\}}$	$0.51897^{\{5\}}$	$0.58196^{\{8\}}$	$0.00876^{\{1\}}$	$0.59179^{\{9\}}$	$0.54575^{\{6\}}$	$0.56963^{\{7\}}$	$0.65288^{\{10\}}$	$0.07331^{\{2\}}$	$0.22622^{\{3\}}$
		$\hat{\alpha}$	$0.03701^{\{2\}}$	$0.04027^{\{4\}}$	$0.05051^{\{8\}}$	$0.03021^{\{1\}}$	$0.05111^{\{9\}}$	$0.04923^{\{7\}}$	$0.04455^{\{6\}}$	$0.05227^{\{10\}}$	$0.04185^{\{5\}}$	$0.03874^{\{3\}}$
	MSE	$\hat{\theta}$	$0.01867^{\{4\}}$	$0.02681^{\{6\}}$	$0.03953^{\{9\}}$	$0.00205^{\{1\}}$	$0.0423^{\{10\}}$	$0.03404^{\{8\}}$	$0.02856^{\{7\}}$	$0.02506^{\{5\}}$	$0.00842^{\{2\}}$	$0.01472^{\{3\}}$
		$\hat{k}$	$0.35282^{\{4\}}$	$0.41819^{\{5\}}$	$0.50993^{\{8\}}$	$0.00764^{\{1\}}$	$0.52482^{\{9\}}$	$0.44969^{\{6\}}$	$0.45689^{\{7\}}$	$0.57658^{\{10\}}$	$0.06869^{\{2\}}$	$0.17934^{\{3\}}$
		$\hat{\alpha}$	$0.00246^{\{2\}}$	$0.00304^{\{5\}}$	$0.00436^{\{9\}}$	$0.00153^{\{1\}}$	$0.00443^{\{10\}}$	$0.00412^{\{7\}}$	$0.00335^{\{6\}}$	$0.00423^{\{8\}}$	$0.00292^{\{4\}}$	$0.00276^{\{3\}}$
	MRE	$\hat{\theta}$	$0.31187^{\{4\}}$	$0.374^{\{5\}}$	$0.46553^{\{9\}}$	$0.10135^{\{1\}}$	$0.48068^{\{10\}}$	$0.43594^{\{8\}}$	$0.41091^{\{7\}}$	$0.37728^{\{6\}}$	$0.17948^{\{2\}}$	$0.23892^{\{3\}}$
		$\hat{k}$	$0.91952^{\{4\}}$	$1.03794^{\{5\}}$	$1.16391^{\{8\}}$	$0.01752^{\{1\}}$	$1.18357^{\{9\}}$	$1.0915^{\{6\}}$	$1.13926^{\{7\}}$	$1.30576^{\{10\}}$	$0.14662^{\{2\}}$	$0.45244^{\{3\}}$
		$\hat{\alpha}$	$0.04935^{\{2\}}$	$0.05369^{\{4\}}$	$0.06734^{\{8\}}$	$0.04028^{\{1\}}$	$0.06815^{\{9\}}$	$0.06564^{\{7\}}$	$0.0594^{\{6\}}$	$0.0697^{\{10\}}$	$0.0558^{\{5\}}$	$0.05166^{\{3\}}$
	$\sum Ranks$		$30^{\{4\}}$	$44^{\{5\}}$	$76^{\{9\}}$	$9^{\{1\}}$	$85^{\{10\}}$	$63^{\{7\}}$	$60^{\{6\}}$	$75^{\{8\}}$	$26^{\{2\}}$	$27^{\{3\}}$

Table 6: Numerical values of the PGD simulation for $\theta = 0.2$, $k = 1.2$, and $\alpha = 0.9$.

n	Mea.	$\widehat{Est.}$	MLE	ADE	CVME	MPSE	OLSE	RTADE	WLSE	LTADE	MSADE	MSALDE
30	BIAS	$\hat{\theta}$	$0.19033^{\{3\}}$	$0.25544^{\{6\}}$	$0.21025^{\{5\}}$	$0.34281^{\{10\}}$	$0.28372^{\{7\}}$	$0.14505^{\{2\}}$	$0.28952^{\{9\}}$	$0.28864^{\{8\}}$	$0.13824^{\{1\}}$	$0.20542^{\{4\}}$
		$\hat{k}$	$0.71908^{\{3\}}$	$0.8983^{\{6\}}$	$0.83681^{\{5\}}$	$0.9447^{\{10\}}$	$0.92892^{\{7\}}$	$0.74395^{\{4\}}$	$0.94054^{\{9\}}$	$0.92939^{\{8\}}$	$0.2586^{\{1\}}$	$0.48444^{\{2\}}$
		$\hat{\alpha}$	$0.13456^{\{3\}}$	$0.15006^{\{5\}}$	$0.16005^{\{6\}}$	$0.16583^{\{7\}}$	$0.17142^{\{9\}}$	$0.13074^{\{1\}}$	$0.16703^{\{8\}}$	$0.18016^{\{10\}}$	$0.13346^{\{2\}}$	$0.1391^{\{4\}}$
	MSE	$\hat{\theta}$	$0.16931^{\{3\}}$	$0.29393^{\{6\}}$	$0.18286^{\{4\}}$	$0.50761^{\{10\}}$	$0.34001^{\{8\}}$	$0.08595^{\{2\}}$	$0.34515^{\{9\}}$	$0.32224^{\{7\}}$	$0.07425^{\{1\}}$	$0.22873^{\{5\}}$
		$\hat{k}$	$0.94467^{\{4\}}$	$1.38303^{\{6\}}$	$1.05927^{\{5\}}$	$1.80301^{\{10\}}$	$1.39224^{\{7\}}$	$0.76905^{\{2\}}$	$1.46422^{\{9\}}$	$1.42652^{\{8\}}$	$0.24219^{\{1\}}$	$0.77736^{\{3\}}$
		$\hat{\alpha}$	$0.03039^{\{3\}}$	$0.03616^{\{5\}}$	$0.04114^{\{6\}}$	$0.04501^{\{8\}}$	$0.04515^{\{9\}}$	$0.02722^{\{1\}}$	$0.04388^{\{7\}}$	$0.05183^{\{10\}}$	$0.02992^{\{2\}}$	$0.03242^{\{4\}}$
	MRE	$\hat{\theta}$	$0.95164^{\{3\}}$	$1.27721^{\{6\}}$	$1.05127^{\{5\}}$	$1.71406^{\{10\}}$	$1.41862^{\{7\}}$	$0.72526^{\{2\}}$	$1.4476^{\{9\}}$	$1.44321^{\{8\}}$	$0.69122^{\{1\}}$	$1.02711^{\{4\}}$
		$\hat{k}$	$0.59923^{\{3\}}$	$0.74858^{\{6\}}$	$0.69735^{\{5\}}$	$0.78725^{\{10\}}$	$0.7741^{\{7\}}$	$0.61996^{\{4\}}$	$0.78378^{\{9\}}$	$0.77449^{\{8\}}$	$0.2155^{\{1\}}$	$0.4037^{\{2\}}$
		$\hat{\alpha}$	$0.14951^{\{3\}}$	$0.16673^{\{5\}}$	$0.17783^{\{6\}}$	$0.18426^{\{7\}}$	$0.19047^{\{9\}}$	$0.14527^{\{1\}}$	$0.18559^{\{8\}}$	$0.20018^{\{10\}}$	$0.14829^{\{2\}}$	$0.15455^{\{4\}}$
		$\sum Ranks$	$28^{\{3\}}$	$51^{\{6\}}$	$47^{\{5\}}$	$82^{\{10\}}$	$70^{\{7\}}$	$19^{\{2\}}$	$77^{\{8.5\}}$	$77^{\{8.5\}}$	$12^{\{1\}}$	$32^{\{4\}}$
70	BIAS	$\hat{\theta}$	$0.12816^{\{4\}}$	$0.1615^{\{6\}}$	$0.15803^{\{5\}}$	$0.17545^{\{8\}}$	$0.17946^{\{10\}}$	$0.12302^{\{2\}}$	$0.16427^{\{7\}}$	$0.1793^{\{9\}}$	$0.09715^{\{1\}}$	$0.12563^{\{3\}}$
		$\hat{k}$	$0.61492^{\{3\}}$	$0.74746^{\{7\}}$	$0.75036^{\{8\}}$	$0.67901^{\{4\}}$	$0.77227^{\{9\}}$	$0.68958^{\{5\}}$	$0.74733^{\{6\}}$	$0.79249^{\{10\}}$	$0.23065^{\{1\}}$	$0.40529^{\{2\}}$
		$\hat{\alpha}$	$0.0912^{\{1\}}$	$0.1049^{\{5\}}$	$0.11395^{\{8\}}$	$0.10766^{\{6\}}$	$0.12054^{\{10\}}$	$0.09762^{\{2\}}$	$0.11025^{\{7\}}$	$0.12027^{\{9\}}$	$0.0979^{\{4\}}$	$0.09769^{\{3\}}$
	MSE	$\hat{\theta}$	$0.05921^{\{3\}}$	$0.08628^{\{7\}}$	$0.07312^{\{5\}}$	$0.12714^{\{10\}}$	$0.09587^{\{8\}}$	$0.04202^{\{2\}}$	$0.08365^{\{6\}}$	$0.10255^{\{9\}}$	$0.0329^{\{1\}}$	$0.07145^{\{4\}}$
		$\hat{k}$	$0.58159^{\{3\}}$	$0.75488^{\{8\}}$	$0.72456^{\{5\}}$	$0.74983^{\{7\}}$	$0.76942^{\{9\}}$	$0.60355^{\{4\}}$	$0.73955^{\{6\}}$	$0.83978^{\{10\}}$	$0.15619^{\{1\}}$	$0.37018^{\{2\}}$
		$\hat{\alpha}$	$0.0142^{\{1\}}$	$0.01848^{\{5\}}$	$0.02042^{\{7\}}$	$0.02064^{\{8\}}$	$0.02314^{\{9\}}$	$0.01516^{\{2\}}$	$0.01967^{\{6\}}$	$0.02328^{\{10\}}$	$0.01614^{\{3\}}$	$0.01653^{\{4\}}$
	MRE	$\hat{\theta}$	$0.64078^{\{4\}}$	$0.80751^{\{6\}}$	$0.79015^{\{5\}}$	$0.87724^{\{8\}}$	$0.89729^{\{10\}}$	$0.61509^{\{2\}}$	$0.82135^{\{7\}}$	$0.89652^{\{9\}}$	$0.48577^{\{1\}}$	$0.62817^{\{3\}}$
		$\hat{k}$	$0.51244^{\{3\}}$	$0.62289^{\{7\}}$	$0.6253^{\{8\}}$	$0.56584^{\{4\}}$	$0.64356^{\{9\}}$	$0.57465^{\{5\}}$	$0.62277^{\{6\}}$	$0.66041^{\{10\}}$	$0.19221^{\{1\}}$	$0.33774^{\{2\}}$
		$\hat{\alpha}$	$0.10133^{\{1\}}$	$0.11655^{\{5\}}$	$0.12662^{\{8\}}$	$0.11962^{\{6\}}$	$0.13394^{\{10\}}$	$0.10846^{\{2\}}$	$0.1225^{\{7\}}$	$0.13363^{\{9\}}$	$0.10877^{\{4\}}$	$0.10855^{\{3\}}$
		$\sum Ranks$	$23^{\{2\}}$	$56^{\{5\}}$	$59^{\{7\}}$	$61^{\{8\}}$	$84^{\{9\}}$	$26^{\{3.5\}}$	$58^{\{6\}}$	$85^{\{10\}}$	$17^{\{1\}}$	$26^{\{3.5\}}$
100	BIAS	$\hat{\theta}$	$0.11203^{\{2\}}$	$0.13274^{\{5\}}$	$0.13679^{\{7\}}$	$0.14035^{\{8\}}$	$0.16025^{\{10\}}$	$0.11517^{\{4\}}$	$0.13633^{\{6\}}$	$0.14478^{\{9\}}$	$0.08428^{\{1\}}$	$0.11414^{\{3\}}$
		$\hat{k}$	$0.57429^{\{3\}}$	$0.69538^{\{7\}}$	$0.7083^{\{8\}}$	$0.60957^{\{4\}}$	$0.73704^{\{10\}}$	$0.66222^{\{5\}}$	$0.6953^{\{6\}}$	$0.72461^{\{9\}}$	$0.22189^{\{1\}}$	$0.39656^{\{2\}}$
		$\hat{\alpha}$	$0.07711^{\{1\}}$	$0.08896^{\{5\}}$	$0.09713^{\{8\}}$	$0.09071^{\{6\}}$	$0.10616^{\{10\}}$	$0.08739^{\{4\}}$	$0.09268^{\{7\}}$	$0.09933^{\{9\}}$	$0.08334^{\{2\}}$	$0.08731^{\{3\}}$
	MSE	$\hat{\theta}$	$0.04261^{\{3\}}$	$0.0568^{\{7\}}$	$0.0521^{\{4\}}$	$0.07587^{\{9\}}$	$0.0771^{\{10\}}$	$0.03463^{\{2\}}$	$0.0543^{\{6\}}$	$0.05956^{\{8\}}$	$0.0241^{\{1\}}$	$0.0539^{\{5\}}$
		$\hat{k}$	$0.50995^{\{3\}}$	$0.65349^{\{8\}}$	$0.65198^{\{7\}}$	$0.5945^{\{5\}}$	$0.71059^{\{10\}}$	$0.56559^{\{4\}}$	$0.63728^{\{6\}}$	$0.69042^{\{9\}}$	$0.13693^{\{1\}}$	$0.33286^{\{2\}}$
		$\hat{\alpha}$	$0.01051^{\{1\}}$	$0.0135^{\{5\}}$	$0.01524^{\{8\}}$	$0.01508^{\{7\}}$	$0.01848^{\{10\}}$	$0.01223^{\{3\}}$	$0.0143^{\{6\}}$	$0.01576^{\{9\}}$	$0.01159^{\{2\}}$	$0.01339^{\{4\}}$
	MRE	$\hat{\theta}$	$0.56017^{\{2\}}$	$0.66372^{\{5\}}$	$0.68393^{\{7\}}$	$0.70174^{\{8\}}$	$0.80123^{\{10\}}$	$0.57584^{\{4\}}$	$0.68167^{\{6\}}$	$0.72391^{\{9\}}$	$0.42141^{\{1\}}$	$0.57068^{\{3\}}$
		$\hat{k}$	$0.47857^{\{3\}}$	$0.57949^{\{7\}}$	$0.59025^{\{8\}}$	$0.50797^{\{4\}}$	$0.6142^{\{10\}}$	$0.55185^{\{5\}}$	$0.57941^{\{6\}}$	$0.60384^{\{9\}}$	$0.18491^{\{1\}}$	$0.33047^{\{2\}}$
		$\hat{\alpha}$	$0.08568^{\{1\}}$	$0.09885^{\{5\}}$	$0.10793^{\{8\}}$	$0.10079^{\{6\}}$	$0.11796^{\{10\}}$	$0.0971^{\{4\}}$	$0.10297^{\{7\}}$	$0.11036^{\{9\}}$	$0.09259^{\{2\}}$	$0.09701^{\{3\}}$
		$\sum Ranks$	$19^{\{2\}}$	$54^{\{5\}}$	$65^{\{8\}}$	$57^{\{7\}}$	$90^{\{10\}}$	$35^{\{4\}}$	$56^{\{6\}}$	$80^{\{9\}}$	$12^{\{1\}}$	$27^{\{3\}}$
150	BIAS	$\hat{\theta}$	$0.095^{\{3\}}$	$0.10904^{\{6\}}$	$0.11971^{\{9\}}$	$0.10059^{\{4\}}$	$0.12692^{\{10\}}$	$0.10269^{\{5\}}$	$0.11112^{\{7\}}$	$0.11813^{\{8\}}$	$0.07157^{\{1\}}$	$0.09127^{\{2\}}$
		$\hat{k}$	$0.53059^{\{4\}}$	$0.63298^{\{6\}}$	$0.66666^{\{8\}}$	$0.51678^{\{3\}}$	$0.67516^{\{9\}}$	$0.62817^{\{5\}}$	$0.63955^{\{7\}}$	$0.68851^{\{10\}}$	$0.22038^{\{1\}}$	$0.37888^{\{2\}}$
		$\hat{\alpha}$	$0.06452^{\{1\}}$	$0.07315^{\{5\}}$	$0.08072^{\{9\}}$	$0.07^{\{2\}}$	$0.08489^{\{10\}}$	$0.07443^{\{6\}}$	$0.07645^{\{7\}}$	$0.07879^{\{8\}}$	$0.07122^{\{3\}}$	$0.07275^{\{4\}}$
	MSE	$\hat{\theta}$	$0.02955^{\{3\}}$	$0.03613^{\{7\}}$	$0.04152^{\{9\}}$	$0.03605^{\{6\}}$	$0.04838^{\{10\}}$	$0.02686^{\{2\}}$	$0.0336^{\{5\}}$	$0.0379^{\{8\}}$	$0.01615^{\{1\}}$	$0.03224^{\{4\}}$
		$\hat{k}$	$0.44308^{\{4\}}$	$0.56131^{\{7\}}$	$0.59891^{\{8\}}$	$0.44168^{\{3\}}$	$0.61265^{\{9\}}$	$0.5256^{\{5\}}$	$0.55746^{\{6\}}$	$0.62665^{\{10\}}$	$0.11414^{\{1\}}$	$0.27554^{\{2\}}$
		$\hat{\alpha}$	$0.00759^{\{1\}}$	$0.00945^{\{5\}}$	$0.01104^{\{9\}}$	$0.00923^{\{3\}}$	$0.01227^{\{10\}}$	$0.00925^{\{4\}}$	$0.00996^{\{7\}}$	$0.01037^{\{8\}}$	$0.00847^{\{2\}}$	$0.00955^{\{6\}}$
	MRE	$\hat{\theta}$	$0.47501^{\{3\}}$	$0.54522^{\{6\}}$	$0.59857^{\{9\}}$	$0.50295^{\{4\}}$	$0.63458^{\{10\}}$	$0.51345^{\{5\}}$	$0.55559^{\{7\}}$	$0.59064^{\{8\}}$	$0.35783^{\{1\}}$	$0.45635^{\{2\}}$
		$\hat{k}$	$0.44216^{\{4\}}$	$0.52748^{\{6\}}$	$0.55555^{\{8\}}$	$0.43065^{\{3\}}$	$0.56263^{\{9\}}$	$0.52348^{\{5\}}$	$0.53296^{\{7\}}$	$0.57376^{\{10\}}$	$0.18365^{\{1\}}$	$0.31573^{\{2\}}$
		$\hat{\alpha}$	$0.07169^{\{1\}}$	$0.08128^{\{5\}}$	$0.08969^{\{9\}}$	$0.07778^{\{2\}}$	$0.09432^{\{10\}}$	$0.0827^{\{6\}}$	$0.08494^{\{7\}}$	$0.08755^{\{8\}}$	$0.07913^{\{3\}}$	$0.08084^{\{4\}}$
		$\sum Ranks$	$24^{\{2\}}$	$53^{\{6\}}$	$78^{\{8.5\}}$	$30^{\{4\}}$	$87^{\{10\}}$	$43^{\{5\}}$	$60^{\{7\}}$	$78^{\{8.5\}}$	$14^{\{1\}}$	$28^{\{3\}}$
250	BIAS	$\hat{\theta}$	$0.07031^{\{3\}}$	$0.07813^{\{5\}}$	$0.0941^{\{9\}}$	$0.07232^{\{4\}}$	$0.09648^{\{10\}}$	$0.08536^{\{7\}}$	$0.08431^{\{6\}}$	$0.08763^{\{8\}}$	$0.06325^{\{1\}}$	$0.0701^{\{2\}}$
		$\hat{k}$	$0.46386^{\{4\}}$	$0.57061^{\{5\}}$	$0.61433^{\{8\}}$	$0.43574^{\{3\}}$	$0.61758^{\{9\}}$	$0.58132^{\{7\}}$	$0.57518^{\{6\}}$	$0.62223^{\{10\}}$	$0.20038^{\{1\}}$	$0.35785^{\{2\}}$
		$\hat{\alpha}$	$0.0492^{\{1\}}$	$0.0536^{\{3\}}$	$0.06337^{\{9\}}$	$0.05235^{\{2\}}$	$0.06473^{\{10\}}$	$0.05957^{\{8\}}$	$0.05628^{\{5\}}$	$0.05912^{\{7\}}$	$0.05782^{\{6\}}$	$0.05477^{\{4\}}$
	MSE	$\hat{\theta}$	$0.013^{\{2\}}$	$0.01725^{\{4\}}$	$0.0238^{\{9\}}$	$0.01749^{\{6.5\}}$	$0.02594^{\{10\}}$	$0.01887^{\{8\}}$	$0.01749^{\{6.5\}}$	$0.01737^{\{5\}}$	$0.01263^{\{1\}}$	$0.01724^{\{3\}}$
		$\hat{k}$	$0.35609^{\{4\}}$	$0.48251^{\{7\}}$	$0.52675^{\{8\}}$	$0.35005^{\{3\}}$	$0.53038^{\{9\}}$	$0.4748^{\{5\}}$	$0.4779^{\{6\}}$	$0.53493^{\{10\}}$	$0.12826^{\{1\}}$	$0.23922^{\{2\}}$
		$\hat{\alpha}$	$0.00439^{\{1\}}$	$0.00526^{\{2\}}$	$0.00694^{\{9\}}$	$0.00536^{\{3\}}$	$0.00738^{\{10\}}$	$0.00624^{\{8\}}$	$0.00555^{\{4\}}$	$0.00581^{\{7\}}$	$0.00572^{\{5.5\}}$	$0.00572^{\{5.5\}}$
	MRE	$\hat{\theta}$	$0.35157^{\{3\}}$	$0.39063^{\{5\}}$	$0.47052^{\{9\}}$	$0.3616^{\{4\}}$	$0.48242^{\{10\}}$	$0.42682^{\{7\}}$	$0.42153^{\{6\}}$	$0.43817^{\{8\}}$	$0.31623^{\{1\}}$	$0.35052^{\{2\}}$
		$\hat{k}$	$0.38655^{\{4\}}$	$0.47551^{\{5\}}$	$0.51194^{\{8\}}$	$0.36312^{\{3\}}$	$0.51465^{\{9\}}$	$0.48443^{\{7\}}$	$0.47932^{\{6\}}$	$0.51852^{\{10\}}$	$0.17443^{\{1\}}$	$0.29821^{\{2\}}$
		$\hat{\alpha}$	$0.05467^{\{1\}}$	$0.05955^{\{3\}}$	$0.07041^{\{9\}}$	$0.05817^{\{2\}}$	$0.07192^{\{10\}}$	$0.06618^{\{8\}}$	$0.06253^{\{5\}}$	$0.06569^{\{7\}}$	$0.06425^{\{6\}}$	$0.06086^{\{4\}}$
		$\sum Ranks$	$23^{\{1\}}$	$39^{\{5\}}$	$78^{\{9\}}$	$30.5^{\{4\}}$	$87^{\{10\}}$	$65^{\{7\}}$	$50.5^{\{6\}}$	$72^{\{8\}}$	$23.5^{\{2\}}$	$26.5^{\{3\}}$

Table 7: Numerical values of the PGD simulation for $\theta = 1.5$, $k = 0.25$, and $\alpha = 1.5$.

n	Mea.	$\widehat{Est.}$	MLE	ADE	CVME	MPSE	OLSE	RTADE	WLSE	LTADE	MSADE	MSALDE
30	BIAS	$\hat{\theta}$	$0.39865^{\{2\}}$	$0.51479^{\{6\}}$	$0.54992^{\{7\}}$	$0.41497^{\{4\}}$	$0.56405^{\{8\}}$	$0.49898^{\{5\}}$	$0.56877^{\{9\}}$	$0.65377^{\{10\}}$	$0.26397^{\{1\}}$	$0.40666^{\{3\}}$
		$\hat{k}$	$0.37927^{\{2\}}$	$0.54978^{\{7\}}$	$0.54937^{\{6\}}$	$0.4719^{\{4\}}$	$0.62634^{\{10\}}$	$0.52726^{\{5\}}$	$0.62017^{\{9\}}$	$0.59021^{\{8\}}$	$0.26025^{\{1\}}$	$0.44806^{\{3\}}$
		$\hat{\alpha}$	$0.20564^{\{3\}}$	$0.19023^{\{1\}}$	$0.21628^{\{8\}}$	$0.20789^{\{4\}}$	$0.2081^{\{6\}}$	$0.20794^{\{5\}}$	$0.19973^{\{2\}}$	$0.57164^{\{10\}}$	$0.21384^{\{7\}}$	$0.21672^{\{9\}}$
	MSE	$\hat{\theta}$	$0.35523^{\{2\}}$	$0.57749^{\{6\}}$	$0.7166^{\{8\}}$	$0.39132^{\{3\}}$	$0.75187^{\{9\}}$	$0.55496^{\{5\}}$	$0.71403^{\{7\}}$	$0.77454^{\{10\}}$	$0.21591^{\{1\}}$	$0.45168^{\{4\}}$
		$\hat{k}$	$0.53666^{\{2\}}$	$0.95779^{\{7\}}$	$1.16749^{\{9\}}$	$0.66765^{\{3\}}$	$1.30282^{\{10\}}$	$0.88596^{\{6\}}$	$1.13786^{\{8\}}$	$0.7946^{\{5\}}$	$0.31973^{\{1\}}$	$0.76925^{\{4\}}$
		$\hat{\alpha}$	$0.07242^{\{7\}}$	$0.05817^{\{1\}}$	$0.08619^{\{9\}}$	$0.06432^{\{2\}}$	$0.06995^{\{4\}}$	$0.07128^{\{5\}}$	$0.06567^{\{3\}}$	$1.38559^{\{10\}}$	$0.07431^{\{8\}}$	$0.07153^{\{6\}}$
	MRE	$\hat{\theta}$	$0.26577^{\{2\}}$	$0.34319^{\{6\}}$	$0.36661^{\{7\}}$	$0.27665^{\{4\}}$	$0.37603^{\{8\}}$	$0.33266^{\{5\}}$	$0.37918^{\{9\}}$	$0.43584^{\{10\}}$	$0.17598^{\{1\}}$	$0.2711^{\{3\}}$
		$\hat{k}$	$1.51706^{\{2\}}$	$2.1991^{\{7\}}$	$2.19747^{\{6\}}$	$1.8876^{\{4\}}$	$2.50537^{\{10\}}$	$2.10906^{\{5\}}$	$2.48069^{\{9\}}$	$2.36086^{\{8\}}$	$1.04101^{\{1\}}$	$1.79225^{\{3\}}$
		$\hat{\alpha}$	$0.13709^{\{3\}}$	$0.12682^{\{1\}}$	$0.14419^{\{8\}}$	$0.13859^{\{4\}}$	$0.13873^{\{6\}}$	$0.13863^{\{5\}}$	$0.13315^{\{2\}}$	$0.38109^{\{10\}}$	$0.14256^{\{7\}}$	$0.14448^{\{9\}}$
	$\sum Ranks$		$25^{\{1\}}$	$42^{\{4\}}$	$68^{\{8\}}$	$32^{\{3\}}$	$71^{\{9\}}$	$46^{\{6\}}$	$58^{\{7\}}$	$81^{\{10\}}$	$28^{\{2\}}$	$44^{\{5\}}$
70	BIAS	$\hat{\theta}$	$0.31477^{\{3\}}$	$0.39061^{\{7\}}$	$0.38683^{\{5\}}$	$0.31416^{\{2\}}$	$0.42466^{\{9\}}$	$0.3906^{\{6\}}$	$0.40746^{\{8\}}$	$0.5016^{\{10\}}$	$0.21014^{\{1\}}$	$0.35523^{\{4\}}$
		$\hat{k}$	$0.32839^{\{2\}}$	$0.41404^{\{6\}}$	$0.39748^{\{5\}}$	$0.35935^{\{3\}}$	$0.44388^{\{9\}}$	$0.41979^{\{7\}}$	$0.42892^{\{8\}}$	$0.48296^{\{10\}}$	$0.23047^{\{1\}}$	$0.39056^{\{4\}}$
		$\hat{\alpha}$	$0.14773^{\{7\}}$	$0.13446^{\{1\}}$	$0.14285^{\{4\}}$	$0.14488^{\{5\}}$	$0.14102^{\{3\}}$	$0.14848^{\{8\}}$	$0.1362^{\{2\}}$	$0.31633^{\{10\}}$	$0.14657^{\{6\}}$	$0.15951^{\{9\}}$
	MSE	$\hat{\theta}$	$0.22597^{\{2\}}$	$0.32212^{\{6\}}$	$0.35489^{\{8\}}$	$0.24593^{\{3\}}$	$0.39725^{\{9\}}$	$0.31618^{\{4\}}$	$0.34876^{\{7\}}$	$0.48076^{\{10\}}$	$0.14092^{\{1\}}$	$0.32165^{\{5\}}$
		$\hat{k}$	$0.30556^{\{2\}}$	$0.4459^{\{6\}}$	$0.49991^{\{8\}}$	$0.36082^{\{3\}}$	$0.55611^{\{10\}}$	$0.44143^{\{4\}}$	$0.47326^{\{7\}}$	$0.53094^{\{9\}}$	$0.22578^{\{1\}}$	$0.44354^{\{5\}}$
		$\hat{\alpha}$	$0.03465^{\{8\}}$	$0.02791^{\{1\}}$	$0.03311^{\{5\}}$	$0.03162^{\{4\}}$	$0.03133^{\{3\}}$	$0.03416^{\{7\}}$	$0.02883^{\{2\}}$	$0.25562^{\{10\}}$	$0.03386^{\{6\}}$	$0.03914^{\{9\}}$
	MRE	$\hat{\theta}$	$0.20984^{\{3\}}$	$0.26041^{\{7\}}$	$0.25789^{\{5\}}$	$0.20944^{\{2\}}$	$0.2831^{\{9\}}$	$0.2604^{\{6\}}$	$0.27164^{\{8\}}$	$0.3344^{\{10\}}$	$0.14009^{\{1\}}$	$0.23682^{\{4\}}$
		$\hat{k}$	$1.31354^{\{2\}}$	$1.65617^{\{6\}}$	$1.58991^{\{5\}}$	$1.43739^{\{3\}}$	$1.77553^{\{9\}}$	$1.67916^{\{7\}}$	$1.71566^{\{8\}}$	$1.93183^{\{10\}}$	$0.92188^{\{1\}}$	$1.56223^{\{4\}}$
		$\hat{\alpha}$	$0.09848^{\{7\}}$	$0.08964^{\{1\}}$	$0.09524^{\{4\}}$	$0.09658^{\{5\}}$	$0.09402^{\{3\}}$	$0.09898^{\{8\}}$	$0.0908^{\{2\}}$	$0.21089^{\{10\}}$	$0.09771^{\{6\}}$	$0.10634^{\{9\}}$
	$\sum Ranks$		$36^{\{3\}}$	$41^{\{4\}}$	$49^{\{5\}}$	$30^{\{2\}}$	$64^{\{9\}}$	$57^{\{8\}}$	$52^{\{6\}}$	$89^{\{10\}}$	$24^{\{1\}}$	$53^{\{7\}}$
100	BIAS	$\hat{\theta}$	$0.29897^{\{3\}}$	$0.35942^{\{9\}}$	$0.34142^{\{4\}}$	$0.26356^{\{2\}}$	$0.35079^{\{6\}}$	$0.3423^{\{5\}}$	$0.35896^{\{8\}}$	$0.46651^{\{10\}}$	$0.19741^{\{1\}}$	$0.35588^{\{7\}}$
		$\hat{k}$	$0.30727^{\{3\}}$	$0.38326^{\{8\}}$	$0.34874^{\{4\}}$	$0.29363^{\{2\}}$	$0.37443^{\{6\}}$	$0.35272^{\{5\}}$	$0.38069^{\{7\}}$	$0.45798^{\{10\}}$	$0.21528^{\{1\}}$	$0.39944^{\{9\}}$
		$\hat{\alpha}$	$0.12845^{\{7\}}$	$0.12111^{\{2\}}$	$0.12328^{\{3\}}$	$0.12384^{\{4\}}$	$0.12541^{\{5\}}$	$0.13845^{\{8\}}$	$0.11839^{\{1\}}$	$0.27236^{\{10\}}$	$0.12715^{\{6\}}$	$0.14087^{\{9\}}$
	MSE	$\hat{\theta}$	$0.18713^{\{3\}}$	$0.26535^{\{5\}}$	$0.26993^{\{7\}}$	$0.18384^{\{2\}}$	$0.27462^{\{8\}}$	$0.21979^{\{4\}}$	$0.26952^{\{6\}}$	$0.41321^{\{10\}}$	$0.12227^{\{1\}}$	$0.29982^{\{9\}}$
		$\hat{k}$	$0.23127^{\{2\}}$	$0.35979^{\{7\}}$	$0.35312^{\{5\}}$	$0.25614^{\{3\}}$	$0.37912^{\{8\}}$	$0.26341^{\{4\}}$	$0.35372^{\{6\}}$	$0.47608^{\{10\}}$	$0.18215^{\{1\}}$	$0.41237^{\{9\}}$
		$\hat{\alpha}$	$0.02565^{\{7\}}$	$0.02234^{\{2\}}$	$0.02451^{\{5\}}$	$0.02367^{\{3\}}$	$0.02447^{\{4\}}$	$0.02901^{\{8\}}$	$0.02171^{\{1\}}$	$0.17345^{\{10\}}$	$0.0255^{\{6\}}$	$0.03019^{\{9\}}$
	MRE	$\hat{\theta}$	$0.19931^{\{3\}}$	$0.23961^{\{9\}}$	$0.22761^{\{4\}}$	$0.1757^{\{2\}}$	$0.23386^{\{6\}}$	$0.2282^{\{5\}}$	$0.23931^{\{8\}}$	$0.31101^{\{10\}}$	$0.13161^{\{1\}}$	$0.23725^{\{7\}}$
		$\hat{k}$	$1.2291^{\{3\}}$	$1.53306^{\{8\}}$	$1.39496^{\{4\}}$	$1.17451^{\{2\}}$	$1.49773^{\{6\}}$	$1.41089^{\{5\}}$	$1.52276^{\{7\}}$	$1.83194^{\{10\}}$	$0.8611^{\{1\}}$	$1.59776^{\{9\}}$
		$\hat{\alpha}$	$0.08563^{\{7\}}$	$0.08074^{\{2\}}$	$0.08219^{\{3\}}$	$0.08256^{\{4\}}$	$0.08361^{\{5\}}$	$0.0923^{\{8\}}$	$0.07893^{\{1\}}$	$0.18157^{\{10\}}$	$0.08476^{\{6\}}$	$0.09391^{\{9\}}$
	$\sum Ranks$		$38^{\{3\}}$	$52^{\{6.5\}}$	$39^{\{4\}}$	$24^{\{1.5\}}$	$54^{\{8\}}$	$52^{\{6.5\}}$	$45^{\{5\}}$	$90^{\{10\}}$	$24^{\{1.5\}}$	$77^{\{9\}}$
150	BIAS	$\hat{\theta}$	$0.27788^{\{3\}}$	$0.31689^{\{6\}}$	$0.31072^{\{5\}}$	$0.21965^{\{2\}}$	$0.32485^{\{8\}}$	$0.31893^{\{7\}}$	$0.30926^{\{4\}}$	$0.39575^{\{10\}}$	$0.18625^{\{1\}}$	$0.33863^{\{9\}}$
		$\hat{k}$	$0.29491^{\{3\}}$	$0.33484^{\{7\}}$	$0.32051^{\{4\}}$	$0.23451^{\{2\}}$	$0.34198^{\{8\}}$	$0.33041^{\{6\}}$	$0.32606^{\{5\}}$	$0.38985^{\{10\}}$	$0.19857^{\{1\}}$	$0.36986^{\{9\}}$
		$\hat{\alpha}$	$0.11864^{\{7\}}$	$0.10693^{\{1\}}$	$0.11344^{\{6\}}$	$0.1075^{\{2\}}$	$0.10927^{\{4\}}$	$0.12551^{\{8\}}$	$0.10768^{\{3\}}$	$0.22849^{\{10\}}$	$0.11285^{\{5\}}$	$0.13053^{\{9\}}$
	MSE	$\hat{\theta}$	$0.15744^{\{3\}}$	$0.20246^{\{7\}}$	$0.2013^{\{6\}}$	$0.14045^{\{2\}}$	$0.2203^{\{8\}}$	$0.18165^{\{4\}}$	$0.19057^{\{5\}}$	$0.28278^{\{10\}}$	$0.10805^{\{1\}}$	$0.25843^{\{9\}}$
		$\hat{k}$	$0.19323^{\{3\}}$	$0.25693^{\{7\}}$	$0.25117^{\{6\}}$	$0.17837^{\{2\}}$	$0.28179^{\{8\}}$	$0.20787^{\{4\}}$	$0.23715^{\{5\}}$	$0.30872^{\{9\}}$	$0.14395^{\{1\}}$	$0.32224^{\{10\}}$
		$\hat{\alpha}$	$0.0216^{\{7\}}$	$0.01734^{\{1\}}$	$0.02048^{\{5\}}$	$0.0179^{\{3\}}$	$0.01857^{\{4\}}$	$0.02385^{\{8\}}$	$0.01759^{\{2\}}$	$0.12105^{\{10\}}$	$0.02062^{\{6\}}$	$0.02566^{\{9\}}$
	MRE	$\hat{\theta}$	$0.18525^{\{3\}}$	$0.21126^{\{6\}}$	$0.20715^{\{5\}}$	$0.14643^{\{2\}}$	$0.21657^{\{8\}}$	$0.21262^{\{7\}}$	$0.20617^{\{4\}}$	$0.26383^{\{10\}}$	$0.12417^{\{1\}}$	$0.22576^{\{9\}}$
		$\hat{k}$	$1.17963^{\{3\}}$	$1.33937^{\{7\}}$	$1.28206^{\{4\}}$	$0.93802^{\{2\}}$	$1.36793^{\{8\}}$	$1.32164^{\{6\}}$	$1.30425^{\{5\}}$	$1.55941^{\{10\}}$	$0.79427^{\{1\}}$	$1.47943^{\{9\}}$
		$\hat{\alpha}$	$0.07909^{\{7\}}$	$0.07128^{\{1\}}$	$0.07563^{\{6\}}$	$0.07166^{\{2\}}$	$0.07285^{\{4\}}$	$0.08367^{\{8\}}$	$0.07179^{\{3\}}$	$0.15232^{\{10\}}$	$0.07523^{\{5\}}$	$0.08702^{\{9\}}$
	$\sum Ranks$		$39^{\{4\}}$	$43^{\{5\}}$	$47^{\{6\}}$	$19^{\{1\}}$	$60^{\{8\}}$	$58^{\{7\}}$	$36^{\{3\}}$	$89^{\{10\}}$	$22^{\{2\}}$	$82^{\{9\}}$
250	BIAS	$\hat{\theta}$	$0.25092^{\{3\}}$	$0.27934^{\{8\}}$	$0.26736^{\{5\}}$	$0.14555^{\{1\}}$	$0.27747^{\{7\}}$	$0.26651^{\{4\}}$	$0.27347^{\{6\}}$	$0.34017^{\{10\}}$	$0.17047^{\{2\}}$	$0.301^{\{9\}}$
		$\hat{k}$	$0.26835^{\{3\}}$	$0.29507^{\{8\}}$	$0.27311^{\{4\}}$	$0.15134^{\{1\}}$	$0.28767^{\{6\}}$	$0.28357^{\{5\}}$	$0.29215^{\{7\}}$	$0.3383^{\{10\}}$	$0.18266^{\{2\}}$	$0.32906^{\{9\}}$
		$\hat{\alpha}$	$0.10298^{\{7\}}$	$0.09768^{\{6\}}$	$0.09718^{\{4\}}$	$0.07824^{\{1\}}$	$0.09767^{\{5\}}$	$0.1114^{\{8\}}$	$0.09633^{\{3\}}$	$0.18221^{\{10\}}$	$0.09207^{\{2\}}$	$0.11484^{\{9\}}$
	MSE	$\hat{\theta}$	$0.11918^{\{4\}}$	$0.14629^{\{7\}}$	$0.1423^{\{6\}}$	$0.07558^{\{1\}}$	$0.14871^{\{8\}}$	$0.11612^{\{3\}}$	$0.14148^{\{5\}}$	$0.207^{\{10\}}$	$0.09031^{\{2\}}$	$0.19491^{\{9\}}$
		$\hat{k}$	$0.1447^{\{4\}}$	$0.17809^{\{8\}}$	$0.16967^{\{5\}}$	$0.09165^{\{1\}}$	$0.17672^{\{6\}}$	$0.13103^{\{3\}}$	$0.17736^{\{7\}}$	$0.22592^{\{9\}}$	$0.11428^{\{2\}}$	$0.23803^{\{10\}}$
		$\hat{\alpha}$	$0.01604^{\{7\}}$	$0.01432^{\{4\}}$	$0.01465^{\{6\}}$	$0.01019^{\{1\}}$	$0.01438^{\{5\}}$	$0.01811^{\{8\}}$	$0.014^{\{3\}}$	$0.07173^{\{10\}}$	$0.01364^{\{2\}}$	$0.01984^{\{9\}}$
	MRE	$\hat{\theta}$	$0.16728^{\{3\}}$	$0.18623^{\{8\}}$	$0.17824^{\{5\}}$	$0.09703^{\{1\}}$	$0.18498^{\{7\}}$	$0.17767^{\{4\}}$	$0.18231^{\{6\}}$	$0.22678^{\{10\}}$	$0.11365^{\{2\}}$	$0.20067^{\{9\}}$
		$\hat{k}$	$1.07339^{\{3\}}$	$1.1803^{\{8\}}$	$1.09245^{\{4\}}$	$0.60537^{\{1\}}$	$1.15067^{\{6\}}$	$1.13428^{\{5\}}$	$1.16859^{\{7\}}$	$1.35319^{\{10\}}$	$0.73065^{\{2\}}$	$1.31624^{\{9\}}$
		$\hat{\alpha}$	$0.06865^{\{7\}}$	$0.06512^{\{6\}}$	$0.06479^{\{4\}}$	$0.05216^{\{1\}}$	$0.06511^{\{5\}}$	$0.07426^{\{8\}}$	$0.06422^{\{3\}}$	$0.12147^{\{10\}}$	$0.06138^{\{2\}}$	$0.07656^{\{9\}}$
	$\sum Ranks$		$41^{\{3\}}$	$63^{\{8\}}$	$43^{\{4\}}$	$9^{\{1\}}$	$55^{\{7\}}$	$48^{\{6\}}$	$47^{\{5\}}$	$89^{\{10\}}$	$18^{\{2\}}$	$82^{\{9\}}$

Table 8: Numerical values of the PGD simulation for $\theta = 2.0$, $k = 2.5$, and $\alpha = 0.5$.

n	Mea.	$\widehat{Est.}$	MLE	ADE	CVME	MPSE	OLSE	RTADE	WLSE	LTADE	MSADE	MSALDE
30	BIAS	$\hat{\theta}$	$0.68641^{\{5\}}$	$0.67819^{\{4\}}$	$0.89874^{\{8\}}$	$0.64778^{\{3\}}$	$0.9267^{\{9\}}$	$1.0066^{\{10\}}$	$0.87473^{\{6\}}$	$0.88315^{\{7\}}$	$0.50833^{\{1\}}$	$0.53425^{\{2\}}$
		$\hat{k}$	$1.05377^{\{5\}}$	$1.05034^{\{4\}}$	$1.40682^{\{7\}}$	$0.99692^{\{3\}}$	$1.43881^{\{9\}}$	$1.56529^{\{10\}}$	$1.33514^{\{6\}}$	$1.40914^{\{8\}}$	$0.73837^{\{1\}}$	$0.82327^{\{2\}}$
		$\hat{\alpha}$	$0.08368^{\{6\}}$	$0.07658^{\{4\}}$	$0.10353^{\{9\}}$	$0.05802^{\{2\}}$	$0.08802^{\{7\}}$	$0.09485^{\{8\}}$	$0.0832^{\{5\}}$	$0.10587^{\{10\}}$	$0.07211^{\{3\}}$	$0.05735^{\{1\}}$
	MSE	$\hat{\theta}$	$0.83033^{\{5\}}$	$0.82821^{\{4\}}$	$1.1658^{\{8\}}$	$0.75697^{\{3\}}$	$1.24545^{\{9\}}$	$1.44951^{\{10\}}$	$1.14533^{\{7\}}$	$1.11712^{\{6\}}$	$0.56523^{\{1\}}$	$0.6413^{\{2\}}$
		$\hat{k}$	$2.07884^{\{4\}}$	$2.11181^{\{5\}}$	$3.02022^{\{8\}}$	$1.90557^{\{3\}}$	$3.17024^{\{9\}}$	$3.82206^{\{10\}}$	$2.87884^{\{6\}}$	$3.00749^{\{7\}}$	$1.46234^{\{1\}}$	$1.70225^{\{2\}}$
		$\hat{\alpha}$	$0.01644^{\{7\}}$	$0.01391^{\{5\}}$	$0.0209^{\{9\}}$	$0.00724^{\{1\}}$	$0.01482^{\{6\}}$	$0.01672^{\{8\}}$	$0.01291^{\{4\}}$	$0.02346^{\{10\}}$	$0.00972^{\{3\}}$	$0.00778^{\{2\}}$
	MRE	$\hat{\theta}$	$0.34321^{\{5\}}$	$0.33909^{\{4\}}$	$0.44937^{\{8\}}$	$0.32389^{\{3\}}$	$0.46335^{\{9\}}$	$0.5033^{\{10\}}$	$0.43737^{\{6\}}$	$0.44158^{\{7\}}$	$0.25417^{\{1\}}$	$0.26712^{\{2\}}$
		$\hat{k}$	$0.42151^{\{5\}}$	$0.42014^{\{4\}}$	$0.56273^{\{7\}}$	$0.39877^{\{3\}}$	$0.57552^{\{9\}}$	$0.62612^{\{10\}}$	$0.53406^{\{6\}}$	$0.56366^{\{8\}}$	$0.29535^{\{1\}}$	$0.32931^{\{2\}}$
		$\hat{\alpha}$	$0.16736^{\{6\}}$	$0.15316^{\{4\}}$	$0.20707^{\{9\}}$	$0.11603^{\{2\}}$	$0.17604^{\{7\}}$	$0.18969^{\{8\}}$	$0.1664^{\{5\}}$	$0.21173^{\{10\}}$	$0.14422^{\{3\}}$	$0.11469^{\{1\}}$
		$\sum Ranks$	$48^{\{5\}}$	$38^{\{4\}}$	$73^{\{7.5\}}$	$23^{\{3\}}$	$74^{\{9\}}$	$84^{\{10\}}$	$51^{\{6\}}$	$73^{\{7.5\}}$	$15^{\{1\}}$	$16^{\{2\}}$
70	BIAS	$\hat{\theta}$	$0.38937^{\{5\}}$	$0.35492^{\{3\}}$	$0.6773^{\{8\}}$	$0.34136^{\{2\}}$	$0.67818^{\{9\}}$	$0.73042^{\{10\}}$	$0.62083^{\{6\}}$	$0.62949^{\{7\}}$	$0.36695^{\{4\}}$	$0.3331^{\{1\}}$
		$\hat{k}$	$0.57748^{\{5\}}$	$0.52603^{\{3\}}$	$1.00068^{\{9\}}$	$0.51511^{\{2\}}$	$0.9953^{\{8\}}$	$1.08324^{\{10\}}$	$0.92772^{\{6\}}$	$0.96283^{\{7\}}$	$0.55581^{\{4\}}$	$0.50521^{\{1\}}$
		$\hat{\alpha}$	$0.04467^{\{5\}}$	$0.03954^{\{3\}}$	$0.07057^{\{9\}}$	$0.03115^{\{1\}}$	$0.06894^{\{7\}}$	$0.07284^{\{10\}}$	$0.06075^{\{6\}}$	$0.06942^{\{8\}}$	$0.0417^{\{4\}}$	$0.03196^{\{2\}}$
	MSE	$\hat{\theta}$	$0.37645^{\{5\}}$	$0.31563^{\{3\}}$	$0.67497^{\{8\}}$	$0.3039^{\{2\}}$	$0.67904^{\{9\}}$	$0.793^{\{10\}}$	$0.57574^{\{6\}}$	$0.59681^{\{7\}}$	$0.31583^{\{4\}}$	$0.30082^{\{1\}}$
		$\hat{k}$	$0.89576^{\{5\}}$	$0.71499^{\{1\}}$	$1.56196^{\{9\}}$	$0.73637^{\{3\}}$	$1.53132^{\{8\}}$	$1.86001^{\{10\}}$	$1.33279^{\{6\}}$	$1.50102^{\{7\}}$	$0.82738^{\{4\}}$	$0.72904^{\{2\}}$
		$\hat{\alpha}$	$0.00611^{\{5\}}$	$0.00508^{\{4\}}$	$0.00954^{\{9\}}$	$0.00304^{\{1\}}$	$0.00872^{\{7\}}$	$0.00998^{\{10\}}$	$0.00706^{\{6\}}$	$0.00922^{\{8\}}$	$0.00365^{\{3\}}$	$0.00348^{\{2\}}$
	MRE	$\hat{\theta}$	$0.19469^{\{5\}}$	$0.17746^{\{3\}}$	$0.33865^{\{8\}}$	$0.17068^{\{2\}}$	$0.33909^{\{9\}}$	$0.36521^{\{10\}}$	$0.31042^{\{6\}}$	$0.31475^{\{7\}}$	$0.18347^{\{4\}}$	$0.16655^{\{1\}}$
		$\hat{k}$	$0.23099^{\{5\}}$	$0.21041^{\{3\}}$	$0.40027^{\{9\}}$	$0.20604^{\{2\}}$	$0.39812^{\{8\}}$	$0.4333^{\{10\}}$	$0.37109^{\{6\}}$	$0.38513^{\{7\}}$	$0.22232^{\{4\}}$	$0.20208^{\{1\}}$
		$\hat{\alpha}$	$0.08934^{\{5\}}$	$0.07908^{\{3\}}$	$0.14114^{\{9\}}$	$0.06231^{\{1\}}$	$0.13788^{\{7\}}$	$0.14567^{\{10\}}$	$0.1215^{\{6\}}$	$0.13883^{\{8\}}$	$0.0834^{\{4\}}$	$0.06392^{\{2\}}$
		$\sum Ranks$	$45^{\{5\}}$	$26^{\{3\}}$	$78^{\{9\}}$	$16^{\{2\}}$	$72^{\{8\}}$	$90^{\{10\}}$	$54^{\{6\}}$	$66^{\{7\}}$	$35^{\{4\}}$	$13^{\{1\}}$
100	BIAS	$\hat{\theta}$	$0.26088^{\{4\}}$	$0.24428^{\{1\}}$	$0.5659^{\{8\}}$	$0.24738^{\{2\}}$	$0.57232^{\{9\}}$	$0.64628^{\{10\}}$	$0.50496^{\{6\}}$	$0.56189^{\{7\}}$	$0.34338^{\{5\}}$	$0.25452^{\{3\}}$
		$\hat{k}$	$0.37157^{\{3\}}$	$0.35078^{\{1\}}$	$0.81104^{\{7\}}$	$0.36542^{\{2\}}$	$0.82814^{\{9\}}$	$0.94636^{\{10\}}$	$0.7277^{\{6\}}$	$0.81868^{\{8\}}$	$0.50452^{\{5\}}$	$0.38331^{\{4\}}$
		$\hat{\alpha}$	$0.02861^{\{4\}}$	$0.02574^{\{3\}}$	$0.05982^{\{9\}}$	$0.02239^{\{1\}}$	$0.05542^{\{7\}}$	$0.06508^{\{10\}}$	$0.04883^{\{6\}}$	$0.05879^{\{8\}}$	$0.03928^{\{5\}}$	$0.02367^{\{2\}}$
	MSE	$\hat{\theta}$	$0.21059^{\{4\}}$	$0.19318^{\{1\}}$	$0.47902^{\{8\}}$	$0.19815^{\{2\}}$	$0.50328^{\{9\}}$	$0.63382^{\{10\}}$	$0.40902^{\{6\}}$	$0.47549^{\{7\}}$	$0.28267^{\{5\}}$	$0.21017^{\{3\}}$
		$\hat{k}$	$0.4637^{\{3\}}$	$0.4279^{\{1\}}$	$1.03391^{\{7\}}$	$0.45659^{\{2\}}$	$1.09981^{\{9\}}$	$1.43409^{\{10\}}$	$0.90927^{\{6\}}$	$1.08117^{\{8\}}$	$0.66863^{\{5\}}$	$0.50608^{\{4\}}$
		$\hat{\alpha}$	$0.00318^{\{4\}}$	$0.00258^{\{3\}}$	$0.00672^{\{9\}}$	$0.002^{\{1\}}$	$0.00567^{\{7\}}$	$0.00833^{\{10\}}$	$0.00448^{\{6\}}$	$0.00655^{\{8\}}$	$0.00416^{\{5\}}$	$0.00216^{\{2\}}$
	MRE	$\hat{\theta}$	$0.13044^{\{4\}}$	$0.12214^{\{1\}}$	$0.28295^{\{8\}}$	$0.12369^{\{2\}}$	$0.28616^{\{9\}}$	$0.32314^{\{10\}}$	$0.25248^{\{6\}}$	$0.28095^{\{7\}}$	$0.17169^{\{5\}}$	$0.12726^{\{3\}}$
		$\hat{k}$	$0.14863^{\{3\}}$	$0.14031^{\{1\}}$	$0.32442^{\{7\}}$	$0.14617^{\{2\}}$	$0.33126^{\{9\}}$	$0.37854^{\{10\}}$	$0.29108^{\{6\}}$	$0.32747^{\{8\}}$	$0.20181^{\{5\}}$	$0.15332^{\{4\}}$
		$\hat{\alpha}$	$0.05721^{\{4\}}$	$0.05148^{\{3\}}$	$0.11963^{\{9\}}$	$0.04478^{\{1\}}$	$0.11084^{\{7\}}$	$0.13015^{\{10\}}$	$0.09767^{\{6\}}$	$0.11758^{\{8\}}$	$0.07857^{\{5\}}$	$0.04733^{\{2\}}$
		$\sum Ranks$	$33^{\{4\}}$	$15^{\{1.5\}}$	$72^{\{8\}}$	$15^{\{1.5\}}$	$75^{\{9\}}$	$90^{\{10\}}$	$54^{\{6\}}$	$69^{\{7\}}$	$45^{\{5\}}$	$27^{\{3\}}$
150	BIAS	$\hat{\theta}$	$0.15549^{\{3\}}$	$0.14704^{\{2\}}$	$0.47431^{\{9\}}$	$0.14019^{\{1\}}$	$0.46006^{\{8\}}$	$0.51142^{\{10\}}$	$0.42005^{\{6\}}$	$0.45043^{\{7\}}$	$0.27066^{\{5\}}$	$0.16384^{\{4\}}$
		$\hat{k}$	$0.22275^{\{3\}}$	$0.20304^{\{2\}}$	$0.67238^{\{9\}}$	$0.19823^{\{1\}}$	$0.64712^{\{8\}}$	$0.72561^{\{10\}}$	$0.59108^{\{6\}}$	$0.64169^{\{7\}}$	$0.37254^{\{5\}}$	$0.23117^{\{4\}}$
		$\hat{\alpha}$	$0.01712^{\{4\}}$	$0.01683^{\{2\}}$	$0.04609^{\{7\}}$	$0.01241^{\{1\}}$	$0.04691^{\{8\}}$	$0.05031^{\{10\}}$	$0.03901^{\{6\}}$	$0.04728^{\{9\}}$	$0.03307^{\{5\}}$	$0.0169^{\{3\}}$
	MSE	$\hat{\theta}$	$0.10863^{\{3\}}$	$0.10409^{\{2\}}$	$0.35274^{\{9\}}$	$0.09317^{\{1\}}$	$0.3331^{\{8\}}$	$0.41139^{\{10\}}$	$0.27391^{\{6\}}$	$0.31463^{\{7\}}$	$0.19272^{\{5\}}$	$0.12297^{\{4\}}$
		$\hat{k}$	$0.23048^{\{3\}}$	$0.20546^{\{2\}}$	$0.73107^{\{9\}}$	$0.19361^{\{1\}}$	$0.68367^{\{8\}}$	$0.86376^{\{10\}}$	$0.55859^{\{6\}}$	$0.67892^{\{7\}}$	$0.41086^{\{5\}}$	$0.25772^{\{4\}}$
		$\hat{\alpha}$	$0.00153^{\{3\}}$	$0.00166^{\{4\}}$	$0.00416^{\{9\}}$	$0.00091^{\{1\}}$	$0.00413^{\{8\}}$	$0.00512^{\{10\}}$	$0.00296^{\{5\}}$	$0.00402^{\{7\}}$	$0.00377^{\{6\}}$	$0.00123^{\{2\}}$
	MRE	$\hat{\theta}$	$0.07775^{\{3\}}$	$0.07352^{\{2\}}$	$0.23715^{\{9\}}$	$0.07009^{\{1\}}$	$0.23003^{\{8\}}$	$0.25571^{\{10\}}$	$0.21003^{\{6\}}$	$0.22521^{\{7\}}$	$0.13533^{\{5\}}$	$0.08192^{\{4\}}$
		$\hat{k}$	$0.0891^{\{3\}}$	$0.08121^{\{2\}}$	$0.26895^{\{9\}}$	$0.07929^{\{1\}}$	$0.25885^{\{8\}}$	$0.29024^{\{10\}}$	$0.23643^{\{6\}}$	$0.25668^{\{7\}}$	$0.14901^{\{5\}}$	$0.09247^{\{4\}}$
		$\hat{\alpha}$	$0.03424^{\{4\}}$	$0.03365^{\{2\}}$	$0.09218^{\{7\}}$	$0.02481^{\{1\}}$	$0.09383^{\{8\}}$	$0.10063^{\{10\}}$	$0.07803^{\{6\}}$	$0.09457^{\{9\}}$	$0.06614^{\{5\}}$	$0.03379^{\{3\}}$
		$\sum Ranks$	$29^{\{3\}}$	$20^{\{2\}}$	$77^{\{9\}}$	$9^{\{1\}}$	$72^{\{8\}}$	$90^{\{10\}}$	$53^{\{6\}}$	$67^{\{7\}}$	$46^{\{5\}}$	$32^{\{4\}}$
250	BIAS	$\hat{\theta}$	$0.06276^{\{4\}}$	$0.06025^{\{1\}}$	$0.36003^{\{9\}}$	$0.06054^{\{2\}}$	$0.34474^{\{7\}}$	$0.40634^{\{10\}}$	$0.32146^{\{6\}}$	$0.35544^{\{8\}}$	$0.17883^{\{5\}}$	$0.06055^{\{3\}}$
		$\hat{k}$	$0.08466^{\{4\}}$	$0.08238^{\{2\}}$	$0.50324^{\{9\}}$	$0.0834^{\{3\}}$	$0.47774^{\{7\}}$	$0.55528^{\{10\}}$	$0.44517^{\{6\}}$	$0.49596^{\{8\}}$	$0.23422^{\{5\}}$	$0.08164^{\{1\}}$
		$\hat{\alpha}$	$0.00721^{\{4\}}$	$0.0064^{\{2\}}$	$0.03396^{\{8\}}$	$0.0056^{\{1\}}$	$0.03335^{\{7\}}$	$0.03756^{\{9\}}$	$0.03014^{\{6\}}$	$0.03923^{\{10\}}$	$0.02116^{\{5\}}$	$0.00651^{\{3\}}$
	MSE	$\hat{\theta}$	$0.03576^{\{2\}}$	$0.03544^{\{1\}}$	$0.2063^{\{9\}}$	$0.03784^{\{4\}}$	$0.19356^{\{7\}}$	$0.26847^{\{10\}}$	$0.15874^{\{6\}}$	$0.20613^{\{8\}}$	$0.1256^{\{5\}}$	$0.03737^{\{3\}}$
		$\hat{k}$	$0.0656^{\{2\}}$	$0.06491^{\{1\}}$	$0.41363^{\{8\}}$	$0.08037^{\{4\}}$	$0.38115^{\{7\}}$	$0.52546^{\{10\}}$	$0.31186^{\{6\}}$	$0.41671^{\{9\}}$	$0.23066^{\{5\}}$	$0.07075^{\{3\}}$
		$\hat{\alpha}$	$0.00052^{\{4\}}$	$0.00042^{\{2\}}$	$0.00218^{\{8\}}$	$0.00034^{\{1\}}$	$0.00202^{\{6.5\}}$	$0.00281^{\{10\}}$	$0.00158^{\{5\}}$	$0.00257^{\{9\}}$	$0.00202^{\{6.5\}}$	$0.00043^{\{3\}}$
	MRE	$\hat{\theta}$	$0.03138^{\{4\}}$	$0.03012^{\{1\}}$	$0.18001^{\{9\}}$	$0.03027^{\{2.5\}}$	$0.17237^{\{7\}}$	$0.20317^{\{10\}}$	$0.16073^{\{6\}}$	$0.17772^{\{8\}}$	$0.08942^{\{5\}}$	$0.03027^{\{2.5\}}$
		$\hat{k}$	$0.03387^{\{4\}}$	$0.03295^{\{2\}}$	$0.20129^{\{9\}}$	$0.03336^{\{3\}}$	$0.1911^{\{7\}}$	$0.22211^{\{10\}}$	$0.17807^{\{6\}}$	$0.19838^{\{8\}}$	$0.09369^{\{5\}}$	$0.03266^{\{1\}}$
		$\hat{\alpha}$	$0.01441^{\{4\}}$	$0.0128^{\{2\}}$	$0.06792^{\{8\}}$	$0.01121^{\{1\}}$	$0.0667^{\{7\}}$	$0.07513^{\{9\}}$	$0.06027^{\{6\}}$	$0.07846^{\{10\}}$	$0.04232^{\{5\}}$	$0.01302^{\{3\}}$
		$\sum Ranks$	$32^{\{4\}}$	$14^{\{1\}}$	$77^{\{8\}}$	$21.5^{\{2\}}$	$62.5^{\{7\}}$	$88^{\{10\}}$	$53^{\{6\}}$	$78^{\{9\}}$	$46.5^{\{5\}}$	$22.5^{\{3\}}$

Table 9: Numerical values of the PGD simulation for $\theta = 0.5$, $k = 1.5$, and $\alpha = 2.5$.

n	Mea.	$\widehat{Est.}$	MLE	ADE	CVME	MPSE	OLSE	RTADE	WLSE	LTADE	MSADE	MSALDE
30	BIAS	$\hat{\theta}$	$0.38242^{\{2\}}$	$0.54166^{\{7\}}$	$0.52674^{\{5\}}$	$0.6097^{\{8\}}$	$0.67283^{\{10\}}$	$0.52859^{\{6\}}$	$0.63371^{\{9\}}$	$0.48975^{\{3\}}$	$0.33305^{\{1\}}$	$0.4898^{\{4\}}$
		$\hat{k}$	$0.85665^{\{2\}}$	$1.16018^{\{7\}}$	$1.15991^{\{6\}}$	$1.19179^{\{8\}}$	$1.35032^{\{10\}}$	$1.15534^{\{5\}}$	$1.27672^{\{9\}}$	$1.12234^{\{4\}}$	$0.60399^{\{1\}}$	$0.90601^{\{3\}}$
		$\hat{\alpha}$	$0.42151^{\{2\}}$	$0.48617^{\{6\}}$	$0.46463^{\{5\}}$	$0.4923^{\{7\}}$	$0.53337^{\{10\}}$	$0.43717^{\{3\}}$	$0.52343^{\{9\}}$	$0.5112^{\{8\}}$	$0.37548^{\{1\}}$	$0.45308^{\{4\}}$
	MSE	$\hat{\theta}$	$0.4092^{\{2\}}$	$0.74692^{\{5\}}$	$0.78359^{\{6\}}$	$0.88812^{\{8\}}$	$1.23521^{\{10\}}$	$0.81818^{\{7\}}$	$0.9586^{\{9\}}$	$0.57516^{\{3\}}$	$0.38798^{\{1\}}$	$0.66622^{\{4\}}$
		$\hat{k}$	$1.51856^{\{2\}}$	$2.43628^{\{5\}}$	$2.59493^{\{7\}}$	$2.50059^{\{6\}}$	$3.76395^{\{10\}}$	$2.607^{\{8\}}$	$2.92356^{\{9\}}$	$2.06313^{\{4\}}$	$1.05837^{\{1\}}$	$1.78033^{\{3\}}$
		$\hat{\alpha}$	$0.28367^{\{2\}}$	$0.34345^{\{6\}}$	$0.32068^{\{4\}}$	$0.36124^{\{7\}}$	$0.40574^{\{9\}}$	$0.30142^{\{3\}}$	$0.38686^{\{8\}}$	$0.41734^{\{10\}}$	$0.23155^{\{1\}}$	$0.32278^{\{5\}}$
	MRE	$\hat{\theta}$	$0.76485^{\{2\}}$	$1.08332^{\{7\}}$	$1.05347^{\{5\}}$	$1.21939^{\{8\}}$	$1.34566^{\{10\}}$	$1.05717^{\{6\}}$	$1.26742^{\{9\}}$	$0.9795^{\{3\}}$	$0.6661^{\{1\}}$	$0.9796^{\{4\}}$
		$\hat{k}$	$0.5711^{\{2\}}$	$0.77345^{\{7\}}$	$0.77327^{\{6\}}$	$0.79452^{\{8\}}$	$0.90021^{\{10\}}$	$0.77023^{\{5\}}$	$0.85114^{\{9\}}$	$0.74823^{\{4\}}$	$0.40266^{\{1\}}$	$0.60401^{\{3\}}$
		$\hat{\alpha}$	$0.16861^{\{2\}}$	$0.19447^{\{6\}}$	$0.18585^{\{5\}}$	$0.19692^{\{7\}}$	$0.21335^{\{10\}}$	$0.17487^{\{3\}}$	$0.20937^{\{9\}}$	$0.20448^{\{8\}}$	$0.15019^{\{1\}}$	$0.18123^{\{4\}}$
	$\sum Ranks$		$18^{\{2\}}$	$56^{\{7\}}$	$49^{\{6\}}$	$67^{\{8\}}$	$89^{\{10\}}$	$46^{\{4\}}$	$80^{\{9\}}$	$47^{\{5\}}$	$9^{\{1\}}$	$34^{\{3\}}$
70	BIAS	$\hat{\theta}$	$0.27755^{\{2\}}$	$0.32299^{\{5\}}$	$0.32695^{\{6\}}$	$0.39841^{\{10\}}$	$0.39527^{\{9\}}$	$0.35632^{\{8\}}$	$0.34141^{\{7\}}$	$0.3066^{\{3\}}$	$0.23548^{\{1\}}$	$0.31744^{\{4\}}$
		$\hat{k}$	$0.67782^{\{3\}}$	$0.74039^{\{4\}}$	$0.7946^{\{6\}}$	$0.82619^{\{9\}}$	$0.8986^{\{10\}}$	$0.80748^{\{7\}}$	$0.75003^{\{5\}}$	$0.81299^{\{8\}}$	$0.46464^{\{1\}}$	$0.64873^{\{2\}}$
		$\hat{\alpha}$	$0.30513^{\{2\}}$	$0.33961^{\{5\}}$	$0.35183^{\{8\}}$	$0.37375^{\{9\}}$	$0.3927^{\{10\}}$	$0.34417^{\{6\}}$	$0.35131^{\{7\}}$	$0.3394^{\{4\}}$	$0.28615^{\{1\}}$	$0.31663^{\{3\}}$
	MSE	$\hat{\theta}$	$0.16235^{\{2\}}$	$0.21019^{\{4\}}$	$0.224^{\{5\}}$	$0.32223^{\{8\}}$	$0.33496^{\{10\}}$	$0.32581^{\{9\}}$	$0.23997^{\{6\}}$	$0.18402^{\{3\}}$	$0.16014^{\{1\}}$	$0.25086^{\{7\}}$
		$\hat{k}$	$0.79002^{\{3\}}$	$0.88505^{\{4\}}$	$1.01734^{\{7\}}$	$1.07377^{\{8\}}$	$1.27994^{\{10\}}$	$1.1909^{\{9\}}$	$0.92141^{\{5\}}$	$1.01137^{\{6\}}$	$0.4847^{\{1\}}$	$0.78597^{\{2\}}$
		$\hat{\alpha}$	$0.14383^{\{2\}}$	$0.1742^{\{4\}}$	$0.18673^{\{8\}}$	$0.20885^{\{9\}}$	$0.22701^{\{10\}}$	$0.18458^{\{7\}}$	$0.18028^{\{6\}}$	$0.17947^{\{5\}}$	$0.13965^{\{1\}}$	$0.17025^{\{3\}}$
	MRE	$\hat{\theta}$	$0.55511^{\{2\}}$	$0.64598^{\{5\}}$	$0.6539^{\{6\}}$	$0.79683^{\{10\}}$	$0.79054^{\{9\}}$	$0.71263^{\{8\}}$	$0.68281^{\{7\}}$	$0.6132^{\{3\}}$	$0.47096^{\{1\}}$	$0.63487^{\{4\}}$
		$\hat{k}$	$0.45188^{\{3\}}$	$0.49359^{\{4\}}$	$0.52973^{\{6\}}$	$0.55079^{\{9\}}$	$0.59906^{\{10\}}$	$0.53832^{\{7\}}$	$0.50002^{\{5\}}$	$0.542^{\{8\}}$	$0.30976^{\{1\}}$	$0.43249^{\{2\}}$
		$\hat{\alpha}$	$0.12205^{\{2\}}$	$0.13584^{\{5\}}$	$0.14073^{\{8\}}$	$0.1495^{\{9\}}$	$0.15708^{\{10\}}$	$0.13767^{\{6\}}$	$0.14053^{\{7\}}$	$0.13576^{\{4\}}$	$0.11446^{\{1\}}$	$0.12665^{\{3\}}$
	$\sum Ranks$		$21^{\{2\}}$	$40^{\{4\}}$	$60^{\{7\}}$	$81^{\{9\}}$	$88^{\{10\}}$	$67^{\{8\}}$	$55^{\{6\}}$	$44^{\{5\}}$	$9^{\{1\}}$	$30^{\{3\}}$
100	BIAS	$\hat{\theta}$	$0.2466^{\{2\}}$	$0.27262^{\{4\}}$	$0.30183^{\{7\}}$	$0.32147^{\{8\}}$	$0.34189^{\{10\}}$	$0.32708^{\{9\}}$	$0.29051^{\{5\}}$	$0.26738^{\{3\}}$	$0.20685^{\{1\}}$	$0.3^{\{6\}}$
		$\hat{k}$	$0.59251^{\{3\}}$	$0.66561^{\{4\}}$	$0.74558^{\{8\}}$	$0.67171^{\{5\}}$	$0.75917^{\{9\}}$	$0.76199^{\{10\}}$	$0.68518^{\{6\}}$	$0.69513^{\{7\}}$	$0.40982^{\{1\}}$	$0.59189^{\{2\}}$
		$\hat{\alpha}$	$0.27614^{\{2\}}$	$0.29416^{\{3\}}$	$0.31982^{\{9\}}$	$0.31149^{\{8\}}$	$0.34329^{\{10\}}$	$0.30203^{\{6\}}$	$0.30971^{\{7\}}$	$0.30028^{\{5\}}$	$0.25062^{\{1\}}$	$0.29877^{\{4\}}$
	MSE	$\hat{\theta}$	$0.11954^{\{2\}}$	$0.14291^{\{4\}}$	$0.19068^{\{6\}}$	$0.21175^{\{8\}}$	$0.24253^{\{9\}}$	$0.26779^{\{10\}}$	$0.14932^{\{5\}}$	$0.12441^{\{3\}}$	$0.1186^{\{1\}}$	$0.21167^{\{7\}}$
		$\hat{k}$	$0.61763^{\{2\}}$	$0.71973^{\{4\}}$	$0.91444^{\{8\}}$	$0.73815^{\{6\}}$	$0.9476^{\{9\}}$	$1.02668^{\{10\}}$	$0.72937^{\{5\}}$	$0.75853^{\{7\}}$	$0.37038^{\{1\}}$	$0.65586^{\{3\}}$
		$\hat{\alpha}$	$0.11823^{\{2\}}$	$0.13292^{\{3\}}$	$0.15427^{\{9\}}$	$0.15149^{\{8\}}$	$0.17559^{\{10\}}$	$0.14361^{\{6\}}$	$0.14185^{\{5\}}$	$0.13663^{\{4\}}$	$0.10705^{\{1\}}$	$0.14828^{\{7\}}$
	MRE	$\hat{\theta}$	$0.49321^{\{2\}}$	$0.54525^{\{4\}}$	$0.60365^{\{7\}}$	$0.64293^{\{8\}}$	$0.68378^{\{10\}}$	$0.65416^{\{9\}}$	$0.58103^{\{5\}}$	$0.53477^{\{3\}}$	$0.4137^{\{1\}}$	$0.6^{\{6\}}$
		$\hat{k}$	$0.39501^{\{3\}}$	$0.44374^{\{4\}}$	$0.49705^{\{8\}}$	$0.44781^{\{5\}}$	$0.50611^{\{9\}}$	$0.50799^{\{10\}}$	$0.45678^{\{6\}}$	$0.46342^{\{7\}}$	$0.27321^{\{1\}}$	$0.3946^{\{2\}}$
		$\hat{\alpha}$	$0.11046^{\{2\}}$	$0.11766^{\{3\}}$	$0.12793^{\{9\}}$	$0.1246^{\{8\}}$	$0.13732^{\{10\}}$	$0.12081^{\{6\}}$	$0.12388^{\{7\}}$	$0.12011^{\{5\}}$	$0.10025^{\{1\}}$	$0.11951^{\{4\}}$
	$\sum Ranks$		$20^{\{2\}}$	$33^{\{3\}}$	$71^{\{8\}}$	$64^{\{7\}}$	$86^{\{10\}}$	$76^{\{9\}}$	$51^{\{6\}}$	$44^{\{5\}}$	$9^{\{1\}}$	$41^{\{4\}}$
150	BIAS	$\hat{\theta}$	$0.20376^{\{2\}}$	$0.22971^{\{5\}}$	$0.25039^{\{8\}}$	$0.24397^{\{6\}}$	$0.26617^{\{9\}}$	$0.26618^{\{10\}}$	$0.22919^{\{4\}}$	$0.22237^{\{3\}}$	$0.19287^{\{1\}}$	$0.25013^{\{7\}}$
		$\hat{k}$	$0.4983^{\{2\}}$	$0.55241^{\{5\}}$	$0.62486^{\{8\}}$	$0.53261^{\{4\}}$	$0.63468^{\{9\}}$	$0.66456^{\{10\}}$	$0.55307^{\{6\}}$	$0.59418^{\{7\}}$	$0.39133^{\{1\}}$	$0.52523^{\{3\}}$
		$\hat{\alpha}$	$0.22743^{\{1\}}$	$0.26072^{\{7\}}$	$0.27816^{\{8\}}$	$0.25014^{\{4\}}$	$0.28029^{\{10\}}$	$0.27819^{\{9\}}$	$0.24895^{\{3\}}$	$0.25575^{\{5\}}$	$0.22797^{\{2\}}$	$0.2596^{\{6\}}$
	MSE	$\hat{\theta}$	$0.0781^{\{1\}}$	$0.09274^{\{3\}}$	$0.11074^{\{6\}}$	$0.11969^{\{7\}}$	$0.13144^{\{8\}}$	$0.14541^{\{10\}}$	$0.09861^{\{5\}}$	$0.08688^{\{2\}}$	$0.09468^{\{4\}}$	$0.13941^{\{9\}}$
		$\hat{k}$	$0.45246^{\{2\}}$	$0.52667^{\{5\}}$	$0.6476^{\{8\}}$	$0.49321^{\{3\}}$	$0.66161^{\{9\}}$	$0.73722^{\{10\}}$	$0.53711^{\{6\}}$	$0.59999^{\{7\}}$	$0.31946^{\{1\}}$	$0.50141^{\{4\}}$
		$\hat{\alpha}$	$0.08149^{\{1\}}$	$0.10338^{\{6\}}$	$0.11785^{\{10\}}$	$0.09994^{\{4\}}$	$0.11734^{\{9\}}$	$0.11619^{\{8\}}$	$0.09366^{\{3\}}$	$0.10209^{\{5\}}$	$0.08792^{\{2\}}$	$0.11185^{\{7\}}$
	MRE	$\hat{\theta}$	$0.40752^{\{2\}}$	$0.45941^{\{5\}}$	$0.50078^{\{8\}}$	$0.48793^{\{6\}}$	$0.53234^{\{9\}}$	$0.53236^{\{10\}}$	$0.45839^{\{4\}}$	$0.44474^{\{3\}}$	$0.38575^{\{1\}}$	$0.50025^{\{7\}}$
		$\hat{k}$	$0.3322^{\{2\}}$	$0.36827^{\{5\}}$	$0.41657^{\{8\}}$	$0.35507^{\{4\}}$	$0.42312^{\{9\}}$	$0.44304^{\{10\}}$	$0.36871^{\{6\}}$	$0.39612^{\{7\}}$	$0.26089^{\{1\}}$	$0.35015^{\{3\}}$
		$\hat{\alpha}$	$0.09097^{\{1\}}$	$0.10429^{\{7\}}$	$0.11126^{\{8\}}$	$0.10006^{\{4\}}$	$0.11212^{\{10\}}$	$0.11128^{\{9\}}$	$0.09958^{\{3\}}$	$0.1023^{\{5\}}$	$0.09119^{\{2\}}$	$0.10384^{\{6\}}$
	$\sum Ranks$		$14^{\{1\}}$	$48^{\{6\}}$	$72^{\{8\}}$	$42^{\{4\}}$	$82^{\{9\}}$	$86^{\{10\}}$	$40^{\{3\}}$	$44^{\{5\}}$	$15^{\{2\}}$	$52^{\{7\}}$
250	BIAS	$\hat{\theta}$	$0.16202^{\{1\}}$	$0.16951^{\{3\}}$	$0.20108^{\{8\}}$	$0.18616^{\{6\}}$	$0.20902^{\{9\}}$	$0.21365^{\{10\}}$	$0.17568^{\{4\}}$	$0.18335^{\{5\}}$	$0.16524^{\{2\}}$	$0.19602^{\{7\}}$
		$\hat{k}$	$0.37514^{\{2\}}$	$0.41545^{\{5\}}$	$0.47727^{\{8\}}$	$0.38633^{\{3\}}$	$0.51795^{\{9\}}$	$0.52296^{\{10\}}$	$0.40251^{\{4\}}$	$0.4743^{\{7\}}$	$0.33952^{\{1\}}$	$0.43819^{\{6\}}$
		$\hat{\alpha}$	$0.18974^{\{1\}}$	$0.19207^{\{2\}}$	$0.22647^{\{8\}}$	$0.21213^{\{7\}}$	$0.23243^{\{10\}}$	$0.22867^{\{9\}}$	$0.19897^{\{4\}}$	$0.2029^{\{5\}}$	$0.19569^{\{3\}}$	$0.20536^{\{6\}}$
	MSE	$\hat{\theta}$	$0.04488^{\{1\}}$	$0.04866^{\{2\}}$	$0.06909^{\{7\}}$	$0.06883^{\{6\}}$	$0.07635^{\{8\}}$	$0.0883^{\{10\}}$	$0.05117^{\{3\}}$	$0.05498^{\{4\}}$	$0.06487^{\{5\}}$	$0.07782^{\{9\}}$
		$\hat{k}$	$0.27424^{\{2\}}$	$0.33679^{\{5\}}$	$0.41223^{\{7\}}$	$0.28768^{\{3\}}$	$0.47177^{\{9\}}$	$0.48462^{\{10\}}$	$0.30569^{\{4\}}$	$0.42377^{\{8\}}$	$0.24349^{\{1\}}$	$0.34919^{\{6\}}$
		$\hat{\alpha}$	$0.05514^{\{1\}}$	$0.05621^{\{2\}}$	$0.077^{\{8\}}$	$0.0707^{\{7\}}$	$0.08137^{\{10\}}$	$0.08113^{\{9\}}$	$0.05947^{\{3\}}$	$0.06318^{\{4\}}$	$0.06669^{\{5\}}$	$0.07046^{\{6\}}$
	MRE	$\hat{\theta}$	$0.32405^{\{1\}}$	$0.33902^{\{3\}}$	$0.40216^{\{8\}}$	$0.37232^{\{6\}}$	$0.41804^{\{9\}}$	$0.4273^{\{10\}}$	$0.35136^{\{4\}}$	$0.3667^{\{5\}}$	$0.33048^{\{2\}}$	$0.39204^{\{7\}}$
		$\hat{k}$	$0.25009^{\{2\}}$	$0.27697^{\{5\}}$	$0.31818^{\{8\}}$	$0.25755^{\{3\}}$	$0.3453^{\{9\}}$	$0.34864^{\{10\}}$	$0.26834^{\{4\}}$	$0.3162^{\{7\}}$	$0.22635^{\{1\}}$	$0.29213^{\{6\}}$
		$\hat{\alpha}$	$0.0759^{\{1\}}$	$0.07683^{\{2\}}$	$0.09059^{\{8\}}$	$0.08485^{\{7\}}$	$0.09297^{\{10\}}$	$0.09147^{\{9\}}$	$0.07959^{\{4\}}$	$0.08116^{\{5\}}$	$0.07828^{\{3\}}$	$0.08214^{\{6\}}$
	$\sum Ranks$		$12^{\{1\}}$	$29^{\{3\}}$	$70^{\{8\}}$	$48^{\{5\}}$	$83^{\{9\}}$	$87^{\{10\}}$	$34^{\{4\}}$	$50^{\{6\}}$	$23^{\{2\}}$	$59^{\{7\}}$

Table 10: Numerical values of the PGD simulation for $\theta = 2.5$, $k = 1.5$, and $\alpha = 2.5$.

n	Mea.	$\widehat{Est.}$	MLE	ADE	CVME	MPSE	OLSE	RTADE	WLSE	LTADE	MSADE	MSALDE
30	BIAS	$\hat{\theta}$	$0.93923^{\{8\}}$	$0.85634^{\{4\}}$	$0.97288^{\{10\}}$	$0.78383^{\{3\}}$	$0.88991^{\{6\}}$	$0.95194^{\{9\}}$	$0.88865^{\{5\}}$	$0.93358^{\{7\}}$	$0.67018^{\{1\}}$	$0.77213^{\{2\}}$
		$\hat{k}$	$1.15822^{\{7\}}$	$1.03411^{\{4\}}$	$1.1675^{\{8\}}$	$0.91233^{\{3\}}$	$1.07246^{\{5\}}$	$1.18694^{\{10\}}$	$1.08699^{\{6\}}$	$1.1785^{\{9\}}$	$0.7726^{\{1\}}$	$0.8833^{\{2\}}$
		$\hat{\alpha}$	$0.49567^{\{8\}}$	$0.44167^{\{4\}}$	$0.5388^{\{9\}}$	$0.35914^{\{1\}}$	$0.45275^{\{6\}}$	$0.46876^{\{7\}}$	$0.4418^{\{5\}}$	$0.7977^{\{10\}}$	$0.42771^{\{3\}}$	$0.40009^{\{2\}}$
	MSE	$\hat{\theta}$	$1.2309^{\{8\}}$	$1.061^{\{4\}}$	$1.36703^{\{10\}}$	$0.89876^{\{2\}}$	$1.13982^{\{5\}}$	$1.27038^{\{9\}}$	$1.17523^{\{7\}}$	$1.16497^{\{6\}}$	$0.76733^{\{1\}}$	$0.90469^{\{3\}}$
		$\hat{k}$	$2.07807^{\{8\}}$	$1.70581^{\{4\}}$	$2.10838^{\{9\}}$	$1.31775^{\{2\}}$	$1.78986^{\{5\}}$	$2.31976^{\{10\}}$	$2.01056^{\{7\}}$	$1.85148^{\{6\}}$	$1.16608^{\{1\}}$	$1.41586^{\{3\}}$
		$\hat{\alpha}$	$0.44133^{\{8\}}$	$0.34419^{\{4\}}$	$0.52611^{\{9\}}$	$0.20548^{\{1\}}$	$0.35043^{\{5\}}$	$0.39264^{\{7\}}$	$0.35387^{\{6\}}$	$2.18571^{\{10\}}$	$0.33535^{\{3\}}$	$0.26867^{\{2\}}$
	MRE	$\hat{\theta}$	$0.37569^{\{8\}}$	$0.34254^{\{4\}}$	$0.38915^{\{10\}}$	$0.31353^{\{3\}}$	$0.35597^{\{6\}}$	$0.38078^{\{9\}}$	$0.35546^{\{5\}}$	$0.37343^{\{7\}}$	$0.26807^{\{1\}}$	$0.30885^{\{2\}}$
		$\hat{k}$	$0.77215^{\{7\}}$	$0.68941^{\{4\}}$	$0.77833^{\{8\}}$	$0.60822^{\{3\}}$	$0.71498^{\{5\}}$	$0.79129^{\{10\}}$	$0.72466^{\{6\}}$	$0.78567^{\{9\}}$	$0.51507^{\{1\}}$	$0.58887^{\{2\}}$
		$\hat{\alpha}$	$0.19827^{\{8\}}$	$0.17667^{\{4\}}$	$0.21552^{\{9\}}$	$0.14366^{\{1\}}$	$0.1811^{\{6\}}$	$0.18751^{\{7\}}$	$0.17672^{\{5\}}$	$0.31908^{\{10\}}$	$0.17109^{\{3\}}$	$0.16004^{\{2\}}$
	$\sum Ranks$		$70^{\{7\}}$	$36^{\{4\}}$	$82^{\{10\}}$	$19^{\{2\}}$	$49^{\{5\}}$	$78^{\{9\}}$	$52^{\{6\}}$	$74^{\{8\}}$	$15^{\{1\}}$	$20^{\{3\}}$
70	BIAS	$\hat{\theta}$	$0.71623^{\{7\}}$	$0.6438^{\{4\}}$	$0.71014^{\{6\}}$	$0.53655^{\{2\}}$	$0.73008^{\{9\}}$	$0.72043^{\{8\}}$	$0.66158^{\{5\}}$	$0.82648^{\{10\}}$	$0.53274^{\{1\}}$	$0.59957^{\{3\}}$
		$\hat{k}$	$0.84792^{\{7\}}$	$0.78379^{\{5\}}$	$0.84295^{\{6\}}$	$0.61946^{\{2\}}$	$0.85243^{\{8\}}$	$0.86852^{\{9\}}$	$0.7797^{\{4\}}$	$1.01174^{\{10\}}$	$0.59744^{\{1\}}$	$0.68724^{\{3\}}$
		$\hat{\alpha}$	$0.31654^{\{7\}}$	$0.28998^{\{5\}}$	$0.33472^{\{9\}}$	$0.2454^{\{1\}}$	$0.33258^{\{8\}}$	$0.29659^{\{6\}}$	$0.28322^{\{3\}}$	$0.51437^{\{10\}}$	$0.28481^{\{4\}}$	$0.27018^{\{2\}}$
	MSE	$\hat{\theta}$	$0.72691^{\{6\}}$	$0.62417^{\{4\}}$	$0.72897^{\{7\}}$	$0.47121^{\{1\}}$	$0.75513^{\{9\}}$	$0.73648^{\{8\}}$	$0.63182^{\{5\}}$	$0.95416^{\{10\}}$	$0.49207^{\{2\}}$	$0.56362^{\{3\}}$
		$\hat{k}$	$1.05889^{\{6\}}$	$0.94992^{\{5\}}$	$1.06551^{\{7\}}$	$0.6594^{\{1\}}$	$1.12588^{\{9\}}$	$1.11346^{\{8\}}$	$0.90929^{\{4\}}$	$1.45848^{\{10\}}$	$0.66493^{\{2\}}$	$0.78317^{\{3\}}$
		$\hat{\alpha}$	$0.18339^{\{7\}}$	$0.13928^{\{5\}}$	$0.19222^{\{8\}}$	$0.09511^{\{1\}}$	$0.19457^{\{9\}}$	$0.14752^{\{6\}}$	$0.13449^{\{4\}}$	$0.66164^{\{10\}}$	$0.13121^{\{3\}}$	$0.11817^{\{2\}}$
	MRE	$\hat{\theta}$	$0.28649^{\{7\}}$	$0.25752^{\{4\}}$	$0.28405^{\{6\}}$	$0.21462^{\{2\}}$	$0.29203^{\{9\}}$	$0.28817^{\{8\}}$	$0.26463^{\{5\}}$	$0.33059^{\{10\}}$	$0.2131^{\{1\}}$	$0.23983^{\{3\}}$
		$\hat{k}$	$0.56528^{\{7\}}$	$0.52253^{\{5\}}$	$0.56196^{\{6\}}$	$0.41297^{\{2\}}$	$0.56829^{\{8\}}$	$0.57901^{\{9\}}$	$0.5198^{\{4\}}$	$0.67449^{\{10\}}$	$0.39829^{\{1\}}$	$0.45816^{\{3\}}$
		$\hat{\alpha}$	$0.12662^{\{7\}}$	$0.11599^{\{5\}}$	$0.13389^{\{9\}}$	$0.09816^{\{1\}}$	$0.13303^{\{8\}}$	$0.11864^{\{6\}}$	$0.11329^{\{3\}}$	$0.20575^{\{10\}}$	$0.11392^{\{4\}}$	$0.10807^{\{2\}}$
	$\sum Ranks$		$61^{\{6\}}$	$42^{\{5\}}$	$64^{\{7\}}$	$13^{\{1\}}$	$77^{\{9\}}$	$68^{\{8\}}$	$37^{\{4\}}$	$90^{\{10\}}$	$19^{\{2\}}$	$24^{\{3\}}$
100	BIAS	$\hat{\theta}$	$0.6526^{\{8\}}$	$0.5885^{\{4\}}$	$0.67378^{\{9\}}$	$0.50432^{\{1\}}$	$0.64718^{\{7\}}$	$0.64441^{\{6\}}$	$0.58907^{\{5\}}$	$0.69952^{\{10\}}$	$0.50682^{\{2\}}$	$0.55763^{\{3\}}$
		$\hat{k}$	$0.76335^{\{8\}}$	$0.66573^{\{4\}}$	$0.78315^{\{9\}}$	$0.57051^{\{2\}}$	$0.75839^{\{7\}}$	$0.75794^{\{6\}}$	$0.6891^{\{5\}}$	$0.83653^{\{10\}}$	$0.56934^{\{1\}}$	$0.63196^{\{3\}}$
		$\hat{\alpha}$	$0.26711^{\{8\}}$	$0.25509^{\{6\}}$	$0.30498^{\{9\}}$	$0.21484^{\{1\}}$	$0.26114^{\{7\}}$	$0.2536^{\{5\}}$	$0.23801^{\{3\}}$	$0.37479^{\{10\}}$	$0.24791^{\{4\}}$	$0.22332^{\{2\}}$
	MSE	$\hat{\theta}$	$0.62827^{\{8\}}$	$0.51453^{\{4\}}$	$0.66142^{\{9\}}$	$0.41213^{\{1\}}$	$0.59744^{\{7\}}$	$0.57976^{\{6\}}$	$0.51667^{\{5\}}$	$0.7324^{\{10\}}$	$0.4306^{\{2\}}$	$0.48924^{\{3\}}$
		$\hat{k}$	$0.89846^{\{8\}}$	$0.668^{\{3\}}$	$0.93431^{\{9\}}$	$0.55041^{\{1\}}$	$0.85326^{\{7\}}$	$0.80439^{\{6\}}$	$0.72522^{\{5\}}$	$1.0748^{\{10\}}$	$0.56871^{\{2\}}$	$0.67368^{\{4\}}$
		$\hat{\alpha}$	$0.12961^{\{8\}}$	$0.10632^{\{5\}}$	$0.15336^{\{9\}}$	$0.07725^{\{1\}}$	$0.11376^{\{7\}}$	$0.10672^{\{6\}}$	$0.09913^{\{3\}}$	$0.32293^{\{10\}}$	$0.1006^{\{4\}}$	$0.08261^{\{2\}}$
	MRE	$\hat{\theta}$	$0.26104^{\{8\}}$	$0.2354^{\{4\}}$	$0.26951^{\{9\}}$	$0.20173^{\{1\}}$	$0.25887^{\{7\}}$	$0.25776^{\{6\}}$	$0.23563^{\{5\}}$	$0.27981^{\{10\}}$	$0.20273^{\{2\}}$	$0.22305^{\{3\}}$
		$\hat{k}$	$0.5089^{\{8\}}$	$0.44382^{\{4\}}$	$0.5221^{\{9\}}$	$0.38034^{\{2\}}$	$0.50559^{\{7\}}$	$0.50529^{\{6\}}$	$0.4594^{\{5\}}$	$0.55768^{\{10\}}$	$0.37956^{\{1\}}$	$0.4213^{\{3\}}$
		$\hat{\alpha}$	$0.10684^{\{8\}}$	$0.10203^{\{6\}}$	$0.12199^{\{9\}}$	$0.08594^{\{1\}}$	$0.10445^{\{7\}}$	$0.10144^{\{5\}}$	$0.0952^{\{3\}}$	$0.14991^{\{10\}}$	$0.09916^{\{4\}}$	$0.08933^{\{2\}}$
	$\sum Ranks$		$72^{\{8\}}$	$40^{\{5\}}$	$81^{\{9\}}$	$11^{\{1\}}$	$63^{\{7\}}$	$52^{\{6\}}$	$39^{\{4\}}$	$90^{\{10\}}$	$22^{\{2\}}$	$25^{\{3\}}$
150	BIAS	$\hat{\theta}$	$0.54254^{\{7\}}$	$0.49281^{\{5\}}$	$0.55019^{\{9\}}$	$0.41393^{\{1\}}$	$0.52908^{\{6\}}$	$0.54649^{\{8\}}$	$0.49115^{\{4\}}$	$0.64252^{\{10\}}$	$0.4352^{\{2\}}$	$0.46715^{\{3\}}$
		$\hat{k}$	$0.63208^{\{8\}}$	$0.57886^{\{5\}}$	$0.63153^{\{7\}}$	$0.46058^{\{1\}}$	$0.61174^{\{6\}}$	$0.64514^{\{9\}}$	$0.57182^{\{4\}}$	$0.74752^{\{10\}}$	$0.50133^{\{2\}}$	$0.53809^{\{3\}}$
		$\hat{\alpha}$	$0.22174^{\{7\}}$	$0.19692^{\{2\}}$	$0.22551^{\{9\}}$	$0.17187^{\{1\}}$	$0.22225^{\{8\}}$	$0.21232^{\{6\}}$	$0.20433^{\{4\}}$	$0.31608^{\{10\}}$	$0.20534^{\{5\}}$	$0.20053^{\{3\}}$
	MSE	$\hat{\theta}$	$0.46395^{\{9\}}$	$0.37465^{\{4\}}$	$0.45241^{\{8\}}$	$0.29609^{\{1\}}$	$0.42852^{\{6\}}$	$0.43721^{\{7\}}$	$0.38307^{\{5\}}$	$0.64952^{\{10\}}$	$0.32656^{\{2\}}$	$0.34592^{\{3\}}$
		$\hat{k}$	$0.63457^{\{9\}}$	$0.52214^{\{4\}}$	$0.6016^{\{7\}}$	$0.38396^{\{1\}}$	$0.58275^{\{6\}}$	$0.6071^{\{8\}}$	$0.53386^{\{5\}}$	$0.90545^{\{10\}}$	$0.45573^{\{2\}}$	$0.46945^{\{3\}}$
		$\hat{\alpha}$	$0.09438^{\{9\}}$	$0.06791^{\{4\}}$	$0.08849^{\{8\}}$	$0.05103^{\{1\}}$	$0.08323^{\{7\}}$	$0.07571^{\{6\}}$	$0.07337^{\{5\}}$	$0.24015^{\{10\}}$	$0.0678^{\{3\}}$	$0.06552^{\{2\}}$
	MRE	$\hat{\theta}$	$0.21701^{\{7\}}$	$0.19713^{\{5\}}$	$0.22007^{\{9\}}$	$0.16557^{\{1\}}$	$0.21163^{\{6\}}$	$0.21859^{\{8\}}$	$0.19646^{\{4\}}$	$0.25701^{\{10\}}$	$0.17408^{\{2\}}$	$0.18686^{\{3\}}$
		$\hat{k}$	$0.42139^{\{8\}}$	$0.38591^{\{5\}}$	$0.42102^{\{7\}}$	$0.30705^{\{1\}}$	$0.40783^{\{6\}}$	$0.43009^{\{9\}}$	$0.38121^{\{4\}}$	$0.49835^{\{10\}}$	$0.33422^{\{2\}}$	$0.35873^{\{3\}}$
		$\hat{\alpha}$	$0.08869^{\{7\}}$	$0.07877^{\{2\}}$	$0.09021^{\{9\}}$	$0.06875^{\{1\}}$	$0.0889^{\{8\}}$	$0.08493^{\{6\}}$	$0.08173^{\{4\}}$	$0.12643^{\{10\}}$	$0.08214^{\{5\}}$	$0.08021^{\{3\}}$
	$\sum Ranks$		$71^{\{8\}}$	$36^{\{4\}}$	$73^{\{9\}}$	$9^{\{1\}}$	$59^{\{6\}}$	$67^{\{7\}}$	$39^{\{5\}}$	$90^{\{10\}}$	$25^{\{2\}}$	$26^{\{3\}}$
250	BIAS	$\hat{\theta}$	$0.41302^{\{6\}}$	$0.40123^{\{4\}}$	$0.45982^{\{9\}}$	$0.32337^{\{1\}}$	$0.45279^{\{7\}}$	$0.4582^{\{8\}}$	$0.40341^{\{5\}}$	$0.51113^{\{10\}}$	$0.37113^{\{2\}}$	$0.39843^{\{3\}}$
		$\hat{k}$	$0.46939^{\{6\}}$	$0.45903^{\{5\}}$	$0.5231^{\{8\}}$	$0.35651^{\{1\}}$	$0.51411^{\{7\}}$	$0.53984^{\{9\}}$	$0.45518^{\{4\}}$	$0.58743^{\{10\}}$	$0.42006^{\{2\}}$	$0.45481^{\{3\}}$
		$\hat{\alpha}$	$0.15601^{\{4\}}$	$0.15098^{\{3\}}$	$0.17886^{\{9\}}$	$0.13145^{\{1\}}$	$0.17145^{\{7\}}$	$0.17213^{\{8\}}$	$0.16157^{\{6\}}$	$0.22492^{\{10\}}$	$0.15644^{\{5\}}$	$0.15063^{\{2\}}$
	MSE	$\hat{\theta}$	$0.28525^{\{6\}}$	$0.25951^{\{4\}}$	$0.33734^{\{9\}}$	$0.19344^{\{1\}}$	$0.31345^{\{7\}}$	$0.32222^{\{8\}}$	$0.26552^{\{5\}}$	$0.45129^{\{10\}}$	$0.2515^{\{2\}}$	$0.25326^{\{3\}}$
		$\hat{k}$	$0.37273^{\{6\}}$	$0.34124^{\{3\}}$	$0.44265^{\{9\}}$	$0.24228^{\{1\}}$	$0.40186^{\{7\}}$	$0.43607^{\{8\}}$	$0.34278^{\{5\}}$	$0.61226^{\{10\}}$	$0.34227^{\{4\}}$	$0.33478^{\{2\}}$
		$\hat{\alpha}$	$0.05025^{\{6\}}$	$0.04182^{\{3\}}$	$0.055^{\{9\}}$	$0.0292^{\{1\}}$	$0.05094^{\{7\}}$	$0.0512^{\{8\}}$	$0.04703^{\{5\}}$	$0.12227^{\{10\}}$	$0.04257^{\{4\}}$	$0.03826^{\{2\}}$
	MRE	$\hat{\theta}$	$0.16521^{\{6\}}$	$0.16049^{\{4\}}$	$0.18393^{\{9\}}$	$0.12935^{\{1\}}$	$0.18112^{\{7\}}$	$0.18328^{\{8\}}$	$0.16137^{\{5\}}$	$0.20445^{\{10\}}$	$0.14845^{\{2\}}$	$0.15937^{\{3\}}$
		$\hat{k}$	$0.31292^{\{6\}}$	$0.30602^{\{5\}}$	$0.34873^{\{8\}}$	$0.23768^{\{1\}}$	$0.34274^{\{7\}}$	$0.35989^{\{9\}}$	$0.30345^{\{4\}}$	$0.39162^{\{10\}}$	$0.28004^{\{2\}}$	$0.30321^{\{3\}}$
		$\hat{\alpha}$	$0.0624^{\{4\}}$	$0.06039^{\{3\}}$	$0.07154^{\{9\}}$	$0.05258^{\{1\}}$	$0.06858^{\{7\}}$	$0.06885^{\{8\}}$	$0.06463^{\{6\}}$	$0.08997^{\{10\}}$	$0.06258^{\{5\}}$	$0.06025^{\{2\}}$
	$\sum Ranks$		$50^{\{6\}}$	$34^{\{4\}}$	$79^{\{9\}}$	$9^{\{1\}}$	$63^{\{7\}}$	$74^{\{8\}}$	$45^{\{5\}}$	$90^{\{10\}}$	$28^{\{3\}}$	$23^{\{2\}}$

Table 11: Numerical values of the PGD simulation for $\theta = 0.9$, $k = 0.75$, and $\alpha = 1.5$.

n	Mea.	$\widehat{Est.}$	MLE	ADE	CVME	MPSE	OLSE	RTADE	WLSE	LTADE	MSADE	MSALDE
30	BIAS	$\hat{\theta}$	$0.58439^{\{3\}}$	$0.5961^{\{4\}}$	$0.66069^{\{9\}}$	$0.63149^{\{7\}}$	$0.71647^{\{10\}}$	$0.62649^{\{6\}}$	$0.65894^{\{8\}}$	$0.61168^{\{5\}}$	$0.42206^{\{1\}}$	$0.53852^{\{2\}}$
		$\hat{k}$	$1.01641^{\{6\}}$	$0.96645^{\{4\}}$	$1.08116^{\{9\}}$	$0.96451^{\{3\}}$	$1.11592^{\{10\}}$	$1.01073^{\{5\}}$	$1.0639^{\{8\}}$	$1.05118^{\{7\}}$	$0.57634^{\{1\}}$	$0.76996^{\{2\}}$
		$\hat{\alpha}$	$0.22655^{\{2\}}$	$0.22225^{\{1\}}$	$0.25631^{\{9\}}$	$0.24311^{\{6\}}$	$0.25261^{\{8\}}$	$0.2276^{\{3\}}$	$0.24665^{\{7\}}$	$0.32263^{\{10\}}$	$0.23531^{\{5\}}$	$0.23306^{\{4\}}$
	MSE	$\hat{\theta}$	$0.71287^{\{2\}}$	$0.7287^{\{4\}}$	$0.91571^{\{8\}}$	$0.83496^{\{6\}}$	$1.01676^{\{10\}}$	$0.91105^{\{7\}}$	$0.92174^{\{9\}}$	$0.77534^{\{5\}}$	$0.53045^{\{1\}}$	$0.72236^{\{3\}}$
		$\hat{k}$	$1.71554^{\{4\}}$	$1.62462^{\{3\}}$	$2.16201^{\{9\}}$	$1.81706^{\{6\}}$	$2.22168^{\{10\}}$	$2.05349^{\{7\}}$	$2.14458^{\{8\}}$	$1.79919^{\{5\}}$	$1.16974^{\{1\}}$	$1.5625^{\{2\}}$
		$\hat{\alpha}$	$0.08408^{\{4\}}$	$0.07781^{\{1\}}$	$0.10529^{\{9\}}$	$0.08435^{\{5\}}$	$0.09623^{\{8\}}$	$0.08233^{\{3\}}$	$0.09258^{\{7\}}$	$0.18482^{\{10\}}$	$0.0876^{\{6\}}$	$0.082^{\{2\}}$
	MRE	$\hat{\theta}$	$0.64932^{\{3\}}$	$0.66233^{\{4\}}$	$0.7341^{\{9\}}$	$0.70165^{\{7\}}$	$0.79607^{\{10\}}$	$0.6961^{\{6\}}$	$0.73216^{\{8\}}$	$0.67964^{\{5\}}$	$0.46896^{\{1\}}$	$0.59835^{\{2\}}$
		$\hat{k}$	$1.35521^{\{6\}}$	$1.28861^{\{4\}}$	$1.44155^{\{9\}}$	$1.28601^{\{3\}}$	$1.48789^{\{10\}}$	$1.34764^{\{5\}}$	$1.41853^{\{8\}}$	$1.40158^{\{7\}}$	$0.76845^{\{1\}}$	$1.02662^{\{2\}}$
		$\hat{\alpha}$	$0.15103^{\{2\}}$	$0.14817^{\{1\}}$	$0.17087^{\{9\}}$	$0.16207^{\{6\}}$	$0.16841^{\{8\}}$	$0.15173^{\{3\}}$	$0.16443^{\{7\}}$	$0.21509^{\{10\}}$	$0.15687^{\{5\}}$	$0.15538^{\{4\}}$
	$\sum Ranks$		$32^{\{4\}}$	$26^{\{3\}}$	$80^{\{9\}}$	$49^{\{6\}}$	$84^{\{10\}}$	$45^{\{5\}}$	$70^{\{8\}}$	$64^{\{7\}}$	$22^{\{1\}}$	$23^{\{2\}}$
70	BIAS	$\hat{\theta}$	$0.40185^{\{3\}}$	$0.40082^{\{2\}}$	$0.45616^{\{7\}}$	$0.42308^{\{5\}}$	$0.49656^{\{10\}}$	$0.49304^{\{9\}}$	$0.42716^{\{6\}}$	$0.4811^{\{8\}}$	$0.304^{\{1\}}$	$0.40193^{\{4\}}$
		$\hat{k}$	$0.73063^{\{5\}}$	$0.69854^{\{4\}}$	$0.79882^{\{7\}}$	$0.66646^{\{3\}}$	$0.82215^{\{9\}}$	$0.80953^{\{8\}}$	$0.7346^{\{6\}}$	$0.8538^{\{10\}}$	$0.44242^{\{1\}}$	$0.62368^{\{2\}}$
		$\hat{\alpha}$	$0.16315^{\{1\}}$	$0.16327^{\{2\}}$	$0.17466^{\{7\}}$	$0.16697^{\{4\}}$	$0.18543^{\{9\}}$	$0.17702^{\{8\}}$	$0.17361^{\{6\}}$	$0.22623^{\{10\}}$	$0.16658^{\{3\}}$	$0.17267^{\{5\}}$
	MSE	$\hat{\theta}$	$0.30147^{\{2\}}$	$0.30346^{\{3\}}$	$0.42333^{\{8\}}$	$0.35751^{\{6\}}$	$0.48861^{\{9\}}$	$0.50832^{\{10\}}$	$0.35298^{\{4\}}$	$0.41798^{\{7\}}$	$0.25256^{\{1\}}$	$0.35366^{\{5\}}$
		$\hat{k}$	$0.77348^{\{3\}}$	$0.73315^{\{2\}}$	$1.02302^{\{7\}}$	$0.79956^{\{5\}}$	$1.09177^{\{9\}}$	$1.11616^{\{10\}}$	$0.83668^{\{6\}}$	$1.05331^{\{8\}}$	$0.56882^{\{1\}}$	$0.78444^{\{4\}}$
		$\hat{\alpha}$	$0.03995^{\{1\}}$	$0.04058^{\{2\}}$	$0.04781^{\{8\}}$	$0.04131^{\{3\}}$	$0.05092^{\{9\}}$	$0.04715^{\{7\}}$	$0.04698^{\{6\}}$	$0.09205^{\{10\}}$	$0.04333^{\{4\}}$	$0.04523^{\{5\}}$
	MRE	$\hat{\theta}$	$0.4465^{\{3\}}$	$0.44536^{\{2\}}$	$0.50684^{\{7\}}$	$0.47009^{\{5\}}$	$0.55173^{\{10\}}$	$0.54782^{\{9\}}$	$0.47462^{\{6\}}$	$0.53456^{\{8\}}$	$0.33778^{\{1\}}$	$0.44659^{\{4\}}$
		$\hat{k}$	$0.97418^{\{5\}}$	$0.93139^{\{4\}}$	$1.0651^{\{7\}}$	$0.88862^{\{3\}}$	$1.0962^{\{9\}}$	$1.07938^{\{8\}}$	$0.97946^{\{6\}}$	$1.1384^{\{10\}}$	$0.58989^{\{1\}}$	$0.83157^{\{2\}}$
		$\hat{\alpha}$	$0.10876^{\{1\}}$	$0.10885^{\{2\}}$	$0.11644^{\{7\}}$	$0.11132^{\{4\}}$	$0.12362^{\{9\}}$	$0.11801^{\{8\}}$	$0.11574^{\{6\}}$	$0.15082^{\{10\}}$	$0.11105^{\{3\}}$	$0.11511^{\{5\}}$
	$\sum Ranks$		$24^{\{3\}}$	$23^{\{2\}}$	$65^{\{7\}}$	$38^{\{5\}}$	$83^{\{10\}}$	$77^{\{8\}}$	$52^{\{6\}}$	$81^{\{9\}}$	$16^{\{1\}}$	$36^{\{4\}}$
100	BIAS	$\hat{\theta}$	$0.37478^{\{6\}}$	$0.36944^{\{4\}}$	$0.41492^{\{8\}}$	$0.36739^{\{3\}}$	$0.44452^{\{10\}}$	$0.42126^{\{9\}}$	$0.37176^{\{5\}}$	$0.40226^{\{7\}}$	$0.28807^{\{1\}}$	$0.35353^{\{2\}}$
		$\hat{k}$	$0.67306^{\{6\}}$	$0.66037^{\{4\}}$	$0.72899^{\{8\}}$	$0.56453^{\{2\}}$	$0.75797^{\{10\}}$	$0.71676^{\{7\}}$	$0.66208^{\{5\}}$	$0.73865^{\{9\}}$	$0.43402^{\{1\}}$	$0.57423^{\{3\}}$
		$\hat{\alpha}$	$0.15132^{\{6\}}$	$0.1414^{\{1\}}$	$0.15734^{\{8\}}$	$0.1484^{\{2\}}$	$0.15529^{\{7\}}$	$0.15897^{\{9\}}$	$0.14905^{\{4\}}$	$0.18566^{\{10\}}$	$0.15104^{\{5\}}$	$0.14843^{\{3\}}$
	MSE	$\hat{\theta}$	$0.24512^{\{2\}}$	$0.25066^{\{3\}}$	$0.33403^{\{8\}}$	$0.26963^{\{5\}}$	$0.37899^{\{10\}}$	$0.34707^{\{9\}}$	$0.26089^{\{4\}}$	$0.2814^{\{7\}}$	$0.22479^{\{1\}}$	$0.27208^{\{6\}}$
		$\hat{k}$	$0.63393^{\{5\}}$	$0.63371^{\{4\}}$	$0.80391^{\{8\}}$	$0.59042^{\{2\}}$	$0.89504^{\{10\}}$	$0.81099^{\{9\}}$	$0.65371^{\{6\}}$	$0.75701^{\{7\}}$	$0.5174^{\{1\}}$	$0.63296^{\{3\}}$
		$\hat{\alpha}$	$0.03443^{\{5\}}$	$0.03097^{\{1\}}$	$0.03851^{\{9\}}$	$0.03298^{\{2\}}$	$0.03682^{\{7\}}$	$0.03797^{\{8\}}$	$0.03402^{\{3\}}$	$0.05964^{\{10\}}$	$0.0351^{\{6\}}$	$0.03421^{\{4\}}$
	MRE	$\hat{\theta}$	$0.41642^{\{6\}}$	$0.41048^{\{4\}}$	$0.46102^{\{8\}}$	$0.40821^{\{3\}}$	$0.49391^{\{10\}}$	$0.46806^{\{9\}}$	$0.41307^{\{5\}}$	$0.44695^{\{7\}}$	$0.32008^{\{1\}}$	$0.39281^{\{2\}}$
		$\hat{k}$	$0.89741^{\{6\}}$	$0.88049^{\{4\}}$	$0.97198^{\{8\}}$	$0.7527^{\{2\}}$	$1.01063^{\{10\}}$	$0.95568^{\{7\}}$	$0.88278^{\{5\}}$	$0.98487^{\{9\}}$	$0.5787^{\{1\}}$	$0.76564^{\{3\}}$
		$\hat{\alpha}$	$0.10088^{\{6\}}$	$0.09427^{\{1\}}$	$0.10489^{\{8\}}$	$0.09894^{\{2\}}$	$0.10353^{\{7\}}$	$0.10598^{\{9\}}$	$0.09936^{\{4\}}$	$0.12377^{\{10\}}$	$0.10069^{\{5\}}$	$0.09895^{\{3\}}$
	$\sum Ranks$		$48^{\{6\}}$	$26^{\{3\}}$	$73^{\{7\}}$	$23^{\{2\}}$	$81^{\{10\}}$	$76^{\{8.5\}}$	$41^{\{5\}}$	$76^{\{8.5\}}$	$22^{\{1\}}$	$29^{\{4\}}$
150	BIAS	$\hat{\theta}$	$0.31596^{\{3\}}$	$0.31784^{\{4\}}$	$0.37434^{\{10\}}$	$0.27316^{\{2\}}$	$0.36418^{\{9\}}$	$0.36411^{\{8\}}$	$0.32498^{\{6\}}$	$0.34173^{\{7\}}$	$0.26146^{\{1\}}$	$0.32309^{\{5\}}$
		$\hat{k}$	$0.58953^{\{6\}}$	$0.57809^{\{4\}}$	$0.66676^{\{10\}}$	$0.43899^{\{2\}}$	$0.64391^{\{7\}}$	$0.64798^{\{8\}}$	$0.5881^{\{5\}}$	$0.66571^{\{9\}}$	$0.40163^{\{1\}}$	$0.52433^{\{3\}}$
		$\hat{\alpha}$	$0.12816^{\{3\}}$	$0.1248^{\{2\}}$	$0.13865^{\{9\}}$	$0.11572^{\{1\}}$	$0.13496^{\{8\}}$	$0.12919^{\{5\}}$	$0.12865^{\{4\}}$	$0.1585^{\{10\}}$	$0.13037^{\{6\}}$	$0.13255^{\{7\}}$
	MSE	$\hat{\theta}$	$0.15959^{\{2\}}$	$0.17716^{\{4\}}$	$0.25053^{\{10\}}$	$0.15171^{\{1\}}$	$0.23734^{\{8\}}$	$0.24609^{\{9\}}$	$0.18182^{\{5\}}$	$0.19232^{\{6\}}$	$0.17101^{\{3\}}$	$0.21419^{\{7\}}$
		$\hat{k}$	$0.46852^{\{3\}}$	$0.47401^{\{4\}}$	$0.62542^{\{10\}}$	$0.36723^{\{1\}}$	$0.58829^{\{7\}}$	$0.62334^{\{9\}}$	$0.4854^{\{5\}}$	$0.59746^{\{8\}}$	$0.38535^{\{2\}}$	$0.50011^{\{6\}}$
		$\hat{\alpha}$	$0.02424^{\{3\}}$	$0.02362^{\{2\}}$	$0.0296^{\{9\}}$	$0.02055^{\{1\}}$	$0.02696^{\{8\}}$	$0.02585^{\{5\}}$	$0.02514^{\{4\}}$	$0.0434^{\{10\}}$	$0.02621^{\{6\}}$	$0.02683^{\{7\}}$
	MRE	$\hat{\theta}$	$0.35107^{\{3\}}$	$0.35316^{\{4\}}$	$0.41593^{\{10\}}$	$0.30351^{\{2\}}$	$0.40465^{\{9\}}$	$0.40456^{\{8\}}$	$0.36109^{\{6\}}$	$0.3797^{\{7\}}$	$0.29051^{\{1\}}$	$0.35899^{\{5\}}$
		$\hat{k}$	$0.78603^{\{6\}}$	$0.77079^{\{4\}}$	$0.88901^{\{10\}}$	$0.58532^{\{2\}}$	$0.85854^{\{7\}}$	$0.86398^{\{8\}}$	$0.78413^{\{5\}}$	$0.88761^{\{9\}}$	$0.53551^{\{1\}}$	$0.69911^{\{3\}}$
		$\hat{\alpha}$	$0.08544^{\{3\}}$	$0.0832^{\{2\}}$	$0.09244^{\{9\}}$	$0.07715^{\{1\}}$	$0.08997^{\{8\}}$	$0.08613^{\{5\}}$	$0.08577^{\{4\}}$	$0.10567^{\{10\}}$	$0.08691^{\{6\}}$	$0.08836^{\{7\}}$
	$\sum Ranks$		$32^{\{4\}}$	$30^{\{3\}}$	$87^{\{10\}}$	$13^{\{1\}}$	$71^{\{8\}}$	$65^{\{7\}}$	$44^{\{5\}}$	$76^{\{9\}}$	$27^{\{2\}}$	$50^{\{6\}}$
250	BIAS	$\hat{\theta}$	$0.2643^{\{4\}}$	$0.26031^{\{3\}}$	$0.3078^{\{9\}}$	$0.22152^{\{2\}}$	$0.30862^{\{10\}}$	$0.2939^{\{7\}}$	$0.28233^{\{6\}}$	$0.29842^{\{8\}}$	$0.21444^{\{1\}}$	$0.268^{\{5\}}$
		$\hat{k}$	$0.51495^{\{5\}}$	$0.49119^{\{4\}}$	$0.56836^{\{8\}}$	$0.35625^{\{2\}}$	$0.57362^{\{9\}}$	$0.5442^{\{7\}}$	$0.52328^{\{6\}}$	$0.57636^{\{10\}}$	$0.34268^{\{1\}}$	$0.45263^{\{3\}}$
		$\hat{\alpha}$	$0.10507^{\{4\}}$	$0.10514^{\{5\}}$	$0.11647^{\{8\}}$	$0.09687^{\{1\}}$	$0.11661^{\{9\}}$	$0.1111^{\{7\}}$	$0.11102^{\{6\}}$	$0.1335^{\{10\}}$	$0.0984^{\{2\}}$	$0.10456^{\{3\}}$
	MSE	$\hat{\theta}$	$0.10719^{\{2\}}$	$0.10751^{\{3\}}$	$0.15387^{\{9\}}$	$0.10997^{\{4\}}$	$0.15866^{\{10\}}$	$0.14563^{\{8\}}$	$0.12962^{\{5\}}$	$0.13995^{\{7\}}$	$0.10371^{\{1\}}$	$0.13551^{\{6\}}$
		$\hat{k}$	$0.35108^{\{5\}}$	$0.32697^{\{3\}}$	$0.4329^{\{8\}}$	$0.27488^{\{2\}}$	$0.44497^{\{9\}}$	$0.4058^{\{7\}}$	$0.37426^{\{6\}}$	$0.45161^{\{10\}}$	$0.25564^{\{1\}}$	$0.3379^{\{4\}}$
		$\hat{\alpha}$	$0.01606^{\{3\}}$	$0.01639^{\{4\}}$	$0.02005^{\{9\}}$	$0.01439^{\{1\}}$	$0.01969^{\{8\}}$	$0.01833^{\{6\}}$	$0.01841^{\{7\}}$	$0.0288^{\{10\}}$	$0.01576^{\{2\}}$	$0.01724^{\{5\}}$
	MRE	$\hat{\theta}$	$0.29367^{\{4\}}$	$0.28924^{\{3\}}$	$0.342^{\{9\}}$	$0.24613^{\{2\}}$	$0.34291^{\{10\}}$	$0.32656^{\{7\}}$	$0.3137^{\{6\}}$	$0.33158^{\{8\}}$	$0.23827^{\{1\}}$	$0.29778^{\{5\}}$
		$\hat{k}$	$0.6866^{\{5\}}$	$0.65492^{\{4\}}$	$0.75782^{\{8\}}$	$0.475^{\{2\}}$	$0.76483^{\{9\}}$	$0.7256^{\{7\}}$	$0.69771^{\{6\}}$	$0.76847^{\{10\}}$	$0.45691^{\{1\}}$	$0.6035^{\{3\}}$
		$\hat{\alpha}$	$0.07005^{\{4\}}$	$0.07009^{\{5\}}$	$0.07765^{\{8\}}$	$0.06458^{\{1\}}$	$0.07774^{\{9\}}$	$0.07407^{\{7\}}$	$0.07401^{\{6\}}$	$0.089^{\{10\}}$	$0.0656^{\{2\}}$	$0.0697^{\{3\}}$
	$\sum Ranks$		$36^{\{4\}}$	$34^{\{3\}}$	$76^{\{8\}}$	$17^{\{2\}}$	$83^{\{9.5\}}$	$63^{\{7\}}$	$54^{\{6\}}$	$83^{\{9.5\}}$	$12^{\{1\}}$	$37^{\{5\}}$

Table 12: Numerical values of the PGD simulation for $\theta = 1.5$, $k = 0.15$, and $\alpha = 0.9$.

n	Mea.	$\widehat{Est.}$	MLE	ADE	CVME	MPSE	OLSE	RTADE	WLSE	LTADE	MSADE	MSALDE
30	BIAS	$\hat{\theta}$	$0.33765^{\{3\}}$	$0.42926^{\{5\}}$	$0.47714^{\{7\}}$	$0.35674^{\{4\}}$	$0.47474^{\{6\}}$	$0.48237^{\{8\}}$	$0.48326^{\{9\}}$	$0.6893^{\{10\}}$	$0.21973^{\{1\}}$	$0.31049^{\{2\}}$
		$\hat{k}$	$0.263^{\{2\}}$	$0.40435^{\{6\}}$	$0.40122^{\{5\}}$	$0.3725^{\{4\}}$	$0.45401^{\{8\}}$	$0.4425^{\{7\}}$	$0.45957^{\{9\}}$	$0.61184^{\{10\}}$	$0.18532^{\{1\}}$	$0.30064^{\{3\}}$
		$\hat{\alpha}$	$0.12046^{\{2\}}$	$0.11544^{\{1\}}$	$0.13143^{\{8\}}$	$0.12612^{\{5\}}$	$0.13041^{\{7\}}$	$0.13018^{\{6\}}$	$0.12415^{\{3\}}$	$0.38283^{\{10\}}$	$0.13638^{\{9\}}$	$0.12437^{\{4\}}$
	MSE	$\hat{\theta}$	$0.25463^{\{2\}}$	$0.39536^{\{5\}}$	$0.52462^{\{8\}}$	$0.29585^{\{4\}}$	$0.52913^{\{9\}}$	$0.50583^{\{6\}}$	$0.51172^{\{7\}}$	$0.98991^{\{10\}}$	$0.13554^{\{1\}}$	$0.25521^{\{3\}}$
		$\hat{k}$	$0.32529^{\{2\}}$	$0.52018^{\{5\}}$	$0.6221^{\{6\}}$	$0.43563^{\{4\}}$	$0.7326^{\{9\}}$	$0.6308^{\{7\}}$	$0.65601^{\{8\}}$	$1.1982^{\{10\}}$	$0.15812^{\{1\}}$	$0.36002^{\{3\}}$
		$\hat{\alpha}$	$0.02576^{\{5\}}$	$0.02124^{\{1\}}$	$0.03024^{\{9\}}$	$0.0237^{\{2\}}$	$0.02704^{\{6\}}$	$0.0273^{\{7\}}$	$0.02469^{\{4\}}$	$0.73584^{\{10\}}$	$0.03011^{\{8\}}$	$0.02425^{\{3\}}$
	MRE	$\hat{\theta}$	$0.2251^{\{3\}}$	$0.28618^{\{5\}}$	$0.3181^{\{7\}}$	$0.23783^{\{4\}}$	$0.3165^{\{6\}}$	$0.32158^{\{8\}}$	$0.32218^{\{9\}}$	$0.45954^{\{10\}}$	$0.14649^{\{1\}}$	$0.20699^{\{2\}}$
		$\hat{k}$	$1.75332^{\{2\}}$	$2.69567^{\{6\}}$	$2.67481^{\{5\}}$	$2.48336^{\{4\}}$	$3.02676^{\{8\}}$	$2.95^{\{7\}}$	$3.0638^{\{9\}}$	$4.07896^{\{10\}}$	$1.23544^{\{1\}}$	$2.00425^{\{3\}}$
		$\hat{\alpha}$	$0.13385^{\{2\}}$	$0.12827^{\{1\}}$	$0.14604^{\{8\}}$	$0.14014^{\{5\}}$	$0.14489^{\{7\}}$	$0.14464^{\{6\}}$	$0.13795^{\{3\}}$	$0.42537^{\{10\}}$	$0.15153^{\{9\}}$	$0.13819^{\{4\}}$
	$\sum Ranks$		$23^{\{1\}}$	$35^{\{4\}}$	$63^{\{8\}}$	$36^{\{5\}}$	$66^{\{9\}}$	$62^{\{7\}}$	$61^{\{6\}}$	$90^{\{10\}}$	$32^{\{3\}}$	$27^{\{2\}}$
70	BIAS	$\hat{\theta}$	$0.24514^{\{2\}}$	$0.31942^{\{5\}}$	$0.33095^{\{6\}}$	$0.25261^{\{3\}}$	$0.34618^{\{7\}}$	$0.34789^{\{8\}}$	$0.34972^{\{9\}}$	$0.50612^{\{10\}}$	$0.17794^{\{1\}}$	$0.26644^{\{4\}}$
		$\hat{k}$	$0.21105^{\{2\}}$	$0.29149^{\{6\}}$	$0.28596^{\{5\}}$	$0.26655^{\{3\}}$	$0.31514^{\{7\}}$	$0.32202^{\{8\}}$	$0.33411^{\{9\}}$	$0.46841^{\{10\}}$	$0.16688^{\{1\}}$	$0.26914^{\{4\}}$
		$\hat{\alpha}$	$0.08224^{\{2\}}$	$0.07968^{\{1\}}$	$0.08946^{\{5\}}$	$0.08809^{\{4\}}$	$0.09096^{\{6\}}$	$0.09869^{\{9\}}$	$0.08568^{\{3\}}$	$0.209^{\{10\}}$	$0.09324^{\{8\}}$	$0.09113^{\{7\}}$
	MSE	$\hat{\theta}$	$0.12359^{\{2\}}$	$0.20424^{\{5\}}$	$0.22686^{\{6\}}$	$0.15537^{\{3\}}$	$0.24384^{\{8\}}$	$0.23505^{\{7\}}$	$0.24767^{\{9\}}$	$0.48191^{\{10\}}$	$0.09105^{\{1\}}$	$0.17607^{\{4\}}$
		$\hat{k}$	$0.11793^{\{2\}}$	$0.20897^{\{4\}}$	$0.21906^{\{7\}}$	$0.18887^{\{3\}}$	$0.255^{\{8\}}$	$0.21251^{\{6\}}$	$0.28346^{\{9\}}$	$0.48684^{\{10\}}$	$0.10379^{\{1\}}$	$0.21202^{\{5\}}$
		$\hat{\alpha}$	$0.01091^{\{2\}}$	$0.00997^{\{1\}}$	$0.01282^{\{6\}}$	$0.01175^{\{4\}}$	$0.01264^{\{5\}}$	$0.01457^{\{9\}}$	$0.01132^{\{3\}}$	$0.12138^{\{10\}}$	$0.0134^{\{8\}}$	$0.01329^{\{7\}}$
	MRE	$\hat{\theta}$	$0.16343^{\{2\}}$	$0.21295^{\{5\}}$	$0.22064^{\{6\}}$	$0.1684^{\{3\}}$	$0.23078^{\{7\}}$	$0.23193^{\{8\}}$	$0.23315^{\{9\}}$	$0.33742^{\{10\}}$	$0.11863^{\{1\}}$	$0.17763^{\{4\}}$
		$\hat{k}$	$1.40698^{\{2\}}$	$1.9433^{\{6\}}$	$1.90643^{\{5\}}$	$1.777^{\{3\}}$	$2.10091^{\{7\}}$	$2.14678^{\{8\}}$	$2.22739^{\{9\}}$	$3.12271^{\{10\}}$	$1.11254^{\{1\}}$	$1.79429^{\{4\}}$
		$\hat{\alpha}$	$0.09138^{\{2\}}$	$0.08854^{\{1\}}$	$0.0994^{\{5\}}$	$0.09787^{\{4\}}$	$0.10106^{\{6\}}$	$0.10965^{\{9\}}$	$0.0952^{\{3\}}$	$0.23223^{\{10\}}$	$0.1036^{\{8\}}$	$0.10125^{\{7\}}$
	$\sum Ranks$		$18^{\{1\}}$	$34^{\{4\}}$	$51^{\{6\}}$	$30^{\{2.5\}}$	$61^{\{7\}}$	$72^{\{9\}}$	$63^{\{8\}}$	$90^{\{10\}}$	$30^{\{2.5\}}$	$46^{\{5\}}$
100	BIAS	$\hat{\theta}$	$0.22343^{\{3\}}$	$0.27286^{\{5\}}$	$0.28746^{\{6\}}$	$0.20584^{\{2\}}$	$0.30929^{\{8\}}$	$0.33422^{\{9\}}$	$0.30172^{\{7\}}$	$0.43374^{\{10\}}$	$0.15952^{\{1\}}$	$0.26197^{\{4\}}$
		$\hat{k}$	$0.20198^{\{2\}}$	$0.25513^{\{4\}}$	$0.2576^{\{5\}}$	$0.20982^{\{3\}}$	$0.29298^{\{8\}}$	$0.31101^{\{9\}}$	$0.28396^{\{7\}}$	$0.40914^{\{10\}}$	$0.15395^{\{1\}}$	$0.2755^{\{6\}}$
		$\hat{\alpha}$	$0.07036^{\{2\}}$	$0.06885^{\{1\}}$	$0.07773^{\{5\}}$	$0.07225^{\{3\}}$	$0.07948^{\{7\}}$	$0.08792^{\{9\}}$	$0.07482^{\{4\}}$	$0.17341^{\{10\}}$	$0.07923^{\{6\}}$	$0.08294^{\{8\}}$
	MSE	$\hat{\theta}$	$0.10766^{\{3\}}$	$0.14874^{\{4\}}$	$0.16344^{\{5\}}$	$0.10718^{\{2\}}$	$0.19199^{\{8\}}$	$0.20495^{\{9\}}$	$0.17889^{\{7\}}$	$0.34977^{\{10\}}$	$0.07153^{\{1\}}$	$0.17411^{\{6\}}$
		$\hat{k}$	$0.1021^{\{2\}}$	$0.14427^{\{4\}}$	$0.15506^{\{5\}}$	$0.12147^{\{3\}}$	$0.21242^{\{9\}}$	$0.17622^{\{6\}}$	$0.17933^{\{7\}}$	$0.34386^{\{10\}}$	$0.07684^{\{1\}}$	$0.21069^{\{8\}}$
		$\hat{\alpha}$	$0.00796^{\{2\}}$	$0.00756^{\{1\}}$	$0.00925^{\{5\}}$	$0.0083^{\{3\}}$	$0.00963^{\{6\}}$	$0.01135^{\{9\}}$	$0.00858^{\{4\}}$	$0.07348^{\{10\}}$	$0.01008^{\{7\}}$	$0.0109^{\{8\}}$
	MRE	$\hat{\theta}$	$0.14895^{\{3\}}$	$0.18191^{\{5\}}$	$0.19164^{\{6\}}$	$0.13722^{\{2\}}$	$0.20619^{\{8\}}$	$0.22281^{\{9\}}$	$0.20115^{\{7\}}$	$0.28916^{\{10\}}$	$0.10635^{\{1\}}$	$0.17465^{\{4\}}$
		$\hat{k}$	$1.34655^{\{2\}}$	$1.70085^{\{4\}}$	$1.71736^{\{5\}}$	$1.3988^{\{3\}}$	$1.95319^{\{8\}}$	$2.07341^{\{9\}}$	$1.89304^{\{7\}}$	$2.72763^{\{10\}}$	$1.02633^{\{1\}}$	$1.83664^{\{6\}}$
		$\hat{\alpha}$	$0.07818^{\{2\}}$	$0.0765^{\{1\}}$	$0.08637^{\{5\}}$	$0.08028^{\{3\}}$	$0.08831^{\{7\}}$	$0.09769^{\{9\}}$	$0.08313^{\{4\}}$	$0.19268^{\{10\}}$	$0.08804^{\{6\}}$	$0.09215^{\{8\}}$
	$\sum Ranks$		$21^{\{1\}}$	$29^{\{4\}}$	$47^{\{5\}}$	$24^{\{2\}}$	$69^{\{8\}}$	$78^{\{9\}}$	$54^{\{6\}}$	$90^{\{10\}}$	$25^{\{3\}}$	$58^{\{7\}}$
150	BIAS	$\hat{\theta}$	$0.1926^{\{3\}}$	$0.24545^{\{5\}}$	$0.26674^{\{7\}}$	$0.17074^{\{2\}}$	$0.27777^{\{8\}}$	$0.30574^{\{9\}}$	$0.2634^{\{6\}}$	$0.39494^{\{10\}}$	$0.14301^{\{1\}}$	$0.24086^{\{4\}}$
		$\hat{k}$	$0.18053^{\{3\}}$	$0.22984^{\{4\}}$	$0.24629^{\{5\}}$	$0.17349^{\{2\}}$	$0.25923^{\{8\}}$	$0.28904^{\{9\}}$	$0.24982^{\{6\}}$	$0.37815^{\{10\}}$	$0.14414^{\{1\}}$	$0.25187^{\{7\}}$
		$\hat{\alpha}$	$0.06185^{\{3\}}$	$0.06153^{\{1\}}$	$0.06797^{\{5\}}$	$0.06159^{\{2\}}$	$0.07049^{\{7\}}$	$0.08092^{\{9\}}$	$0.06677^{\{4\}}$	$0.14685^{\{10\}}$	$0.06834^{\{6\}}$	$0.07151^{\{8\}}$
	MSE	$\hat{\theta}$	$0.07644^{\{2\}}$	$0.11513^{\{4\}}$	$0.13828^{\{6\}}$	$0.07959^{\{3\}}$	$0.14573^{\{7\}}$	$0.16755^{\{9\}}$	$0.12961^{\{5\}}$	$0.29423^{\{10\}}$	$0.05726^{\{1\}}$	$0.14576^{\{8\}}$
		$\hat{k}$	$0.07042^{\{2\}}$	$0.10249^{\{4\}}$	$0.12177^{\{5\}}$	$0.08406^{\{3\}}$	$0.13916^{\{7\}}$	$0.14351^{\{8\}}$	$0.12394^{\{6\}}$	$0.29916^{\{10\}}$	$0.0619^{\{1\}}$	$0.16941^{\{9\}}$
		$\hat{\alpha}$	$0.00608^{\{1\}}$	$0.0061^{\{2\}}$	$0.00714^{\{5\}}$	$0.00618^{\{3\}}$	$0.00754^{\{7\}}$	$0.00939^{\{9\}}$	$0.00675^{\{4\}}$	$0.04876^{\{10\}}$	$0.00734^{\{6\}}$	$0.00831^{\{8\}}$
	MRE	$\hat{\theta}$	$0.1284^{\{3\}}$	$0.16363^{\{5\}}$	$0.17782^{\{7\}}$	$0.11383^{\{2\}}$	$0.18518^{\{8\}}$	$0.20383^{\{9\}}$	$0.1756^{\{6\}}$	$0.26329^{\{10\}}$	$0.09534^{\{1\}}$	$0.16057^{\{4\}}$
		$\hat{k}$	$1.20351^{\{3\}}$	$1.53228^{\{4\}}$	$1.64191^{\{5\}}$	$1.15658^{\{2\}}$	$1.72822^{\{8\}}$	$1.92692^{\{9\}}$	$1.66548^{\{6\}}$	$2.52098^{\{10\}}$	$0.96092^{\{1\}}$	$1.67912^{\{7\}}$
		$\hat{\alpha}$	$0.06872^{\{3\}}$	$0.06837^{\{1\}}$	$0.07552^{\{5\}}$	$0.06843^{\{2\}}$	$0.07832^{\{7\}}$	$0.08991^{\{9\}}$	$0.07419^{\{4\}}$	$0.16317^{\{10\}}$	$0.07594^{\{6\}}$	$0.07946^{\{8\}}$
	$\sum Ranks$		$23^{\{2\}}$	$30^{\{4\}}$	$50^{\{6\}}$	$21^{\{1\}}$	$67^{\{8\}}$	$80^{\{9\}}$	$47^{\{5\}}$	$90^{\{10\}}$	$24^{\{3\}}$	$63^{\{7\}}$
250	BIAS	$\hat{\theta}$	$0.17447^{\{3\}}$	$0.20452^{\{5\}}$	$0.24567^{\{7\}}$	$0.11444^{\{1\}}$	$0.25323^{\{8\}}$	$0.2924^{\{9\}}$	$0.22953^{\{6\}}$	$0.3336^{\{10\}}$	$0.14195^{\{2\}}$	$0.20204^{\{4\}}$
		$\hat{k}$	$0.16822^{\{3\}}$	$0.19586^{\{4\}}$	$0.22789^{\{7\}}$	$0.11398^{\{1\}}$	$0.23698^{\{8\}}$	$0.2811^{\{9\}}$	$0.22167^{\{6\}}$	$0.32115^{\{10\}}$	$0.14187^{\{2\}}$	$0.21089^{\{5\}}$
		$\hat{\alpha}$	$0.05201^{\{2\}}$	$0.05339^{\{3\}}$	$0.06156^{\{7\}}$	$0.04486^{\{1\}}$	$0.06218^{\{8\}}$	$0.07326^{\{9\}}$	$0.05938^{\{6\}}$	$0.11952^{\{10\}}$	$0.05777^{\{5\}}$	$0.05745^{\{4\}}$
	MSE	$\hat{\theta}$	$0.06205^{\{3\}}$	$0.07772^{\{4\}}$	$0.10725^{\{7\}}$	$0.04157^{\{1\}}$	$0.1144^{\{8\}}$	$0.14545^{\{9\}}$	$0.09106^{\{5\}}$	$0.2028^{\{10\}}$	$0.05789^{\{2\}}$	$0.09878^{\{6\}}$
		$\hat{k}$	$0.05799^{\{2\}}$	$0.06797^{\{4\}}$	$0.09051^{\{6\}}$	$0.04061^{\{1\}}$	$0.10076^{\{7\}}$	$0.12891^{\{9\}}$	$0.08194^{\{5\}}$	$0.20189^{\{10\}}$	$0.05831^{\{3\}}$	$0.10689^{\{8\}}$
		$\hat{\alpha}$	$0.00439^{\{2\}}$	$0.00456^{\{3\}}$	$0.00557^{\{6\}}$	$0.00358^{\{1\}}$	$0.00558^{\{7\}}$	$0.00744^{\{9\}}$	$0.00527^{\{4\}}$	$0.02993^{\{10\}}$	$0.00536^{\{5\}}$	$0.00565^{\{8\}}$
	MRE	$\hat{\theta}$	$0.11632^{\{3\}}$	$0.13635^{\{5\}}$	$0.16378^{\{7\}}$	$0.07629^{\{1\}}$	$0.16882^{\{8\}}$	$0.19493^{\{9\}}$	$0.15302^{\{6\}}$	$0.2224^{\{10\}}$	$0.09464^{\{2\}}$	$0.1347^{\{4\}}$
		$\hat{k}$	$1.12144^{\{3\}}$	$1.30574^{\{4\}}$	$1.51924^{\{7\}}$	$0.75985^{\{1\}}$	$1.57989^{\{8\}}$	$1.87398^{\{9\}}$	$1.47782^{\{6\}}$	$2.14098^{\{10\}}$	$0.94582^{\{2\}}$	$1.40596^{\{5\}}$
		$\hat{\alpha}$	$0.05779^{\{2\}}$	$0.05932^{\{3\}}$	$0.0684^{\{7\}}$	$0.04984^{\{1\}}$	$0.06909^{\{8\}}$	$0.0814^{\{9\}}$	$0.06598^{\{6\}}$	$0.1328^{\{10\}}$	$0.06419^{\{5\}}$	$0.06383^{\{4\}}$
	$\sum Ranks$		$23^{\{2\}}$	$35^{\{4\}}$	$61^{\{7\}}$	$9^{\{1\}}$	$70^{\{8\}}$	$81^{\{9\}}$	$50^{\{6\}}$	$90^{\{10\}}$	$28^{\{3\}}$	$48^{\{5\}}$

Table 13: Partial and overall ranks of the techniques for estimation of the PGD using various values of parameters.

Parameter	n	MLE	ADE	CVME	MPSE	OLSE	RTADE	WLSE	LTADE	MSADE	MSALDE
$\theta = 0.25,\ k = 0.5,\ \alpha = 0.75$	30	4.0	5.0	8.0	6.0	10.0	3.0	9.0	7.0	1.0	2.0
	70	3.0	6.0	7.0	4.0	10.0	5.0	9.0	8.0	1.0	2.0
	100	4.0	5.0	7.5	1.0	10.0	6.0	9.0	7.5	2.0	3.0
	150	4.0	5.0	9.0	1.0	10.0	7.0	6.0	8.0	2.0	3.0
	250	4.0	5.0	9.0	1.0	10.0	7.0	6.0	8.0	2.0	3.0
$\theta = 0.2,\ k = 1.2,\ \alpha = 0.9$	30	3.0	6.0	5.0	10.0	7.0	2.0	8.5	8.5	1.0	4.0
	70	2.0	5.0	7.0	8.0	9.0	3.5	6.0	10.0	1.0	3.5
	100	2.0	5.0	8.0	7.0	10.0	4.0	6.0	9.0	1.0	3.0
	150	2.0	6.0	8.5	4.0	10.0	5.0	7.0	8.5	1.0	3.0
	250	1.0	5.0	9.0	4.0	10.0	7.0	6.0	8.0	2.0	3.0
$\theta = 1.5,\ k = 0.25,\ \alpha = 1.5$	30	1.0	4.0	8.0	3.0	9.0	6.0	7.0	10.0	2.0	5.0
	70	3.0	4.0	5.0	2.0	9.0	8.0	6.0	10.0	1.0	7.0
	100	3.0	6.5	4.0	1.5	8.0	6.5	5.0	10.0	1.5	9.0
	150	4.0	5.0	6.0	1.0	8.0	7.0	3.0	10.0	2.0	9.0
	250	3.0	8.0	4.0	1.0	7.0	6.0	5.0	10.0	2.0	9.0
$\theta = 2.0,\ k = 2.5,\ \alpha = 0.5$	30	5.0	4.0	7.5	3.0	9.0	10.0	6.0	7.5	1.0	2.0
	70	5.0	3.0	9.0	2.0	8.0	10.0	6.0	7.0	4.0	1.0
	100	4.0	1.5	8.0	1.5	9.0	10.0	6.0	7.0	5.0	3.0
	150	3.0	2.0	9.0	1.0	8.0	10.0	6.0	7.0	5.0	4.0
	250	4.0	1.0	8.0	2.0	7.0	10.0	6.0	9.0	5.0	3.0
$\theta = 0.5,\ k = 1.5,\ \alpha = 2.5$	30	2.0	7.0	6.0	8.0	10.0	4.0	9.0	5.0	1.0	3.0
	70	2.0	4.0	7.0	9.0	10.0	8.0	6.0	5.0	1.0	3.0
	100	2.0	3.0	8.0	7.0	10.0	9.0	6.0	5.0	1.0	4.0
	150	1.0	6.0	8.0	4.0	9.0	10.0	3.0	5.0	2.0	7.0
	250	1.0	3.0	8.0	5.0	9.0	10.0	4.0	6.0	2.0	7.0
$\theta = 2.5,\ k = 1.5,\ \alpha = 2.5$	30	7.0	4.0	10.0	2.0	5.0	9.0	6.0	8.0	1.0	3.0
	70	6.0	5.0	7.0	1.0	9.0	8.0	4.0	10.0	2.0	3.0
	100	8.0	5.0	9.0	1.0	7.0	6.0	4.0	10.0	2.0	3.0
	150	8.0	4.0	9.0	1.0	6.0	7.0	5.0	10.0	2.0	3.0
	250	6.0	4.0	9.0	1.0	7.0	8.0	5.0	10.0	3.0	2.0
$\theta = 0.9,\ k = 0.75,\ \alpha = 1.5$	30	4.0	3.0	9.0	6.0	10.0	5.0	8.0	7.0	1.0	2.0
	70	3.0	2.0	7.0	5.0	10.0	8.0	6.0	9.0	1.0	4.0
	100	6.0	3.0	7.0	2.0	10.0	8.5	5.0	8.5	1.0	4.0
	150	4.0	3.0	10.0	1.0	8.0	7.0	5.0	9.0	2.0	6.0
	250	4.0	3.0	8.0	2.0	9.5	7.0	6.0	9.5	1.0	5.0
$\theta = 1.5,\ k = 0.15,\ \alpha = 0.9$	30	1.0	4.0	8.0	5.0	9.0	7.0	6.0	10.0	3.0	2.0
	70	1.0	4.0	6.0	2.5	7.0	9.0	8.0	10.0	2.5	5.0
	100	1.0	4.0	5.0	2.0	8.0	9.0	6.0	10.0	3.0	7.0
	150	2.0	4.0	6.0	1.0	8.0	9.0	5.0	10.0	3.0	7.0
	250	2.0	4.0	7.0	1.0	8.0	9.0	6.0	10.0	3.0	5.0
$\sum$ Ranks		135.0	171.0	300.5	130.5	347.5	290.5	241.5	337.0	80.0	166.5
Overall Rank		3	5	8	2	10	7	6	9	1	4

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 24, No. 1, 2026 (New Series), pp 27–52
http://www.ssca.org.in/journal

An Inferential Study of Two Kumaraswamy Populations under Joint Ranked Set Sampling

Mahesh K. Bhingikar and D. P. Raykundaliya

Department of Statistics
Sardar Patel University, Vallabh Vidyanagar, Anand, Gujarat, India

Received: 24 January 2025; Revised: 02 May 2025; Accepted: 04 May 2025

Abstract

In this study, we have derived the expression for Maximum Likelihood (ML) estimates and likelihood ratio test (LRT) for two Kumaraswamy populations under Joint Ranked Set Sampling (JRSS), Joint Modified Minimum Ranked Set Sampling (JMnRSS), Joint Modified Maximum Ranked Set Sampling (JMxRSS), and Joint Simple Random Sampling (JSRS). The performance of the ML estimates is evaluated using Root Mean Square Error (RMSE) and the bias criterion. LRT is conducted to test the equality of the first shape parameters of two Kumaraswamy populations when the other two shape parameters are known. The power of the LRT is determined to compare the test performance under the mentioned sampling schemes. This simulations for ML and LRT results are obtained using Monte Carlo simulations in R Studio. We found that all joint ranked sampling schemes demonstrate superior performance compared to JSRS in both ML estimation and LRT. However, the JRSS exhibited the best results in LRT when the other two shape parameters are known, while the JMnRSS excelled in ML estimation. Additionally, we provide an illustration using real-life lung cancer data to highlight these findings.

Key words: Kumaraswamy distribution; Joint modified minimum ranked set sampling; Joint modified maximum ranked set sampling; Maximum likelihood estimates.

AMS Subject Classifications: 62D05,62F07, 62F10, 62F03

The video recording of the paper made under the SSCA's Online Lecture series is available at the Youtube channel URL `https://youtu.be/xp3vRT40tWg`.

1. Introduction

According to World Health Organization (2023), lung cancer is one of the leading causes of cancer-related mortality worldwide, accounting for a significant proportion of all cancer deaths. It's costly to diagnose lung cancer in India: $Rs.15,000 - Rs.1,50,000$ and in the USA: \$2,000–\$20,000. To reduce the cost of testing for lung cancer in the laboratory, it is possible to determine the probability of cancer by using modeling based on auxiliary

Corresponding Author: D. P. Raykundaliya
Email: dp_raykundaliya@spuvvn.edu

variables. The probability of cancer can be calculated by logistic regression, Probit Model, Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Decision Trees (*e.g.*, CART), Random Forests, Neural Networks (SLPs, DNNs), *etc.* We can use the probability (proportion) calculated from modeling to analyze the lung cancer population based on a sample. If the population is finite and the goal is to estimate the population proportion, we can apply sampling theory methods. If our sample size is small relative to the population and we aim to predict the proportion while considering greater variation in the data, we should assume an infinite population. So, to analyze the proportion data in the range of 0 to 1, we may have two choices between Beta-I distribution and Kumaraswamy distribution. According to the available literature, the Kumaraswamy distribution has additional characteristics compared to the Beta-I distribution, such as the fact that Kumaraswamy variables exhibit closeness under exponentiation and linear transformation (Mitnik, 2013). Therefore, to analyze the proportion data, it is advisable to use the Kumaraswamy distribution. If we wish to analyze the proportion of lung cancer by fitting the Kumaraswamy distribution, we can address only the questions related to lung cancer in the overall population, not separately for males and females. To do so, we need a sample of the male population of size m_1 and a sample of the female population of size m_2 for the study. As a result, both the cost and the duration of the study increase. To avoid these issues, we can use joint censoring schemes used in reliability theory (Ashour and Abo-Kasem, 2014; Ding and Gui, 2023). This scheme involves selecting a combined total of m units from the population. To analyze the situation, assume we choose $m_1 = 12$ patients from the male population and $m_2 = 5$ patients from the female population, we may find that the estimates related to the female population are not as efficient as those for the male population. However, by using joint censoring with a total sample size of $m = 17$ (*i.e.*, $m_1 + m_2 = m$), we can still proceed with our analysis. Joint censoring refers to the need for a combined number of sample units rather than individual sample units from each population. This method is applicable when both populations are independent and relatively similar, which is true in this case. To increase efficiency and reduce cost and time, popular schemes are also available in the literature called ranked set sampling (RSS) schemes (McIntyre, 1952; Chen *et al.*, 2004; Wolfe, 2004; Akgul and Senoglu, 2017; Abbasi and Shahd, 2017; Koyuncu, 2018).

Therefore, analysis of lung cancer data, we believe that the various joint ranked set sampling schemes are well-suited compared to JSRS. The aim of any study is to increase efficiency with lesser cost and a smaller sample size. As discussed previously, it is possible to reduce the sample size by using the various joint sampling schemes. Initially various JRSS schemes used to estimate the parameter of the exponential distribution by Raykundaliya and Patel (2022b) and to estimate the parameters of the Rayleigh distribution using JRSS by Patel *et al.* (2022). Raykundaliya and Patel (2022a) used the joint percentile Ranked Set Sampling (JPRSS) scheme to estimate the parameters of the exponential distribution and recently Raykundaliya and Bhingikar (2025) also used this sampling scheme to estimate the parameters for the Weibull population using JRSS. Therefore, we use the JRSS schemes to estimate the parameters by ML method for the Kumaraswamy population. Rest of paper is arranged as follows. Section 2 - 4 derive the expressions required for ML estimates as well as for the Fisher information (I) matrix under JRSS, JMnRSS, JMxRSS, and JSRS. Section 5 discusses the LRT for the JRSS, JMnRSS, JMxRSS, and JSRS. In Section 6, we have conducted the simulation for ML estimates of parameters of two Kumaraswamy populations using the mentioned sampling schemes. Further, we discuss the application of

the Kumaraswamy population for lung cancer data using various joint ranked set sampling. Finally, some concluding remarks are given in Section 7.

2. Joint Ranked Set Sampling (JRSS)

In this Section, we have derived the ML estimates of the parameters of two Kumaraswamy populations under JRSS.

2.1. PDF and CDF of Kumaraswamy populations

Let $X_1, X_2, ..., X_{m_1}$ be r.s (random sample) drawn from Kumaraswamy (say it population-I) whose pdf and cdf are given as follows,

$$f(x; a_1, b_1) = \begin{cases} a_1 b_1 x^{a_1-1}(1-x^{a_1})^{b_1-1}, & \text{if } 0 < x < 1 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$F(x; a_1, b_1) = 1 - (1-x^{a_1})^{b_1} \tag{2}$$

where a_1, b_1 are non negative shape parameters
Let $Y_1, Y_2, ..., Y_{m_1}$ be r.s drawn from Kumaraswamy (say it population-II) whose pdf and cdf are given as follows,

$$g(y; a_2, b_2) = \begin{cases} a_2 b_2 y^{a_2-1}(1-y^{a_2})^{b_2-1}, & \text{if }, 0 < y < 1 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$G(y; a_2, b_2) = 1 - (1-y^{a_2})^{b_2} \tag{4}$$

where a_2, b_2 are non negative shape parameters

2.2. Algorithm of JRSS

Algorithm 1 The algorithm to select JRSS from two populations

1: Select m_1 observations from a population-I and m_2 observations from population-II randomly and combine these observations to create a joint sample of size $m = m_1 + m_2$.
2: Arrange all the observations from the joint sample in ascending order.
3: To get m sets of m sizes, repeat 1 and 2, m times.
4: To get a joint ranked set sample of size m, select i^{th} minimum from i^{th} set. where $i = 1, 2, ..., m$.
5: Repeat 1 to 4 to increase the size of the joint ranked set sample k times *i.e.* $n = mk$.

2.3. Maximum likelihood estimation under JRSS

The likelihood function according to the algorithm 1 based on the observations $w_{ij}, i = 1, 2, 3, ..., m; j = 1, 2, 3, ..., k$ obtained in the joint rank set sample is given as follows,

$$L(a_1, b_1, a_2, b_2|\underline{w}) = \prod_{j=1}^{k}\prod_{i=1}^{m} \frac{m!}{(i-1)!(m-i)!}[F_x(w_{ij})]^{a_{ij}}[1-F_x(w_{ij})]^{b_{ij}}[f_x(w_{ij})]^{z_{ij}}[G_y(w_{ij})]^{c_{ij}} \times [1-G_y(w_{ij})]^{d_{ij}}[g_y(w_{ij})]^{1-z_{ij}} \tag{5}$$

where w_{ij} represents the i^{th} order element from the i^{th} joint sample in the j^{th} cycle, where $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, k$. Likelihood under JRSS for Kumaraswamy populations is obtained by substituting equations (1),(2),(3), and (4) in (5).

$$L(a_1, b_1, a_2, b_2|\underline{w}) = \prod_{j=1}^{k}\prod_{i=1}^{m} C\Big[1-\left(1-w_{ij}^{a_1}\right)^{b_1}\Big]^{a_{ij}}\Big[\left(1-w_{ij}^{a_1}\right)^{b_1}\Big]^{b_{ij}}\Big[a_1b_1w_{ij}^{a_1-1}\left(1-w_{ij}^{a_1}\right)^{b_1-1}\Big]^{z_{ij}}$$
$$\Big[1-\left(1-w_{ij}^{a_2}\right)^{b_2}\Big]^{c_{ij}}\Big[\left(1-w_{ij}^{a_2}\right)^{b_2}\Big]^{d_{ij}}\Big[a_2b_2w_{ij}^{a_2-1}\left(1-w_{ij}^{a_2}\right)^{b_2-1}\Big]^{1-z_{ij}} \tag{6}$$

where,

$$z_{ij} = \begin{cases} 1 & \text{if } w_{ij} \text{ is from first population-I(X)} \\ 0 & \text{if } w_{ij} \text{ is from second population-II(Y)}, \end{cases} \qquad i = 1, 2, ..., m; j = 1, 2, ..., k$$

$C = \frac{m!}{(i-1)!(m-i)!}$
$w_{ij} = i^{th}$ minimum of i^{th} set of combined sample in j^{th} cycle.
a_{ij}= Number of x's observation less than or equal to w_{ij} in i^{th} set of combined samples in j^{th} cycle.
b_{ij}= Number of x's observation greater than to w_{ij} in i^{th} set of combined samples in j^{th} cycle.
c_{ij}= Number of y's observation less than or equal to w_{ij} in i^{th} set of combined samples in j^{th} cycle.
d_{ij}= Number of y's observation greater than to w_{ij} in i^{th} set of combined samples in j^{th} cycle.
The log-likelihood function for the Kumaraswamy distribution under JRSS can be expressed by taking the logarithm of the equation (6) based on the observations w_{ij}, where $i = 1, 2, 3, ..., m$ and $j = 1, 2, 3, ..., k$.

$$\begin{aligned} l = log\left(L(a_1, b_1, a_2, b_2|\underline{w})\right) &= \sum_{i=1}^{m}\sum_{j=1}^{k} a_{ij}\ln\left(1-(1-{w_{ij}}^{a_1})^{b_1}\right) + b_1\sum_{i=1}^{m}\sum_{j=1}^{k} b_{ij}\ln\left(1-{w_{ij}}^{a_1}\right) \\ &+ (b_1-1)\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\ln\left(1-{w_{ij}}^{a_1}\right) + \sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\ln\left(a_1b_1{w_{ij}}^{a_1-1}\right) \\ &+ \sum_{i=1}^{m}\sum_{j=1}^{k} c_{ij}\ln\left(1-(1-{w_{ij}}^{a_2})^{b_2}\right) + b_2\sum_{i=1}^{m}\sum_{j=1}^{k} d_{ij}\ln\left(1-{w_{ij}}^{a_2}\right) \\ &+ (b_2-1)\sum_{i=1}^{m}\sum_{j=1}^{k}(1-z_{ij})\ln\left(1-{w_{ij}}^{a_2}\right) \\ &+ \sum_{i=1}^{m}\sum_{j=1}^{k}(1-z_{ij})\ln\left(a_2b_2{w_{ij}}^{a_2-1}\right) \end{aligned} \tag{7}$$

$$\begin{aligned} \frac{\partial l}{\partial a_1} &= \frac{1}{a_1}\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\left(a_1\ln\left(w_{ij}\right)+1\right) - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{{w_{ij}}^{a_1}\ln\left(w_{ij}\right)\left(\left(b_1-1\right)z_{ij}+b_1b_{ij}\right)}{1-{w_{ij}}^{a_1}} \\ &+ \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{b_1a_{ij}{w_{ij}}^{a_1}\ln\left(w_{ij}\right)\left(1-{w_{ij}}^{a_1}\right)^{b_1-1}}{1-\left(1-{w_{ij}}^{a_1}\right)^{b_1}} \end{aligned} \tag{8}$$

$$\frac{\partial l}{\partial b_1} = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{z_{ij}}{b_1} + \sum_{i=1}^{m}\sum_{j=1}^{k}\ln\left(1-{w_{ij}}^{a_1}\right)\ z_{ij} - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{a_{ij}\ln\left(1-{w_{ij}}^{a_1}\right)\left(1-{w_{ij}}^{a_1}\right)^{b_1}}{1-\left(1-{w_{ij}}^{a_1}\right)^{b_1}} + \sum_{i=1}^{m}\sum_{j=1}^{k} b_{ij}\ln\left(1-{w_{ij}}^{a_1}\right) \tag{9}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial a_1^2} &= \frac{1}{a_1}\sum_{i=1}^{m}\sum_{j=1}^{k} a_1 z_{ij}\left(\ln^2(w_{ij}) + 2b_1\ln(w_{ij})\right) - \frac{1}{a_1}\sum_{i=1}^{m}\sum_{j=1}^{k}\ln(w_{ij})\ z_{ij}\left(a_1\ln(w_{ij})+1\right)\\ &- \frac{1}{a_1^2}\sum_{i=1}^{m}\sum_{j=1}^{k} a_1 z_{ij}\left(\ln(w_{ij})+1\right) - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(b_1-1)\,{w_{ij}}^{a_1}\ln^2(w_{ij})\,z_{ij} + b_1 b_{ij}{w_{ij}}^{a_1}\ln^2(w_{ij})}{1-{w_{ij}}^{a_1}}\\ &- \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(b_1-1)\,{w_{ij}}^{a_1}\ln(w_{ij})\left({w_{ij}}^{a_1}\ln(w_{ij})\,z_{ij} - b_1 b_{ij}{w_{ij}}^{a_1}\ln(w_{ij})\right)}{\left(1-{w_{ij}}^{a_1}\right)^2}\\ &- \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{b_1^2 a_{ij}{w_{ij}}^{2a_1}\ln^2(w_{ij})\left(1-{w_{ij}}^{a_1}\right)^{2b_1-2}}{\left(1-\left(1-{w_{ij}}^{a_1}\right)^{b_1}\right)^2}\\ &- \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(b_1-1)\,b_1 a_{ij}{w_{ij}}^{2a_1}\ln^2(w_{ij})\left(1-{w_{ij}}^{a_1}\right)^{b_1-2}}{1-\left(1-{w_{ij}}^{a_1}\right)^{b_1}}\\ &+ \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{b_1 a_{ij}{w_{ij}}^{a_1}\ln^2(w_{ij})\left(1-{w_{ij}}^{a_1}\right)^{b_1-1}}{1-\left(1-{w_{ij}}^{a_1}\right)^{b_1}}\end{aligned} \tag{10}$$

$$\frac{\partial^2 l}{\partial b_1^2} = -\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{z_{ij}}{{b_1}^2} - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{a_{ij}\ln^2\left(1-{w_{ij}}^{a_1}\right)\left(1-{w_{ij}}^{a_1}\right)^{b_1}}{\left(\left(1-{w_{ij}}^{a_1}\right)^{b_1}-1\right)^2} \tag{11}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial a_1 \partial b_1} &= -\frac{1}{a_1 b_1}\sum_{i=1}^{m}\sum_{j=1}^{k} a_1 z_{ij}\left(\ln(w_{ij})+1\right)\\ &+ \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{b_1 a_{ij}{w_{ij}}^{a_1}\ln(w_{ij})\ln\left(1-{w_{ij}}^{a_1}\right)\left(1-{w_{ij}}^{a_1}\right)^{b_1-1}}{1-\left(1-{w_{ij}}^{a_1}\right)^{b_1}}\\ &+ \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{b_1 a_{ij}{w_{ij}}^{a_1}\ln(w_{ij})\ln\left(1-{w_{ij}}^{a_1}\right)\left(1-{w_{ij}}^{a_1}\right)^{2b_1-1}}{\left(1-\left(1-{w_{ij}}^{a_1}\right)^{b_1}\right)^2}\\ &+ \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{a_1{w_{ij}}^{1-a_1}\left({w_{ij}}^{a_1-1}\ln(w_{ij}) + {w_{ij}}^{a_1-1}\right)z_{ij}}{a_1 b_1} - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{{w_{ij}}^{a_1}\ln(w_{ij})\ z_{ij}}{1-{w_{ij}}^{a_1}}\\ &+ \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{a_{ij}{w_{ij}}^{a_1}\ln(w_{ij})\left(1-{w_{ij}}^{a_1}\right)^{b_1-1}}{1-\left(1-{w_{ij}}^{a_1}\right)^{b_1}} - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{b_{ij}{w_{ij}}^{a_1}\ln(w_{ij})}{1-w_{ij}z_{ij}^{a_1}}\end{aligned} \tag{12}$$

$$\frac{\partial l}{\partial a_2} = \frac{1}{a_2}\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\left(a_2\ln(w_{ij})+1\right) - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(b_2-1)\,{w_{ij}}^{a_2}\ln(w_{ij})\,z_{ij} - b_2 d_{ij}{w_{ij}}^{a_2}\ln(w_{ij})}{1-{w_{ij}}^{a_2}} + b_2\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{c_{ij}{w_{ij}}^{a_2}\ln(w_{ij})\left(1-{w_{ij}}^{a_2}\right)^{b_2-1}}{1-\left(1-{w_{ij}}^{a_2}\right)^{b_2}} \tag{13}$$

$$\frac{\partial l}{\partial b_2} = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{z_{ij}}{b_2} + \sum_{i=1}^{m}\sum_{j=1}^{k}\ln\left(1-{w_{ij}}^{a_2}\right)\ z_{ij} - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{c_{ij}\ln\left(1-w_{ij}^{a_2}\right)\left(1-w_{ij}^{a_2}\right)^{b_2}}{1-\left(1-w_{ij}^{a_2}\right)^{b_2}} + \sum_{i=1}^{m}\sum_{j=1}^{k} d_{ij}\ln\left(1-{w_{ij}}^{a_2}\right) \tag{14}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial a_2^2} &= \frac{1}{a_2}\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\left(a_2\ln^2(w_{ij}) + 2\ln(w_{ij})\right) - \frac{1}{a_2}\sum_{i=1}^{m}\sum_{j=1}^{k}\ln(w_{ij})\ z_{ij}\left(a_2\ln(w_{ij})+1\right)\\ &- \frac{1}{a_2^2}\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\left(a_2\ln(w_{ij})+1\right) - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(b_2-1)\,{w_{ij}}^{a_2}\ln^2(w_{ij})\,z_{ij} - b_2 d_{ij}{w_{ij}}^{a_2}\ln^2(w_{ij})}{1-{w_{ij}}^{a_2}}\\ &- \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{{w_{ij}}^{a_2}\ln(w_{ij})\left((b_2-1)\,{w_{ij}}^{a_2}\ln(w_{ij})\,z_{ij} - b_2 d_{ij}{w_{ij}}^{a_2}\ln(w_{ij})\right)}{\left(1-{w_{ij}}^{a_2}\right)^2}\\ &- \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{b_2^2 c_{ij}{w_{ij}}^{2a_2}\ln^2(w_{ij})\left(1-{w_{ij}}^{a_2}\right)^{2b_2-2}}{\left(1-\left(1-{w_{ij}}^{a_2}\right)^{b_2}\right)^2}\\ &- \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(b_2-1)\,b_2 c_{ij}{w_{ij}}^{2a_2}\ln^2(w_{ij})\left(1-{w_{ij}}^{a_2}\right)^{b_2-2}}{1-\left(1-{w_{ij}}^{a_2}\right)^{b_2}}\\ &+ \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{b_2 c_{ij}{w_{ij}}^{a_2}\ln^2(w_{ij})\left(1-{w_{ij}}^{a_2}\right)^{b_2-1}}{1-\left(1-{w_{ij}}^{a_2}\right)^{b_2}}\end{aligned} \tag{15}$$

$$\frac{\partial^2 l}{\partial b_2^2} = -\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{z_{ij}}{{b_2}^2} - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{c_{ij}\ln^2\left(1-{w_{ij}}^{a_2}\right)\left(1-{w_{ij}}^{a_2}\right)^{b_2}}{\left(\left(1-{w_{ij}}^{a_2}\right)^{b_2}-1\right)^2} \tag{16}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial a_2 \partial b_2} &= -\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{{w_{ij}}^{1-a_2} z_{ij}\left({w_{ij}}^{a_2-1}\ln(w_{ij})\ a_2 b_2 + {w_{ij}}^{a_2-1} b_2\right)}{a_2 b_2^2}\\ &+ \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{c_{ij}{w_{ij}}^{a_2}\ln(w_{ij})\ln\left(1-{w_{ij}}^{a_2}\right)\left(1-{w_{ij}}^{a_2}\right)^{b_2-1} b_2}{1-\left(1-{w_{ij}}^{a_2}\right)^{b_2}}\\ &+ \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{c_{ij}{w_{ij}}^{a_2}\ln(w_{ij})\ln\left(1-{w_{ij}}^{a_2}\right)\left(1-{w_{ij}}^{a_2}\right)^{2b_2-1} b_2}{\left(1-\left(1-{w_{ij}}^{a_2}\right)^{b_2}\right)^2}\\ &+ \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{{w_{ij}}^{1-a_2}\left({w_{ij}}^{a_2-1}\ln(w_{ij})\ a_2 + {w_{ij}}^{a_2-1}\right) z_{ij}}{a_2 b_2} - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{{w_{ij}}^{a_2}\ln(w_{ij})\ z_{ij}}{1-{w_{ij}}^{a_2}}\\ &+ \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{c_{ij}{w_{ij}}^{a_2}\ln(w_{ij})\left(1-{w_{ij}}^{a_2}\right)^{b_2-1}}{1-\left(1-{w_{ij}}^{a_2}\right)^{b_2}} - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{d_{ij}{w_{ij}}^{a_2}\ln(w_{ij})}{1-w_{ij}}\end{aligned} \tag{17}$$

To determine the ML estimates for the parameters of the Kumaraswamy distribution under the JRSS, solve the equations (8), (9), (13), and (14) iteratively. Fisher information matrix (I) for the parameters of two Kumaraswamy populations under JRSS can be written as

follows,

$$I(a_1,b_1,a_2,b_2)=\begin{bmatrix} I_{a_1a_1} & I_{a_1b_1} & I_{a_1a_2} & I_{a_1b_2} \\ I_{b_1a_1} & I_{b_1b_1} & I_{b_1a_2} & I_{b_1b_2} \\ I_{a_2a_1} & I_{a_2b_1} & I_{a_2a_2} & I_{a_2b_2} \\ I_{b_2a_1} & I_{b_2b_1} & I_{b_2a_2} & I_{b_2b_2} \end{bmatrix} = -E\begin{bmatrix} \frac{\partial^2 l}{\partial a_1^2} & \frac{\partial^2 l}{\partial a_1 \partial b_1} & 0 & 0 \\ \frac{\partial^2 l}{\partial b_1 \partial a_1} & \frac{\partial^2 l}{\partial b_1^2} & 0 & 0 \\ 0 & 0 & \frac{\partial^2 l}{\partial a_2^2} & \frac{\partial^2 l}{\partial a_2 \partial b_2} \\ 0 & 0 & \frac{\partial^2 l}{\partial b_2 \partial a_2} & \frac{\partial^2 l}{\partial b_2^2} \end{bmatrix} \tag{18}$$

3. Joint Modified Minimum Ranked Set Sampling (JMnRSS)

In this Section, we studied the estimation of parameters of the Kumaraswamy population under JMnRSS.

Algorithm 2 The algorithm to select JMnRSS sample from two populations

1: Select m_1 observations from population-I and m_2 observations from population-II randomly and combine these observations to create a joint sample of size $m = m_1 + m_2$.
2: Arrange all these observations of the joint sample in ascending order.
3: To get m joint sets of sizes m, repeat 1 and 2 m times.
4: To get a JMnRSS sample of size m, select the first minimum from all m sets.
5: Repeat 1 to 4 to increase the size of the JMnRSS sample k times *i.e.* $n = mk$.

3.1. Maximum likelihood estimation under JMnRSS

The likelihood function according to algorithm 2 based on the observations $w_{ij}, i = 1,2,3,...,m; j = 1,2,3,...,k$ can be obtained by putting $a_{ij} = c_{ij} = 0$ for all $i = 1,2,3,...,m; j = 1,2,...,k$ in equation (7)

$$\begin{aligned} l &= log\left(L(a_1,b_1,a_2,b_2|\underline{w})\right) \\ &= b_1\sum_{i=1}^{m}\sum_{j=1}^{k} b_{ij}\ln\left(1-{w_{ij}}^{a_1}\right) + (b_1-1)\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\ln\left(1-{w_{ij}}^{a_1}\right) \\ &+ \sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\ln\left(a_1b_1{w_{ij}}^{a_1-1}\right) + b_2\sum_{i=1}^{m}\sum_{j=1}^{k} d_{ij}\ln\left(1-{w_{ij}}^{a_2}\right) \\ &+ (b_2-1)\sum_{i=1}^{m}\sum_{j=1}^{k}(1-z_{ij})\ln\left(1-{w_{ij}}^{a_2}\right) + \sum_{i=1}^{m}\sum_{j=1}^{k}(1-z_{ij})\ln\left(a_2b_2{w_{ij}}^{a_2-1}\right) \end{aligned} \tag{19}$$

$$\frac{\partial l}{\partial a_1} = \frac{1}{a_1}\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\left(a_1\ln\left(w_{ij}\right)+1\right) - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{{w_{ij}}^{a_1}\ln\left(w_{ij}\right)\left(\left(b_1-1\right)z_{ij}+b_1b_{ij}\right)}{1-{w_{ij}}^{a_1}} \tag{20}$$

$$\frac{\partial l}{\partial b_1} = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{z_{ij}}{b_1} + \sum_{i=1}^{m}\sum_{j=1}^{k}\ln\left(1-{w_{ij}}^{a_1}\right)\, z_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{k} b_{ij}\ln\left(1-{w_{ij}}^{a_1}\right) \tag{21}$$

$$\frac{\partial^2 l}{\partial a_1^2} = \frac{1}{a_1}\sum_{i=1}^{m}\sum_{j=1}^{k} a_1 z_{ij}\left(\ln^2(w_{ij}) + 2b_1\ln(w_{ij})\right) - \frac{1}{a_1}\sum_{i=1}^{m}\sum_{j=1}^{k}\ln(w_{ij})\ z_{ij}\left(a_1\ln(w_{ij})+1\right)$$
$$- \frac{1}{a_1^2}\sum_{i=1}^{m}\sum_{j=1}^{k} a_1 z_{ij}\left(\ln(w_{ij})+1\right) - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(b_1-1)\,{w_{ij}}^{a_1}\ln^2(w_{ij})\,z_{ij} + b_1 b_{ij}{w_{ij}}^{a_1}\ln^2(w_{ij})}{1-{w_{ij}}^{a_1}}$$
$$- \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(b_1-1)\,{w_{ij}}^{a_1}\ln(w_{ij})\left({w_{ij}}^{a_1}\ln(w_{ij})\,z_{ij} - b_1 b_{ij}{w_{ij}}^{a_1}\ln(w_{ij})\right)}{\left(1-{w_{ij}}^{a_1}\right)^2} \tag{22}$$

$$\frac{\partial^2 l}{\partial b_1^2} = -\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{z_{ij}}{{b_1}^2} \tag{23}$$

$$\frac{\partial^2 l}{\partial a_1 \partial b_1} = -\frac{1}{a_1 b_1}\sum_{i=1}^{m}\sum_{j=1}^{k} a_1 z_{ij}\left(\ln(w_{ij})+1\right) + \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{a_1{w_{ij}}^{1-a_1}\left({w_{ij}}^{a_1-1}\ln(w_{ij}) + {w_{ij}}^{a_1-1}\right)z_{ij}}{a_1 b_1}$$
$$- \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{{w_{ij}}^{a_1}\ln(w_{ij})\ z_{ij}}{1-{w_{ij}}^{a_1}} - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{b_{ij}{w_{ij}}^{a_1}\ln(w_{ij})}{1-w_{ij}} \tag{24}$$

$$\frac{\partial l}{\partial a_2} = \frac{1}{a_2}\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\left(a_2\ln(w_{ij})+1\right) - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(b_2-1)\,{w_{ij}}^{a_2}\ln(w_{ij})\,z_{ij} - b_2 d_{ij}{w_{ij}}^{a_2}\ln(w_{ij})}{1-{w_{ij}}^{a_2}} \tag{25}$$

$$\frac{\partial l}{\partial b_2} = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{z_{ij}}{b_2} + \sum_{i=1}^{m}\sum_{j=1}^{k}\ln\left(1-{w_{ij}}^{a_2}\right)\ z_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{k} d_{ij}\ln\left(1-{w_{ij}}^{a_2}\right) \tag{26}$$

$$\frac{\partial^2 l}{\partial a_2^2} = \frac{1}{a_2}\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\left(a_2\ln^2(w_{ij}) + 2\ln(w_{ij})\right) - \frac{1}{a_2}\sum_{i=1}^{m}\sum_{j=1}^{k}\ln(w_{ij})\ z_{ij}\left(a_2\ln(w_{ij})+1\right)$$
$$- \frac{1}{a_2^2}\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\left(a_2\ln(w_{ij})+1\right) - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(b_2-1)\,{w_{ij}}^{a_2}\ln^2(w_{ij})\,z_{ij} - b_2 d_{ij}{w_{ij}}^{a_2}\ln^2(w_{ij})}{1-{w_{ij}}^{a_2}}$$
$$- \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{{w_{ij}}^{a_2}\ln(w_{ij})\left((b_2-1)\,{w_{ij}}^{a_2}\ln(w_{ij})\,z_{ij} - b_2 d_{ij}{w_{ij}}^{a_2}\ln(w_{ij})\right)}{\left(1-{w_{ij}}^{a_2}\right)^2} \tag{27}$$

$$\frac{\partial^2 l}{\partial^2 b_2} = -\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{z_{ij}}{{b_2}^2} \tag{28}$$

$$\frac{\partial^2 l}{\partial a_2 \partial b_2} = -\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{{w_{ij}}^{1-a_2} z_{ij}\left({w_{ij}}^{a_2-1}\ln(w_{ij})\ a_2 b_2 + {w_{ij}}^{a_2-1} b_2\right)}{a_2 b_2^2}$$
$$+ \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{{w_{ij}}^{1-a_2}\left({w_{ij}}^{a_2-1}\ln(w_{ij})\ a_2 + {w_{ij}}^{a_2-1}\right)z_{ij}}{a_2 b_2} - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{{w_{ij}}^{a_2}\ln(w_{ij})\ z_{ij}}{1-{w_{ij}}^{a_2}}$$
$$- \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{d_{ij}{w_{ij}}^{a_2}\ln(w_{ij})}{1-w_{ij}} \tag{29}$$

To determine the ML estimates for the parameters of the Kumaraswamy distribution under the JMnRSS, solve the equations (20), (21), (25), and (26) iteratively. Fisher information matrix (I) for the parameters of two Kumaraswamy populations under JMnRSS can be written as follows,

$$I(a_1,b_1,a_2,b_2)=\begin{bmatrix} I_{a_1a_1} & I_{a_1b_1} & I_{a_1a_2} & I_{a_1b_2} \\ I_{b_1a_1} & I_{b_1b_1} & I_{b_1a_2} & I_{b_1b_2} \\ I_{a_2a_1} & I_{a_2b_1} & I_{a_2a_2} & I_{a_2b_2} \\ I_{b_2a_1} & I_{b_2b_1} & I_{b_2a_2} & I_{b_2b_2} \end{bmatrix} \tag{30}$$

$$=-E\begin{bmatrix} \frac{\partial^2 l}{\partial a_1^2} & \frac{\partial^2 l}{\partial a_1 \partial b_1} & 0 & 0 \\ \frac{\partial^2 l}{\partial b_1 \partial a_1} & \frac{\partial^2 l}{\partial b_1^2} & 0 & 0 \\ 0 & 0 & \frac{\partial^2 l}{\partial a_2^2} & \frac{\partial^2 l}{\partial a_2 \partial b_2} \\ 0 & 0 & \frac{\partial^2 l}{\partial b_2 \partial a_2} & \frac{\partial^2 l}{\partial b_2^2} \end{bmatrix} \tag{31}$$

4. Joint Modified Maximum Ranked Set Sampling (JMxRSS)

In this Section, we studied the estimation of parameters of the Kumaraswamy population under JMxRSS. The algorithm to select JMxRSS sample from two populations is as follows.

Algorithm 3 The algorithm to select JMxRSS sample from two populations

1: Select m_1 observation from a population-I and m_2 observation from a population-II randomly and combine these observations to create a joint sample of size $m = m_1 + m_2$.
2: Arrange all these observations of the joint sample in ascending order.
3: To get m joint sets of sizes m, repeat the 1 and 2 m times.
4: To have a JMxRSS sample of size m, select the maximum from all m sets.
5: Repeat 1 to 4 to increase the size of the JMxRSS sample k times *i.e.* $n = mk$.

4.1. Maximum likelihood estimation under JMxRSS

The likelihood function according to the algorithm 3 discuss above based on the observations $w_{ij}, i = 1,2,3,...,m; j = 1,2,3,...,k$ can be obtained by putting $b_{ij} = d_{ij} = 0$ for all $i = 1,2,3,...,m; j = 1,2,...,k$ in equation (6)

$$l = log\left(L(a_1,b_1,a_2,b_2|\underline{w})\right) = \sum_{i=1}^{m}\sum_{j=1}^{k} a_{ij}\ln\left(1-(1-{w_{ij}}^{a_1})^{b_1}\right) + (b_1-1)$$
$$\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\ln\left(1-{w_{ij}}^{a_1}\right) + \sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\ln\left(a_1b_1{w_{ij}}^{a_1-1}\right) + \sum_{i=1}^{m}\sum_{j=1}^{k} c_{ij}\ln\left(1-(1-{w_{ij}}^{a_2})^{b_2}\right)$$
$$+ (b_2-1)\sum_{i=1}^{m}\sum_{j=1}^{k}(1-z_{ij})\ln\left(1-{w_{ij}}^{a_2}\right) + \sum_{i=1}^{m}\sum_{j=1}^{k}(1-z_{ij})\ln\left(a_2b_2{w_{ij}}^{a_2-1}\right) \tag{32}$$

$$\frac{\partial l}{\partial a_1} = \frac{1}{a_1}\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\left(a_1\ln\left(w_{ij}\right)+1\right) + \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{b_1a_{ij}{w_{ij}}^{a_1}\ln\left(w_{ij}\right)\left(1-{w_{ij}}^{a_1}\right)^{b_1-1}}{1-\left(1-{w_{ij}}^{a_1}\right)^{b_1}} \tag{33}$$

$$\frac{\partial l}{\partial b_1} = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{z_{ij}}{b_1} + \sum_{i=1}^{m}\sum_{j=1}^{k}\ln\left(1-{w_{ij}}^{a_1}\right)\ z_{ij} - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{a_{ij}\ln\left(1-{w_{ij}}^{a_1}\right)\left(1-{w_{ij}}^{a_1}\right)^{b_1}}{1-\left(1-{w_{ij}}^{a_1}\right)^{b_1}} \tag{34}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial a_1^2} = {} & \frac{1}{a_1}\sum_{i=1}^{m}\sum_{j=1}^{k}a_1 z_{ij}\left(\ln^2(w_{ij}) + 2b_1\ln(w_{ij})\right) - \frac{1}{a_1}\sum_{i=1}^{m}\sum_{j=1}^{k}\ln(w_{ij})\ z_{ij}\left(a_1\ln(w_{ij})+1\right) \\ & - \frac{1}{a_1^2}\sum_{i=1}^{m}\sum_{j=1}^{k}a_1 z_{ij}\left(\ln(w_{ij})+1\right) - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(b_1-1)\,{w_{ij}}^{a_1}\ln^2(w_{ij})\,z_{ij}}{1-{w_{ij}}^{a_1}} \\ & - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(b_1-1)\,{w_{ij}}^{2a_1}\ln^2(w_{ij})\,z_{ij}}{\left(1-{w_{ij}}^{a_1}\right)^2} - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{b_1^2 a_{ij}{w_{ij}}^{2a_1}\ln^2(w_{ij})\left(1-{w_{ij}}^{a_1}\right)^{2b_1-2}}{\left(1-\left(1-{w_{ij}}^{a_1}\right)^{b_1}\right)^2} \\ & - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(b_1-1)\,b_1 a_{ij}{w_{ij}}^{2a_1}\ln^2(w_{ij})\left(1-{w_{ij}}^{a_1}\right)^{b_1-2}}{1-\left(1-{w_{ij}}^{a_1}\right)^{b_1}} \\ & + \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{b_1 a_{ij}{w_{ij}}^{a_1}\ln^2(w_{ij})\left(1-{w_{ij}}^{a_1}\right)^{b_1-1}}{1-\left(1-{w_{ij}}^{a_1}\right)^{b_1}}\end{aligned} \tag{35}$$

$$\frac{\partial^2 l}{\partial b_1^2} = -\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{z_{ij}}{{b_1}^2} - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{a_{ij}\ln^2\left(1-{w_{ij}}^{a_1}\right)\left(1-{w_{ij}}^{a_1}\right)^{b_1}}{\left(\left(1-{w_{ij}}^{a_1}\right)^{b_1}-1\right)^2} \tag{36}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial a_1 \partial b_1} = {} & -\frac{1}{a_1 b_1}\sum_{i=1}^{m}\sum_{j=1}^{k}a_1 z_{ij}\left(\ln(w_{ij})+1\right) \\ & + \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{b_1 a_{ij}{w_{ij}}^{a_1}\ln(w_{ij})\ln\left(1-{w_{ij}}^{a_1}\right)\left(1-{w_{ij}}^{a_1}\right)^{b_1-1}}{1-\left(1-{w_{ij}}^{a_1}\right)^{b_1}} \\ & + \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{b_1 a_{ij}{w_{ij}}^{a_1}\ln(w_{ij})\ln\left(1-{w_{ij}}^{a_1}\right)\left(1-{w_{ij}}^{a_1}\right)^{2b_1-1}}{\left(1-\left(1-{w_{ij}}^{a_1}\right)^{b_1}\right)^2} \\ & + \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{a_1{w_{ij}}^{1-a_1}\left({w_{ij}}^{a_1-1}\ln(w_{ij}) + {w_{ij}}^{a_1-1}\right)z_{ij}}{a_1 b_1} - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{{w_{ij}}^{a_1}\ln(w_{ij})\ z_{ij}}{1-{w_{ij}}^{a_1}} \\ & + \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{a_{ij}{w_{ij}}^{a_1}\ln(w_{ij})\left(1-{w_{ij}}^{a_1}\right)^{b_1-1}}{1-\left(1-{w_{ij}}^{a_1}\right)^{b_1}}\end{aligned} \tag{37}$$

$$\begin{aligned}\frac{\partial l}{\partial a_2} = {} & \frac{1}{a_2}\sum_{i=1}^{m}\sum_{j=1}^{k}z_{ij}\left(a_2\ln(w_{ij})+1\right) - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(b_2-1)\,{w_{ij}}^{a_2}\ln(w_{ij})\,z_{ij}}{1-{w_{ij}}^{a_2}} \\ & + b_2\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{c_{ij}{w_{ij}}^{a_2}\ln(w_{ij})\left(1-{w_{ij}}^{a_2}\right)^{b_2-1}}{1-\left(1-{w_{ij}}^{a_2}\right)^{b_2}}\end{aligned} \tag{38}$$

$$\frac{\partial l}{\partial b_2} = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{z_{ij}}{b_2} + \sum_{i=1}^{m}\sum_{j=1}^{k}\ln\left(1-{w_{ij}}^{a_2}\right)\ z_{ij} - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{c_{ij}\ln\left(1-{w_{ij}}^{a_2}\right)\left(1-{w_{ij}}^{a_2}\right)^{b_2}}{1-\left(1-{w_{ij}}^{a_2}\right)^{b_2}} \tag{39}$$

$$\frac{\partial^2 l}{\partial a_2^2} = \frac{1}{a_2}\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\left(a_2 \ln^2(w_{ij}) + 2\ln(w_{ij})\right) - \frac{1}{a_2}\sum_{i=1}^{m}\sum_{j=1}^{k} \ln(w_{ij})\ z_{ij}\left(a_2\ln(w_{ij})+1\right)$$
$$- \frac{1}{a_2^2}\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\left(a_2\ln(w_{ij})+1\right) - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(b_2-1)\,{w_{ij}}^{a_2}\ln^2(w_{ij})\,z_{ij}}{1-{w_{ij}}^{a_2}}$$
$$- \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(b_2-1)\,{w_{ij}}^{2a_2}\ln^2(w_{ij})\,z_{ij}}{\left(1-{w_{ij}}^{a_2}\right)^2} - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{b_2^2 c_{ij}{w_{ij}}^{2a_2}\ln^2(w_{ij})\left(1-{w_{ij}}^{a_2}\right)^{2b_2-2}}{\left(1-\left(1-{w_{ij}}^{a_2}\right)^{b_2}\right)^2}$$
$$- \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(b_2-1)\,b_2 c_{ij}{w_{ij}}^{2a_2}\ln^2(w_{ij})\left(1-{w_{ij}}^{a_2}\right)^{b_2-2}}{1-\left(1-{w_{ij}}^{a_2}\right)^{b_2}}$$
$$+ \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{b_2 c_{ij}{w_{ij}}^{a_2}\ln^2(w_{ij})\left(1-{w_{ij}}^{a_2}\right)^{b_2-1}}{1-\left(1-{w_{ij}}^{a_2}\right)^{b_2}} \tag{40}$$

$$\frac{\partial^2 l}{\partial b_2^2} = -\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{z_{ij}}{{b_2}^2} - \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{c_{ij}\ln^2\left(1-{w_{ij}}^{a_2}\right)\left(1-{w_{ij}}^{a_2}\right)^{b_2}}{\left(\left(1-{w_{ij}}^{a_2}\right)^{b_2}-1\right)^2} \tag{41}$$

$$\frac{\partial^2 l}{\partial a_2 \partial b_2} = -\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{{w_{ij}}^{1-a_2} z_{ij}\left({w_{ij}}^{a_2-1}\ln(w_{ij})\ a_2 b_2 + {w_{ij}}^{a_2-1} b_2\right)}{a_2 b_2^2}$$
$$+ \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{c_{ij}{w_{ij}}^{a_2}\ln(w_{ij})\ln\left(1-{w_{ij}}^{a_2}\right)\left(1-{w_{ij}}^{a_2}\right)^{b_2-1} b_2}{1-\left(1-{w_{ij}}^{a_2}\right)^{b_2}}$$
$$+ \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{c_{ij}{w_{ij}}^{a_2}\ln(w_{ij})\ln\left(1-{w_{ij}}^{a_2}\right)\left(1-{w_{ij}}^{a_2}\right)^{2b_2-1} b_2}{\left(1-\left(1-{w_{ij}}^{a_2}\right)^{b_2}\right)^2}$$
$$+ \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{{w_{ij}}^{1-a_2}\left({w_{ij}}^{a_2-1}\ln(w_{ij})\ a_2 + {w_{ij}}^{a_2-1}\right) z_{ij}}{a_2 b_2}$$
$$- \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{{w_{ij}}^{a_2}\ln(w_{ij})\ z_{ij}}{1-{w_{ij}}^{a_2}} + \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{c_{ij}{w_{ij}}^{a_2}\ln(w_{ij})\left(1-{w_{ij}}^{a_2}\right)^{b_2-1}}{1-\left(1-{w_{ij}}^{a_2}\right)^{b_2}} \tag{42}$$

To determine the ML estimates for the parameters of the Kumaraswamy distribution under the JMxRSS, solve the equations (33), (34), (38), and (39) iteratively. Fisher information matrix (I) for the parameters of two Kumaraswamy populations under JMxRSS can be written as follows,

$$I(a_1,b_1,a_2,b_2) = \begin{bmatrix} I_{a_1a_1} & I_{a_1b_1} & I_{a_1a_2} & I_{a_1b_2} \\ I_{b_1a_1} & I_{b_1b_1} & I_{b_1a_2} & I_{b_1b_2} \\ I_{a_2a_1} & I_{a_2b_1} & I_{a_2a_2} & I_{a_2b_2} \\ I_{b_2a_1} & I_{b_2b_1} & I_{b_2a_2} & I_{b_2b_2} \end{bmatrix} = -E\begin{bmatrix} \frac{\partial^2 l}{\partial a_1^2} & \frac{\partial^2 l}{\partial a_1 \partial b_1} & 0 & 0 \\ \frac{\partial^2 l}{\partial b_1 \partial a_1} & \frac{\partial^2 l}{\partial b_1^2} & 0 & 0 \\ 0 & 0 & \frac{\partial^2 l}{\partial a_2^2} & \frac{\partial^2 l}{\partial a_2 \partial b_2} \\ 0 & 0 & \frac{\partial^2 l}{\partial b_2 \partial a_2} & \frac{\partial^2 l}{\partial b_2^2} \end{bmatrix} \tag{43}$$

4.2. Maximum likelihood estimation under Joint Simple Random Sampling (JSRS)

Likelihood under JSRS for the sample size $n_1 = m_1 k$ of population-I and sample size $n_2 = m_2 k$ is written as follows,

$$L(a_1, b_1, a_2, b_2|x, y) = \prod_{1=1}^{n_1} \prod_{i=1}^{n_2} f(x_i) g(y_i) = \prod_{1=1}^{n_1} f(x_i) \prod_{i=1}^{n_2} g(y_i) \tag{44}$$

Substituting equations (1) and (3) into equation (44) and applying the logarithm, we obtain the following equation.

$$\begin{aligned} l &= \log(L(a_1, b_1, a_2, b_2|x, y)) = \sum_{i=1}^{n_1} \log(f(x_i; a_1, b_1)) + \sum_{i=1}^{n_2} \log(g(y_i; a_2, b_2)) \\ l &= n_1 \log(a_1 b_1) + (a_1 - 1) \sum_{i=1}^{n_1} \log(x_i) + (b_1 - 1) \sum_{i=1}^{n_1} \log(1 - x_i^{a_1}) \\ &\quad + n_2 \log(a_2 b_2) + (a_2 - 1) \sum_{i=1}^{n_2} \log(y_i) + (b_2 - 1) \sum_{i=1}^{n_2} \log(1 - y_i^{a_2}) \end{aligned} \tag{45}$$

$$\frac{\partial l}{\partial a_1} = \frac{n_1}{a_1} - (b_1 - 1) \sum_{i=1}^{n_1} \frac{x_i^{a_1} \ln(x_i)}{1 - x_i^{a_1}} + \sum_{i=1}^{n_1} \ln(x_i) \tag{46}$$

$$\frac{\partial^2 l}{\partial a_1^2} = -\frac{n_1}{{a_1}^2} - (b_1 - 1) \sum_{i=1}^{n_1} \frac{x_i^{a_1} \ln^2(x_i)}{1 - x_i^{a_1}} - (b_1 - 1) \sum_{i=1}^{n_1} \frac{x_i^{2a_1} \ln^2(x_i)}{(1 - x_i^{a_1})^2} \tag{47}$$

$$\frac{\partial^2 l}{\partial a_1 \partial b_1} = \sum_{i=1}^{n_1} \frac{x_i^{a_1} \ln(x_i)}{x_i^{a_1} - 1} \tag{48}$$

$$\frac{\partial l}{\partial b_1} = \frac{n_1}{b_1} + \sum_{i=1}^{n_1} \ln(1 - x_i^{a_1}) \tag{49}$$

$$\frac{\partial^2 l}{\partial^2 b_1} = -\frac{n_1}{b_1^2} \tag{50}$$

$$\frac{\partial l}{\partial a_2} = \frac{n_2}{a_2} - (b_2 - 1) \sum_{i=1}^{n_2} \frac{y_i^{a_2} \ln(y_i)}{1 - y_i^{a_2}} + \sum_{i=1}^{n_2} \ln(y_i) \tag{51}$$

$$\frac{\partial^2 l}{\partial a_2^2} = -\frac{n_2}{{a_2}^2} - (b_2 - 1) \sum_{i=1}^{n_2} \frac{y_i^{a_2} \ln^2(y_i)}{1 - y_i^{a_2}} - (b_2 - 1) \sum_{i=1}^{n_2} \frac{y_i^{2a_2} \ln^2(y_i)}{(1 - y_i^{a_2})^2} \tag{52}$$

$$\frac{\partial^2 l}{\partial a_2 \partial b_2} = \sum_{i=1}^{n_2} \frac{y_i^{a_2} \ln(y_i)}{y_i^{a_2} - 1} \tag{53}$$

$$\frac{\partial l}{\partial b_2} = \frac{n_2}{b_2} + \sum_{i=1}^{n_2} \ln(1 - y_i^{a_2}) \tag{54}$$

$$\frac{\partial^2 l}{\partial b_2^2} = -\frac{n_2}{b_2^2} \tag{55}$$

To determine the ML estimates for the parameters of the Kumaraswamy distribution under the JSRS, solve the equations (46), (49), (51), and (54) iteratively. Fisher information matrix (I) for the parameters of two Kumaraswamy populations under JSRS can be written

as follows,

$$I(a_1,b_1,a_2,b_2)=\begin{bmatrix} I_{a_1a_1} & I_{a_1b_1} & I_{a_1a_2} & I_{a_1b_2} \\ I_{b_1a_1} & I_{b_1b_1} & I_{b_1a_2} & I_{b_1b_2} \\ I_{a_2a_1} & I_{a_2b_1} & I_{a_2a_2} & I_{a_2b_2} \\ I_{b_2a_1} & I_{b_2b_1} & I_{b_2a_2} & I_{b_2b_2} \end{bmatrix} = -E\begin{bmatrix} \frac{\partial^2 l}{\partial a_1^2} & \frac{\partial^2 l}{\partial a_1 \partial b_1} & 0 & 0 \\ \frac{\partial^2 l}{\partial b_1 \partial a_1} & \frac{\partial^2 l}{\partial b_1^2} & 0 & 0 \\ 0 & 0 & \frac{\partial^2 l}{\partial a_2^2} & \frac{\partial^2 l}{\partial a_2 \partial b_2} \\ 0 & 0 & \frac{\partial^2 l}{\partial b_2 \partial a_2} & \frac{\partial^2 l}{\partial b_2^2} \end{bmatrix} \tag{56}$$

5. Likelihood ratio test

In this Section, we derive the likelihood ratio test to compare the two Kumaraswamy distributions for equality of the first shape parameter when the second shape parameter is known, considering various joint ranked set sampling methods.

5.1. Likelihood ratio test for JRSS

$H_0 : a_1 = a_2 = a$ Vs H_1:$a_1 \neq a_2$, when b_1 and b_2 known.
The log-likelihood under H_0 for JRSS is expressed as follows:

$$\begin{aligned} l = log\left(L(a_1,b_1,a_2,b_2|\underline{w})\right) &= \sum_{i=1}^{m}\sum_{j=1}^{k} a_{ij}\ln\left(1-(1-{w_{ij}}^{a_1})^{b_1}\right) + b_1\sum_{i=1}^{m}\sum_{j=1}^{k} b_{ij}\ln\left(1-{w_{ij}}^{a_1}\right) \\ &+ (b_1-1)\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\ln\left(1-{w_{ij}}^{a_1}\right) + \sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\ln\left(a_1b_1{w_{ij}}^{a_1-1}\right) \\ &+ \sum_{i=1}^{m}\sum_{j=1}^{k} c_{ij}\ln\left(1-(1-{w_{ij}}^{a_2})^{b_2}\right) + b_2\sum_{i=1}^{m}\sum_{j=1}^{k} d_{ij}\ln\left(1-{w_{ij}}^{a_2}\right) \\ &+ (b_2-1)\sum_{i=1}^{m}\sum_{j=1}^{k}(1-z_{ij})\ln\left(1-{w_{ij}}^{a_2}\right) \\ &+ \sum_{i=1}^{m}\sum_{j=1}^{k}(1-z_{ij})\ln\left(a_2b_2{w_{ij}}^{a_2-1}\right) \end{aligned} \tag{57}$$

Differentiate (57) with respect to a and equate to zero.

$$\begin{aligned} \frac{\partial l}{\partial a} &= \frac{1}{a}\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\left(a\ln(w_{ij})+1\right) + \frac{1}{a}\sum_{i=1}^{m}\sum_{j=1}^{k}(1-z_{ij})\left(a\ln(w_{ij})+1\right) \\ &- (b_1-1)\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{w_{ij}^a\ln(w_{ij})\,z_{ij}}{1-w_{ij}^a} - (b_2-1)\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{w_{ij}^a\ln(w_{ij})\,(1-z_{ij})}{1-w_{ij}^a} \\ &+ b_1\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{a_{ij}w_{ij}^a\ln(w_{ij})\left(1-w_{ij}^a\right)^{b_1-1}}{1-\left(1-w_{ij}^a\right)^{b_1}} - b_1\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{b_{ij}w_{ij}^a\ln(w_{ij})}{1-w_{ij}^a} \\ &+ b_2\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{c_{ij}w_{ij}^a\ln(w_{ij})\left(1-w_{ij}^a\right)^{b_2-1}}{1-\left(1-w_{ij}^a\right)^{b_2}} - b_2\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{d_{ij}w_{ij}^a\ln(w_{ij})}{1-w_{ij}^a} \end{aligned} \tag{58}$$

To get MLE of a under H_0 solve the (58) numerically. MLEs of a_1 and a_2 obtained by numerically solving the equations (8), (9), (13), and (14). LRT test statistic is determined by putting the MLEs of a_1, a_2, and a into following equation.

$$\lambda_{JRSS}(\underline{w}) = \frac{\sup\limits_{a} L(a|\underline{w})}{\sup\limits_{a_1,a_2} L(a_1,a_2|\underline{w})}$$

LRT reject H_0 when $\lambda_{JRSS}(w) > \chi^2_{1,\alpha}$, where α is the level of significance.

5.2. Likelihood ratio test for JMnRSS

$H_0 : a_1 = a_2 = a \quad Vs \quad H_1 : a_1 \neq a_2, \quad$ when b_1 and b_2 known.
The MLE of a under H_0 for JMnRSS can be obtained by setting $a_{ij} = c_{ij} = 0$ in equation (57) and differentiating with respect to a.

$$\begin{aligned}\frac{\partial l}{\partial a} &= \frac{1}{a}\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\left(a\ln\left(w_{ij}\right)+1\right) + \frac{1}{a}\sum_{i=1}^{m}\sum_{j=1}^{k}\left(1-z_{ij}\right)\left(a\ln\left(w_{ij}\right)+1\right)\\ &- (b_1-1)\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{w_{ij}^a\ln\left(w_{ij}\right)z_{ij}}{1-w_{ij}^a} - (b_2-1)\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{w_{ij}^a\ln\left(w_{ij}\right)\left(1-z_{ij}\right)}{1-w_{ij}^a}\\ &- b_1\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{b_{ij}w_{ij}^a\ln\left(w_{ij}\right)}{1-w_{ij}^a} - b_2\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{d_{ij}w_{ij}^a\ln\left(w_{ij}\right)}{1-w_{ij}^a}\end{aligned} \tag{59}$$

To get MLE of a under H_0, solve the (59) numerically. MLEs of a_1 and a_2 obtained by numerically solving the equations (20), (21), (25), and (26). LRT test statistic is determined by putting the MLEs of a_1, a_2, and a into following equation.

$$\lambda_{JMnRSS}(\underline{W}) = \frac{\sup\limits_{a} L(a|\underline{w})}{\sup\limits_{a_1,a_2} L(a_1,a_2|\underline{w})}$$

LRT reject H_0 when $\lambda_{JMnRSS}(w) > \chi^2_{1,\alpha}$ where α is the level of significance.

5.3. Likelihood ratio test for JMxRSS

$H_0 : a_1 = a_2 = a \quad$ Vs $\quad H_1{:}a_1 \neq a_2, \quad$ when b_1 and b_2 known.
The MLE of a under H_0 for JMxRRS can be obtained by setting $b_{ij} = d_{ij} = 0$ in equation (57) and differentiating with respect to a.

$$\begin{aligned}\frac{\partial l}{\partial a} &= \frac{1}{a}\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}\left(a\ln\left(w_{ij}\right)+1\right) + \frac{1}{a}\sum_{i=1}^{m}\sum_{j=1}^{k}\left(1-z_{ij}\right)\left(a\ln\left(w_{ij}\right)+1\right)\\ &- (b_1-1)\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{w_{ij}^a\ln\left(w_{ij}\right)z_{ij}}{1-w_{ij}^a} - (b_2-1)\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{w_{ij}^a\ln\left(w_{ij}\right)\left(1-z_{ij}\right)}{1-w_{ij}^a}\\ &+ b_1\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{a_{ij}w_{ij}^a\ln\left(w_{ij}\right)\left(1-w_{ij}^a\right)^{b_1-1}}{1-\left(1-w_{ij}^a\right)^{b_1}} + b_2\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{c_{ij}w_{ij}^a\ln\left(w_{ij}\right)\left(1-w_{ij}^a\right)^{b_2-1}}{1-\left(1-w_{ij}^a\right)^{b_2}}\end{aligned} \tag{60}$$

To get MLE of a under H_0 Solve the (60) numerically. MLEs of a_1 and a_2 also obtained by solving the equations (33), (34), (38), and (39) numerically. LRT test statistic is determined by putting the MLEs of a_1, a_2, and a into following equation.

$$\lambda_{JMxRSS}(\underline{w}) = \frac{\overset{sup}{a} L(a|\underline{w})}{\underset{a_1,a_2}{sup} L(a_1, a_2|\underline{w})}$$

LRT reject H_0 when $\lambda_{JMxRSS}(w) > \chi^2_{1,\alpha}$ where α is the level of significance.

5.4. Likelihood ratio test for JSRS

$H_0 : a_1 = a_2 = a$ Vs H_1:$a_1 \neq a_2$, when b_1 and b_2 known.
The MLE under H_0 can be obtained by substituting $a_1 = a_2 = a$ in equation (45) and differentiate with respect to a.

$$\frac{\partial l}{\partial a} = \frac{n_1 + n_2}{a} - (b_1 - 1)\sum_{i=1}^{n_1} \frac{x_i^a \ln(x_i)}{1 - x_i^a} - (b_2 - 1)\sum_{i=1}^{n_2} \frac{y_i^a \ln(y_i)}{1 - y_i^a} + \sum_{i=1}^{n_2} \ln(y) + \sum_{i=1}^{n_2} \ln(x_i) \quad (61)$$

MLEs of a_1 and a_2 can be found by solving the equations (46), (49),(51), and (54). LRT test statistic is determined by putting the MLEs of a_1, a_2, and a into following equation.

$$\lambda_{JSRS}(x, y) = \frac{\overset{sup}{a} L(a|x, y)}{\underset{a_1,a_2}{sup} L(a_1, a_2|x, y)}$$

LRT reject H_0 when $\lambda_{JSRS}(x, y) > \chi^2_{1,\alpha}$ where α is the level of significance.

6. Simulation results and applications

In this Section, we present theoretical simulation results and also analyze a real data set.

6.1. Simulation results

In this Section, we evaluate the efficiency of ML estimates of the Kumaraswamy distribution parameters under JRSS, JMnRSS, JMxRSS, and JSRS. To evaluate the efficiency of our ML estimates, we employed Monte Carlo simulation techniques utilizing R Studio. For this specifically, we used the "mle" function for optimization and the Conjugate Gradient (CG) method given in R Studio. We compared these ML estimates using RMSE and bias calculated by 100 repetitions of various sample sizes of the first Kumaraswamy population $m_1 = (3, 4, 5)$, and from the second Kumaraswamy population $m_2 = (4, 5, 6)$, for different cycles $k = 2, 3, 4$. Additionally, we simulate the power of the LRT for the mentioned joint sampling schemes based on the algorithm 5. These results were also performed for various values of $m_1 = (3, 4, 5)$, $m_2 = (4, 5, 6)$, and $k = (2, 3, 4)$ with 1000 repetitions to compare the power of the test, ensuring at least three digits after the decimal point.

From Table (1) (see Annexure), we observe that with an increase in the total sample size ($m = m_1 + m_2$) and the number of cycles k, the RMSE and bias of the ML estimates of the parameters for the Kumaraswamy population decrease for all joint ranked sampling schemes. We observed that all joint ranked sampling schemes performed better than the

Algorithm 4 MLE and MSE Calculation for JRSS / JMnRSS/ JMxRSS

Initialization: Set parameters for the Kumaraswamy population-I: (a_1, b_1) and sample size m_1.
Set parameters for the Kumaraswamy population-II: (a_2, b_2) and sample size m_2.
Ensure the total sample size is $m = m_1 + m_2$.
Specify the number of cycles k and repetitions B.
for each sample size m_1 **do**
  for each sample size m_2 **do**
    for each cycle size k **do**
      while $i <= 1$ to B **do**
        Assign (a_1, b_1, a_2, b_2) to vector $Para$.
        Generate $m_1 \times k$ random observation from the Kumaraswamy population-I and $m_2 \times k$ observation from the Kumaraswamy population-II.
        Select a sample using the JRSS / JMnRSS / JMxRSS algorithms 1, 2, 3
        Construct the likelihood function $\mathcal{L}(Para|\mathrm{W})$.
        Maximize $\mathcal{L}(Para)$ to estimate $\hat{para}_i$. using maxLik function in R and initial values of the parameters are found by the method of moments Dey *et al.* (2018).
        Compute RMSE: $\mathrm{RMSE}_i = \sqrt{\left(\hat{Para}_i - Para\right)^2}$.
      end while
      ML estimates: $\hat{Para} = \frac{1}{B}\sum_{i=1}^{B} \hat{Para}_i$
      bias= $\hat{Para} - Para$
      $\mathrm{RMSE} = \frac{1}{B}\sum_{i=1}^{B} \mathrm{RMSE}_i$.
    end for
  end for
end for
Output: Average MLE and RMSE for each n.

Algorithm 5 Algorithm for power calculation of LRT using JRSS/ JMxRSS/ JMnRSS/ JSRS

Step 1: Determine LRT test Statistic $\lambda(w)$ for$(m = m_1 + m_2, k)$ using one of the joint sampling schemes JRSS/ JMxRSS/ JMnRSS/ JSRS (for JSRS $n_1 = m_1 k, n_2 = m_2 k$).
Step 2: Choose level of significance ($\alpha = 0.05$) and determine type-II error ($p_value >= 0.05$ when H_1 is True).
Step 3: Repeat step.1 and Step.2, B=1000 times and calculate p(type-II error) $= \frac{\text{No. of times p_value >=0.05}}{B}$.
Step 4: Determine the power of the test, power$= 1 - p$(type-II error).
Step 5: Repeat above procedure for different combination of (m_1, m_2, k).

JSRS scheme. Additionally, the simulation study for ML estimation shows that JMnRSS outperforms both JRSS and JMxRSS. In general, Table (2) (see Annexure) shows that as the difference between the parameters of the Kumaraswamy distribution increases, the RMSE and bias also increase. Table (3) (see Annexure) gives the power of the LRT for testing the equality of parameters (*i.e.*, $a_1 = a_2 = a$) when b_1 and b_2 are known, and the result indicates that JRSS has better power compared to other joint sampling schemes.

6.2. Application

In this Section, we delve into a practical example that illustrates the application of the Kumaraswamy distribution in estimating the probability of lung cancer within a specific population.

In this study, we have taken a lung cancer data set (click here) from Kaggle, which includes the following variables: Gender, Age, Smoking, Yellow fingers, Anxiety, Peer pressure, Chronic disease, Fatigue, Allergy, Wheezing, Alcohol consuming, Coughing of breath, Swallowing difficulty, Chest pain, lung cancer. Lung cancer is a binary (yes/no) variable. Hence, we used the Probit model for binary regression to classify individuals based on the explanatory variable as either having or not having lung cancer. Further, we checked the significance of the variable and found that Smoking, Yellow fingers, Peer pressure, Chronic disease, Fatigue, Allergy, Coughing of breath, and Swallowing difficulty, all these variables contribute significantly. For the given lung cancer data, we found that the Probit link function is suitable and also assessed the adequacy of the model using the hoslem.test, with a *p-value* of 0.9391, which is greater than $\alpha = 0.05$, indicating that the model fits well. Based on the Probit model, we predict the probability of lung cancer using the data. Then, we classify the predicted probabilities into two groups, male and female, and check the fit of the Kumaraswamy distribution using the goodness-of-fit test (Anderson-Darling test). All these results are provided in Table (4) (see Annexure) for the whole population and the separated male and female populations. For this fitting, we estimate the parameters under SRS using the ML method, with initial values obtained by the method of moments. So, in Table (4), the first column represents the groups (combined, male, and female), and the second column represents the initial estimates. The third column represents the ML estimates used for fitting the distribution using SRS, along with the corresponding test statistic value of the Anderson-Darling test and its *p-value*. In the fourth, fifth, and sixth columns, we estimate the parameters under JRSS, JMnRSS, and JMxRSS (which means considering the male and female populations as a joint sample) respectively. So, estimates for the combined sample under JRSS, JMnRSS, and JMxRSS are obtained by averaging the estimates from the male and female populations. The last three columns show the error in the estimates, or the difference between the estimates under SRS and JRSS, JMnRSS, and JMxRSS. So, the last three columns represent the difference in the estimates when we estimate parameters separately for the male and female populations, as well as based on a joint sample. The purpose of Table (4) is to compare the error in the estimates for the combined sample based on the male and female populations, with those estimated based on the joint sample, considering the average. From Figure (1) and (2), we can easily observe the distributional structure for the male, female, and combined populations. In Table (5) (see Annexure), we also provide the values of ML estimates, RMSE, and bias using methods such as JRSS, JSRS, JMnRSS, and JMxRSS. We do this by following the steps of Algorithm 4. From Table (5), we observe

that JMnRSS performs better than JRSS, JMxRSS, and JSRS, which also aligns with the theoretical simulation study.

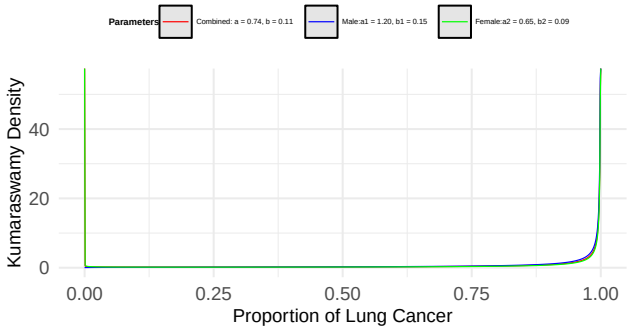


Figure 1: Density Plot for Lung Cancer

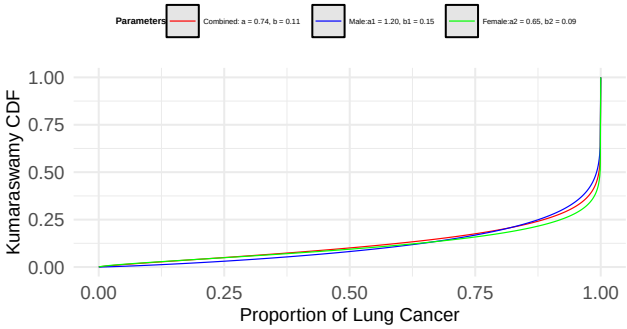


Figure 2: CDF Plot for Lung Cancer

7. Conclusion

Our study indicates that the JMnRSS scheme outperforms the JRSS, JMxRSS, and JSRS methods in terms of RMSE for ML estimation of the Kumaraswamy population. We also observed the behaviour of ML estimates of the Kumaraswamy population under all joint sampling schemes, and according to the simulation tables, as the total sample size $m = m_1 + m_2$ and cycle k increase, RMSE and bias decrease. As the difference (d) in the parameters of the two Kumaraswamy populations increases, the RMSE and bias also increase. We recommend using JMnRSS schemes for the estimation of parameters of two Kumaraswamy distributions. In the real scenario, we observed that JMnRSS is more suitable for predicting the probability of lung cancer compared to JRSS, JMxRSS, and JSRS. More research is needed in the area of joint ranked set sampling, including methods like joint percentile sampling. Additionally, Bayesian estimation can be applied to estimate the parameters of the Kumaraswamy distribution using various joint sampling.

Acknowledgements

The first author is thankful to SHODH, Gujarat for providing financial support under the state government of Gujarat (2022016452). The authors are grateful to the Chief Editor and anonymous referees for their valuable suggestions to improve our research paper.

Conflict of interest

The authors declare no conflict of interest.

References

Abbasi, A. M. and Shahd, M. Y. (2017). Estimation of population mean and median using double robust truncation based ranked set sampling. *Pakistan Journal of Statistics and Operation Research*, **13**, 379–394.

Akgul, F. G. and Senoglu, B. (2017). Estimation of using modifications of ranked set sampling for Weibull distribution. *Pakistan Journal of Statistics and Operation Research*, **13**, 931–958.

Ashour, S. and Abo-Kasem, O. (2014). Parameter estimation for two Weibull populations under joint type II censored scheme. *International Journal of Engineering*, **5**, 31–36.

Chen, Z., Bai, Z., and Sinha, B. K. (2004). *Ranked Set Sampling: Theory and Applications*, volume 176. Springer.

Dey, S., Mazucheli, J., and Nadarajah, S. (2018). Kumaraswamy distribution: different methods of estimation. *Computational and Applied Mathematics*, **37**, 2094–2111.

Ding, L. and Gui, W. (2023). Statistical inference of two gamma distributions under the joint type-ii censoring scheme. *Mathematics*, **11**, 2003.

Koyuncu, N. (2018). Regression estimators in ranked set, median ranked set and neoteric ranked set sampling. *Pakistan Journal of Statistics and Operation Research*, **14**, 89–94.

McIntyre, G. (1952). A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research*, **3**, 385–390.

Mitnik, P. A. (2013). New properties of the Kumaraswamy distribution. *Communications in Statistics-Theory and Methods*, **42**, 741–755.

Patel, M. N., Muralidharan, K., and Bavagosai, P. (2022). Comparing two Rayleigh populations using joint ranked set sampling. *International Journal of Mathematics and Statistics*, **23**, 385–390.

Raykundaliya, D. and Bhingikar, M. (2025). Estimation of parameters of two Weibull populations under the joint ranked set Sampling . *Revista Investigación Operacional*, **46**, 439–472.

Raykundaliya, D. and Patel, M. (2022a). Estimation for two exponential populations based on joint percentile rank set sampling. *International Journal of Statistics and Reliability Engineering*, **9**, 282–292.

Raykundaliya, D. and Patel, M. (2022b). Estimation for two exponential populations based on joint rank set sampling. *Revista Investigación Operacional*, **43**, 544–559.

Wolfe, D. A. (2004). Ranked set sampling: an approach to more efficient data collection. *Statistical Science*, **19**, 636–643.

World Health Organization (2023). Lung cancer. `https://www.who.int/news-room/fact-sheets/detail/lung-cancer`. Accessed: 2025-01-16.

ANNEXURE

Table 1: Simulation Result of MLE of Kumaraswamy parameters $a_1 = 1.5, b_1 = 8$ and $a_2 = 1.65, b_2 = 8.8$

			JRSS				JMnRSS				JMxRSS				JSRS			
m_1	m_2	k	$\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$	$\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$	$\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$	$\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$
			RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE
			Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias
3	4	2	1.52042	8.74694	1.67808	9.4451	1.5071	8.3475	1.6543	8.8928	1.6105	9.4153	1.725	10.2453	1.70427	11.48768	1.82056	11.82987
			0.20423	2.53662	0.17952	2.16132	0.1801	2.3415	0.1776	2.3886	0.3301	2.823	0.3112	2.9758	0.37004	4.69534	0.34346	4.32441
			0.02042	0.74694	0.02808	0.6451	0.0071	0.3475	0.0043	0.0928	0.1105	1.4153	0.075	1.4453	0.20427	3.48768	0.17056	3.02987
		3	1.48903	8.08997	1.65839	9.17293	1.4781	8.1915	1.6839	9.3758	1.535	8.7852	1.6761	9.4917	1.59852	10.16494	1.72934	10.64433
			0.15559	1.61661	0.15126	1.83125	0.1503	2.4149	0.1275	1.8764	0.2918	2.535	0.2893	2.4968	0.32773	3.52876	0.29465	3.55104
			-0.01097	0.08997	0.00839	0.37293	-0.0219	0.1915	0.0339	0.5758	0.035	0.7852	0.0261	0.6917	0.09852	2.16494	0.07934	1.84433
		4	1.52597	8.46756	1.62412	8.74326	1.4869	7.9214	1.6448	9.0335	1.569	9.0162	1.6414	9.0979	1.55196	9.22715	1.71098	10.45493
			0.15298	1.6016	0.12414	1.38213	0.1444	1.4674	0.1245	1.3401	0.2402	2.1269	0.2198	1.9701	0.22626	2.82601	0.25781	3.27956
			0.02597	0.46756	-0.02588	-0.05674	-0.0131	-0.0786	-0.0052	0.2335	0.069	1.0162	-0.0086	0.2979	0.05196	1.22715	0.06098	1.65493
	5	2	1.47201	8.14049	1.61921	8.88119	1.5433	8.6529	1.634	8.8785	1.5259	9.0356	1.6417	9.3497	1.62578	10.32224	1.69332	10.33858
			0.18974	2.13464	0.16257	1.79604	0.1723	1.9977	0.1638	2.594	0.3601	2.9143	0.3254	2.8226	0.36054	3.95396	0.29393	3.84344
			-0.02799	0.14049	-0.03079	0.08119	0.0433	0.6529	-0.016	0.0785	0.0259	1.0356	-0.0083	0.5497	0.12578	2.32224	0.04332	1.53858
		3	1.48583	8.25521	1.64063	8.85063	1.5005	8.313	1.656	9.226	1.4668	8.3746	1.7127	9.5877	1.62752	10.10057	1.73163	10.22245
			0.13487	1.54567	0.14191	1.72893	0.127	1.7582	0.1526	1.995	0.3294	2.4214	0.2326	2.2042	0.29952	3.56592	0.28102	3.34221
			-0.01417	0.25521	-0.00937	0.05063	0.0005	0.313	0.006	0.426	-0.0332	0.3746	0.0627	0.7877	0.12752	2.10057	0.08163	1.42245
		4	1.49401	8.14764	1.6395	9.04624	1.4864	8.127	1.6355	9.1881	1.5185	8.3664	1.659	8.9808	1.60941	9.50905	1.69059	10.23561
			0.13804	1.31315	0.11549	1.40191	0.1122	1.4002	0.1027	1.6387	0.1928	1.2865	0.1691	1.33	0.25071	3.11481	0.20955	2.90981
			-0.00599	0.14764	-0.0105	0.24624	-0.0136	0.127	-0.0145	0.3881	0.0185	0.3664	0.009	0.1808	0.10941	1.50905	0.04059	1.43561
	6	2	1.53153	8.35245	1.65106	9.24652	1.5197	8.4047	1.6463	9.2382	1.4556	8.2586	1.7266	10.0508	1.55013	10.02691	1.68686	10.37209
			0.1506	1.23993	0.11308	1.63126	0.1698	2.1377	0.1496	1.9254	0.3666	2.6342	0.2819	2.6538	0.2737	3.28596	0.2671	3.53977
			0.03153	0.35245	0.00106	0.44652	0.0197	0.4047	-0.0037	0.4382	-0.0444	0.2586	0.0766	1.2508	0.05013	2.02691	0.03686	1.57209
		3	1.49639	8.33889	1.63636	8.90609	1.5043	8.261	1.6567	8.9116	1.4939	8.342	1.6673	9.1017	1.50812	9.01897	1.74196	10.75446
			0.14747	1.58555	0.11467	1.34916	0.1231	1.3795	0.1006	1.355	0.1855	1.2142	0.1839	1.4168	0.31534	3.16649	0.22777	3.07345
			-0.00361	0.33889	-0.01364	0.10609	0.0043	0.261	0.0067	0.1116	-0.0061	0.342	0.0173	0.3017	0.00812	1.01897	0.09196	1.95446
		4	1.51102	8.38895	1.63615	8.81763	1.541	8.4263	1.625	8.7421	1.4675	8.0014	1.6893	9.2976	1.58103	9.28695	1.71101	10.162
			0.10626	1.22692	0.10878	1.29116	0.1269	1.7628	0.1031	1.7009	0.2157	1.6695	0.2149	1.8116	0.25318	2.77328	0.21167	2.76845
			0.01102	0.38895	-0.01385	0.01763	0.041	0.4263	-0.025	-0.0579	-0.0325	0.0014	0.0393	0.4976	0.08103	1.28695	0.06101	1.362

Table 1: Continued

m_1	m_2	k	JRSS				JMnRSS				JMxRSS				JSRS			
			$\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$	$\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$	$\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$	$\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$
			RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE
			Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias
4	4	2	1.49286	8.25801	1.67274	9.53145	1.4909	8.2363	1.6479	9.1722	1.5208	8.7169	1.716	9.8436	1.66583	10.7926	1.756	10.28901
			0.18434	2.27303	0.15789	1.73823	0.1784	2.4963	0.1703	2.2305	0.3614	2.9985	0.3175	2.8136	0.34001	4.36076	0.34107	3.5746
			-0.00714	0.25801	0.02274	0.73145	-0.0091	0.2363	-0.0021	0.3722	0.0208	0.7169	0.066	1.0436	0.16583	2.7926	0.106	1.48901
		3	1.46499	8.02776	1.64601	8.91012	1.5	8.3285	1.6341	9.1258	1.5666	9.0351	1.6795	9.4076	1.54263	9.69631	1.70524	10.3902
			0.1326	1.67665	0.13566	1.44047	0.1285	1.826	0.145	1.8077	0.2881	2.6854	0.2653	2.3074	0.21483	3.1611	0.27166	3.50769
			-0.03501	0.02776	-0.00399	0.11012	0.1285	0.3285	-0.0159	0.3258	0.0666	1.0351	0.0295	0.6076	0.04263	1.69631	0.05524	1.5902
		4	1.53286	8.71851	1.65888	9.13196	1.4724	7.9228	1.6728	9.3161	1.5382	8.8496	1.689	9.3796	1.57266	9.55059	1.68745	9.82742
			0.15238	1.89434	0.13485	1.73994	0.1328	1.7914	0.1093	1.3466	0.2522	2.2657	0.2335	2.0799	0.22858	2.79425	0.26297	3.1497
			0.03286	0.71851	0.00888	0.33196	-0.0276	-0.0772	0.0228	0.5161	0.0382	0.8496	0.039	0.5796	0.07266	1.55059	0.03745	1.02742
	5	2	1.53649	8.8251	1.63839	8.90336	1.5105	8.6123	1.6744	9.099	1.5353	8.9401	1.74	9.8085	1.60176	9.96134	1.71982	10.96764
			0.15339	1.8379	0.14193	1.61145	0.1393	2.1509	0.1612	1.7171	0.2898	2.569	0.2689	2.5554	0.27897	3.13417	0.24793	3.66735
			0.03649	0.8251	-0.01161	0.10336	0.0105	0.6123	0.0244	0.299	0.0353	0.9401	0.09	1.0085	0.10176	1.96134	0.06982	2.16764
		3	1.51083	8.38981	1.66195	9.18154	1.5067	8.4458	1.6802	9.115	1.5758	9.0887	1.6723	9.5127	1.58959	9.83401	1.68042	9.95384
			0.13429	1.39775	0.12176	1.42818	0.1316	1.728	0.1159	1.2251	0.2998	2.6519	0.2308	1.9946	0.26432	3.20095	0.23664	2.88007
			0.01083	0.38981	0.01195	0.38154	0.0067	0.4458	0.0302	0.315	0.0758	1.0887	0.0223	0.7127	0.08959	1.83401	0.03042	1.15384
		4	1.50682	8.13379	1.61718	8.66516	1.5071	8.5482	1.6511	8.7686	1.5045	8.1788	1.6547	9.0448	1.53221	9.20624	1.68242	9.36452
			0.12553	1.41001	0.12633	1.37759	0.1073	1.7723	0.1081	1.4952	0.1966	1.4313	0.1728	1.4345	0.20655	2.7503	0.23708	2.65352
			0.00682	0.13379	-0.03282	-0.13484	0.0071	0.5482	0.0011	-0.0314	0.0045	0.1788	0.0047	0.2448	0.03221	1.20624	0.03242	0.56452
	6	2	1.52455	8.46861	1.64938	8.91842	1.4945	8.5106	1.6556	9.4685	1.5893	9.0071	1.7052	9.6838	1.59438	10.08091	1.72498	10.18676
			0.15115	1.58216	0.11672	1.68982	0.149	2.1655	0.1366	2.0424	0.3316	2.65	0.2774	2.4755	0.3062	3.69095	0.26459	3.29431
			0.02455	0.46861	-0.00062	0.11842	-0.0055	0.5106	0.0056	0.6685	0.0893	1.0071	0.0552	0.8838	0.09438	2.08091	0.07498	1.38676
		3	1.48871	8.28029	1.64183	8.95378	1.5043	8.4693	1.6437	8.6977	1.5302	8.5058	1.7057	9.4572	1.5782	9.97349	1.74419	10.459
			0.11894	1.36256	0.11325	1.52915	0.0993	1.5849	0.1107	1.3881	0.2198	1.5732	0.2154	1.9084	0.24481	3.46441	0.24894	3.13182
			-0.01129	0.28029	-0.00817	0.15378	0.0043	0.4693	-0.0063	-0.1023	0.0302	0.5058	0.0557	0.6572	0.0782	1.97349	0.09419	1.659
		4	1.49008	8.12498	1.63823	8.83481	1.4869	7.9199	1.6483	9.0197	1.4993	8.4833	1.6761	9.3754	1.51529	8.76422	1.70084	9.82427
			0.11396	1.40001	0.10384	1.3087	0.1162	1.4423	0.1031	1.5729	0.1941	1.7086	0.1837	1.617	0.20438	2.69504	0.19635	2.62445
			-0.00992	0.12498	-0.01177	0.03481	-0.0131	-0.0801	-0.0017	0.2197	-0.0007	0.4833	0.0261	0.5754	0.01529	0.76422	0.05084	1.02427

Table 1: Continued

m_1	m_2	k	JRSS				JMnRSS				JMxRSS				JSRS			
			$\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$	$\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$	$\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$	$\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$
			RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE
			Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias
5	4	2	1.48816	8.27453	1.66497	8.9578	1.5227	8.7664	1.6813	8.7665	1.6	9.3217	1.6564	9.4102	1.59855	9.93289	1.73751	10.6201
			0.12018	1.56676	0.12897	1.09736	0.1301	2.0904	0.1614	1.6084	0.2888	2.6692	0.3089	2.7612	0.26986	3.49422	0.29757	3.35348
			-0.01184	0.27453	0.01497	0.1578	0.0227	0.7664	0.0313	-0.0335	0.1	1.3217	0.0064	0.6102	0.09855	1.93289	0.08751	1.8201
		3	1.49904	8.26458	1.64212	8.85833	1.4819	7.7353	1.648	8.7964	1.5608	9.1097	1.7019	9.7415	1.60716	10.0496	1.77386	11.01777
			0.12487	1.55068	0.12663	1.36741	0.1233	1.4287	0.1274	1.1432	0.2362	2.1425	0.2959	2.5743	0.26576	3.4895	0.27533	3.58305
			-0.00096	0.26458	-0.00788	0.05833	-0.0181	-0.2647	-0.002	-0.0036	0.0608	1.1097	0.0519	0.9415	0.10716	2.0496	0.12386	2.21777
		4	1.50033	8.30283	1.66359	8.99673	1.5122	8.725	1.68	9.5934	1.5431	8.5502	1.6501	8.9983	1.53403	9.26047	1.75299	10.80342
			0.10876	1.40566	0.12898	1.53228	0.1184	2.2533	0.1047	1.6158	0.2314	1.8853	0.1753	1.5094	0.21066	2.92393	0.23196	3.12815
			0.00033	0.30283	0.01359	0.19673	0.0122	0.725	0.03	0.7934	0.0431	0.5502	0.0001	0.1983	0.03403	1.26047	0.10299	2.00342
	5	2	1.47711	7.85813	1.63547	8.78846	1.5055	8.2648	1.6518	9.2239	1.5891	9.3453	1.7194	9.7755	1.63624	10.45833	1.70631	10.60979
			0.1315	1.36304	0.13285	1.33712	0.1203	1.7857	0.1141	1.8308	0.305	2.7984	0.2442	2.2355	0.30936	3.70549	0.27717	3.60894
			-0.02289	-0.14187	-0.01453	-0.01154	0.0055	0.2648	0.0018	0.4239	0.0891	1.3453	0.0694	0.9755	0.13624	2.45833	0.05631	1.80979
		3	1.53389	8.76255	1.65234	9.0756	1.5126	8.4426	1.6603	9.2183	1.56	8.8246	1.6635	9.3273	1.5746	9.55388	1.7006	10.68063
			0.13852	1.68649	0.12719	1.60717	0.1027	1.8195	0.1151	1.7292	0.2973	2.4951	0.2351	2.0674	0.2255	3.01462	0.23814	3.00685
			0.03389	0.76255	0.00234	0.2756	0.0126	0.4426	0.0103	0.4183	0.06	0.8246	0.0135	0.5273	0.0746	1.55388	0.0506	1.88063
		4	1.49763	8.11658	1.64213	8.85041	1.5327	8.9147	1.6572	9.015	1.4898	8.0667	1.6636	9.0496	1.57873	9.72109	1.69001	9.75746
			0.09675	1.18347	0.08422	1.04815	0.0962	1.7947	0.1222	1.6516	0.1732	1.3127	0.1714	1.3468	0.23349	3.07959	0.20892	2.54929
			-0.00237	0.11658	-0.00787	0.05041	0.0327	0.9147	0.0072	0.215	-0.0102	0.0667	0.0136	0.2496	0.07873	1.72109	0.04001	0.95746
	6	2	1.45535	7.88926	1.65669	9.00768	1.522	8.8238	1.6791	9.2446	1.5775	9.0796	1.6785	9.4493	1.60715	10.43475	1.72228	10.41793
			0.15067	1.77778	0.13747	1.72694	0.1273	2.0251	0.1215	1.66	0.3161	2.4909	0.3163	2.6428	0.28067	3.68499	0.25893	3.48022
			-0.04465	-0.11074	0.00669	0.20768	0.022	0.8238	0.0291	0.4446	0.0775	1.0796	0.0285	0.6493	0.10715	2.43475	0.07228	1.61793
		3	1.47545	7.86434	1.62126	8.72511	1.4818	8.0136	1.6558	8.8154	1.5533	8.4833	1.6483	9.1125	1.59503	9.75264	1.69195	10.41235
			0.1182	1.28614	0.10698	1.20663	0.1206	1.7415	0.1074	1.0814	0.2293	1.6834	0.1923	1.4398	0.22472	2.81341	0.24977	3.30357
			-0.02455	-0.13566	-0.02874	-0.07489	-0.0182	0.0136	0.0058	0.0154	0.0533	0.4833	-0.0017	0.3125	0.09503	1.75264	0.04195	1.61235
		4	1.52193	8.35878	1.63265	8.75807	1.4997	8.4966	1.6487	8.7416	1.497	8.1003	1.673	9.1451	1.59803	9.80179	1.66894	9.73596
			0.10536	1.26196	0.09571	1.17483	0.0832	1.3748	0.0985	1.2126	0.1861	1.408	0.1719	1.43	0.23598	2.92769	0.18036	2.46181
			0.02193	0.35878	-0.01735	-0.04193	-0.0003	0.4966	-0.0013	-0.0584	-0.003	0.1003	0.023	0.3451	0.09803	1.80179	0.01894	0.93596

Table 2: Simulation Result of MLE of Kumaraswamy parameters for change in difference d between parameters $a_1 = 1.5, b_1 = 8$ and $a_2 = d \times a_1, b_2 = d \times b_1$

		$d = 1.1$				$d = 1.3$			
$m_1 = 3, m_2 = 4$	k	$\hat{a}_1$ RMSE Bias	$\hat{b}_1$ RMSE Bias	$\hat{a}_2$ RMSE Bias	$\hat{b}_2$ RMSE Bias	$\hat{a}_1$ RMSE Bias	$\hat{b}_1$ RMSE Bias	$\hat{a}_2$ RMSE Bias	$\hat{b}_2$ RMSE Bias
JRSS	2	1.52042 0.20423 0.02042	8.74694 2.53662 0.74694	1.67808 0.17952 0.02808	9.4451 2.16132 0.6451	1.5291 0.23957 0.0291	9.17915 2.97852 1.17915	1.9949 0.23274 0.0449	11.49791 3.13233 1.09791
	3	1.48903 0.15559 -0.01097	8.08997 1.61661 0.08997	1.65839 0.15126 0.00839	9.17293 1.83125 0.37293	1.49837 0.17681 -0.00163	8.5496 2.19129 0.5496	1.97007 0.17885 0.02007	11.09915 2.30199 0.69915
	4	1.52597 0.15298 0.02597	8.46756 1.6016 0.46756	1.62412 0.12414 -0.02588	8.74326 1.38213 -0.05674	1.51706 0.15302 0.01706	8.35237 1.65903 0.35237	1.92898 0.13385 -0.02102	10.36807 1.59539 -0.03193
JMnRSS	2	1.5071 0.1801 0.0071	8.3475 2.3415 0.3475	1.6543 0.1776 0.0043	8.8928 2.3886 0.0928	1.499 0.2181 -0.001	8.3764 2.8794 0.3764	1.9692 0.2134 0.0192	10.6688 2.0536 0.2688
	3	1.4781 0.1503 -0.0219	8.1915 2.4149 0.1915	1.6839 0.1275 0.0339	9.3758 1.8764 0.5758	1.4968 0.1433 -0.0032	8.364 2.0406 0.364	1.9436 0.1723 -0.0064	10.2982 1.88 -0.1018
	4	1.4869 0.1444 -0.0131	7.9214 1.4674 -0.0786	1.6448 0.1245 -0.0052	9.0335 1.3401 0.2335	1.5142 0.1194 0.0142	8.5056 1.6181 0.5056	1.9409 0.1243 -0.0091	10.9441 1.242 0.5441
JMxRSS	2	1.6105 0.3301 0.1105	9.4153 2.823 1.4153	1.725 0.3112 0.075	10.2453 2.9758 1.4453	1.5455 0.331 0.0455	8.7778 2.7325 0.7778	1.9005 0.3005 -0.0495	10.712 2.77 0.312
	3	1.535 0.2918 0.035	8.7852 2.535 0.7852	1.6761 0.2893 0.0261	9.4917 2.4968 0.6917	1.5749 0.3347 0.0749	8.9757 2.743 0.9757	2.0615 0.3207 0.1115	11.7672 3.1144 1.3672
	4	1.569 0.2402 0.069	9.0162 2.1269 1.0162	1.6414 0.2198 -0.0086	9.0979 1.9701 0.2979	1.5316 0.2831 0.0316	8.595 2.3068 0.595	1.9448 0.2157 -0.0052	10.7263 2.2432 0.3263
JSRS	2	1.70427 0.37004 0.20427	11.48768 4.69534 3.48768	1.82056 0.34346 0.17056	11.82987 4.32441 3.02987	1.66509 0.41662 0.16509	11.07387 5.32529 3.07387	2.07283 0.38875 0.12283	13.03948 4.95431 2.63948
	3	1.59852 0.32773 0.09852	10.16494 3.52876 2.16494	1.72934 0.29465 0.07934	10.64433 3.55104 1.84433	1.54414 0.23431 0.04414	9.31753 3.11811 1.31753	2.02295 0.24579 0.07295	11.61412 2.92413 1.21412
	4	1.55196 0.22626 0.05196	9.22715 2.82601 1.22715	1.71098 0.25781 0.06098	10.45493 3.27956 1.65493	1.56442 0.23593 0.06442	9.45817 3.14709 1.45817	2.02109 0.24714 0.07109	11.83446 3.31259 1.43446

Table 3: Simulation Result of power of LRT test for equality of parameters of Kumaraswamy populations $a_1 = a_2 = a$, when b_1 and b_2 are known

			JRSS		JMnRSS		JMxRSS		JSRS	
m_1	m_2	k	$\lambda(W)$	**Power**	$\lambda(W)$	**Power**	$\lambda(W)$	**Power**	$\lambda(W)$	**Power**
3	4	2	0.605946	0.109	0.634739	0.1	0.640683	0.097	0.683245	0.071
		3	0.560637	0.17	0.602584	0.127	0.597573	0.118	0.66023	0.087
		4	0.511984	0.2	0.55775	0.172	0.596695	0.125	0.654486	0.089
	5	2	0.561658	0.148	0.616627	0.123	0.624577	0.122	0.683656	0.066
		3	0.527265	0.193	0.581667	0.157	0.612644	0.113	0.665368	0.077
		4	0.480041	0.251	0.518879	0.194	0.563974	0.17	0.651837	0.096
	6	2	0.579479	0.151	0.604477	0.103	0.61886	0.118	0.677142	0.08
		3	0.507203	0.205	0.553525	0.17	0.577514	0.133	0.661561	0.086
		4	0.45349	0.276	0.510304	0.21	0.542722	0.172	0.647128	0.106
4	4	2	0.59251	0.146	0.610892	0.12	0.619238	0.122	0.690272	0.074
		3	0.510793	0.213	0.57404	0.14	0.593448	0.117	0.653123	0.09
		4	0.462144	0.254	0.514641	0.204	0.551085	0.165	0.643129	0.08
	5	2	0.548538	0.195	0.600239	0.115	0.602536	0.141	0.655993	0.086
		3	0.490216	0.243	0.553969	0.179	0.571588	0.155	0.657353	0.079
		4	0.408267	0.307	0.517983	0.2	0.53759	0.181	0.641473	0.1
	6	2	0.51064	0.217	0.576494	0.144	0.570677	0.142	0.661429	0.079
		3	0.450653	0.281	0.527459	0.188	0.543078	0.175	0.645314	0.089
		4	0.373734	0.379	0.468754	0.256	0.518053	0.208	0.621441	0.111
5	4	2	0.548698	0.159	0.602322	0.126	0.617442	0.121	0.676036	0.072
		3	0.46877	0.25	0.531177	0.177	0.572671	0.15	0.647266	0.091
		4	0.428425	0.294	0.49494	0.209	0.538261	0.18	0.640677	0.097
	5	2	0.519791	0.214	0.572252	0.144	0.590827	0.131	0.677568	0.077
		3	0.449507	0.277	0.513034	0.201	0.547614	0.177	0.638682	0.095
		4	0.361311	0.366	0.464509	0.241	0.497757	0.239	0.620769	0.113
	6	2	0.475895	0.231	0.564788	0.162	0.598058	0.133	0.669942	0.082
		3	0.393947	0.341	0.50123	0.226	0.530053	0.192	0.644247	0.099
		4	0.31849	0.439	0.450126	0.285	0.484016	0.217	0.624834	0.117

Table 4: Results for lung-cancer dataset

	SRS		JRSS	JMnRSS	JMxRSS	Error_JRSS_SRS	Error_JMnRSS_SRS	Error_JMxRSS_SRS
	Initial Value	ML	ML	ML	ML	ML	ML	ML
Combined Population of male ans female								
a	0.67105	0.749813	1.0726105	0.9197033	1.2180716	0.161253245	-0.1698899	-0.4682582
b	0.110138	0.117571	0.1352246	0.16985445	0.1392217	0.000629332	-0.05228335	-0.0216506
Test-Statistics(AD)		0.85645						
$p-value$		0.442						
Male population								
a_1	1.0782	1.209441	1.35477	1.1726907	1.499672	0.07649091	0.0367507	-0.2902306
b_1	0.128353	0.150617	0.1600692	0.1998558	0.1381828	0.001005943	-0.0492385	0.0124345
Test-Statistics(AD)		1.5376						
$p-value$		0.1678						
Female Population								
a_2	0.42208	0.65685	0.790451	0.6667159	0.9364712	0.135694615	-0.00986429	-0.27961959
b_2	0.09864	0.09809	0.11038	0.1398531	0.1402606	0.000137789	-0.04175813	-0.04216563
Test-Statistics(AD)		2.5825						
$p-value$		0.04493						

Table 5: Result of MLE Kumaraswamy parameters (a_1, b_1) and (a_2, b_2)

m_1	m_2	k	JRSS				JMnRSS				JMxRSS				JSRS			
			$\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$	$\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$	$\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$	$\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$
			RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE
			Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias	Bias
5	4	2	2.19255	0.16734	0.76217	0.09424	1.5109	0.1998	0.806	0.1542	7.449	0.1502	55.0472	0.1175	5.01078	0.18472	1.97813	0.10901
			1.23843	0.02735	0.45422	0.01486	0.5805	0.0882	0.232	0.0711	6.3491	0.0255	54.3904	0.0202	4.08135	0.0537	1.50313	0.02366
			0.98311	0.01672	0.10532	-0.0039	0.3015	0.0492	0.1491	0.0561	6.2396	-0.0004	54.3903	0.0194	3.80134	0.0341	1.32128	0.01092
		3	1.58935	0.15835	0.66964	0.09419	1.3191	0.1647	0.7887	0.1456	1.039	0.1469	8.4275	0.1153	2.46861	0.16429	0.99721	0.10256
			0.66195	0.02429	0.29454	0.01267	0.3176	0.0416	0.1803	0.053	0.7655	0.0172	7.7707	0.0172	1.56602	0.03721	0.50017	0.01493
			0.37991	0.00773	0.01279	-0.0039	0.1097	0.0141	0.1318	0.0475	-0.1704	-0.0037	7.7706	0.0172	1.25917	0.01367	0.34036	0.00447
		4	1.56245	0.1558	0.57733	0.09586	1.1889	0.1484	0.7675	0.1386	1.1218	0.1419	2.2007	0.1117	2.01851	0.16664	0.82286	0.10514
			0.56322	0.01677	0.25792	0.00969	0.2295	0.032	0.1503	0.0437	0.4233	0.0109	1.5438	0.0144	1.06235	0.03341	0.30901	0.01675
			0.35301	0.00518	-0.0795	-0.0022	-0.0205	-0.0022	0.1106	0.0405	-0.0876	-0.0087	1.5438	0.0136	0.80907	0.01602	0.16601	0.00705
	5	2	2.09839	0.16378	0.71042	0.09722	1.3139	0.1683	0.8624	0.171	3.5029	0.1511	8.4418	0.1171	4.94195	0.19485	2.3303	0.11005
			1.10089	0.02768	0.3219	0.01217	0.3283	0.0607	0.2439	0.0779	2.6262	0.0132	7.785	0.021	4.03999	0.06423	1.88627	0.02353
			0.88895	0.01316	0.05357	-0.0009	0.1045	0.0177	0.2055	0.0729	2.2935	0.0005	7.7849	0.019	3.73251	0.04423	1.67345	0.01196
		3	1.68475	0.16108	0.56554	0.09429	1.191	0.1503	0.8027	0.1519	6.4043	0.1458	11.4464	0.117	2.61316	0.17801	0.9468	0.10569
			0.64269	0.01964	0.23871	0.01013	0.2495	0.0419	0.1776	0.0573	5.5168	0.0146	10.7895	0.0189	1.66118	0.05019	0.48115	0.01968
			0.47531	0.01046	-0.0913	-0.0038	-0.0184	-0.0003	0.1458	0.0538	5.1949	-0.0048	10.7895	0.0189	1.40372	0.02739	0.28995	0.0076
		4	1.54937	0.15834	0.58177	0.09218	1.2054	0.1568	0.7744	0.1475	2.5569	0.1451	8.523	0.1112	1.64065	0.16345	0.96234	0.10367
			0.52591	0.01789	0.2347	0.00965	0.2028	0.035	0.1417	0.0528	1.8386	0.0089	7.8662	0.0131	0.73474	0.02887	0.43613	0.01551
			0.33993	0.00772	-0.0751	-0.0059	-0.004	0.0062	0.1175	0.0494	1.3475	-0.0055	7.8661	0.0131	0.43121	0.01283	0.30549	0.00558
	6	2	1.94568	0.16245	0.59578	0.09202	1.2016	0.1605	0.8603	0.1695	2.3226	0.1444	78.9649	0.1235	3.50634	0.1853	2.85091	0.11353
			1.00226	0.02282	0.24649	0.01011	0.2592	0.0532	0.2417	0.077	1.7632	0.0154	78.308	0.0254	2.68982	0.06214	2.33475	0.0278
			0.73624	0.01183	-0.0611	-0.0061	-0.0078	0.0099	0.2034	0.0714	1.1132	-0.0062	78.308	0.0254	2.2969	0.03468	2.19406	0.01544
		3	1.59066	0.16091	0.62907	0.09494	1.1199	0.1526	0.8318	0.1611	1.9973	0.1491	23.9325	0.1176	2.30441	0.17859	1.39773	0.10379
			0.61624	0.02058	0.24399	0.00924	0.209	0.0433	0.1956	0.067	1.2251	0.0162	23.2757	0.0195	1.38949	0.04091	0.8689	0.01592
			0.38122	0.01029	-0.0278	-0.0032	-0.0895	0.002	0.1749	0.063	0.7879	-0.0015	23.2756	0.0195	1.09497	0.02797	0.74088	0.0057
		4	1.5049	0.15607	0.54227	0.09282	1.1487	0.1593	0.8176	0.1605	6.8924	0.1447	14.2493	0.1218	2.08889	0.16881	0.94778	0.10326
			0.49538	0.01608	0.20588	0.00787	0.2034	0.0387	0.1674	0.0631	6.0897	0.0151	13.5924	0.0237	1.13437	0.03105	0.45022	0.01319
			0.29546	0.00545	-0.1146	-0.0053	-0.0607	0.0087	0.1607	0.0624	5.683	-0.0059	13.5924	0.0237	0.87945	0.01819	0.29093	0.00517

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 24, No. 1, 2026 (New Series), pp 53–78
http://www.ssca.org.in/journal

A Family of Additive-Multiplicative Frailty Models using the Inverse Gaussian as Frailty Distribution

Alok D. Dabade

Department of Statistics
University of Mumbai, Mumbai

Received: 31 March 2024; Revised: 05 June 2025; Accepted: 07 June 2025

Abstract

Traditionally, frailty models are built with the assumption that frailty influences the baseline hazard function in a multiplicative manner, known as a multiplicative frailty model. An alternate model available in the literature is the additive frailty model, in which the frailty variable is linearly related to the baseline hazard function. Although both models are popular, there is a need for a model that incorporates both multiplicative and additive models, especially in epidemiological data where neither the multiplicative nor the additive model adequately describes the data. This paper aims to address this gap by developing a family of models that include both multiplicative and additive frailty models under the inverse Gaussian frailty distribution. The paper also discusses the inference procedure for estimating model parameters using the MCMC method and applies the proposed model to real-life datasets.

Keywords: Additive frailty; Multiplicative frailty; Additive-Multiplicative frailty; Inverse Gaussian distribution; MCMC.

AMS Subject Classifications: 62C10, 62F15, 62N02.

The video recording of the paper made under the SSCA's Online Lecture series is available at the Youtube channel URL `https://youtu.be/aL-Dy1mhwnA`.

1. Introduction

The proportional hazard model becomes the workhorse of survival analysis. In these models, we include explanatory variables or covariates to study the effect of these variables on the hazard function. The covariates can have a multiplicative or additive effect on the hazard function. The resulting models are commonly referred to as multiplicative and additive hazard models, respectively. These models are popular because of their ease of interpretation and inference-making. Despite their popularity, Lin and Ying (1997) argued that a compromise between these models is desirable. Furthermore, to illustrate the need for compromise between these models, Lin and Ying (1997) referenced the British doctor's

Corresponding Author: Alok D. Dabade
Email: addabade@yahoo.co.in

study by Breslow and Day (1987). The data concern the effects of smoking on mortality. The data suggested that the difference between the two hazard functions increases over time, whereas their ratio decreases. Thus, neither additive nor multiplicative model adequately describes the data. Aranda-Ordaz (1983) proposed a Box-Cox type transformation family,

$$\frac{h^p(t \mid \underline{x}) - 1}{p} = h_0(t) + \underline{\beta}^T \underline{x}$$

in which $p \to 0$ corresponds to multiplicative and $p = 1$ corresponds to additive hazard models. In addition, Lin and Ying (1995) proposed an alternate extension given by

$$h(t \mid \underline{x}_a, \underline{x}_m) = e^{\underline{\beta}^T \underline{x}_m} \left\{ h_0(t) + \gamma^T \underline{x}_a \right\}$$

The covariates in the additive part $\underline{x}_a$ can be the same as, related to, or different from that of the multiplicative part $\underline{x}_m$.

In proportional hazard models, the assumption is that the population is homogeneous. However, heterogeneity can arise from unknown factors that influence the hazard function. For example, in a family disease study, the risk of disease occurrence varies according to genetic predisposition or shared environmental factors. Vaupel *et al.* (1979) were the first to introduce the term frailty to describe the population's heterogeneity. Duchateau and Janseen (2008), Hanagal (2019), Hougaard (2000), and Wienke (2011) are good references for the frailty models.

Several frailty models are currently being developed. The most common approach to define frailty models is to assume that frailty interacts multiplicatively with the baseline hazard function, called the multiplicative frailty model. Following the introduction of additive models by Aalen (1980), many additive frailty models have been developed in the literature as an alternative to the multiplicative frailty models. Recent advances in additive frailty models include contributions from Silva and Amarmal-Turkman (2004), Hanagal and Pandey (2016, 2017) and Hanagal (2022). In this model, frailty acts additively with the baseline hazards function.

In the case of frailty models, frailty represents the effect of some unknown factors that affect the hazard function. These unknown factors can affect the hazard function either additively or multiplicatively. For example, Kheiri *et al.* (2005) analyzed corneal transplant data in which the event of interest is graft rejection. The unknown causes that increase the rate of graft rejection may involve recipient-related, donor-related, surgery-related, or environmental-related factors. Some of these factors may have an additive effect, and others may have a multiplicative effect. Aalen and Tretli (1999) analyzed data on testicular cancer. Men who develop cancer after the hormonal process of puberty has started, receive damage during a critical period of their fetal development. The damage may result from the mother's exposure to environmental factors, an excessive estrogenic burden, or genetic factors. These causes may additively or multiplicatively affect the hazard function. However, it is challenging to determine which factors have an additive effect and which ones have a multiplicative effect, as frailty is unobservable. The usual approach is to fit the additive, multiplicative and additive-multiplicative models and select the best among them.

According to the literature review, no previous study has considered incorporating frailty with both additive and multiplicative effects into a model. This research aims to

bridge this gap. This paper introduces the concept of the additive-multiplicative frailty model, where the frailty random variable is assumed to follow an inverse Gaussian distribution. The paper is structured as follows: Section 2 presents the general univariate additive and multiplicative frailty models. Further, the Section continues to introduce additive-multiplicative frailty models. Section 3 explores the inverse Gaussian distribution, as well as additive, multiplicative, and additive-multiplicative inverse Gaussian frailty models. Section 4 discusses the baseline distribution. Section 5 outlines the Bayesian inferential procedure. Sections 6 and 7 are respectively dedicated to the simulation study and real-life data analysis. Finally, Section 8 concludes the paper with a summary of the findings.

2. Univariate frailty models

2.1. Univariate multiplicative frailty models

Let T be a lifetime random variable and Z_m be a non-negative frailty random variable. The conditional hazard function under the multiplicative frailty (MF) model for given frailty $Z_m = z_m$ and known covariates $\underline{X}_m = \underline{x}_m$ at time $t > 0$ is given by

$$h_m(t \mid \underline{x}_m, z_m) = \begin{cases} z_m h_0(t)\eta_m & ; \quad t > 0 \\ 0 & ; \quad \text{otherwise} \end{cases} \tag{1}$$

where $h_0(t)$ is the baseline hazard function; $\eta_m = e^{\underline{\gamma}^T \underline{x}_m}$ is the link function and $\underline{\gamma}$ is a vector of regression coefficients. Using the relation between the survival function and the hazard function, the conditional survival function for given frailty is

$$S_m(t \mid \underline{x}_m, z_m) = \begin{cases} exp\left(-z_m H_0(t)\eta_m\right) & ; \quad t > 0 \\ 1 & ; \quad \text{otherwise} \end{cases} \tag{2}$$

where $H_0(t)$ is the cumulative baseline hazard function at time $t > 0$. Suppose that the frailty random variable Z_m follows a continuous distribution defined over the positive half part of the real line with probability density function $f_{Z_m}(\cdot)$ and Laplace transform $L_{Z_m}(\cdot)$. Integrating over the range of the frailty random variable, the survival function of the lifetime random variable T is,

$$S_m(t \mid \underline{x}_m) = \begin{cases} L_{Z_m}(H_0(t)\eta_m) & ; \quad t > 0 \\ 1 & ; \quad \text{otherwise} \end{cases} \tag{3}$$

The probability density function of T is given by

$$\psi_m(t \mid \underline{x}_m) = \begin{cases} h_0(t)\eta_m L^{(1)}_{Z_m}(H_0(t)\eta_m) & ; \quad t > 0 \\ 0 & ; \quad \text{otherwise} \end{cases} \tag{4}$$

where $L^{(1)}_{Z_m}(s) = -\int_0^\infty z e^{-zs} f_{Z_m}(z) dz$.

2.2. Univariate additive frailty models

If Z_a is a frailty random variable that affects the hazard function additively having the probability density function $f_{Z_a}(\cdot)$ and the Laplace transform $L_{Z_a}(\cdot)$, then the conditional

hazard function for given frailty $Z_a = z_a$ and known covariates $\underline{X}_a = \underline{x}_a$ under the additive frailty (AF) model is

$$h_a(t \mid \underline{x}_a, z_a) = \begin{cases} h_0(t) + z_a\eta_a & ; \quad t > 0 \\ 0 & ; \quad \text{otherwise} \end{cases} \tag{5}$$

where $\eta_a = e^{\underline{\beta}^T \underline{x}_a}$ and $\underline{\beta}$ is a vector of regression coefficients. The conditional survival function for given frailty and covariates is

$$S_a(t \mid \underline{x}_a, z_a) = \begin{cases} exp\left(-H_0(t)\right) exp\left(-z_a t\eta_a\right) & ; \quad t > 0 \\ 1 & ; \quad \text{otherwise} \end{cases} \tag{6}$$

After integrating over the range of the frailty random variable, the survival and probability density functions of T are given by

$$S_a(t \mid \underline{x}_a) = \begin{cases} e^{-H_0(t)} L_{Z_a}(t\eta_a) & ; \quad t > 0 \\ 1 & ; \quad \text{otherwise} \end{cases} \tag{7}$$

$$\psi_a(t \mid \underline{x}_a) = \begin{cases} h_0(t) S_a(t) + \eta_a e^{-H_0(t)} L^{(1)}_{Z_a}(t\eta_a) & ; \quad t > 0 \\ 0 & ; \quad \text{otherwise} \end{cases} \tag{8}$$

The MF and AF models can be extended as shared and correlated for multivariate cases. In shared frailty models, the frailty variable is shared by all individuals in a family. In contrast, in correlated frailty models, frailty variables associated with all individuals in a family are correlated. This article considers only univariate frailty models, so shared and correlated frailty models are not discussed further.

2.3. Univariate additive-multiplicative frailty models

Suppose $\underline{Z} = (Z_a, Z_m)$ is a frailty random vector with joint probability density function $f_{\underline{Z}}(\cdot, \cdot)$ and Laplace transform $L_{\underline{Z}}(\cdot, \cdot)$ in which Z_a and Z_m are non-negative frailty random variables that act additively and multiplicatively on the hazard function, respectively. Furthermore, suppose $\underline{X}_a$ and $\underline{X}_m$ are vectors of covariates acting additively and multiplicatively on the hazards function. The conditional hazard function given $\underline{z} = (z_a, z_m)$ and $\underline{x} = (\underline{x}_a, \underline{x}_m)$ under the additive-multiplicative frailty (AMF) model is

$$h(t \mid \underline{x}, \underline{z}) = \begin{cases} z_a\eta_a + z_m h_0(t)\eta_m & ; \quad t > 0 \\ 0 & ; \quad \text{otherwise} \end{cases} \tag{9}$$

The conditional survival function given $\underline{x}$ and $\underline{z}$ is

$$S(t \mid \underline{x}, \underline{z}) = \begin{cases} exp\left[-\left(z_a t\eta_a + z_m H_0(t)\eta_m\right)\right] & ; \quad t > 0 \\ 0 & ; \quad \text{otherwise} \end{cases} \tag{10}$$

The survival and probability density functions of the lifetime random variable T at $t > 0$ are, respectively,

$$S(t \mid \underline{x}) = \begin{cases} L_{\underline{Z}}\left(t\eta_a, H_0(t)\eta_m\right) & ; \quad t > 0 \\ 1 & ; \quad \text{otherwise} \end{cases} \tag{11}$$

$$\psi(t \mid \underline{x}) = \begin{cases} \eta_a L^{(a)}_{\underline{Z}}(t\eta_a, H_0(t)\eta_m) + h_0(t)\eta_m L^{(m)}_{\underline{Z}}(t\eta_a, H_0(t)\eta_m) & ; \quad t > 0 \\ 0 & ; \quad \text{otherwise} \end{cases} \tag{12}$$

where $L_{\underline{Z}}^{(i)}(s_1, s_2) = -\int_0^\infty \int_0^\infty z_i e^{-[z_a s_1 + z_m s_2]} f_{\underline{Z}}(z_a, z_m)\, dz_a dz_m \;;\; i = a, m.$

If Z_a has a degenerate distribution at zero, then Equation (9) reduces to Equation (1), and hence, AMF model reduces to the MF model similarly if Z_m is degenerate at one, then the AMF model reduces to the AF model. In particular, if $\eta_m = 1$ then Equation (9) reduces to Equation (5).

3. Inverse Gaussian frailty models

3.1. Inverse Gaussian distribution

The gamma distribution is the most common and simple frailty distribution due to its mathematical convenience. However, it has some drawbacks (*see* Kheiri *et al.* (2007)). For example, it may weaken the effect of covariates. Another choice of frailty distribution that can be considered is the inverse Gaussian distribution, as it shares many striking similarities with the most popular statistical distribution, the normal distribution (*see* Chhikara *et al.* 1986). Hougaard (1984) introduced the inverse Gaussian as a frailty distribution.

Let a continuous random variable X follows inverse Gaussian (IG) distribution with parameters μ and α, denoted by $IG(\mu, \alpha)$, then the probability density function of X is

$$f(x) = \begin{cases} \sqrt{\dfrac{\alpha}{2\pi x^3}}\; exp\left[-\dfrac{\alpha(x-\mu)^2}{2x\mu^2}\right] & ;\; x > 0, \alpha > 0, \mu > 0 \\ 0 & ;\; otherwise. \end{cases}$$

Hougaard (2000) given a re-parameterization of $f(\cdot)$ from an exponential family point of view as;

$$f(x) = \sqrt{\frac{\psi}{2\pi}} exp\left(\sqrt{\psi\phi}\right) x^{-3/2} exp\left\{-\frac{\phi x}{2} - \frac{\psi}{2x}\right\} \;\; ;\; x > 0$$

where $\psi = \alpha$ and $\phi = \frac{\alpha}{\mu^2}$.

The Laplace transform of inverse Gaussian distribution is,

$$L_X(s) = exp\left[\frac{\alpha}{\mu}\left(1 - \sqrt{1 + \frac{2\mu^2 s}{\alpha}}\right)\right];\;\; s \geq 0 \tag{13}$$

Differentiating the Laplace transform with respect to s and evaluating at $s = 0$, the first- and second-order moments are

$$E\,[X] = \mu \;\text{ and }\; \text{var}\,[\text{X}] = \frac{\mu^3}{\alpha}.$$

Frailty models suffer from the identifiability problem. So, the expected value of the frailty random variable is always restricted to one. For inverse Gaussian frailty models, the restriction on parameters is $\mu = 1$, under this restriction, the variance of X is, $\sigma^2 = \dfrac{1}{\alpha}$. The

Laplace transform and probability density function of X then reduce to

$$\begin{aligned} f(x) &= \begin{cases} \sqrt{\frac{1}{2\pi\sigma^2 x^3}} exp\left[-\frac{(x-1)^2}{2x\sigma^2}\right] & ;\ x>0, \sigma^2>0 \\ 0 & ;\ otherwise. \end{cases} \\ L_Z(s) &= exp\left[\frac{1}{\sigma^2}\left(1-\sqrt{1+2\sigma^2 s}\right)\right];\ \ s\geq 0 \end{aligned} \tag{14}$$

3.2. Additive and multiplicative IG frailty models

Suppose the multiplicative and additive frailty random variables $Z_i \sim IG(\mu_i, \alpha_i)$, $i = a, m$. Under the restriction for identifiability $\mu_a = 1$ and $\mu_m = 1$ the Laplace transform of Z_a and Z_m is as given in Equation (14). Substituting this Laplace transform (Equation (14)) in Equations (3) and (7), we get the survival function of the lifetime random variable T under the multiplicative inverse Gaussian frailty (MIGF) and additive inverse Gaussian frailty (AIGF) models as

$$S_m(t \mid \underline{x}_m) = \begin{cases} exp\left[\frac{1}{\sigma_m^2}\left(1-\sqrt{1+2\sigma_m^2 H_0(t)\eta_m}\right)\right] & ;\ t>0 \\ 1 & ;\ \text{otherwise} \end{cases} \tag{15}$$

$$S_a(t \mid \underline{x}_a) = \begin{cases} exp\left(-H_0(t)\right) exp\left[\frac{1}{\sigma_a^2}\left(1-\sqrt{1+2\sigma_a^2 t\eta_a}\right)\right] & ;\ t>0 \\ 1 & ;\ \text{otherwise} \end{cases} \tag{16}$$

The respective probability density functions are given by

$$f_m(t \mid \underline{x}_m) = \begin{cases} \frac{h_0(t) S_m(t \mid \underline{x}_m)\eta_m}{\sqrt{1+2\sigma_m^2 H_0(t)\eta_m}} & ;\ t>0 \\ 0 & ;\ \text{otherwise} \end{cases}$$

$$f_a(t \mid \underline{x}_a) = \begin{cases} S_a(t \mid \underline{x}_a)\left\{h_0(t) + \frac{\eta_a}{\sqrt{1+2\sigma_a^2 t\eta_a}}\right\} & ;\ t>0 \\ 0 & ;\ \text{otherwise} \end{cases}$$

3.3. Additive-Multiplicative IG frailty models

First, to obtain the additive-multiplicative inverse Gaussian frailty (AMIGF) model, assume that frailty random variables Z_a and Z_m are independent but not identical. Suppose $Z_i \sim IG(\mu_i, \alpha_i)$, $i = a, m$. Again, under the restriction for identifiability, we assume $\mu_a = \mu_m = 1$ and Laplace transform of Z_m and Z_a is given by Equation (14). Using Equation (11) and the fact that if X and Y are independent random variables, then $L_{(X,Y)}(s_1, s_2) = L_X(s_1) \cdot L_Y(s_2)$, the survival function of lifetime random variable is given by

$$S(t \mid \underline{x}) = \begin{cases} L_{Z_a}(t\eta_a) L_{Z_m}(H_0(t)\eta_m) & ;\ t>0 \\ 1 & ;\ \text{otherwise} \end{cases} \tag{17}$$

Substituting the Laplace transform of IG distribution (14) the survival function (17) becomes

$$S(t \mid \underline{x}) = \begin{cases} exp\left[\frac{1}{\sigma_a^2}\left(1 - D_a(t \mid \underline{x}_a)\right)\right] \cdot exp\left[\frac{1}{\sigma_m^2}\left(1 - D_m(t \mid \underline{x}_m)\right)\right] & ; \quad t > 0 \\ 1 & ; \quad \text{otherwise} \end{cases} \tag{18}$$

and the probability density function is given by

$$\psi(t \mid \underline{x}) = \begin{cases} S(t)\left\{\frac{\eta_a}{D_a(t \mid \underline{x}_a)} + \frac{h_0(t)\eta_m}{D_m(t \mid \underline{x}_m)}\right\} & ; \quad t > 0 \\ 0 & ; \quad \text{otherwise} \end{cases}$$

where $D_a(t \mid \underline{x}_a) = \sqrt{1 + 2\sigma_a^2 t\eta_a}$, $D_m(t\underline{x}_m) = \sqrt{1 + 2\sigma_m^2 H_0(t)\eta_m}$.

Now, we consider the case where Z_a and Z_m are not independent. To obtain the survival function of the lifetime random variable under the AMIGF model when Z_a and Z_m are not independent, the Laplace transform of frailty random vector $\underline{Z}$ needs to be obtained. In this paper, it is obtained using the concept of trivariate reduction (*see* Kocherlakota and Kocherlakota (1992)). The idea is to create a pair of dependent random variables from three or more possibly independent initial random variables. Let U_i, $i = 0, 1, 2$ be independently distributed continuous random variables, then the trivariate reduction method consists of defining a pair of correlated random variables V_1 and V_2 by relation,

$$V_1 = \tau(U_0, U_1); \qquad V_2 = \tau(U_0, U_2)$$

Define the function $\tau(x, y) = c\,(dx + y)$ so that the variable $V_j = c_j(d_j U_0 + U_j), j = 1, 2$. If $L_{U_j}(\cdot)$ is the Laplace transform of $U_j, j = 0, 1, 2$ then the Laplace transform of V_j and $\underline{V} = (V_1, V_2)$ are easily shown to be respectively.

$$L_{V_j}(s) = L_{U_0}(c_j d_j s) L_{U_j}(c_j s)\ ; \qquad j = 1, 2 \tag{19}$$

$$L_{\underline{V}}(\underline{s}) = L_{U_0}(c_1 d_1 s_1 + c_2 d_2 s_2) L_{U_1}(c_1 s_1) L_{U_2}(c_2 s_2) \tag{20}$$

Now, to obtain the Laplace transform of the frailty vector $\underline{Z}$ using the trivariate reduction method, let Y_0, Y_a, Y_m be independent positive valued random variables following $IG(\mu_i, \alpha_i), i = 0, a, m$ having Laplace transform as given in Equation (13). Define a random variable $Z_j = \frac{\alpha_j}{\alpha_0 + \alpha_j}\left(\frac{\alpha_0}{\alpha_j} Y_0 + Y_j\right)$, $j = a, m$. Putting $c_j = \frac{\alpha_j}{\alpha_0 + \alpha_j}; d_j = \frac{\alpha_0}{\alpha_j}$ and using Equation (19) the Laplace transform of the frailty random variable Z_j, $j = a, m$ is given by

$$L_{Z_j}(s) = exp\left[\frac{\alpha_0 + \alpha_j}{\mu}\left(1 - \sqrt{1 + \frac{2\mu^2 s}{\alpha_0 + \alpha_j}}\right)\right]; \quad s \geq 0 \tag{21}$$

Hence, each Z_j marginally follows $IG(\mu, \alpha_0 + \alpha_j)$; $j = a, m$. The moments of Z_j's are given by

$$E\left[Z_j\right] = \mu; \quad var\left[Z_j\right] = \frac{\mu^3}{\alpha_0 + \alpha_j}; \quad j = a, m.$$

Theorem 1: The correlation coefficient between Z_a and Z_m is

$$\rho = \frac{\alpha_0}{\sqrt{\alpha_0 + \alpha_a}\sqrt{\alpha_0 + \alpha_m}}$$

Proof: Given in Appendix I

The Laplace transform of $\underline{Z}$ for $s_1 \geq 0, s_2 \geq 0$ using Equation (20), is given by,

$$\begin{aligned} L_{\underline{Z}}(s_1, s_2) &= exp\left[\frac{\alpha_0}{\mu}\left(1 - \sqrt{1 + 2\mu^2\left(\frac{s_1}{\alpha_0 + \alpha_a} + \frac{s_2}{\alpha_0 + \alpha_m}\right)}\right)\right] \cdot \\ & \quad exp\left[\frac{\alpha_a}{\mu}\left(1 - \sqrt{1 + \frac{2\mu^2 s_1}{\alpha_0 + \alpha_a}}\right)\right] \cdot exp\left[\frac{\alpha_m}{\mu}\left(1 - \sqrt{1 + \frac{2\mu^2 s_2}{\alpha_0 + \alpha_m}}\right)\right] \end{aligned} \tag{22}$$

Restricting Z_j to have the mean one, for the identifiability problem, the restriction on the parameter is $\mu = 1$, so that the Laplace transform in Equation (22) reduces to

$$\begin{aligned} L_{\underline{Z}}(s_1, s_2) &= exp\left[\alpha_0\left(1 - \sqrt{1 + 2\left(\frac{s_1}{\alpha_0 + \alpha_a} + \frac{s_2}{\alpha_0 + \alpha_m}\right)}\right)\right] \cdot exp\left[\alpha_a\left(1 - \sqrt{1 + \frac{2s_1}{\alpha_0 + \alpha_a}}\right)\right] \\ & \quad \cdot exp\left[\alpha_m\left(1 - \sqrt{1 + \frac{2s_2}{\alpha_0 + \alpha_m}}\right)\right]; s_1 \geq 0, s_2 \geq 0 \end{aligned}$$

and from Equation (11), the survival function of the lifetime random variable T is

$$S(t \mid \underline{x}) = \begin{cases} exp\left[\alpha_0\left(1 - \sqrt{G(t \mid \underline{x})}\right)\right] \cdot exp\left[\alpha_a\left(1 - \sqrt{G_a(t \mid \underline{x}_a)}\right)\right] & \\ \qquad exp\left[\alpha_m\left(1 - \sqrt{G_m(t \mid \underline{x}_m)}\right)\right] & ; \quad t > 0 \\ 1 & ; \quad \text{otherwise} \end{cases} \tag{23}$$

where $G(t \mid \underline{x}) = 1 + 2A(t \mid \underline{x}_a) + 2B(t \mid \underline{x}_m)$; $G_a(t \mid \underline{x}_a) = 1 + 2A(t \mid \underline{x}_a)$; $G_m(t \mid \underline{x}_m) = 1 + 2B(t \mid \underline{x}_m)$; $A(t \mid \underline{x}_a) = \frac{t\eta_a}{\alpha_0 + \alpha_a}$, $B(t \mid \underline{x}_m) = \frac{H_0(t)\eta_m}{\alpha_0 + \alpha_m}$ and the probability density function is

$$\psi(t \mid \underline{x}) = \begin{cases} S(t \mid \underline{x})\left\{\frac{\alpha_a}{\alpha_0 + \alpha_a}E_1(t \mid \underline{x}_a) + \frac{\alpha_m}{\alpha_0 + \alpha_m}E_2(t \mid \underline{x}_m) + \alpha_0 E(t \mid \underline{x})\right\} & ; \quad t > 0 \\ 0 & ; \quad \text{otherwise} \end{cases}$$

where $E_1(t \mid \underline{x}_a) = \frac{\eta_a}{\sqrt{1 + 2A(t \mid \underline{x}_a)}}$, $E_2(t \mid \underline{x}_m) = \frac{h_0(t)\eta_m}{\sqrt{1 + 2B(t \mid \underline{x}_m)}}$ and $E(t \mid \underline{x}) = \frac{\left(A^{(1)}(t \mid \underline{x}_a) + B^{(1)}(t \mid \underline{x}_m)\right)}{\sqrt{1 + 2A(t \mid \underline{x}_a) + 2B(t \mid \underline{x}_m)}}$

$A^{(1)}(t \mid \underline{x}_a)$ and $B^{(1)}(t \mid \underline{x}_m)$ are first-order derivatives of $A(t \mid \underline{x}_a)$ and $B(t \mid \underline{x}_m)$ respectively with respect to t.

Furthermore, expressing the parameters α_0, α_a and α_m in terms of σ_a^2, σ_m^2 and ρ we have

$$\alpha_0 = \frac{\rho}{\sigma_a \sigma_m}; \qquad \alpha_a = \frac{1}{\sigma_a^2}\left[1 - \frac{\sigma_a}{\sigma_m}\rho\right]; \qquad \alpha_m = \frac{1}{\sigma_m^2}\left[1 - \frac{\sigma_m}{\sigma_a}\rho\right].$$

Defining $\kappa_0 = \frac{\rho}{\sigma_a \sigma_m}$; $\kappa_a = \left[1 - \frac{\sigma_a}{\sigma_m}\rho\right]$ and $\kappa_m = \left[1 - \frac{\sigma_m}{\sigma_a}\rho\right]$ and rewriting Equation (23) in terms of $\sigma_a^2, \sigma_m^2, \kappa_0, \kappa_a$ and κ_m the survival function of the lifetime random variable T becomes

$$S(t \mid \underline{x}) = \begin{cases} exp\left[\kappa_0 \left(1 - F(t \mid \underline{x})\right)\right] \cdot exp\left[\frac{\kappa_a}{\sigma_a^2}\left(1 - F_a(t \mid \underline{x}_a)\right)\right] & \\ \quad exp\left[\frac{\kappa_m}{\sigma_m^2}\left(1 - F_m(t \mid \underline{x}_m)\right)\right] & ; \quad t > 0 \\ 1 & ; \quad \text{otherwise} \end{cases}$$

where $F(t \mid \underline{x}) = \sqrt{1 + 2\sigma_a^2 \eta_a t + 2\sigma_m^2 \eta_m H_0(t)}$, $F_a(t \mid \underline{x}_a) = \sqrt{1 + 2\sigma_a^2 \eta_a t}$ and $F_m(t \mid \underline{x}_m) = \sqrt{1 + 2\sigma_m^2 \eta_m H_0(t)}$. From Equations (15) and (16) we can write

$$\sqrt{1 + 2\sigma_a^2 \eta_a t} = 1 - \sigma_a^2 \ln\left[S_a(t \mid \underline{x}_a) e^{H_0(t)}\right] \text{ and } \sqrt{1 + 2\sigma_{\mathrm{m}}^2 \eta_{\mathrm{m}} \mathrm{t}} = 1 - \sigma_{\mathrm{m}}^2 \ln \mathrm{S_m(t \mid \underline{x}_m)}$$

Hence, the survival function can be expressed as

$$S(t \mid \underline{x}) = \begin{cases} exp\left\{\kappa_0 \left[1 - \sqrt{D(t \mid \underline{x}) - 1}\right]\right\} \cdot \left[S_a(t \mid \underline{x}_a) e^{H_0(t)}\right]^{\kappa_a} \cdot S_m(t \mid \underline{x}_m)^{\kappa_m} & ; \quad t > 0 \\ 1 & ; \quad \text{otherwise} \end{cases} \tag{24}$$

where $D(t \mid \underline{x}) = \left[1 - \sigma_a^2 \ln\left(S_a(t \mid \underline{x}_a) e^{H_0(t)}\right)\right]^2 + \left[1 - \sigma_m^2 \ln S_m(t \mid \underline{x}_m)\right]^2$.

If we compare the survival functions in Equations (24) and (18) then the first factor and powers to $S_a(t \mid \underline{x}_a), S_m(t \mid \underline{x}_m)$ in Equation (24) are due to the non-independence between Z_a and Z_m. When $\rho = 0$, $\kappa_a = 1 = \kappa_m$ and $\kappa_0 = 0$, Equation (24) reduces to Equation (18).

Thus, the family of AMIGF models is $\left\{S(t, \underline{\theta}),\ t > 0;\ \underline{\theta} = \left(\underline{\tau}, \underline{\omega}, \underline{\beta}\right)\right\}$, where $S(t, \underline{\theta})$ is the survival function given by Equation (24) or (23) and $\underline{\tau}, \underline{\omega}$ and $\underline{\beta}$ are respectively vectors of the baseline parameters, frailty parameters and regression parameters. By considering different baseline distributions, one can develop different AMIGF models.

4. Baseline specification

To complete the form of the survival function this paper considers the baseline hazard function to be the hazard function of the generalized exponential distribution. Gupta and Kundu (1999) suggested that the generalized exponential distribution can be used effectively in analyzing many lifetime data sets, particularly as an alternative to the gamma and Weibull distributions. A continuous random variable X is said to follow the generalized exponential distribution if its survival function is,

$$S_0(x) = \begin{cases} 1 - (1 - e^{-x/\theta_2})^{\theta_1} & ; \ x > 0, \theta_1 > 0, \theta_2 > 0 \\ 1 & ; \ \text{otherwise.} \end{cases}$$

where θ_1 and θ_2 are, respectively, the shape and scale parameters of the distribution. The hazard function and the cumulative hazard function are, respectively,

$$h_0(x) = \begin{cases} \frac{\theta_1 e^{-x/\theta_2}(1-e^{-x/\theta_2})^{\theta_1-1}}{\theta_2\left[1-(1-e^{-x/\theta_2})^{\theta_1}\right]} & ; \ x>0, \theta_1>0, \theta_2>0 \\ 0 & ; \text{ otherwise.} \end{cases}$$

$$H_0(x) = \begin{cases} -ln[1-(1-e^{-x/\theta_2})^{\theta_1}] & ; \ x>0, \theta_1>0, \theta_2>0 \\ 0 & ; \text{ otherwise.} \end{cases}$$

When $\theta_1 = 1$ distribution reduces to the one-parameter exponential distribution and has constant failure rate $\frac{1}{\theta_2}$. When $\theta_1 > 1$, the hazard function is an increasing function of time, and for $\theta_1 < 1$, the hazard function is a decreasing function of time.

5. Bayesian inferential procedure

Suppose n individuals are observed with lifetimes $\underline{t} = (t_1, t_2, \cdots, t_n)$; censoring times $c_1, c_2, \cdots, c_n$ and censoring indicators $\underline{\delta} = (\delta_1, \delta_2, \cdots, \delta_n)$ with

$$\delta_j = \begin{cases} 1 & ; \text{ if j}^{\text{th}} \text{ individual is observed} \\ 0 & ; \text{ if j}^{\text{th}} \text{ individual is censored} \end{cases}$$

Also, suppose that x_{auj} and x_{mvj}, are the observed covariate values of u^{th} additive covariate and v^{th} multiplicative covariate, $u = 1, 2, \cdots, k_a$; $v = 1, 2, \cdots, k_m$; $j = 1, 2, \cdots, n$. Assuming the independence between censoring times and lifetimes, the likelihood function is

$$L\left(\underline{\tau}, \underline{\omega}, \underline{\beta} \mid (\underline{t}, \underline{\delta})\right) = \prod_{j=1}^{n} \left\{\psi(t_j)^{\delta_j} S(t_j)^{1-\delta_j}\right\}$$

The commonly used estimation method is the maximum likelihood method, which involves solving simultaneous likelihood equations, namely the first-order partial derivatives of the log-likelihood function with respect to the parameters. The likelihood equations for the proposed models could not provide an analytical solution. As a result, an iterative procedure such as Newton-Raphson has to be used to solve likelihood equations. Unfortunately, in frailty models, the maximum likelihood method fails to converge to the true parameters due to a large number of parameters and heavy censoring. (*see* Kheiri *et al.* (2005), Hanagal (2021)). Hence, a computational Bayesian approach is adopted to estimate the model parameters in this paper. The joint posterior density function of the parameters $\left(\underline{\tau}, \underline{\omega}, \underline{\beta}\right)$ for given data $\underline{t}, \underline{\delta}$ is given by

$$\pi(\underline{t}, \underline{\delta} | \underline{\tau}, \underline{\omega}, \underline{\beta}) \ \propto \ L(\underline{\tau}, \underline{\omega}, \underline{\beta} | \underline{t}, \underline{\delta}) \times \prod_{i=1}^{n_b} g_i(a_i) \times \prod_{i=1}^{n_f} h_i(b_i) \times \prod_{i=1}^{k} p_i(\beta_i)$$

where $g_i(\cdot), h_i(\cdot), p_i(\cdot)$ are the prior density functions of the baseline, frailty, and regression parameters with known hyperparameters. The number of baseline, frailty, and regression parameters are represented by n_b, n_f and $k = k_a + k_m$, respectively. Here, independence is assumed between all model parameters. The posterior density function of a specific parameter can be obtained by integrating over other parameters.

When no information about the parameters is available, one can choose an empirical approach to determine prior distributions or use non-informative distributions as priors. In this paper, the latter approach is used. A widely used prior distribution for frailty parameters is $Gamma(shape = \phi, scale = \phi)$ with a small choice of ϕ so that the distribution will have mean one and a large variance. For the regression coefficients, $N(0, \sigma^2)$ with a large value of σ^2 is a popular choice. This paper examines two sets of prior distributions to investigate the impact of prior distributions on estimators. The first set is

$$\begin{aligned} \theta_1, \theta_2 &\sim Gamma(shape = 1, scale = 0.0001) \\ \alpha_0, \alpha_1, \alpha_2 &\sim Gamma(shape = 1, scale = 0.0001) \\ \underline{\beta} &\sim Normal(mean = 0, sd = 1000) \end{aligned}$$

and second set based on $U(a, b)$ is

$$\begin{aligned} \theta_1, \theta_2 &\sim U(0, 100) \\ \alpha_0, \alpha_1, \alpha_2 &\sim U(0, 100) \\ \underline{\beta} &\sim U(-50, 50) \end{aligned}$$

The probability density functions of the prior distributions in both sets are flat, providing very little information about the parameters.

A computational Bayesian approach was used to generate two Markov chains, each comprising $N = 50000$ iterations, using the Gibbs sampler and the Metropolis-Hastings algorithm with two sets of prior distributions. A normal transition kernel was considered to generate the chains. The burn-in period (B) was determined using coupling from the past plots, and the convergence of chains to a stationary distribution was monitored using the Gelman-Rubin convergence statistic and the Geweke test. The plots of the sample autocorrelation function were used to determine the autocorrelation lag (k). Once the values of B and k were decided, a pseudorandom sample of size n (where n is less than or equal to $(N - B)/k$) was obtained, and the model parameters were estimated using the sample posterior mean. The posterior variances of the estimators, along with credible intervals, were also obtained.

A Deviance Information Criteria (DIC) was used to compare the fitted models. As Spiegelhalter *et al.* (2002) noted, DIC is a good measure for comparing the Bayesian models, where analysis is carried out via MCMC methods to assess the posterior distribution. Instead, the advantage of the DIC is that it can be easily calculated from a posterior summary, unlike the other information criterion. Log-likelihood and Bayes' factor (*see* Kass and Raftery (1995)) were also used to compare the models.

6. Simulation study

A simulation study was conducted to evaluate the performance of the estimation procedure. In this comprehensive study, data were generated from the AMIGF model with two covariates: one having an additive effect (x_a) and the other having a multiplicative effect (x_m) on the hazard function. Both these variables were generated from a normal distribution with different parameters. The inverse transform technique was used to generate lifetimes, which involves equating the survival function for given frailties z_a, z_m and covariates x_a, x_m

to a random number R. By equating the Equation (10) to R, we get

$$z_a\eta_a t + z_m\eta_m H_0(t) = -log(R) \tag{25}$$

Solving the Equation (25), one can generate lifetimes, but unfortunately, the equation cannot be expressed as an explicit function of R, so it must be generated using the Newton-Raphson iterative procedure. Table 1 provides an algorithm for generating data. The frailties were

Table 1: Simulation algorithm

step 1	Generate x_a, x_m from $N(\mu_a, \sigma_a^2)$ and $N(\mu_m, \sigma_m^2)$ respectively.
step 2	Generate frailties z_a, z_m from $IG(\mu = 1, \alpha_0 + \alpha_a)$ and $IG(\mu = 1, \alpha_0 + \alpha_m)$ respectively.
step 3	Generate lifetime t by solving the Equation (25) using generated values z_a, z_m, x_a, x_m.

generated from IG using the R package 'statmod'. The true values of the parameters were $\theta_1 = 1.0; \theta_2 = 42, \alpha_0 = \alpha_a = \alpha_m = 2.5, \beta_1 = 3.0 = \beta_2$.

The generated data was used to fit three models: AIGF, MIGF, and AMIGF. The AIGF model considered both covariates to have an additive effect, while MIGF considered a multiplicative effect. For the AMIGF model, the actual model (x_a additive and x_m multiplicative) was first fitted (AMIGF I). The second model (AMIGF II) was then fitted by reversing the positions of the covariates (x_a multiplicative and x_m additive) to test the model's sensitivity to the position of the covariates.

The trace plots depicting the parameters of the AMIGF I model are shown in Figure 1. The plots do not show a trend or pattern, suggesting that the parameter values generated are randomly distributed throughout the parameter space. The same observations were made for the other model parameters.

The Gelman-Rubin convergence statistic values were quite close to one for all the models' parameters, and the p-values of the Geweke tests were significant enough to indicate that the chains have attained a stationary distribution. A similar pattern was observed regarding the posterior summary in both chains and with both prior sets. Hence, results are presented here for only one chain and with only one prior set of distributions. The posterior summaries for the AIGF and MIGF models are presented in Table 2, and Table 3 represents the posterior summaries for the AMIGF models. Figures 2 to 4 represent the posterior distribution of the parameters of the AIGF, MIGF, AMIGF I and AMIGF II models. The Table 4 shows the DIC values and log-likelihood for all fitted models. Twice the logarithm of Bayes' factor with AMIGF I as the numerator model and the other three models AIGF, MIGF, and AMIGF II as denominator are, respectively, $16.5830, 22.9156$ and 1.4136. All these measures indicate that the fitted model (AMIGF I) estimates the parameters unbiasedly when the fitted model is the actual model. This performance of the estimation strategy underscores its reliability and ability to fit the data accurately.

7. Data analysis

The AMIGF model presented in this research is demonstrated using two datasets: Kidney Infection (KI) data from McGilchrist and Aisbett (1999) and Acute Myelogenous

Leukemia (AML) data as reported in Miller (1997).

7.1. Kidney infection data

The data pertain to the time it took for an infection to recur after catheter insertion in 38 kidney patients who used portable dialysis equipment. For each patient, the first and second recurrence times (in days) of infection from the time of catheter insertion until it had to be removed due to infection were recorded. It's essential to note that the catheter may have been removed for reasons unrelated to the kidney infection, which should be considered as censoring. The actual data includes the first and second infection times or censoring times for a patient represented by T_1, T_2 respectively, along with censoring indicators and five covariates: age, gender (Male(0), Female(1)), and presence (1) or absence (0) of disease type GN, AN, PKD represented by x_1, x_2, x_3, x_4, and x_5 respectively, where GN, AN and PKD are short forms of Glomerulo Neptiritis, Acute Neptiritis and Polycyatic Kidney Disease. Since this paper addresses a univariate case, only the first infection time is considered for the analysis.

All three models were fitted to the first infection time using a self-written program in R. When fitting the AIGF model; all covariates were considered to have an additive effect, whereas when fitting the MIGF model, they were all considered to have a multiplicative effect. The AMIGF models were attempted by considering different combinations of covariates as either additive or multiplicative to determine whether the effects of covariates were additive or multiplicative. In total, there were 30 such combinations.

The two chains, generated using the Bayesian inference procedure, demonstrated similar results. Additionally, both sets of prior distributions yielded matching results. Therefore, only the results from one chain, using the first set of prior distributions, are discussed here. The Gelman-Rubin convergence statistic values were close to one, indicating convergence and the Geweke test values were relatively small, with corresponding p-values that were significant enough to suggest the chain had reached a stationary distribution.

Table 5 contains the posterior summary for the AIGF and MIGF models. Figures 5 and 6 display the histogram of the posterior distribution for the AIGF and MIGF models, respectively. Table 6 provides the DIC and log-likelihood values for all the fitted 30 AMIGF models, including AIGF and MIGF models. In the case of AMIGF models, the covariates listed in the first column of Table 6 indicate additive covariates, while the remaining covariates were considered multiplicative. Table 6 shows that the MIGF model has a lower DIC than the AIGF model. Although the log-likelihood values are not significantly different, the twice the log of Bayes' factor with MIGF as the numerator model is 17.0077, indicating that the MIGF is a better model. Additionally, it's worth noting that some AMIGF models have less DIC and larger log-likelihood values than the MIGF model. The model with x_2 as an additive covariate stands out among these similar models. Therefore, including x_2 into the model additively may be a better choice than multiplicatively. Furthermore, AMIGF models with x_2 and x_5 as additive covariates and x_1, x_3, and x_4 as multiplicative covariates (AMIGF-25) have a lower DIC value than all other models, including AIGF and MIGF. The log-likelihood value is also slightly higher than all other models, indicating that the AMIGF-25 model performs better than the MIGF model. Moreover, twice the log of Bayes' factor with AMIGF-25 as the numerator model and MIGF as the denominator model is 4.6016,

suggesting that the AMIGF-25 model is the best model for modeling the first recurrence time.

Figure 7 shows the trace plots for the parameters of the best model AMIGF-25. The plot shows the random behavior of generated values. The posterior summary for model AMIGF-25 is provided in Table 7. Figure 8 displays the posterior distribution.

The estimate to standard error ratios for β_2 and β_5 are significantly different from one. Specifically, these ratios are -5.4782 and -1.4486 respectively. This suggests that gender (x_2) and PKD disease type (x_5) have a significant impact on kidney infection development. The negative values of both regression coefficients indicate that they have a reverse effect on infection. Therefore, for female patients, the frequency of infection is lower than for male patients who suffer from PKD.

7.2. Acute myelogenous leukemia data

The study aimed to assess the efficacy of maintenance chemotherapy for AML patients. Once the patients achieved remission through chemotherapy, those who continued the study were randomly divided into two groups. One group received maintenance chemotherapy, while the other did not receive any further treatment. The data include the survival or censoring time (t), censoring indicators (δ), and covariate x, which indicates whether the patient received maintenance chemotherapy (1) or not (0). There were 23 patients in total, with 5 of them censored.

All three models (AIGF, MIGF, and AMIGF) were used to analyze the data. Two AMIGF models were fitted to the data to determine whether the covariate effect is additive or multiplicative. The first model (AMIGF I) considered the covariate to have an additive effect, while the second model (AMIGF II) considered it to have a multiplicative effect. Similar results were observed from both chains and both sets of prior distributions, so the discussion focuses on the results from one chain with the first set of prior distributions.

The Gelman-Rubin convergence statistic values were close to one, the Geweke test values were relatively small, and the corresponding p-values were significant enough to confirm that the chain had attained a stationary distribution. The Figure 9 displays trace plots for the AMIGF I model. A similar random pattern was observed in the trace plots of other models. The posterior summaries for the AIGF and MIGF models are presented in Table 8, and the posterior summaries for AMIGF I and AMIGF II models are presented in Table 9. The posterior distribution for the AIGF, MIGF, AMIGF I, and AMIGF II models are shown in Figures 10 and 11. Table 10 provides DIC and log-likelihood values for all fitted models. The results indicate that the MIGF model fits better than the AIGF model. The twice log of Bayes' factor with MIGF as the numerator model is 5.7282, which confirms the conclusion. While comparing AMIGF models, both models exhibited similar behavior, suggesting no difference in whether the covariate is included in the model as having an additive or multiplicative effect. Furthermore, these models showed slightly larger log-likelihood values and smaller DIC values than the MIGF models. The twice the log of the Bayes' factor with the AMIGF I and AMIGF II models as the numerator and MIGF as the denominator is 8.4439 and 9.1167. These findings collectively indicate that the AMIGF models outperform the AIGF and MIGF models.

The absolute ratio of the estimate to the standard error of β in both models, AMIGF I and AMIGF II, is slightly larger than one. The credible interval contains zero, indicating that the treatment has a significant effect on remission time. However, the negative value of β indicates that maintained chemotherapy does not help to increase the remission period. Further research on testing the significance of β can provide a better understanding of this.

8. Conclusion

The additive and multiplicative hazard models are commonly used regression models to analyze the impact of covariates on failure time. The additive hazard model examines the relationship between covariates and the hazard function in terms of risk difference, while the multiplicative model focuses on the risk ratio. Although these models are straightforward and practical, researchers have identified a need for a compromise model incorporating covariates with both additive and multiplicative effects, as proposed by Lin and Ying (1997). Furthermore, to account for heterogeneity within the population, it is necessary to include frailty variables in the model. These frailty variables may affect the hazard function additively (additive frailty model) or multiplicatively (multiplicative frailty model). The additive and multiplicative frailty models are well-studied in the literature. Since frailty variables are unobservable, one can not decide whether frailty is acting additively or multiplicatively. Some of the unknown factors may affect the hazard additively, while others may affect it multiplicatively. Therefore, it is essential to explore additive-multiplicative frailty models.

This study aims to present a new family of models that incorporates both additive and multiplicative frailty variables. This model can capture the additive and multiplicative relationships between the frailty random variable and the hazard function. In this article, to complete the parametric form, the inverse Gaussian distribution is considered as the frailty distribution, and the generalized exponential distribution, introduced by Gupta and Kundu (1999), is used as the baseline distribution. One can define various AMIGF models by considering more advanced baseline distributions.

The proposed AMIGF model was applied to real-world data, the first infection time of kidney infection data, and acute myelogenous leukemia data. It was then compared with the traditional MIGF and AIGF models using a self-written program in R. Based on the DIC criteria, log-likelihood, and Bayes' factor, the best model that emerged was the AMIGF model. This practical application emphasizes the importance of the proposed model and also highlights the need for frailty to have additive and multiplicative effects.

Further research can involve exploring and comparing different frailty distributions, such as the gamma, power variance, compound Poisson, and compound negative binomial distributions. The AMIGF model can also be expanded to include shared and correlated frailty models.

Acknowledgements

I am indeed grateful to the Editors for their guidance and counsel. I am very grateful to the reviewer for valuable comments and suggestions for improvement.

Conflict of interest

The author has no financial or non-financial conflict of interest to declare for the research work included in this article.

References

Aalen, O. O. (1980). *A Model for Non-parametric Regression Analysis of Counting Processes*. Lecture Notes in Statistics, **2**, Springer. 1–25.

Aalen, O. O. and Tretli, S. (1999). Analyzing incidence of testis cancer by means of a frailty model. *Cancer Causes and Control*, **10**, 285–292.

Aranda-Ordaz, F. J. (1983). An extension of the proportional-hazards model for grouped data. *Biometrics*, **39**, 109–117.

Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research Volume - **II** The Design and Analysis of Cohort Studies*. IARC Science Publications.

Chhikara, R. S. and Folks, J. L. (1986). *The Inverse Gaussian Distribution*. Marcel Dekker, New York.

Duchateu, L. and Janseen, P. (2008). *The Frailty Model*. New York: Springer.

Gupta, R. D. and Kundu, D. (1999). Generalized exponential distributions. *Australian and New Zealand Journal of Statistics*, **41**, 173–188.

Hanagal, D. D. (2019). *Modeling Survival Data Using Frailty Models*. CRC Press.

Hanagal, D. D. (2021). Correlated positive stable frailty models. *Communications in Statistics - Theory and Methods*, **50**, 5617–5633.

Hanagal, D. D. (2022). Compound Poisson shared frailty models based on additive hazards. *Communications in Statistics - Theory and Methods* **52**, 6287–6309.

Hanagal, D. D. and Pandey, A. (2016). Shared gamma frailty models based on additive hazards. *Journal of Indian Society for Probability and Statistics* **17**, 161–184.

Hanagal, D. D. and Pandey, A. (2017). Shared inverse Gaussian frailty models based on additive hazards. *Communications in Statistics - Theory and Methods*, **46**, 11143–11162.

Hougaard, P.(1984). Life table methods for heterogeneous populations. *Biometrika*, **71**, 75–83.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, **90**, 773–795.

Kheiri, S., Meshkani, M. R., and Faghihzadeh, S. (2005). A correlated frailty model for analysing risk factors in bilateral corneal grafts rejection for Keratoconus: A Bayesian approach. *Statistics in Medicine*, **24**, 2681–2693.

Kheiri, S., Kimber, A., and Meshkani M. R. (2007). Bayesian analysis of an inverse Gaussian correlated frailty model. *Computational Statistics and Data analysis*, **51**, 5317–5326.

Kocherlakota, S. and Kocherlakota, K. (1992). *Bivariate Discrete Distributions*. Marcel Dekker.

Lin, D. Y. and Ying, Z. (1995). Semiparametric analysis of general additive-multiplicative hazard models for counting processes. *The Annals of Statistics*, **23**, 1712–1734.

Lin, D. Y. and Ying, Z. (1997). *Additive Hazards Regression Models for Survival Data.* Lecture Notes in Statistics. Proceedings of the First Seattle Symposium in Biostatistics, **123**, 185–198.

Miller, R. G. (1997). *Survival Analysis.* John Wiley & Sons.

McGilchrist, C. A. and Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics*, **47**, 461–466.

Silva, G. L. and Amaral-Turkman, M. A. (2004). Bayesian analysis of an additive survival model with frailty. *Communications in Statistics - Theory and Methods*, **33**, 2517–2533.

Spiegelhalter, D. J, Best, N. G. , Carlin, B. P., and Angelika van der Linde. (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society series B*, **64**, 583–639.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**, 439–454.

Wienke, A. (2011): *Frailty Models in Survival Analysis.* CRC Press.

Appendix I
Proof of Theorem 1

Proof: As $Z_a = c_a(d_aY_0 + Y_a)$ and $Z_m = c_m(d_mY_0 + Y_m)$ we have,

$$E\left[Z_a\right] = c_a(d_aE\left[Y_0\right] + E\left[Y_a\right]) \quad \text{and } \mathrm{E}\left[\mathrm{Z_m}\right] = \mathrm{c_m}(\mathrm{d_m}\mathrm{E}\left[\mathrm{Y_0}\right] + \mathrm{E}\left[\mathrm{Y_m}\right])$$

Consider

$$\begin{aligned} cov\left(Z_a, Z_m\right) &= E\left[\left(Z_a - E(Z_a)\right)\left(Z_m - E(Z_m)\right)\right] \\ &= c_ac_md_ad_mvar(Y_0) + c_ac_md_mcov(Y_0, Y_a) + c_ac_md_acov(Y_0, Y_m) + c_ac_mcov(Y_a, Y_m) \\ &= c_ac_md_ad_mvar(Y_0) \qquad \text{since} \mathrm{Y_0}, \mathrm{Y_a} \text{ and } \mathrm{Y_m} \text{ are independent} \end{aligned}$$

substituting c_j, d_j, $j = a, m$ and $var(Y_0)$ we have

$$cov\left(Z_a, Z_m\right) = \frac{\mu^3\alpha_0}{\left(\alpha_0 + \alpha_a\right)\left(\alpha_0 + \alpha_m\right)}$$

Therefore

$$\rho = \frac{cov\left(Z_a, Z_m\right)}{\sqrt{var\left(Z_a\right)}\sqrt{var\left(Z_m\right)}} = \frac{\alpha_0}{\sqrt{\alpha_0 + \alpha_a}\sqrt{\alpha_0 + \alpha_m}}$$

□

Appendix II: Tables and figures

Table 2: Simulation results for AIGF and MIGF models

Parameter	True values	AIGF				MIGF			
		Mean	Standard error	Credible Interval Lower	Credible Interval Upper	Mean	Standard error	Credible Interval Lower	Credible Interval Upper
θ_1	1.0	1.24	0.73	0.37	2.89	0.30	0.05	0.22	0.42
θ_2	42	55.72	23.36	21.29	97.69	26.66	9.25	20.15	50.22
α_a (α_m)	2.5	3.16	1.17	0.74	4.90	3.63	0.91	1.78	4.95
β_1	3.0	2.84	0.42	2.02	3.64	1.23	0.43	0.54	2.08
β_2	3.0	0.49	0.29	0.02	1.13	0.33	0.26	0.01	0.92

Table 3: Simulation results for AMIGF models

Parameter	True values	AMIGF I				AMIGF II			
		Mean	Standard error	Credible Interval Lower	Credible Interval Upper	Mean	Standard error	Credible Interval Lower	Credible Interval Upper
θ_1	1.0	1.02	0.28	0.57	1.59	0.61	0.12	0.38	0.84
θ_2	42	42.27	19.88	20.56	91.04	44.33	21.46	20.99	96.96
α_0	2.5	2.56	1.43	0.17	4.89	0.67	0.82	0.02	3.21
α_a	2.5	2.54	1.43	0.10	4.84	0.74	0.99	0.02	4.26
α_m	2.5	2.63	1.41	0.21	4.91	2.89	1.35	0.27	4.85
β_1	3.0	3.24	0.35	2.45	3.89	3.62	0.76	2.22	4.91
β_2	3.0	3.61	0.99	1.37	4.94	1.77	0.50	0.77	2.82

Table 4: Model comparison values for simulation study

Model	AIGF	MIGF	AMIGF I	AMIGF II
DIC	7.93	44.29	-2.81	18.48
log likelihood	-2.42	-19.97	3.59	-8.01

Table 5: Posterior summary for AIGF and MIGF models fitted to KI data

Parameter	AIGF Model				MIGF Modle			
	mean	standard error	Credible Intervals Lower	Upper	mean	standard error	Credible Intervals Lower	Upper
θ_1	2.72	1.12	0.93	4.83	0.95	0.21	0.55	1.43
θ_2	203.90	49.43	121.93	294.67	143.57	52.61	101.64	298.31
$\alpha_a(\alpha_m)$	0.03	0.02	0.001	0.071	5.97	2.37	1.25	9.68
β_1	0.42	0.54	-0.53	1.61	-0.15	0.25	-0.58	0.33
β_2	-2.48	0.49	-2.99	-0.99	-1.04	0.51	-2.02	0.14
β_3	-0.46	1.19	-2.71	1.92	1.14	0.63	-0.19	2.19
β_4	0.47	1.09	-1.69	2.66	1.68	0.59	0.36	2.74
β_5	-1.45	1.01	-2.88	0.91	0.39	0.65	-1.07	1.65

Table 6: DIC and log-likelihood values for the models fitted to KI data

Additive covariates	log likelihood	DIC	Additive covariates	log likelihood	DIC
Model: AMIGF					
-	-185.36	390.89	x_1	-188.67	379.82
x_2	-183.99	374.49	x_3	-189.39	383.49
x_4	-189.52	384.76	x_5	-188.68	379.65
x_1, x_2	-183.74	374.89	x_1, x_3	-189.09	382.74
x_1, x_4	-189.16	384.43	x_1, x_5	-187.53	377.95
x_2, x_3	-183.69	374.64	x_2, x_4	-184.14	377.04
x_2, x_5	-180.06	370.13	x_3, x_4	-189.57	386.57
x_3, x_5	-188.75	381.97	x_4, x_5	-189.18	384.69
x_1, x_2, x_3	-183.60	374.95	x_1, x_2, x_4	-184.04	377.33
x_1, x_2, x_5	-183.10	372.70	x_1, x_3, x_4	-189.49	387.26
x_1, x_3, x_5	-188.47	382.09	x_1, x_4, x_5	-188.73	384.06
x_2, x_3, x_4	-183.85	377.11	x_2, x_3, x_5	-182.84	373.71
x_2, x_4, x_5	-183.31	375.69	x_3, x_4, x_5	-189.37	386.46
x_1, x_2, x_3, x_4	-183.78	376.82	x_1, x_2, x_3, x_5	-182.85	373.67
x_1, x_2, x_4, x_5	-183.16	375.93	x_1, x_3, x_4, x_5	-189.09	386.00
x_2, x_3, x_4, x_5	-183.08	375.85	x_1, x_2, x_3, x_4, x_5	-183.21	376.44
Model: MIGF			Model: AIGF		
-	-181.49	374.89	x_1, x_2, x_3, x_4, x_5	-182.68	377.03

Table 7: Posterior summary for AMIGF-25 model fitted to KI data

parameter	mean	standard error	Credible Intervals	
			Lower	Upper
θ_1	2.0454	0.5415	1.0603	2.9363
θ_2	209.8744	50.9649	116.1107	293.6556
α_0	0.0098	0.0053	0.0008	0.0192
α_a	0.0094	0.0055	0.0007	0.0192
α_m	2.0033	1.1084	0.2547	3.8709
β_1	0.0435	0.3386	-0.7293	0.4889
β_2	-2.4608	0.4492	-2.9851	-1.3263
β_3	-0.3596	1.2644	-2.7740	2.1238
β_4	1.0061	1.4008	-2.5124	2.8934
β_5	-1.4067	0.9711	-2.8853	0.7154

Table 8: Posterior summary for AIGF and MIGF models fitted to AML data

parameter	mean	standard error	Credible Intervals	
			Lower	Upper
AIGF Model				
θ_1	2.2076	0.9644	1.1761	4.8805
θ_2	40.7607	15.5808	16.9704	95.1269
α_a	0.0038	0.0039	0.0001	0.0163
β_1	-1.1354	0.8959	-1.9652	0.2658
MIGF Model				
θ_1	3.1361	1.2108	1.1264	4.8691
θ_2	14.4335	4.6274	17.5399	94.0934
α_m	1.7154	1.0976	0.0001	0.0163
β_1	-0.6188	0.3396	-1.9620	0.1611

Table 9: Posterior summary for AMIGF I and AMIGF II models fitted to AML data

parameter	mean	standard error	Credible Intervals Lower	Upper
AMIGF I Model				
θ_1	2.7997	1.0354	1.0509	4.8087
θ_2	40.2723	20.8904	16.6114	93.0445
α_0	0.0042	0.0040	9.46×10^{-5}	1.53×10^{-2}
α_a	0.0043	0.0039	0.0001	0.0156
α_m	2.3590	1.3707	0.1925	4.7937
β_1	-0.7880	0.5790	-1.1886	0.3373
AMIGF II Model				
θ_1	2.8090	1.0433	1.1264	4.8691
θ_2	46.2515	21.5199	17.5399	94.0934
α_0	0.0047	0.0041	0.0001	0.0151
α_a	0.0048	0.0042	0.0001	0.0156
α_m	2.7134	1.3423	0.3410	4.8755
β_1	-1.1019	0.5925	-1.9619	0.1611

Table 10: DIC and log-likelihood values for the models fitted to AML data

Model	AIGF	MIGF	AMIGF I	AMIGF II
log-likelihood	-84.7276	-80.6346	-79.3018	-79.3251
DIC values	173.0994	167.2703	164.5656	164.6036

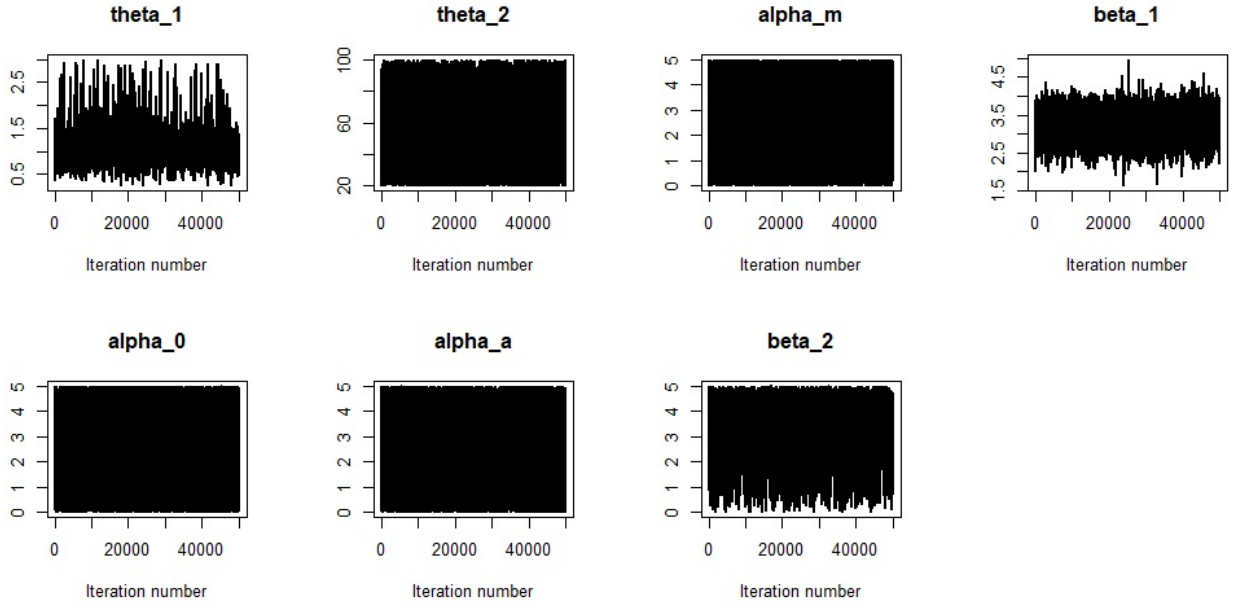


Figure 1: Trace plots of AMIGF I model parameters for simulated data

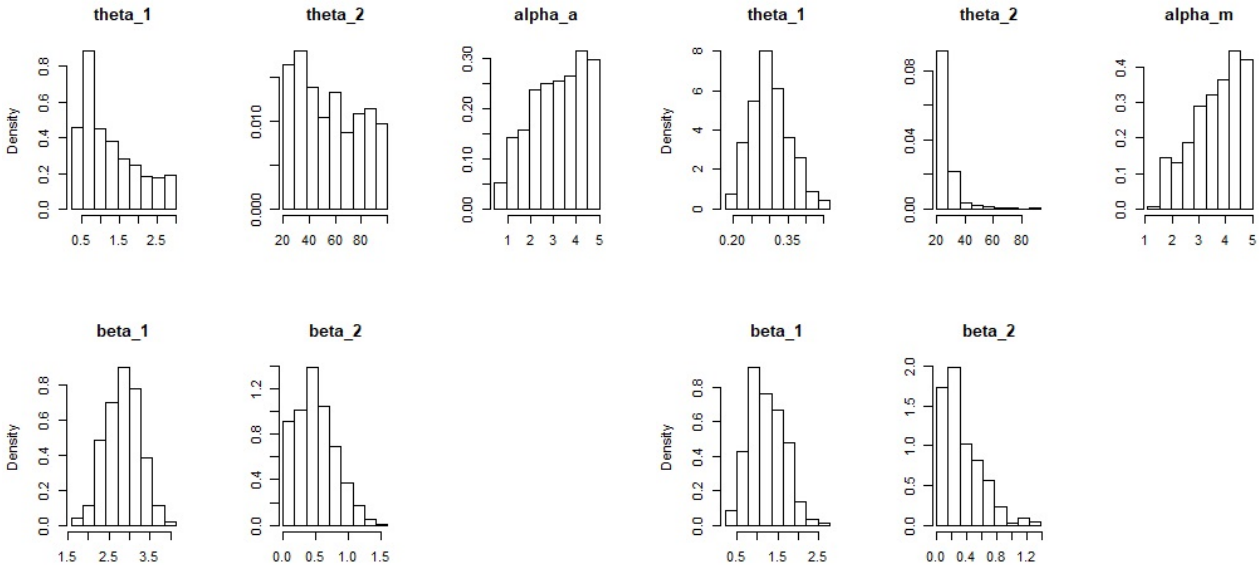


Figure 2: Posterior distribution of AIGF (left) and MIGF (right) model parameters for simulated data

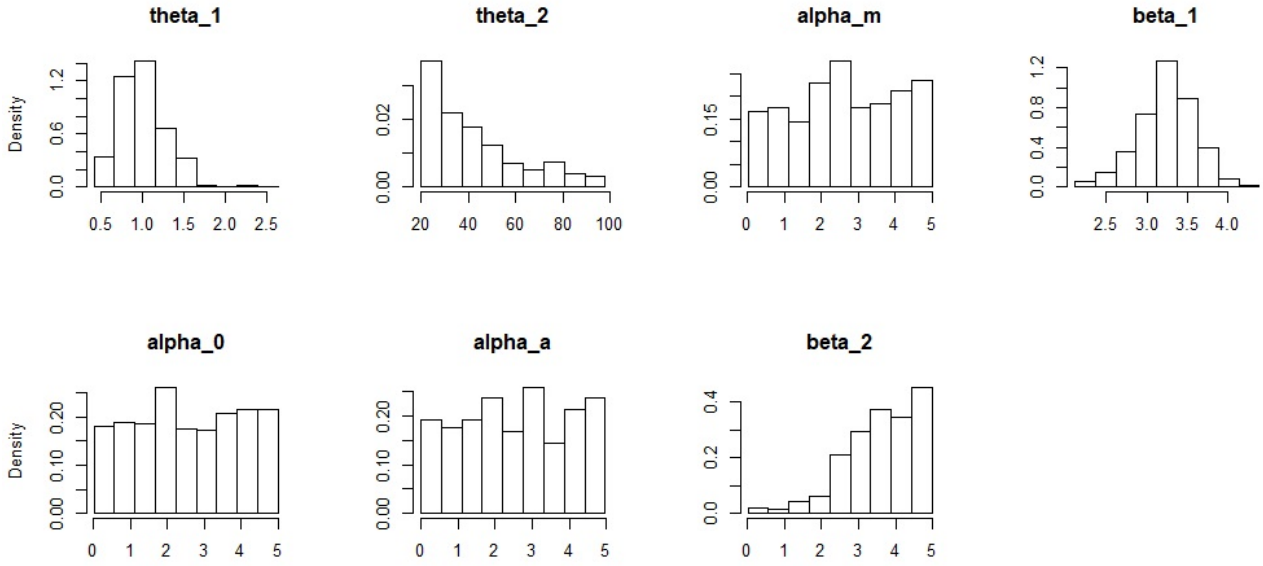


Figure 3: Posterior distribution of AMIGF I model parameters for simulated data

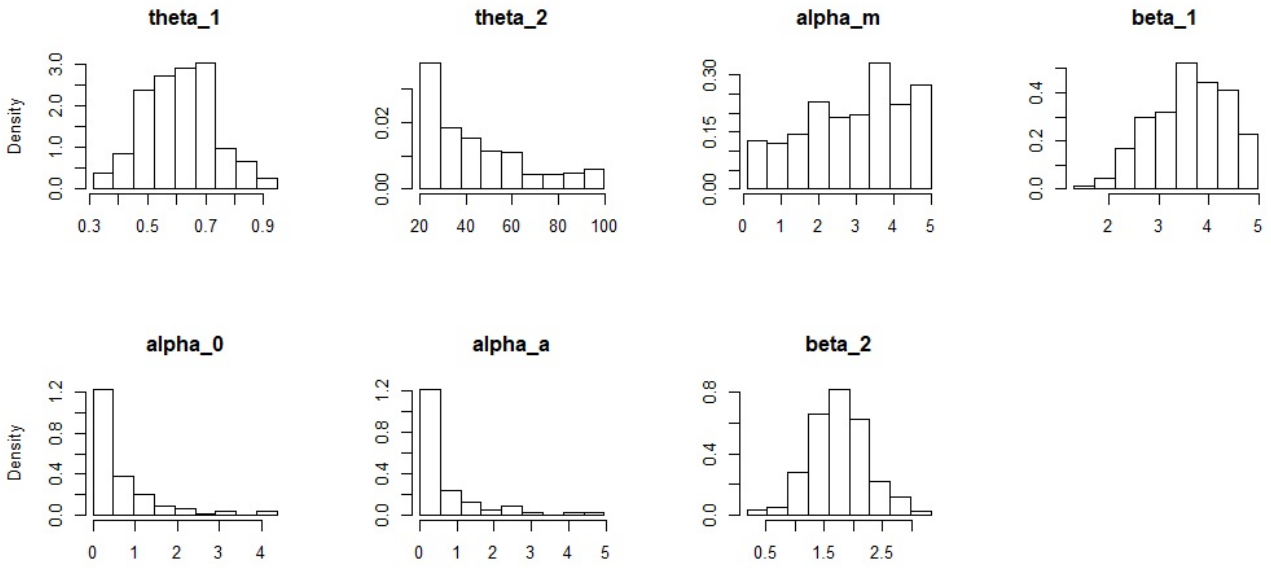


Figure 4: Posterior distribution of AMIGF II model parameters for simulated data

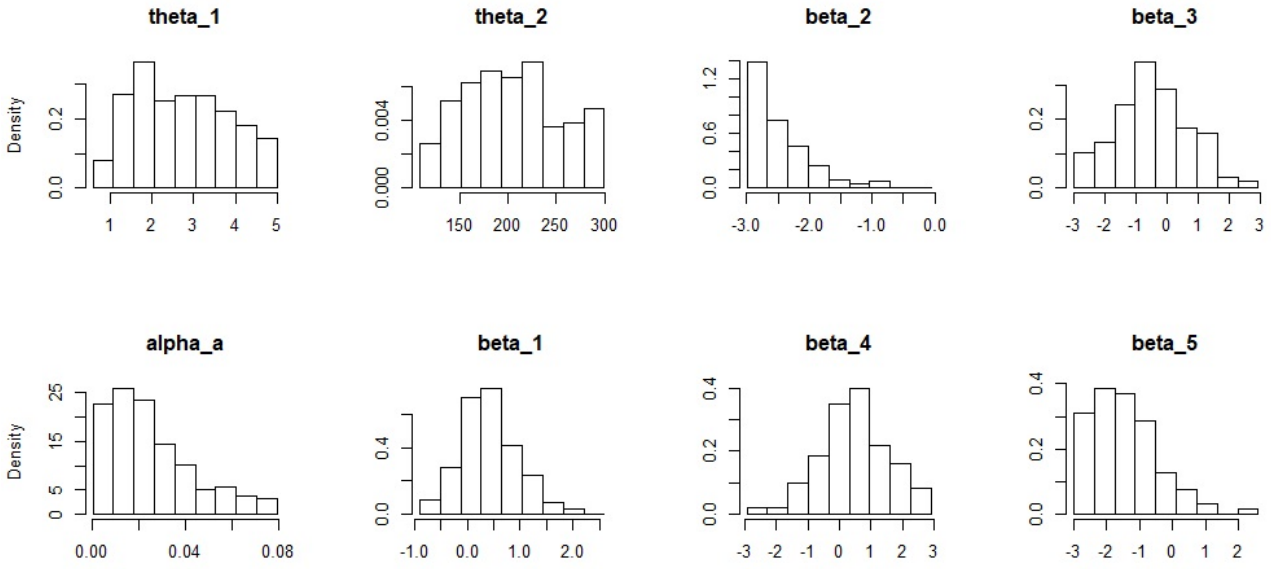


Figure 5: Posterior distribution of AIGF model parameters for KI data

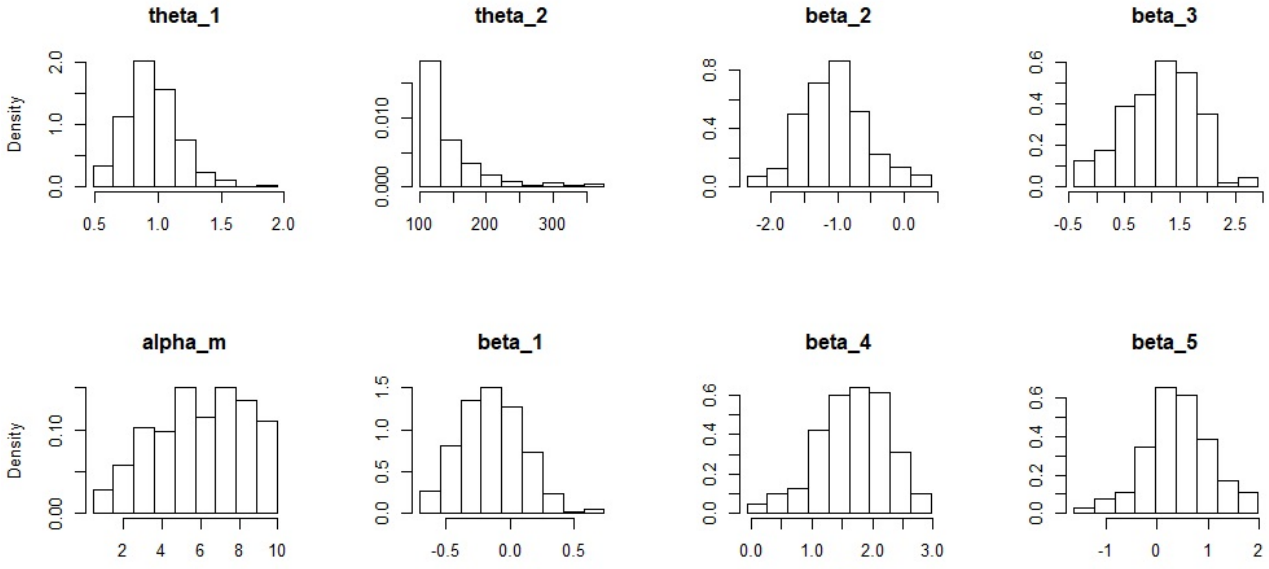


Figure 6: Posterior distribution of MIGF model parameters for KI data

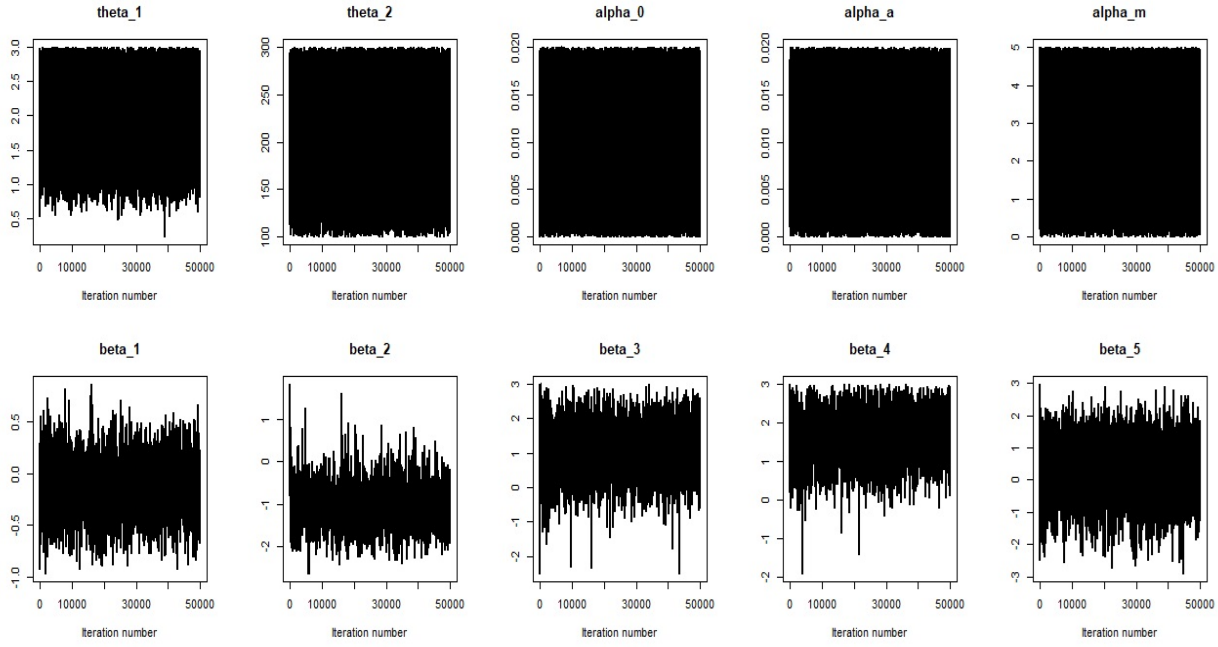


Figure 7: Trace plots of the AMIGF-25 model parameters fitted to KI data

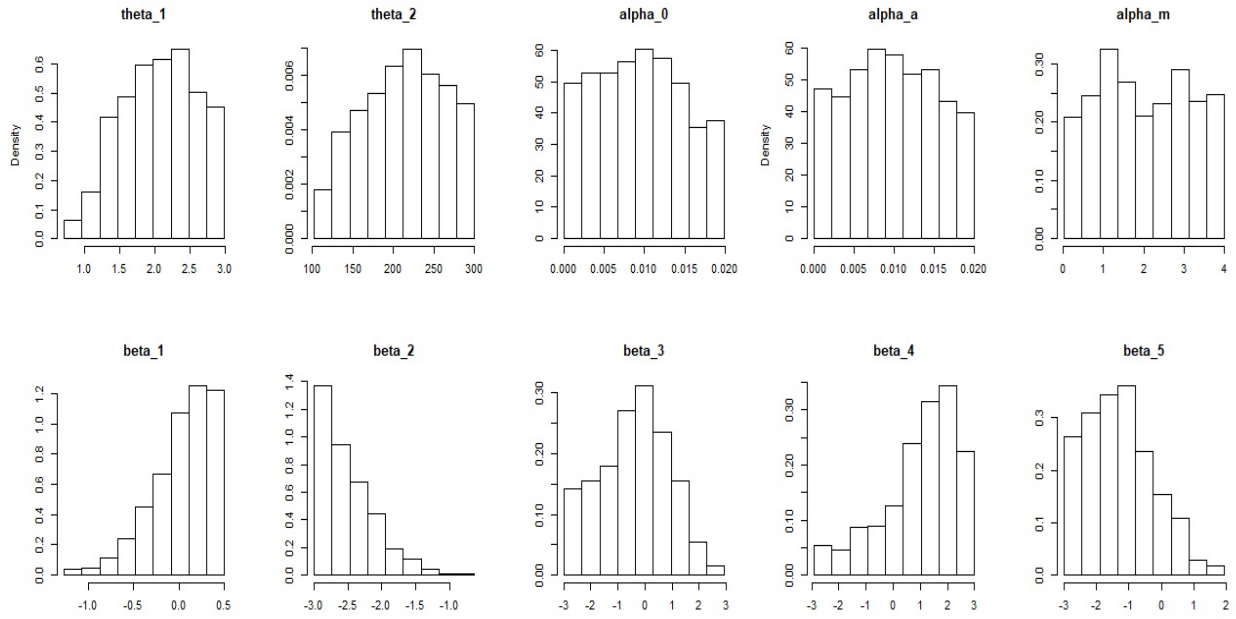


Figure 8: Posterior distribution of AMIGF-25 model parameters for KI data

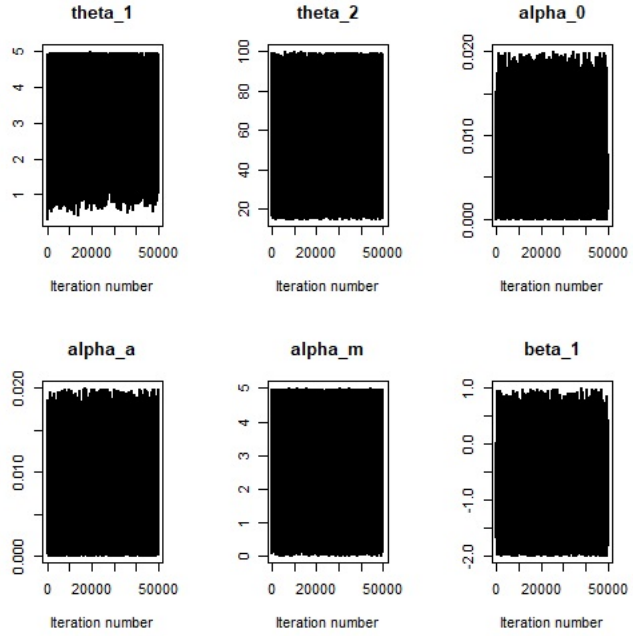


Figure 9: Trace plots of AMIGF I model parameters for AML data

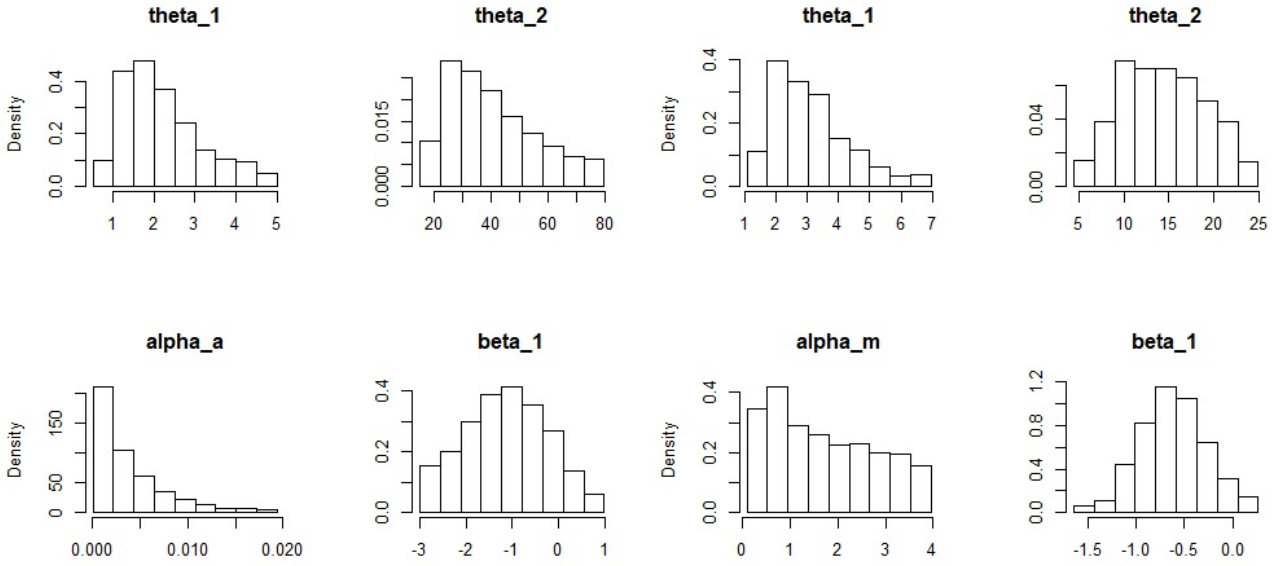


Figure 10: Posterior distribution for AIGF (left), MIGF (right) model parameters for AML data

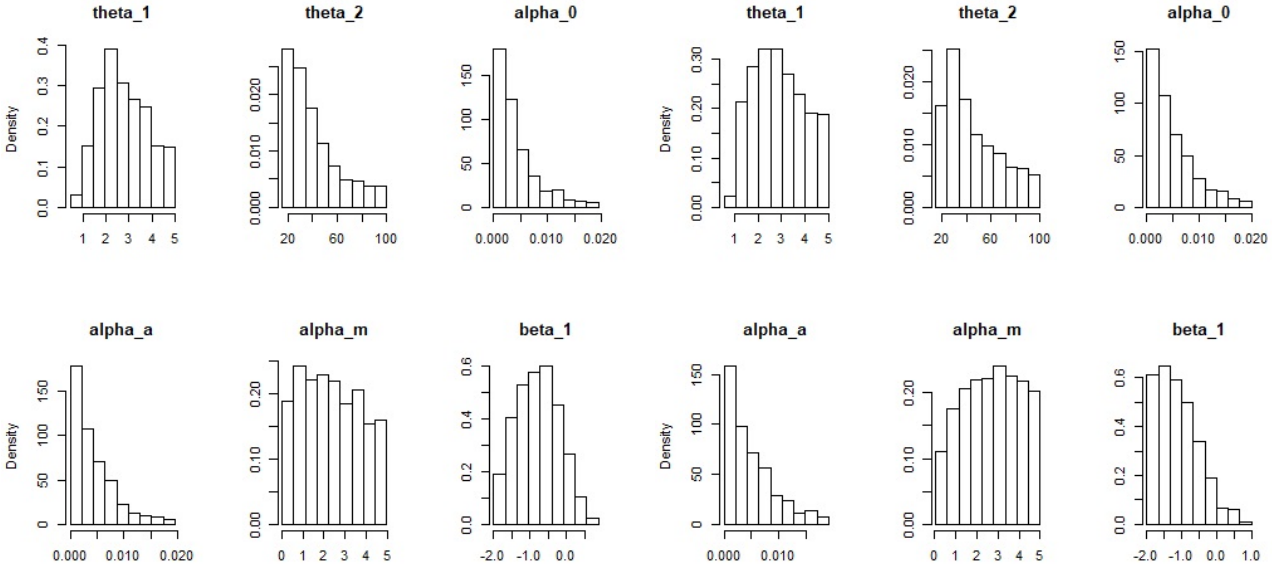


Figure 11: Posterior distribution for AMIGF I (left), AMIGF II (right) model parameters for AML data

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 24, No. 1, 2026 (New Series), pp 79–91
http://www.ssca.org.in/journal

Competing Risks Analysis of Factors Influencing the Runs Scored by Top T20 Batsmen - A Survival Analysis Approach

M. Sathishkumar[1], M. Ramakrishnan[1] and N. Viswanathan[2]

[1]*Department of Mathematics, RKM Vivekananda College, Chennai, India*
[2]*Department of Statistics, Presidency College, Chennai, India.*

Received: 22 August 2024; Revised: 07 June 2025; Accepted: 09 June 2025

Abstract

*T*20 cricket is an exciting format characterized by explosive hitting and strategic play, engaging fans with each boundary and a wicket. A comprehensive dataset of T20 matches is analyzed to understand the factors affecting batsmen's performance in this highly dynamic format. Survival analysis approach is used to study the performance of the batsmen, measured in terms of 'number of runs' taken as the 'innings survival time'. In this context, dismissal of batsmen is taken as the 'event'. The dismissal may be due to getting Bowled, being Caught, LBW, Run out, Stumped or Hit wicket. These different forms of dismissal can be taken as 'competing risks' and this study specifically focuses on identifying factors associated with specific dismissals. In this process, Cumulative Incidence Function (CIF), Cause-Specific Hazard function (CSH), and Fine and Gray's Subdistribution Hazard function (SDH) were used. The results from this analysis offer insights into the game dynamics and aids in player's performance evaluation and strategic decision-making, such as, team composition, batting order and making the choice of 'batting first' or 'chasing'. Data for representative batsmen from the top ICC ranking of T20 game with specific inclusion and exclusion criteria was taken from www.espncricinfo.com as on 28th of April, 2025. The analysis was carried out using R Programming Language (R 4.5.0) with suitable packages.

Key words: Survival analysis; Competing risks; Cumulative incidence function; Cause-specific hazard; Subdistribution hazard.

AMS Subject Classifications: 62N01, 62N02

The video recording of the paper made under the SSCA's Online Lecture series is available at the Youtube channel URL `https://youtu.be/E4rnAoT1f0g`.

1. Introduction

Cricket is becoming one of the most popular sports of the today's world. International cricket games are categorized as Test Cricket, One-day International (ODI) and Twenty20

Corresponding Author: M. Ramakrishnan
Email: mramkey@rkmvc.ac.in

(T20). One-day cricket was introduced in 1960s as an alternative to the Test Cricket characterized by more aggressive batting, colorful uniforms and fewer matches ending in draws. ODI cricket is limited to fifty overs. The biggest event in ODI cricket takes place in every four years when the Men's Cricket World Cup is organized by the International Cricket Council (ICC) which is the global governing body for cricket games. Later in 2003, T20 form of the game was introduced with focus on gaining wider audience and with emphasis on power hitting. Cricket in T20 format is limited to twenty overs for each team. The current study focusses on the game of T20's. Batting is the heart of cricket and bowling is its backbone. Cricket knowledge tells us that batting is more difficult early in a player's innings but becomes easier as players familiarize themselves with the pitch conditions.

Survival Analysis is defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of a particular event of interest. In Competing Risks model, the subject is exposed to more than one possible event of interest, but only one event will occur at any given time. In cricket, a player's dismissal is taken as the event of interest and the different ways in which the dismissal occurs, namely, Bowled, Caught, LBW, Runout and Stumped can be considered as competing risks for the event 'dismissal'. Of these competing risks, one can be taken as the main event and the rest as competing events that prevent the observation of the main outcome or change the probability of its occurrence. The added feature of competing risks data is the presence of failure types, in addition to failure time or survival time.

The developments in the field of survival analysis had the most profound impact on clinical trials are the Kaplan and Meier (1958), for estimating the survival function, the Log-rank statistics by Mantel (1966) for comparing two survival distributions, and the Cox (1972) proportional hazards model for quantifying the effects of covariates on the survival time. The Cox regression model can be used to identify the variables that significantly affect the outcome of interest and present the results in terms of the hazard ratio.

Staden *et al.* (2010) developed alternative batting average measures to address issues with traditional averages, particularly when a batsman ends up with 'not out'. In most of the previous studies such as Kimber and Hansford (1993), Kachoyan and West (2016), Brown (2017) and Saikai and Bhattacharjee (2018) survival abilities of individual batsman were analyzed based on the survival function of the number of balls faced till dismissal. In this study, instead of considering the 'number of balls faced before dismissal', the 'number of runs scored before dismissal' is taken as the life span. Kachoyan and West (2016) described batsman's innings as a lifespan, in that, "when the batsman goes out to bat, he is 'born' and 'lives' for a certain number of balls before he is dismissed". A dismissal was referred to as a batsman's 'death' which is the event of interest. When a batsman was not dismissed during a match, the particular observation was referred to as a censored observation. In this study, instead of considering the 'number of balls before he is dismissed', the 'number of runs scored before he is dismissed' is taken as the life span.

Survival analysis using competing risks data is widely used in medical research and is gaining increasing attention and interest in a variety of research fields. Cumulative Incidence Function (CIF) represents the cumulative probability of an event due to a particular type of cause over time and it is a useful metric for analysing competing risks data (Pintilie (2006)). By treating the Cumulative Incidence Curve (CIC) as a subdistribution hazard Function,

Gray and Jason (1999) provided a methodology for computing the CIF and comparing it across different categories of a variable. Sapir-Pichhadze *et al.* (2016) observed that the Cause-specific and subdistribution hazard models provide complementary information regarding the relationship between exposures and outcomes of interest in the presence of competing events.

Shah *et al.* (2023) studied the survival probabilities of the top-10 ODI batsmen around the world and suggested that it can be used as a new measure for evaluating batsman's ability to survive on crease. Kottearachchi *et al.* (2022) studied the survival abilities of the opening batsmen in one-day international cricket. Preetham *et al.* (2023) suggested a model for predicting the outcome of the IPL matches, in particular, to forecast the score of an innings using machine learning models. Saikai and Bhattacharjee (2018) examined survival ability of batsmen in Indian Premiere League (IPL) 2012. Ramakrishnan *et al.* (2023) studied the performance of Indian Batsmen in the 2023 World Cup Squad using survival analysis approach.

In this paper, in order to identify the factors influencing the survival time of a batsman, measured in terms of the number of runs scored, the Cox Proportional Hazard model under 'single event' and 'competing events' are considered. For competing risks model, both 'Cause-specific' and 'Subdistribution' approaches are employed. Representative batsmen from Top 10 T20 batsmen, as per the ICC rankings, are selected with specific inclusion and exclusion criteria described in the 'Data structure' section. The data and the rankings pertain as on 28$^{\text{th}}$ April, 2025 taken from www.espncricinfo.com. The covariates of the models include the order of batting, the position in which the player is slotted for batting, Match venue, the time of the game, Toss Result, the stage of the match and the Tournament Type. The results from the study clearly point to using the efficacy of competing risks models in identifying the significant factors and the patterns present in them.

2. Methodology

The Cox PH model is used to identify significant factors that influence the runs scored by the batsmen. In this paper, the general Cox PH model and the one pertaining to 'competing risks' are used. Under the considered competing risks models, both Cause-specific and Subdistribution Hazard approaches are used.

2.1. Survival functions and methods

The survival function is of at most priority in the field of survival analysis is defined as the probability of survival beyond time t.

$$S(t) = P(T > t) = 1 - F(t)$$

where T is a random variable denotes the time that the event occurs. The survival function is the complement of the Cumulative Density Function (CDF).

The hazard function $h(t)$ gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

The hazard function and the survival function have a straightforward one-to-one connection.

2.2. Cox proportional hazard model

Cox (1972) proposed the following regression model for the hazard function

$$h\left(t|X, \beta\right) = h_0(t) e^{\sum_1^p \beta_i X_i}$$

where the survival time is denoted by t, the p covariates are denoted as $(X_1, X_2, ..., X_p)$. The coefficients $\beta = (\beta_1, \beta_2, ..., \beta_p)$ quantify the effect of these covariates on the hazard function. Additionally, the term $h_0(t)$ signifies the baseline hazard, serving as a reference line for understanding how the covariates modify the risk of the event occurring over time. This model of the hazard function is used to analyze the survival data and identify the significant covariates influencing the hazard of the event under consideration. The proportional hazards assumption requires that covariates are multiplicatively related to the hazard. This study employs Schoenfeld Residual test to verify the proportional hazard assumption.

2.3. Competing risk models

Competing risk models, in this section, predominantly use Cause-specific and Sub-distribution hazard models. These approaches employ cumulative incidence function for studying the pattern of the risks, which in turn extracts the significant factors, if present in the model.

2.3.1. Cumulative Incidence Function (CIF)

In situations where competing risks are involved, the survival curve approach by the Kaplan-Meier method may not be fully reliable due to the violation of the independence assumption regarding the competing risks. This paved the way to introduce new approaches, one of which is the Cumulative Incidence Function that uses marginal probabilities.

The Cumulative Incidence Function (CIF) for event type c at time t_j which is calculated as the cumulative sum upto time t_j of the incidence probabilities over all event type c failure times

$$CIC_c\left(t_j\right) = \sum_{i=1}^{j} I_c\left(t_i\right) \ = \sum_{i=1}^{j} \widehat{S}(t_{i-1}) \widehat{h}_c(t_i)$$

where $\widehat{h}_c\left(t_j\right) = \frac{m_{cj}}{n_j}$ represents the ratio of number of events for type c that occur at t_j to the number at risk at t_j. Here, $S(t_{j-1})$ is the surviving probability of the prior time t_{j-1}, where *S(t)* represents the general survival curve instead of the Cause-specific survival curve S_c(t). Here, $I_c\left(t_i\right)$ is the incidence function for event type c at time t_i.

2.3.2. Cause-Specific Hazard (CSH) model

The Cause-Specific Hazard (CSH) Model that makes use of the Cox PH method to individually evaluate hazards for every type of failure, treating competing events in the form of censors as well as the others who are censored due to follow-up loss or due to pulling out of the study.

The Cause-specific hazard for event type c under Cox PH model with covariates $X = (X_1,\ X_2,\ ...,\ X_p)$ is defined as

$$h_c(t, X) = h_{0c}(t)e^{\sum_{i=1}^{p} \beta_{ic} X_i}$$

$$h_c(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T_c < t + \Delta t | T_c \geq t)}{\Delta t}$$

and T_c is the random variable that represents the time to failure for events of type c with $c = 1,\ 2,\ 3,\ \ldots,\ K$, β_{ic} represents the regression coefficient of X_i to the event type c and $h_{0c}(t)$ is the baseline hazard for event type c.

2.3.3. Subdistribution Hazard (SDH) function

Gray and Jason (1999) provides a regression model that applies SDH and can directly be used in CIF for computation under competing risks analysis. Under this model, In addition to the cause c's hazard, this CIF for cause c depends on hazards of all other causes, as well. For this approach, the SDH is also defined as

$$h_c^*(t; X) = \lim_{\Delta t \to 0} \left\{ \frac{P\left[t \leq T < t + \Delta t, D = c | T \geq t \cup (T < t \cap D \neq c)\right]}{\Delta t} \right\}$$

As a result, the covariate effect is proportional to the CIF. Proportional hazards assumption was imposed by Fine-Gray on the SDHs and provided estimators and their large sample properties. The SHD approach assumes that the occurrence of a competing event does not influence the rate of the event of interest. The partial likelihood estimation used in the standard Cox model is followed in the CSH and SDH models to estimate the covariate coefficients. However, the difference between CSH and SDH exists only in the risk set. The risk set for the CSH, decreases each time an event of another cause occurs. But these observations remain in the risk set for the SDH.

3. Data structure

The Data on the T20 matches for selected batsmen were taken from `www.espncricinfo.com` as on 28$^{\text{th}}$ April, 2025. The batsmen with top 10 ICC rankings, with the condition that they should have played at least 50 matches were selected. These 10 batsmen are Suryakumar Yadav (India), Jos Buttler (England), Pathum Nissanka (Sri Lanka), Tim Seifert (New Zealand), Babar Azam (Pakistan), Kusal Perera (Sri Lanka), Reeza Hendricks

(South Africa), Mohammad Rizwan (Pakistan), Finn Allen (New Zealand) and Kusal Mendis (Sri Lanka). Of these 10 batsmen, three were selected as a fair representation from this conditional top 10 batsmen. They constitute Jos Buttler (Rank 2), Babar Azam (Rank 5) and Mohammad Rizwan (Rank 8). The selection was made to avoid the extreme ranks, namely Rank 1 and Rank 10.

For the purpose of using competing risks model, runs scored by batsmen are considered as their survival time, represented by the variable Runs. The type of dismissals of a batsman is represented by the variable Dismissal. The dichotomous variable Dismissal_2 takes values '0' and '1', with '1' representing a dismissal, irrespective of its type such as Bowled, Caught or LBW and '0' representing the 'not out' status of the batsman. The covariates included in the Cox PH competing risk models include 'Innings' (Batting First or Chasing) represented by the variable Innings, 'Position' (Top, Middle, Low) represented by the variable Position, 'Match Place' (Home Venue, Home of Opposition, Neutral Venue) represented by the variable Venue, 'Tournament Type' (2 Team Series, 3-4 Teams, 5 or more Teams) represented by the variable Type, 'Time of the match' (Day, Day/Night, Night) represented by the variable Time and 'Toss Result' (Won, Lost) represented the variable Toss. All the included covariates are categorical in nature.

The first few rows of data for the batsman Babar Azam are presented in Table 1.

Table 1: Data for batsman Babar Azam representing first 10 of his matches

S. No	Runs	Dismissal	Innings	Venue	Type	Time	Toss	$Dismissal_2$	Position
1	86	caught	Batting First	Home Venue	2 Team Series	Night	Lost	1	Top
2	45	caught	Batting First	Home Venue	2 Team Series	Night	Won	1	Top
3	48	caught	Batting First	Home Venue	2 Team Series	Night	Lost	1	Top
4	34	not out	Batting First	Home Venue	2 Team Series	Night	Lost	0	Top
5	17	lbw	Batting First	Home Venue	2 Team Series	Day	Lost	1	Top
6	97	not out	Batting First	Home Venue	2 Team Series	Day	Won	0	Top
7	51	caught	Chasing	Home Venue	2 Team Series	Day	Lost	1	Top
8	13	caught	Chasing	Home Venue	2 Team Series	Night	Won	1	Top
9	3	bowled	Chasing	Home Venue	2 Team Series	Night	Lost	1	Top
10	27	caught	Chasing	Home Venue	2 Team Series	Night	Lost	1	Top

4. Empirical analysis

An overview of the details of the selected batsmen regarding their matches is presented in the Table 2.

Table 2: Descriptive statistics of the selected batsmen

Batsman	No. of Matches	Average Run	# Not Out	# Bowled	# Caught	# LBW	# Run Out	# Stumped	# Dismissal
Jos Buttler	123	28.74	23	14	72	7	5	2	100
Babar Azam	121	34.90	15	15	74	8	6	3	106
Rizwan	93	36.71	21	14	44	6	5	3	72

Table 2 indicates that out of 123 matches played by Jos Buttler, he reminded 'not out' in 23 matches and dismissed by 'Caught' in 72 matches followed by 'Bowled' in 14 matches. Similar information for the other two batsmen is also provided in Table 2. From this Table, it is inferred that the dismissal occurs mostly by way of Caught, followed by

Bowled. The other three categories, namely LBW, Runout and Stumped are relatively low in number. The low frequencies in these categories make us take Caught and Bowled as the two competing risks and drop the rest from the competing risks analysis.

A Cox PH model (Model 1) with 'Runs' as survival time and all types of dismissals taken as the event of interest is developed with covariates Innings, Position, Venue, Tournament Type, Time and Toss for all three batsmen separately. The results are consolidated in Table 3.

Table 3: Results of Cox PH model with all types of dismissal considered under one category -Model 1

		Jos Buttler		Babar Azam		Mohammad Riswan	
Covariates	Level	Hazard Ratio	p-value	Hazard Ratio	p-value	Hazard Ratio	p-value
Innings							
	Chasing	1.016	0.952	1.144	0.524	1.311	0.287
Position							
	Middle	1.709	**0.021***	1.083	0.920	2.802	**0.022***
	Bottom					6.275	0.093
Venue							
	Home	1.039	0.883	1.347	0.281	1.575	0.202
	Neutral	1.358	0.590	1.166	0.680	0.498	0.377
Tournament Type							
	Tournament3-4	2.382	0.132	0.749	0.596	2.288	0.195
	Tournament5+	0.595	0.341	1.435	0.319	2.659	0.211
Time							
	Day/Night	1.049	0.864	1.234	0.697	3.562	0.065
	Night	0.835	0.496	1.187	0.480	1.558	0.152
Toss							
	Lost	0.589	**0.050***	1.152	0.511	0.865	0.578

*denotes significance at 5% level

The results of Model 1 presented in Table 3 indicates that for the batsman Jos Buttler, Position and Toss turn out to be statistically significant at 5% level. For batsman Mohammad Rizwan the only covariate that is statistically significant is Position. For batsman Babar Azam, it is observed that none of the covariates is statistically significant. Going by the estimated Hazard Ratio, it is seen that for the batsman Jos Buttler, the hazard of getting out while playing in the middle-order is 1.71 times higher compared to playing in the top-order. It is further observed that, for batsman Jos Buttler, the hazard of getting out when his team loses the toss is 0.41 times lesser compared to when his team wins the toss. For the batsman Mohammad Rizwan, the hazard of getting out while playing in the middle-order is 2.8 times compared to playing in the top-order.

4.1. Schoenfeld residual test for PH assumption

The validity of estimation of Model 1 depends on how it satisfies the Proportional Hazards assumption for all the covariates under consideration. Schoenfeld residual test is carried out for this purpose and the results are presented in Table 4. The p-values from Table 4 indicates that the PH assumption is satisfied for all covariates for Jos Buttler and all but the covariate 'Position' for Babar Azam and Mohammad Rizwan at 5% level. Thus, it is seen that in most of the cases the PH assumption is well satisfied for Model 1 and estimates derived are stable and valid.

Table 4: Schoenfeld residual test results

	Jos Buttler		Babar Azam		Mohammad Riswan	
Covariates	Chi-square	p-value	Chi-square	p-value	Chi-square	p-value
Innings	1.786	0.180	3.606	0.058	0.042	0.837
Position	0.528	0.470	4.229	0.040	8.264	0.016
Venue	2.081	0.350	5.074	0.079	1.646	0.439
Tournament Type	0.976	0.610	0.104	0.949	1.124	0.570
Time	0.511	0.770	2.139	0.343	0.379	0.828
Toss	0.302	0.580	0.196	0.658	1.174	0.279

4.2. Cause-specific hazard model for 'Bowled'

The Cause-specific Hazard Model (Model 2) for the event 'Bowled' with 'Caught' as competing risks is developed and the significance of the selected covariates in terms of their p-values are presented in Table 5. This Table indicates that for batsman Jos Buttler, the Position and Venue are statistically significant at 10% level and their hazard ratios imply that the hazard of getting out while playing in the middle-order is 2.879 times higher compared to playing in the top-order. Further, hazard of getting out while playing in the Home town is 69% less compared to playing in the venue of the opposition team.

Table 5: Cause-specific hazard model for 'Bowled'

		Jos Buttler		Babar Azam		Mohammad Riswan	
Covariates	Level	Hazard Ratio	p-value	Hazard Ratio	p-value	Hazard Ratio	p-value
Innings							
	Chasing	0.638	0.483	0.852	0.769	1.902	0.263
Position							
	Middle	2.879	0.070	3.921	0.231	2.040	0.373
Venue							
	Home	0.309	0.091	0.901	0.874	0.958	0.954
	Neutral	0.546	0.343	1.000	1.000	1.009	0.990
Toss							
	Lost	0.646	0.498	1.058	0.917	1.392	0.571

4.3. Cause-specific hazard model for 'Caught'

The Cause-specific Hazard Model (Model 3) for the event 'Caught' with 'Bowled' as competing risks is developed and the significance of the selected covariates in terms of their p-values are presented in Table 6. This Table indicates that only for the batsman Jos Buttler, the Venue is statistically significant at 10% level and its hazard ratio implies that the hazard of getting out when the team loses the toss is 46% less compared to when the team wins the toss.

4.4. Schoenfeld residual test for PH assumption for Model 2 and Model 3

Schoenfeld residual test is carried out for Model 2 and Model 3 and the results are presented in Table 7. The p-values from Table 7 indicates that the PH assumption is satisfied for all covariates for the three batsmen for Model 2. In Model 3, for Babar Azam and

Table 6: Cause-specific hazard model for 'Caught'

		Jos Buttler		Babar Azam		Mohammad Riswan	
Covariates	Level	Hazard Ratio	p-value	Hazard Ratio	p-value	Hazard Ratio	p-value
Innings							
	Chasing	1.079	0.809	1.178	0.496	1.050	0.876
Position							
	Middle	1.556	0.102	0.689	0.717	2.289	0.140
Venue							
	Home	1.289	0.386	1.539	0.154	1.534	0.337
	Neutral	0.793	0.503	1.575	0.169	1.543	0.356
Toss							
	Lost	0.562	0.067	1.243	0.383	0.815	0.526

Mohammad Rizwan, all the covariates satisfy the PH assumption. For the batsman Bulter, the covariate Innings and Toss does not satisfy the PH assumption. The significance of the PH assumptions is tested at 5% level.

Table 7: *p*-values of Schoenfeld residual test for PH assumption for Model 2 and Model 3

	Bowled			Caught		
Covariates	Butler	Babar	Riswan	Butler	Babar	Riswan
Innings	0.602	0.17	0.35	0.005	0.17	0.455
Position	0.781	0.077	0.61	0.308	0.15	0.147
Toss	0.056	0.53	0.99	0.021	0.84	0.072
Venue	0.959	0.603	0.38	0.135	0.17	0.095
GLOBAL	0.456	0.286	0.69	0.048	0.15	0.112

The validity of PH assumption for batsman Jos Buttler in Model 2 is presented graphically in Figure 1. From this Figure, it is seen that for all the covariates under consideration, the Schoenfeld residuals mostly fall within the estimated boundaries, indicating that the PH assumption is valid for all the covariates in Model 2. Though similar Figures for the other two batsmen were verified, not presented in this text.

4.5. Subdistribution hazard model for 'Bowled'

Similar to Cause-specific Hazard model (Model 4), the subdistribution Hazard model is run for the event 'Bowled' and the associated hazard ratios and p-values for all the three batsmen are presented in Table 8. This Table indicates that only for the batsman Jos Buttler, the Venue is statistically significant at 10% level and its hazard ratio implies the hazard of getting out while playing in the Home Town is 73% less compared to playing in the Opposition Venue.

4.6. Subdistribution hazard model for 'Caught'

Similar to Subdistribution Hazard model (Model 5), the subdistribution Hazard model is run for the event 'Caught' and the associated hazard ratios and p-values for all the three batsmen are presented in Table 9. This Table indicates that only for the batsman

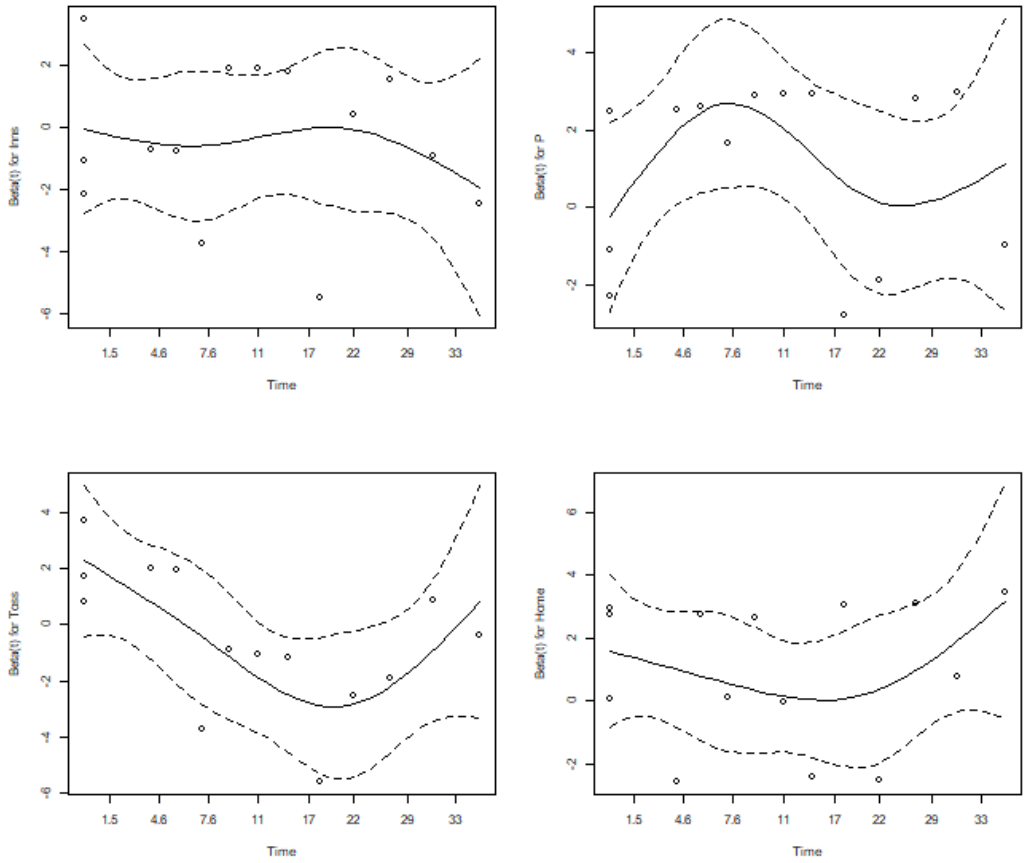


Figure 1: Pictorial representation of PH assumption verification for batsman Jos Buttler in Model 2

Table 8: Subdistribution hazard model for 'Bowled'

		Jos Buttler		Babar Azam		Mohammad Riswan	
Covariates	Level	Hazard Ratio	p-value	Hazard Ratio	p-value	Hazard Ratio	p-value
Innings							
	Chasing	0.647	0.490	0.744	0.610	1.874	0.270
Position							
	Middle	2.507	0.120	3.947	0.170	1.500	0.590
Venue							
	Home	0.276	0.062	0.754	0.670	0.875	0.880
	Neutral	0.570	0.380	0.817	0.760	0.991	0.990
Toss							
	Lost	0.916	0.880	0.921	0.890	1.589	0.480

Jos Buttler, the Venue is statistically significant at 10% level and its hazard ratio implies that the hazard of getting out while playing in the Home Town is 1.63 times compared to playing in the Opposition Venue.

The Schoenfeld residuals for the Subdistributional hazard models for 'Bowled' and

Table 9: Subdistribution hazard model for 'Caught'

		Jos Buttler		Babar Azam		Mohammad Riswan	
Covariates	Level	Hazard Ratio	p-value	Hazard Ratio	p-value	Hazard Ratio	p-value
Innings							
	Chasing	1.085	0.800	1.175	0.480	0.880	0.680
Position							
	Middle	1.026	0.920	0.484	0.450	1.304	0.640
Venue							
	Home	1.630	0.074	1.408	0.260	1.564	0.250
	Neutral	0.962	0.910	1.348	0.360	1.368	0.450
Toss							
	Lost	0.639	0.180	1.091	0.720	0.731	0.310

'Caught' satisfies the proportional hazard assumptions and as such the results from the above models can be taken as valid and stable. The Proportionality assumption for the competing risks regression log(-log(1-F)) can be plotted against log(Runs), where F is the CIF for the event of interest. The Figure 2 illustrates such a plot for the covariates in the batsman Jos Buttler for 'Bowled out'.

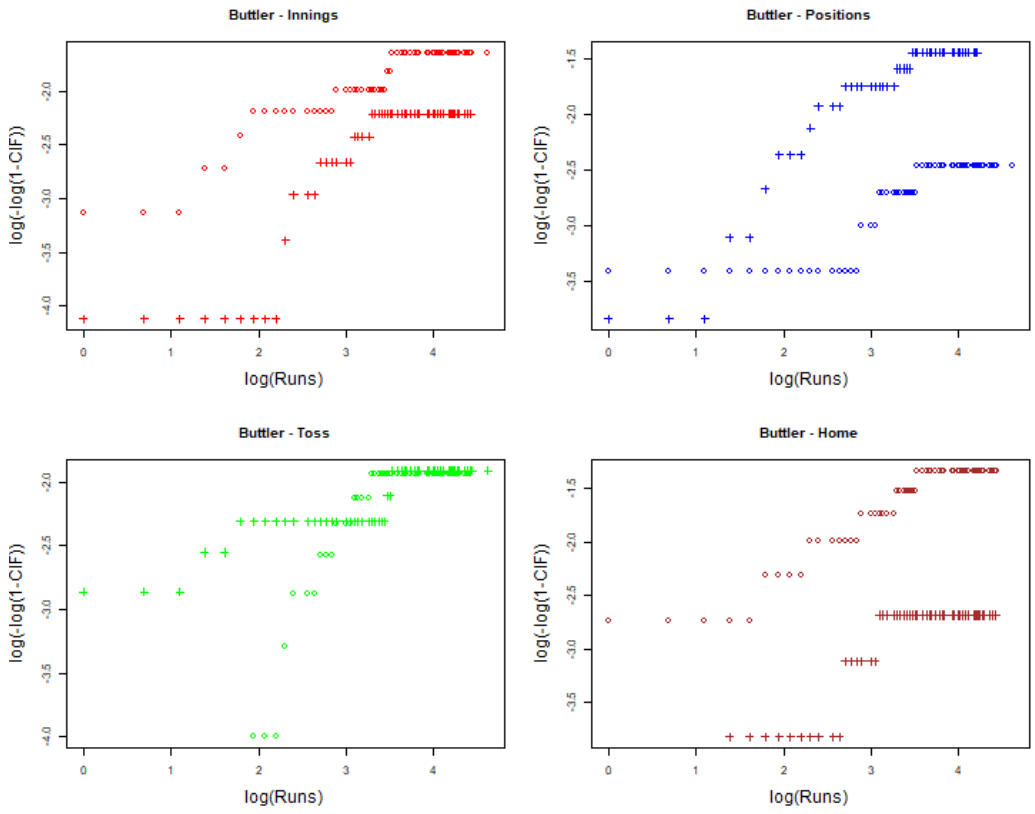


Figure 2: Proportionality of the hazard for all covariates in Model 4 for batsman Jos Buttler being bowled out

5. Conclusion

The runs scored by a batsman is taken as his 'Innings survival time' with covariates order of batting, the position in which the player is slotted for batting, Match venue, the time of the game, Toss Result, the stage of the match and the Tournament Type are considered under competing risk models. A comparison is carried out considering the usual Cox PH model with that of Cause-specific Proportional Hazards model and Subdistributional Proportional Hazards model. The Validity of the estimates derived under each model is verified by their respective Schoenfeld residuals. The General Cox PH model indicates Position and Toss as significant variable, whereas in the Cause-specific Models for the event 'bowled', Position and Venue are significant covariates. In the Cause-specific Model for the event 'caught' toss is a significant covariate. Under Subdistributional model for 'Bowled' and 'Caught', 'Venue' is statistically significant.

On comparing the significant covariates under different models, it is observed that if competing risks are not taken into account certain covariates tend to appear significant. But, when the competing risks are properly accounted for, the real scenario emerges and all the evidences point to the significant covariates 'Position', 'Toss' and 'Venue'. Thus, it is seen that the factors that influence the runs-scored by a batsman are the 'Position' in which he is slotted to play in the game, the result of the Toss and the venue in which the match is played. In particular, batsmen who play in the top order score considerably more runs compared to those who bat in the middle order slots. Further, playing the home venue decreases the hazard of getting out when the event is 'bowled' and the same increases the hazard of getting out when the event considered is 'caught'. Further, it is seen that the covariates 'Innings', 'Match Time', and 'Tournament Type' have no say in the run-scoring pattern of a batsman. This provides useful information about selection of batsmen for a particular venue and that of his batting-order in the game, which in turn increase the team's chances of winning the match.

6. Limitations of the study and future work

This study, though includes a number of covariates, some of them are dropped from the model, mainly because they have a highly imbalanced distribution in their levels. Future studies can include more batsmen in the analysis and in some of them, the distribution may be even and this will help to study the effect of the covariates that are dropped in this paper. Further, additional covariates such as the level of the match, indicating whether it is at league stage or knockout stage could also be included in the study and this may throw more light on the pressure exerted on the batsmen at different stages of the tournament.

Acknowledgements

We are very much grateful to the editor for providing consistent support to bring out tbe article in a compact, complete and readable format. We further express our high degree of appreciation to the reviewer, who had gone through the crude version of this article and helped us to transform it to a more refined structure in content and presentation, by providing finer details for great improvement in its current version.

Conflict of interest

I hereby declare that, to the best of my knowledge, I have no actual, potential, or perceived conflicts of interest related to this article. I understand the importance of objectivity and integrity in my work and will strive to fulfill my duties with the highest standards of professionalism.

References

Brown, P. (2017). Optimising batting partnership strategy in the first innings of a limited overs cricket match. [dissertation]; available from `https://researcharchive.vuw.ac.nz/xmlui/bitstream/handle/10063/6871/thesis_access.pdf?sequence=1`.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, **34**, 187–202.

Gray, R. J. and Jason, P. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, **94**, 469–509.

Kachoyan, B. and West, M. (2016). Cricket as life and death. in proceedings of the 13th australian conference on mathematics and computers in sport. *ANZIAM MathSport*, Melbourne, Victoria, Australia, 85–90.

Kaplan, E. L. and Meier, P. L. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481.

Kimber, A. C. and Hansford, A. R. (1993). A statistical analysis of batting in cricket. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **156**, 443–455.

Kottearachchi, S. S. S., Jayasinghe, C. L., and Silva, R. M. (2022). An investigation of survival abilities of opening batsmen in one-day international cricket. *Studies of Applied Economics*, **40**.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, **50**, 163–170.

Pintilie, M. (2006). *Competing Risks, A Practical Perspective*. John Wiley & Sons Ltd. Chichester, West Sussex, England.

Preetham, HK., Prajwal, R., Prince Kumar., and Naveen Kumar. (2023). Cricket score prediction using machine learning. *International Journal of Innovative Research in Technology*, **9**, 109–114.

Ramakrishnan, M., Viswanathan, N., and Ravanan, R. (2023). Performance of Indian batsmen in the 2023 world cup squad – a survival analysis approach. *Statistica*, **83**, 271–290.

Saikai, H. and Bhattacharjee, D. (2018). Survival ability of Indian and overseas batsmen on the cricket pitch in Indian premier league. *MOJ Sports Medicine*, **2**, 113–116.

Sapir-Pichhadze, R., Pintilie, M., Tinckam, K. J., Laupacis, A., Logan, A. G., Beyene, J., and Kim, S. J. (2016). Survival analysis in the presence of competing risks: The example of waitlisted kidney transplant candidates. *American Journal of Transplantation*, **16**, 1958–1966.

Shah, P., Chaudhari, R. D., and Patel, M. N. (2023). Evaluating batsman using survival analysis. *Statistics and Applications*, **21**, 51–62.

Staden, P. J. V., Meiring, A. T., Steyn, J. A., and Fabris-Rotelli. (2010). Meaning batting averages in cricket. *South African Statistical Journal Proceedings: Peer-reviewed Proceedings of the 52nd Annual Conference of the South African Statistical Association for 2010 (SASA 2010): Congress*.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 24, No. 1, 2026 (New Series), pp 93–108
http://www.ssca.org.in/journal

Understanding North Atlantic Climate Instabilities and Complex Interactions using Data Science

Alka Yadav[1], Sourish Das[2], Anirban Chakraborti[3] and Sudeep Shukla[4]

[1]*School of Computer Science, University of Petroleum and Energy Studies (UPES), Bidholi Campus, Dehradun-248007, India*
[2]*Chennai Mathematical Institute, Chennai-603103, India*
[3]*School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi-110067, India*
[4]*AI 4 Water LTD, Orpington BR6 9QX, United Kingdom*

Received: 12 December 2024; Revised: 25 June 2025; Accepted: 28 June 2025

Abstract

The North Atlantic Oscillation (NAO) index, a measure of sea-level atmospheric pressure variability, holds significant influence over weather patterns in North America and Northern Europe. A negative (positive) NAO value signifies increased cold air outbreaks and storm occurrences (reduced occurrences) in these regions. NAO, a product of multiple climate factors, demonstrates intricate dynamics with sea surface temperature (SST) and sea ice extent (SIE). In this study, we adopt a data-driven approach to explore the complex interplay between NAO, SST, and SIE, revealing a critical instability rooted in positive feedback loops among these climate variables. Our statistical machine learning methodology examines the impacts of melting Arctic SIE and rising SST on NAO, thereby understanding the weather patterns across the North Atlantic region. The skewness analysis yields a negative skewness in NAO across various time intervals, daily, weekly, and monthly. This skewness, coupled with NAO's mean zero stationary nature, accentuates system instability. To capture these dynamics, we formulate a Bayesian Granger causal dynamic linear model, which effectively updates the predictor-dependent variable relationship over time. The findings underscore an impending critical instability, indicative of more frequent occurrences of intensely cold climates in eastern North America and northern Europe, theory signifies a notable climate shift. By delving into the intricate feedback mechanisms of NAO, SST, and SIE, our study enhances our comprehension of climate variability, fostering a more informed perspective on the imminent climate changes that lie ahead.

Key words: Complex systems; North Atlantic Oscillation; Climate change; Bayesian Granger causality; Sea surface temperature; Sea ice extent.

The video recording of the paper made under the SSCA's Online Lecture series is available at the Youtube channel URL `https://youtu.be/_NTHQXf5M4g`.

Corresponding Author: Anirban Chakraborti
Email: anirban@jnu.ac.in

1. Introduction

The complex dynamics of Earth's climate system, influenced profoundly by human activities since the Industrial Revolution, is characterized by complex interactions and feedback mechanisms. These dynamics often lead to critical instabilities in the climate system, stemming from reinforcing (positive) feedback loops that amplify changes and destabilize the system, as highlighted by Halloran *et al.* (2020). In contrast, stabilizing (negative) feedback mechanisms play a crucial role in maintaining the stability of the Earth's climate system, which supports the existence of complex life forms (Kastens *et al.*, 2009). A deeper understanding of these feedback loops, particularly those that exacerbate climate instabilities, is imperative to address the pressing environmental challenges of biodiversity loss, climate change, and ecosystem degradation.

Among the many climate variables, the North Atlantic Oscillation (NAO), sea surface temperature (SST), and Arctic sea ice extent (SIE) are key indicators of climate variability in the North Atlantic region. The NAO, a widely studied atmospheric pressure index, influences weather patterns in eastern North America and northern Europe. Positive phases of the NAO are associated with fewer storms and milder winters, while negative phases bring increased storm activity and colder air outbreaks (Hurrell, 1995; Hurrell and Deser, 2009; Lindsey and Dahlman, 2009; Kwok, 2000). These interactions are further linked to variations in SST and SIE, forming a complex web of atmospheric and oceanic processes. Studies have revealed that NAO phases impact SST by altering wind patterns, heat flux, and ocean circulation, leading to significant regional climate changes (Rogers *et al.*, 2004; Hurrell and Deser, 2009). Simultaneously, changes in SST and SIE influence atmospheric pressure patterns, further modifying the NAO system, and highlighting the interconnected nature of these variables (Eayrs *et al.*, 2019).

The presence of a positive feedback loop among SIE, SST, and NAO has been widely acknowledged. Melting Arctic sea ice reduces the surface albedo, increasing heat absorption by the ocean, which in turn raises SST and accelerates further ice loss (DallÓsto *et al.*, 2017). This loop intensifies atmospheric circulation patterns and impacts NAO variability, as discussed in studies by Pan (2005); Becker and Pauly (1996); Slonosky and Yiou (2001). The NAO, in turn, influences Arctic sea ice recovery, (Warner, 2018), and SST distributions, further reinforcing this feedback mechanism (Miettinen *et al.*, 2011). This dynamic interplay is also evident on sub-seasonal and decadal timescales, as shown by Dai *et al.* (2021); Delworth *et al.* (2016); Parkinson (2000); Bader *et al.* (2011); Kvamstø *et al.* (2004); Kastens *et al.* (2009); Kwok (2000), with significant implications for regional and global climates.

Existing literature provides evidence of these interactions. For instance, the connection between Arctic sea ice and NAO variability has been extensively examined. Researchers have identified statistically significant correlations between autumn Arctic sea ice anomalies and winter NAO phases, which influence weather patterns and climate variability in Europe and North America (Warner, 2018; Horvath *et al.*, 2021). Several scientific research communities have identified a significant decrease in SIE in coming years, see Das *et al.* (2018); Cressey (2007). Moreover, studies suggest that NAO-driven SST anomalies contribute to multi-decadal fluctuations in the Atlantic meridional overturning circulation, affecting Arctic warming and tropical storm activity (Delworth *et al.*, 2016). While many of these insights rely on computationally intensive climate models, they often focus on specific Arctic subre-

gions or limited temporal scales, leaving gaps in the understanding of broader North Atlantic dynamics.

Our research aims to address these gaps by employing advanced data-driven statistical methodologies to investigate the presence of a reinforcing feedback loop involving SIE, SST, and NAO. Unlike traditional climate models, which may assume static relationships between variables over time, our approach incorporates dynamic modeling techniques, such as Bayesian Granger causal dynamic linear models, to account for temporal variations in predictor-dependent relationships (Migon *et al.*, 2010; Das and Dey, 2013). This approach directly addresses criticisms highlighted by Kolstad and Screen (2019) regarding the limitations of static assumptions in conventional statistical models. By adopting this dynamic framework, our study provides a more adaptive and comprehensive analysis of the interconnected North Atlantic climate system.

This study seeks to establish the existence of a positive feedback loop among SIE, SST, and NAO, providing new insights into the instabilities of the North Atlantic climate system. By utilizing advanced statistical models, we aim to contribute to a deeper understanding of the mechanisms driving climate variability. Our findings may help inform strategies to mitigate the challenges posed by rapid Arctic warming and its cascading effects on regional and global climates.

2. Data

2.1. Description

This study utilizes three distinct datasets: (a) daily mean Arctic Sea Ice Extent (SIE), (b) daily mean Sea Surface Temperature (SST) of north Atlantic basin, and (c) daily mean North Atlantic Oscillation (NAO) index. The NAO and SST datasets are sourced from the National Oceanic and Atmospheric Administration (NOAA) website (NSIDC, 2020; NOAA, 2020b). The SIE dataset is obtained from the National Snow and Ice Data Centre's website (NOAA, 2020a). These datasets cover a range of time periods: the NAO data is available from 1950, the SIE data from 1979, and the SST data from 1982. The period under consideration spans from January 1982 to September 2019, covering a duration of 38 years.

2.2. Exploratory analyses

This section encompasses the exploratory data analysis conducted on the Arctic Sea Ice Extent (SIE), Sea Surface Temperature (SST), and North Atlantic Oscillation (NAO).

In Figure 1(a), the gradual reduction in SIE from 1982 to 2019 is evident. Notably, during the summer season, Arctic SIE diminishes from 7.412 square km to 3.340 square km. Figure 1(b) illustrates a consistent upward trajectory in SST over the same timeframe. Complementing this, Figure 1(c) presents the NAO's time series spanning 1979 to 2019. NAO, representing the difference in sea-level air pressure between the Icelandic Low and the Azores, exhibits a stationary process with a mean of zero.

NAO plays a pivotal role in shaping westerly winds, storm tracks, and climatic conditions across the North Atlantic region. In the positive NAO phase, intensified high and low-pressure systems lead to warmer, wetter winters in northern Europe and northeastern

North America. Conversely, the negative NAO phase triggers colder winters and increased Arctic air intrusions in these regions.

Figure 1(c) affirms NAO's mean-zero stationary nature. This was further confirmed through the Augmented Dickey-Fuller test, as indicated by the p-value. Furthermore, the Auto-correlation function (ACF) analysis depicted in Figure 2(a) with a maximum lag of 5000 days (approximately 13 years) showing the long memory, and the Figure 2(b) displays the NAO index's marginal probability distribution, expected to be bell-shaped and symmetric with zero skewness, suggesting stability. However, subsequent empirical evidence, presented later, reveals instability in the system.

Lastly, Table 1 showcases the Hurst exponent of the NAO index, significantly surpassing 0.5. This robustly suggests the presence of long memory within the NAO, underscoring its zero-mean stationary characteristics.

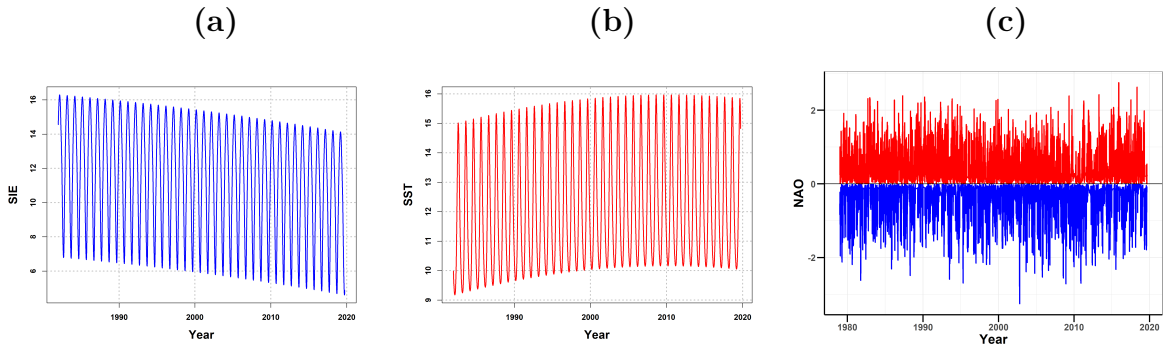


Figure 1: (a) SIE time series plot from 1982 to 2019. (b) SST time series plot from 1982 to 2019. (c) NAO time series plot from 1979 to 2019

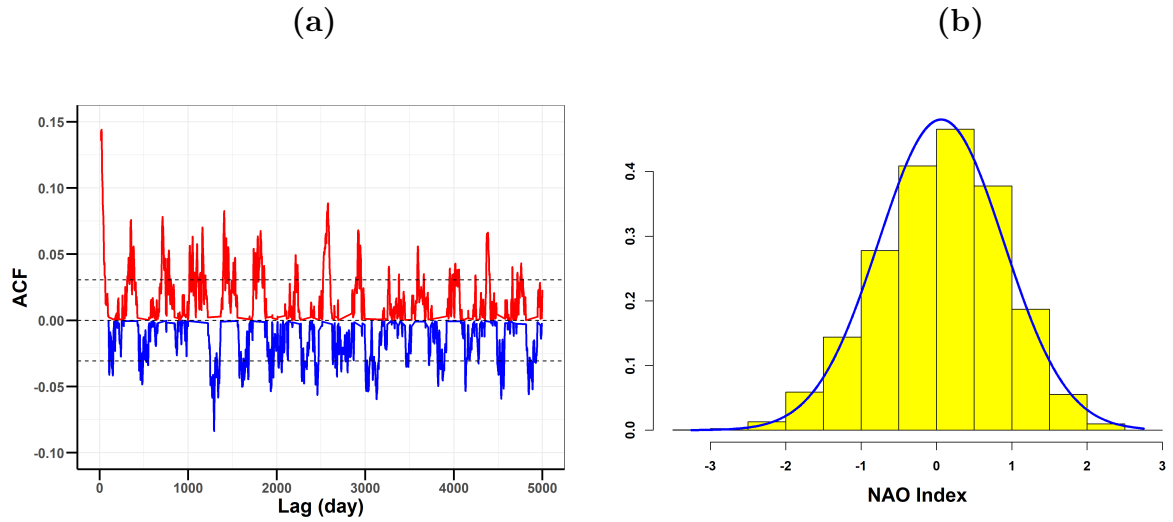


Figure 2: (a) Autocorrelation plot of NAO with a maximum lag of 5000 days (almost 13 years) indicating the existence of long memory.(b) The NAO index's marginal probability distribution from 1979 to 2019. While it is expected to be bell-shaped and symmetric, with zero skewness reflecting stability, the empirical evidence indicates instability in the system

Table 1: The Hurst Exponent values of the NAO index using different methods

Methods	NAO
Simple R/S Hurst estimation	0.73
Corrected R over S Hurst exponent	0.73
Empirical Hurst exponent	0.67
Corrected empirical Hurst exponent	0.66
Theoretical Hurst exponent	0.52

3. Methodology

This section outlines the statistical machine learning (SML) methodology employed to elucidate the feedback loop. Figure 1(a) and Figure 1(b) illustrate the non-stationary nature of SIE and SST. To address this, we introduce the equations in Section (3.1) and (3.2) that capture the trend and seasonality components of SIE and SST. By de-trending the SIE and SST, we obtain the residuals of both variables. These residuals are expected to exhibit characteristics of a zero-mean stationary process. Consequently, we investigate the Granger causality between the residuals of SIE, residuals of SST, and the North Atlantic Oscillation (NAO). It is important to note that the NAO was not subjected to the same filtering process as SST to remove seasonal effects because the NAO is inherently a zero-mean stationary process. This characteristic of the NAO indicates the absence of any intrinsic trend or seasonality in the data, making such filtering unnecessary. The Granger causal model employs auto-regressive time-series techniques to examine the relationships among these variables (Granger, 1969).

3.1. Modeling sea ice extent

Let us consider $x(t)$ as the SIE at time point t. We formulated the modeling of SIE by incorporating both a trend and a seasonal component. The trend component accounts for long-term variations, indicating either an increase or a decrease over time. On the other hand, the seasonal component captures recurring patterns at a fixed and known frequency, based on specific time periods such as the year, week, or day,

$$x(t) = \underbrace{\beta_0 + \beta_1 t + \beta_2 t^2}_{trend} + \underbrace{\left\{ \sum_{i=1}^{K} \alpha_i \sin(i\omega t) + \sum_{i=1}^{K} \gamma_i \cos(i\omega t) \right\}}_{seasonality} + \epsilon(t), \tag{1}$$

where ϵ is error with $\mathbb{E}(\epsilon) = 0$, and $Var(\epsilon) = \sigma^2$. In the seasonality component, we consider the periodicity with $\omega = \frac{2\pi}{365}$; α_i is the coefficient corresponding to the sine of i^{th} harmonics and γ_i is the coefficient corresponding to the cosine of the i^{th} harmonics. Within the model, we take into consideration K harmonics for each period. Now, two questions arise: (i) How do we determine the appropriate value of K? and (ii) When K becomes large, several harmonics may become redundant. To address the first question, we fit the model using various values of K ranging from 1 to K_0, and select the model with the minimum out-of-sample root mean square error (RMSE). Next, we utilize the least absolute shrinkage and selection operator (LASSO) technique to identify the optimal harmonics in model (1). The LASSO method selectively retains the harmonics that demonstrate a statistically significant

impact in minimizing the error (Tibshirani, 1996). A similar technique was successfully used to model long-term memory in climate variables (Yadav *et al.*, 2023, 2024).

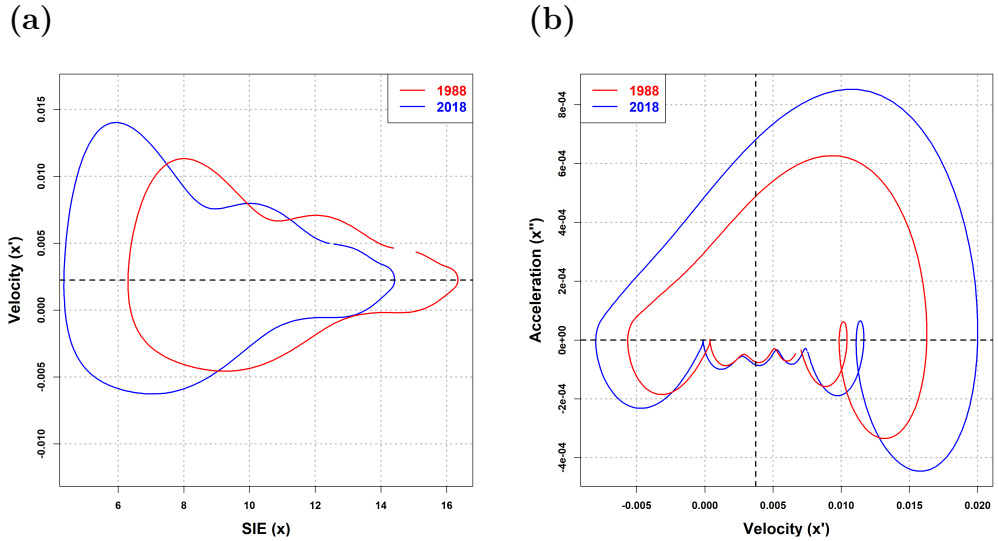


Figure 3: (a) Phase-line plot of SIE $x(t)$ as per Equation (1) vs. velocity of SIE $x'(t)$ as per Equation (8) for the two years 1988 and 2018. (b) The phase-plane plot of the velocity of SIE $x'(t)$ as per Equation (8) vs. acceleration of SIE $x''(t)$ as per Equation (9) for the two years 1988 and 2018

3.2. Modeling sea surface temperature

Let us consider $y(t)$ as the SST at time point t. Like SIE, in Equation (1), we formulated the modeling of SST by incorporating both a trend and a seasonal component,

$$y(t) = \underbrace{\tilde{\beta}_0 + \tilde{\beta}_1 t + \tilde{\beta}_2 t^2}_{trend} + \underbrace{\left\{ \sum_{i=1}^{K} \tilde{\alpha}_i \sin(i\omega t) + \sum_{i=1}^{K} \tilde{\gamma}_i \cos(i\omega t) \right\}}_{seasonality} + \delta(t), \tag{2}$$

where δ is error with $\mathbb{E}(\delta) = 0$, and $Var(\delta) = \sigma^2$. In the seasonality component, we consider the periodicity with $\omega = \frac{2\pi}{365}$; the component and parameters of the model (2) have same interpretation as in the model (1). Therefore, we employ the same strategy to fit the model.

Let us assume that $e(t)$ represents the residual obtained from the best fit of model (1). Henceforth, we will refer to $e(t)$ as Sea Ice Extent Residuals (SIER). Similarly, $d(t)$ denotes the residual obtained from the best fit of model (2). Moving forward, we will refer to $d(t)$ as Sea Surface Temperature Residuals (SSTR). It is anticipated that both $e(t)$ and $d(t)$ will exhibit characteristics of zero-mean stationary processes, similar to the North Atlantic Oscillation (NAO). Therefore, our objective is to explore the potential presence of a feedback loop between the SIER, SSTR and NAO by employing the Granger causality test.

3.3. Formulating feedback loop with Granger causality

In order to examine the evolving causal dynamics between SIER and SSTR, and vice versa, we formulate the following hypothesis utilizing the Granger causal models. The full

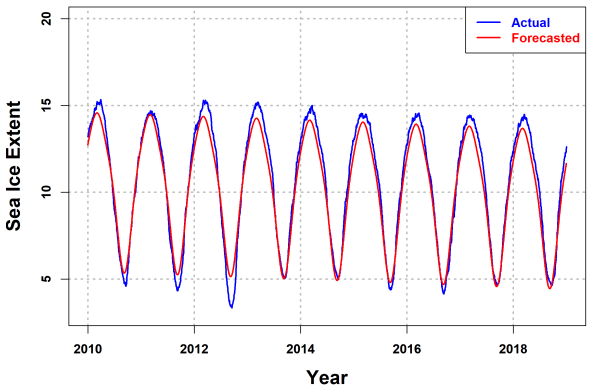


Figure 4: To evaluate the effectiveness of our proposed model (Equations (1, 8, 9)), a machine learning assessment was carried out. The training dataset spanned from 1979 to 2009, while the test dataset encompassed the years 2010 to 2019. The R-square value is -0.985

model incorporates the consideration of one variable being influenced by its own historical memory as well as the lagged values of the other two variables. For instance, the NAO variable is modeled as a function of its own lagged values and the lagged value of both SIER and SSTR, *i.e.*,

$$\begin{aligned} n(t) &= \beta_0 + \beta_1 n(t-1) + \cdots + \beta_k n(t-k) \\ &\quad + \gamma_1 e(t-1) + \cdots + \gamma_k e(t-k) \\ &\quad + \delta_1 d(t-1) + \cdots + \delta_k d(t-k) + \epsilon(t). \end{aligned} \tag{3}$$

Similarly, SIER is modelled as a function of its own lagged values and the lagged value of both NAO and SSTR, *i.e.*,

$$\begin{aligned} e(t) &= \beta_0 + \beta_1 e(t-1) + \cdots + \beta_k e(t-k) \\ &\quad + \gamma_1 n(t-1) + \cdots + \gamma_k n(t-k) \\ &\quad + \delta_1 d(t-1) + \cdots + \delta_k d(t-k) + \epsilon(t). \end{aligned} \tag{4}$$

Finally, SSTR is modelled as a function of its own lagged values and the lagged value of both NAO and SIER, *i.e.*,

$$\begin{aligned} d(t) &= \beta_0 + \beta_1 d(t-1) + \cdots + \beta_k d(t-k) \\ &\quad + \gamma_1 n(t-1) + \cdots + \gamma_k n(t-k) \\ &\quad + \delta_1 e(t-1) + \cdots + \delta_k e(t-k) + \epsilon(t). \end{aligned} \tag{5}$$

The Equations (3, 4, 5) collectively form the model depicting the feedback loop between NAO, SIER, and SSTR. To examine the presence of this feedback loop, we conduct the Granger causality test, such as the null hypothesis says that all $\gamma_i = 0$, *i.e.*,

$$H_0 : \gamma_1 = \cdots = \gamma_k = 0 \quad \text{and} \quad \delta_1 = \cdots = \delta_k = 0, \tag{6}$$

to reject the null hypothesis in our alternate hypothesis we have to check if at least one γ_i or $\delta_i \neq 0$, *i.e.*,

$$H_1 : \text{at least one } \gamma_i \neq 0 \quad \text{or} \quad \delta_i \neq 0. \tag{7}$$

We run this test for all three models, and if all three tests are rejected, it confirms the existence of the feedback loop.

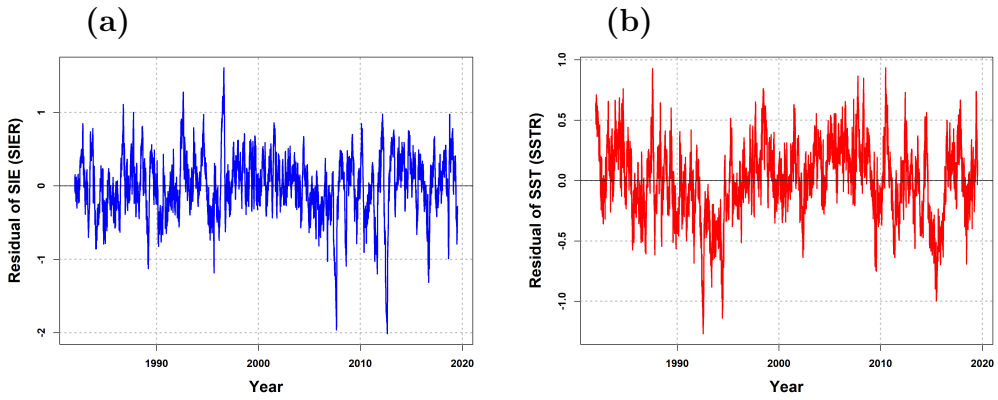


Figure 5: Time series plots of (a) Residual of SIE (SIER), (b) Residual of SST (SSTR)

3.4. Dynamic statistical approach

Kolstad and Screen (2019) has criticized the statistical models for climate research, citing the reason that the relationship between the predictor and dependent variable did not change over time (*i.e.*, regression coefficients remained constant over time). Here, we take into account this criticism by constructing a Bayesian Dynamic Linear Model (BDLM), along the lines of the work by Das and Dey (2013), where the relationship between a predictor and the dependent variable is updated over time by updating the coefficients. First, we consider the first derivative (momentum) and second derivative (acceleration) of the model presented in the above Equation (1) are:

$$x'(t) = \underbrace{\beta_1 + 2\beta_2 t}_{trend} + \underbrace{\omega\left\{\sum_{i=1}^{K}\alpha_i \cos(i\omega t) - \sum_{i=1}^{K}\gamma_i \sin(i\omega t)\right\}}_{seasonality} + \epsilon_1(t) \tag{8}$$

$$x''(t) = \underbrace{2\beta_2}_{trend} + \underbrace{\omega^2\left\{\sum_{i=1}^{K}\alpha_i \sin(i\omega t) + \sum_{i=1}^{K}\gamma_i \cos(i\omega t)\right\}}_{seasonality} + \epsilon_2(t), \tag{9}$$

where ϵ_1 and ϵ_2 are noises corresponding to momentum and acceleration signal. Then we developed the Bayesian dynamic linear models (BDLM) for NAO as follows:

$$n_t = \beta_{0t} + \beta_{1t} n_{t-1} + \gamma_{1t} x'_{t-1} + \epsilon_t, \tag{10}$$

$$\beta_{0t} = \rho_0 \beta_{0,t-1} + \eta_{1t}, \tag{11}$$

$$\beta_{1t} = \rho_1 \beta_{1,t-1} + \eta_{2t}, , \tag{12}$$

$$\gamma_{1t} = \rho_2 \gamma_{1,t-1} + \eta_{3t}, \tag{13}$$

where Equation (10) is known as observation equation; and Equations (11, 12, 13) are known as the system equations; the n_t is NAO at time point t; the x'_t is the momentum of SIE at time point t; ϵ_t is white noise follows $N(0, \sigma^2_\epsilon)$; $\eta_{jt},\ j = 1, 2, 3$ are white noise associated with system equation that follow $N(0, \sigma^2_{\eta_j})$. Note that $\rho = (\rho_0, \rho_1, \rho_2)$ should be bounded, that is, $|\rho_j| < 1$, such that the model would be a stationary process like the NAO index. We know that NAO is a mean stationary process; see Figure 1(c); hence this restriction of ρ_j is required. We present the Equation (10) and Equations (11, 12, 13) in the matrix notation as follows:

$$\begin{aligned} Y_t &= X_t\beta_t + \epsilon_t, \quad \text{observation equation,} \\ \beta_t &= R\beta_{t-1} + \eta_t, \quad \text{system equation,} \end{aligned} \tag{14}$$

where $\epsilon_t \sim N(0, \sigma^2)$, and $\eta_t \sim N(0, Q)$. The Bayesian update solution (aka, Kalman filter) for Equation (14) can be taken from (Das and Dey, 2013; Richard and Singpurwalla, 1983), as follows

$$\begin{aligned} K_t &= R\Sigma_t X_t^T (X_t \Sigma_t X^T + \sigma^2)^{-1}, \\ \hat{\beta}_{t+1} &= R\hat{\beta}_t + K_t(Y_t - X_t\hat{\beta}_t), \\ \Sigma_{t+1} &= R\Sigma_t R^T - K_t X_t \Sigma_t R^T + Q. \end{aligned} \tag{15}$$

In this study, BDLM corresponds to the Equation (1) developed. These models are then updated by the BDLM using a Kalman filter, as presented in Equation (15). Through these Kalman updates, β coefficients are dynamically updated. The model for SIE is as follows:

$$x(t) = \beta_0(t) + \beta_1(t)\sin(\omega t) + \gamma_1(t)\cos(\omega t) + \epsilon, \tag{16}$$

where $x(t)$ is SIE extent at time point t; $\beta_0(t)$, $\beta_1(t)$ and $\gamma_1(t)$ are the dynamic coefficients, updated via Kalman update Equation (15). Note that the model does not have any trend part because if there is any trend in the data, that will be captured automatically in the coefficients. In addition, the BDLM was built in the following way:

$$n(t) = \beta_0(t) + \sum_{k=1}^{K} \beta_k(t) n(t-1) + \sum_{k=1}^{K} \gamma_k(t) n(t-k) + \epsilon(t), \tag{17}$$

where $n(t)$ is NAO and $x'(t-k)$ is the momentum of the SIE at $(t-k)$. Here, the Akaike information criterion type model selection process was used to obtain the optimal choice for the model Equation (17). It may be noted that if the model's coefficients are static, then the Granger causal model is a special case of Equation (17).

4. Analyses and results

4.1. Analyses of SIER, SSTR, and NAO

The modeling framework pertaining to SIE and SST is detailed in sections (3.1) and (3.2). In Figure 3(a,b), which illustrate the phase-plane analysis of SIE, noticeable fluctuations in both the spatial extent and the rate of volume change of SIE become apparent. These figures portray the dynamic interplay through variations in the phase-line distribution, represented by the expanding area beneath the paired curves. Furthermore, a significant

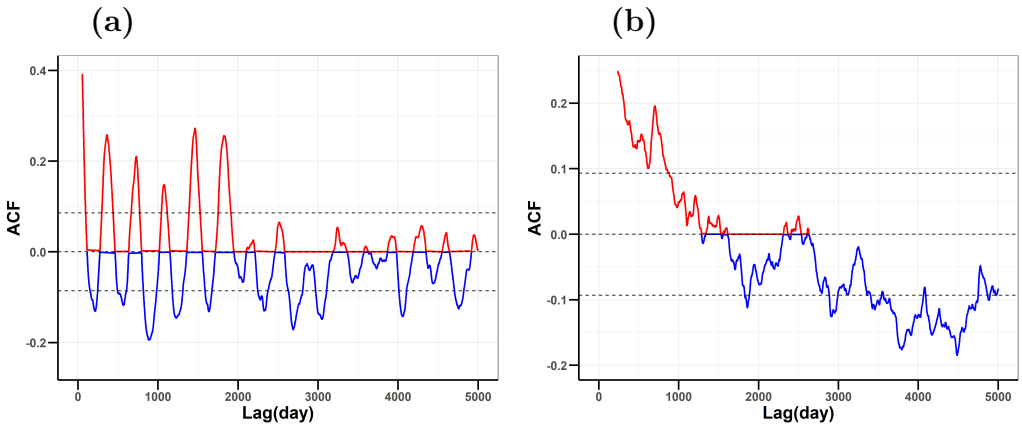


Figure 6: ACF plots of (a) Residual of SIE (SIER), (b) Residual of SST (SSTR)

Table 2: The Hurst Exponent values of SIER and SSTR using different methods

Hurst Exponent	SIER	SSTR
Simple R/S Hurst estimation	0.77	0.81
Corrected R over S Hurst exponent	0.83	0.89
Empirical Hurst exponent	0.82	0.90
Corrected empirical Hurst exponent	0.81	0.89
Theoretical Hurst exponent	0.53	0.52

increase in the rate of sea ice melting between 1988 and 2018 becomes evident, an outcome attributed to the phenomenon of melting SIE.

Moving on to Figure 4, a visualization of the projected and observed trajectories within the test dataset spanning from 2010 to 2019 is presented. Root Mean Square Error (RMSE) values are computed for both the training and test datasets, yielding an RMSE value of 0.36 for the training set and 0.41 for the test set. This underscores the model's robust capacity for generalization in out-of-sample scenarios, as indicated by the high R-squared value of 0.9865. The time series plot for SIER and SSTR is portrayed in Figure 5. Notably, both processes exhibit the characteristic of being mean zero stationary, which is a fundamental requirement for conducting the Granger causal test.

Figure 6 and Table 2 jointly demonstrate the presence of long memory in SIER and SSTR. This inference is substantiated by the significantly elevated Hurst exponent values, exceeding the threshold of 0.5, thus reflecting the underlying memory dynamics in these processes. The correlation matrix for NAO, SIER, and SSTR is presented in Table 3, calculated over a span of 38 years from January 1982 to September 2019. Enclosed within parentheses are the associated p-values, which offer insights into the statistical significance of these correlations. Particularly noteworthy is the robust and statistically significant correlation between SSTR and NAO. Furthermore, a strong correlation is evident between SIER and SSTR. However, it is important to note that the correlation between NAO and SIER appears to be relatively weaker in significance.

Table 3: Over a span of 38 years, from January 1982 to September 2019, the correlation matrix for NAO, SIER, and SSTR is examined. The accompanying p-values enclosed in parentheses provide insights into the significance of the correlations. Notably, the correlation between SSTR and NAO is statistically significant. Similarly, a robust correlation is observed between SIER and SSTR. However, the correlation between NAO and SIER exhibits a relatively weak significance

	NAO	**SIER**	**SSTR**
NAO	1.000	0.016 (0.063)	-0.133 ($< 2.2 * 10^{-16}$)
SIER		1.000	-0.173 ($< 2.2 * 10^{-16}$)
SSTR			1.000

4.2. Granger causal test

We examine the positive feedback loop. The Granger causal models are formulated with null and alternative hypotheses, as depicted in Equations (6) and (7). The null hypothesis asserts that all γ_i coefficients are equal to zero, thereby establishing $H_0 : \gamma_1 = \cdots = \gamma_k = 0$. In contrast, the alternative hypothesis H_1 aims to reject this by examining whether at least one γ_i deviates from zero. Table 4 presents the ANOVA F-test outcomes for the Granger causal models outlined in Equations (3), (4), and (5).

The ANOVA F-test (p-value = 0.0178) effectively refutes the null hypothesis, indicating that SSTR and SIER indeed exert a Granger causal influence on NAO. Similarly, the ANOVA F-test (p-value = 2.16×10^{-6}) dismisses the notion that NAO and SIER lack a Granger causal impact on SSTR. Likewise, the ANOVA F-test (p-value = 2.17×10^{-10}) rejects the null hypothesis that NAO and SSTR do not possess a Granger causal effect on SIER. These compelling outcomes collectively confirm the existence of a feedback loop connecting SIER, SSTR, and NAO.

Continuing the analysis, we employ an Akaike information criterion-based model selection process, and we identify the optimal configuration for the model Equation (17). Notably, if the model coefficients were static, the Granger causal model would represent a special case of Equation (17).

The synthesis of these revelations underscores the presence of a reciprocal feedback loop among NAO, SIER, and SSTR. Subsequently, we move forward to demonstrate the affirmative nature of this loop. Emphasizing the skewness of NAO in Table 5, we offer insight into the bootstrap confidence intervals (C.I.) across different time intervals—daily, weekly, and monthly. In a scenario of stable NAO, a skewness value of zero is expected. However, our findings unveil a negatively skewed distribution, indicating a statistically significant departure from stability. This pronounced outcome vividly underscores the prevailing instability within the NAO dynamics.

Together, these analyses substantiate the existence of a complex feedback loop among NAO, SIER, and SSTR. This discovery not only expands our understanding of climate interdependencies but also reveals a distinctive form of instability inherent within the North Atlantic system.

Table 4: F-values and p-values of different combinations of Granger causal models. Small p-values indicate that there is a feedback loop among NAO, SIER, and SSTR

GC models	F-value	p-value
SSTR + SIER → NAO	2.31	0.0178
NAO + SIER → SSTR	5.546	2.16×10^{-6}
NAO + SSTR → SIER	7.714	2.27×10^{-10}

Table 5: The skewness of NAO along with bootstrap-derived confidence intervals (C.I.) across daily, weekly, and monthly time spans. While an anticipated stable NAO would exhibit a skewness of zero, our observations indicate a negatively skewed distribution. This statistically significant result underscores the presence of instability in NAO

Period	Skewness	C.I.
Daily	-.210	[-0.242, -0.169]
Weekly	-.213	[-0.305, -0.107]
Monthly	-.194	[-0.368, -0.005]

4.3. Dynamic statistical approach

In Figure 7(a), the diminishing trajectory of the dynamic intercept $\beta_0(t)$ for SIE signifies a gradual decline in SIE over time. Correspondingly, Figure 7(b) highlights the ascending trend of the dynamic intercept $\beta_0(t)$ for SST, indicating a progressive increase in SST. Transitioning to Figure 7(c), the dynamic intercept $\beta_0(t)$ for NAO represents a mean-zero stationary process akin to NAO(t).

Further exploration unveils Figure 7(d), illustrating the dynamic coefficient $\beta_1(t)$ in harmony with $\sin \omega t$ for SIE(t). Similarly, Figure 7(e) elucidates the dynamic coefficient $\beta_1(t)$ corresponding to $\sin \omega t$ for SST(t). In Figure 7(f), the portrayal of the dynamic coefficient $\beta_1(t)$ pertains to $NAO(t-1)$ in relation to $NAO(t)$.

Additionally, Figures 7(g), 7(h), and 7(i) provide insights into the dynamic coefficients $\gamma_1(t)$ associated with $\cos \omega t$ for SIE(t), SST(t), and NAO(t), respectively.

Common criticisms of traditional statistical models often underscore the assumption of static relationships between predictors and dependent variables over time (Kolstad and Screen, 2019). To counter this limitation, our study adopts a BDLM as detailed in Section 3.4, drawing inspiration from previous works i.e, Petris *et al.* (2009); Migon *et al.* (2010). This innovative approach ensures an adaptive update of the predictor-dependent variable relationship as time unfolds (see Figure 7).

Furthermore, aligning the findings of Kolstad and Screen (2019), our study substantiates a lagged correlation between NAO and SIE within the Barents-Kara Sea. Remarkably, our research extends beyond this region to encompass the broader North Atlantic and Arctic area, thereby broadening the scope of the observed relationship.

Overall, our methodology not only addresses the limitations of conventional statistical models but also enriches our understanding of the evolving relationships between key climate variables within the dynamic context of the North Atlantic region.

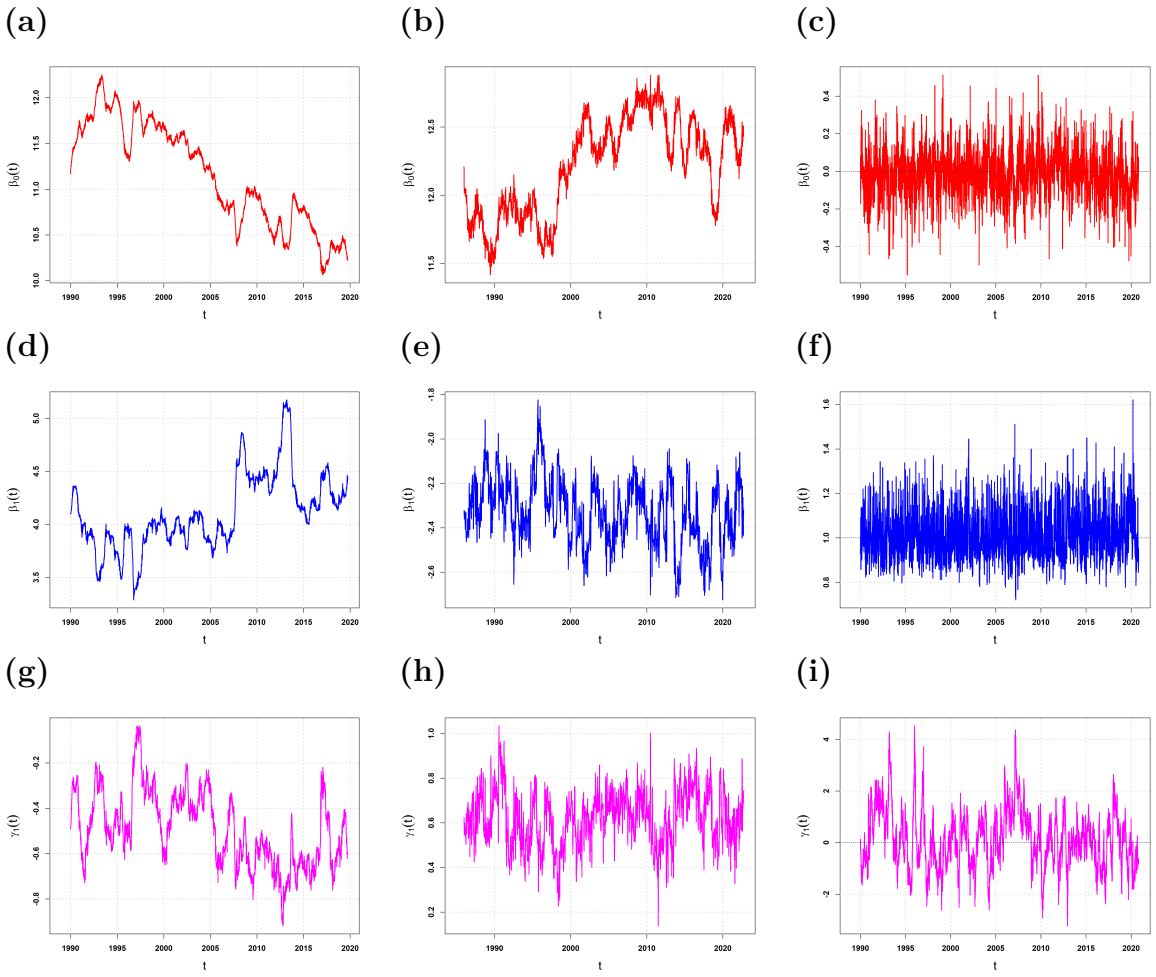


Figure 7: (a) The downward trend in the dynamic intercept $\beta_0(t)$ of SIE indicates that the SIE is shrinking over time. (b) The upward trend in the dynamic intercept $\beta_0(t)$ of SST indicates an increasing trend in SST. (c) The dynamic intercept $\beta_0(t)$ of NAO is the mean-zero stationary process like NAO(t). (d) The dynamic coefficient $\beta_1(t)$ corresponding to $\sin \omega t$ for SIE(t). (e) The dynamic coefficient $\beta_1(t)$ corresponding to $\sin \omega t$ for SST(t). (f) The dynamic coefficient $\beta_1(t)$ corresponding to $NAO(t-1)$ for $NAO(t)$. (g) The dynamic coefficient $\gamma_1(t)$ corresponding to $\cos \omega t$ for SIE(t). (h) The dynamic coefficient $\gamma_1(t)$ corresponding to $\cos \omega t$ for SST(t). (i) The dynamic coefficient $\gamma_1(t)$ corresponding to SST(t) for NAO(t)

5. Concluding remarks

In conclusion, this study looked into the intricate dynamics and identified critical instability driven by positive feedback loops among three pivotal climate variables: melting SIE, rising SST, and the NAO. Employing a generic approach rooted in statistical machine

learning, we pursued a comprehensive analysis that offers valuable insights into climate variability and its implications for the North Atlantic region. Unlike intricate climate forecast models, our methodology embraced a less computationally intensive, yet more universal strategy, facilitating an all-encompassing examination across the vast North Atlantic area. Addressing a central critique highlighted by Kolstad and Screen (2019), our BDLM dynamically updated the predictor-dependent variable relationship, enabling us to overcome static model limitations.

The study's noteworthy revelations are: (i) a mutual Granger causality between SIE and SST, (ii) a mutual Granger causality between SST and NAO, and (iii) an anti-correlation between SST and NAO. This anti-correlation implies that the increasing SST trend is likely to trigger increased occurrences of negative NAO. This aligns with our intriguing finding that the NAO index exhibits negative skewness at various time scales (daily, weekly, and monthly), contrary to its expected mean-zero stationary behavior. Importantly, the negative skewness of the NAO index, despite its mean-zero stationary nature, signals an impending critical instability. This unsettling phenomenon suggests an elevated probability of negative NAO occurrences, foretelling increased bouts of frigid climates in the North Atlantic region, particularly affecting northern Europe and eastern North America. This underscores the significance of this study in predicting a notable climate transformation.

Overall, this research contributes substantively to the understanding of critical instability within intricate climate systems. Leveraging techniques from statistical machine learning and data science for complex systems (Chakrabarti *et al.*, 2023), our study enhances our grasp of the dynamic interplay among vital climate variables, extending our insights into the intricate mechanisms that shape climate patterns.

Data availability statement

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Acknowledgments

We thank the anonymous reviewers for their comments and suggestions that helped improve the manuscript.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Bader, J., Mesquita, M. D., Hodges, K. I., Keenlyside, N., Østerhus, S., and Miles, M. (2011). A review on northern hemisphere sea-ice, storminess and the north Atlantic oscillation: Observations and projected changes. *Atmospheric Research*, **101**, 809–834.

Becker, G. A. and Pauly, M. (1996). Sea surface temperature changes in the North Sea and their causes. *ICES Journal of Marine Science*, **53**, 887–898.

Chakrabarti, A. S., Bakar, K. S., and Chakraborti, A. (2023). *Data Science for Complex Systems*. Cambridge University Press, Cambridge, UK.

Cressey, D. (2007). Arctic sea ice at record low. Nature. `https://doi.org/10.1038/news070917-3`.

Dai, G., Mu, M., and Wang, L. (2021). The influence of sudden Arctic sea-ice thinning on north Atlantic oscillation events. *Atmosphere-Ocean*, **59**, 39–52.

DallÓsto, M., Beddows, D. C. S., Tunved, P., Krejci, R., Ström, J., Hansson, H.-C., Yoon, Y. J., Park, K.-T., Becagli, S., Udisti, R., Onasch, T., OĎowd, C. D., Simó, R., and Harrison, R. M. (2017). Arctic sea ice melt leads to atmospheric new particle formation. *Scientific Reports*, **7**, 3318.

Das, P., Lahiri, A., and Das, S. (2018). Understanding sea ice melting via functional data analysis. *Current Science*, **115**, 920–929.

Das, S. and Dey, D. K. (2013). On dynamic generalized linear models with applications. *Methodology and Computing in Applied Probability*, **15**, 407–421.

Delworth, T. L., Zeng, F., Vecchi, G. A., Yang, X., Zhang, L., and Zhang, R. (2016). The North Atlantic Oscillation as a driver of rapid climate change in the Northern Hemisphere. *Nature Geoscience*, **9**, 509–512.

Eayrs, C., Holland, D., Francis, D., Wagner, T., Kumar, R., and Li, X. (2019). Understanding the seasonal cycle of Antarctic sea ice extent in the context of longer-term variability. *Reviews of Geophysics*, **57**, 1037–1064.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, **37**, 424–438.

Halloran, P. R., Hall, I. R., Menary, M., Reynolds, D. J., Scourse, J., A. Screen, J., Bozzo, A., Dunstone, N., Phipps, S., Schurer, A., P. Sueyoshi, T., Zhou, T., and Garry, F. (2020). Natural drivers of multidecadal Arctic sea ice variability over the last millennium. *Scientific Reports*, **10**, 688.

Horvath, S., Stroeve, J., Rajagopalan, B., and Jahn, A. (2021). Arctic sea ice melt onset favored by an atmospheric pressure pattern reminiscent of the north American-Eurasian Arctic pattern. *Climate Dynamics*, **57**, 1771–1787.

Hurrell, J. W. (1995). Decadal trends in the north Atlantic oscillation: Regional temperatures and precipitation. *Science*, **269**, 676–679.

Hurrell, J. W. and Deser, C. (2009). North atlantic climate variability: The role of the north Atlantic oscillation. *Journal of Marine Systems*, **78**, 28–41.

Kastens, K. A., Manduca, C. A., Cervato, C., Frodeman, R., Goodwin, C., Liben, L. S., Mogk, D. W., Spangler, T. C., Stillings, N. A., and Titus, S. (2009). How geoscientists think and learn. *Eos, Transactions American Geophysical Union*, **90**, 265–266.

Kolstad, E. W. and Screen, J. A. (2019). Nonstationary relationship between autumn Arctic sea ice and the winter north Atlantic oscillation. *Geophysical Research Letters*, **46**, 7583–7591.

Kvamstø, N. G., Skeie, P., and Stephenson, D. B. (2004). Impact of Labrador sea-ice extent on the north Atlantic oscillation. *International Journal of Climatology*, **24**, 603–612.

Kwok, R. (2000). Recent changes in Arctic ocean sea ice motion associated with the north Atlantic oscillation. *Geophysical Research Letters*, **27**, 775–778.

Lindsey, R. and Dahlman, L. (2009). Climate variability: North Atlantic oscillation. NOAA:Climate.gov `https://www.climate.gov/news-features/understanding-climate/climate-variability-north-atlantic-oscillation`.

Miettinen, A., Koç, N., Hall, I. R., Godtliebsen, F., and Divine, D. (2011). North atlantic sea surface temperatures and their relation to the north Atlantic oscillation during the last 230 years. *Climate Dynamics*, **36**, 533–543.

Migon, H. S., Petris, G.and Petrone, S., and Campagnoli, P. (2010). Dynamic linear models with R. *Biometrics*, **66**, 1311–1312.

NOAA (2020a). Daily NAO index since January 1950. Data can be downloaded from: `https://nsidc.org/data/G02135/versions/3`.

NOAA (2020b). NOAA optimum interpolation (OI) sea surface temperature (SST) v2. Data can be downloaded from: `https://psl.noaa.gov/repository/entry/show?entryid=f8d470f4-a072-4c1e-809e-d6116a393818`.

NSIDC (2020). Daily sea ice extent. Data can be downloaded from: `https://psl.noaa.gov/repository/entry/show?entryid=f8d470f4-a072-4c1e-809e-d6116a393818`.

Pan, L.-L. (2005). Observed positive feedback between the NAO and the north Atlantic SSTA tripole. *Geophysical Research Letters*, **32**, L06707.

Parkinson, L. C. (2000). Recent trend reversals in arctic sea ice extents: Possible connections to the north Atlantic oscillation. *Polar Geography*, **24**, 1–12.

Petris, G., Petrone, S., and Campagnoli, P. (2009). *Dynamic Linear Models with R*, volume 38, pages 31–84. Springer, New York, USA.

Richard, M. J. and Singpurwalla, N. D. (1983). Understanding the Kalman filter. *The American Statistician*, **37**, 123–127.

Rogers, J. C., Wang, S.-H., and Bromwich, D. H. (2004). On the role of the NAO in the recent northeastern Atlantic Arctic warming. *Geophysical Research Letters*, **31**, L02201.

Slonosky, V. C. and Yiou, P. (2001). The north Atlantic oscillation and its relationship with near surface temperature. *Geophysical Research Letters*, **28**, 807–810.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 267–288.

Warner, J. L. (2018). Arctic sea ice – a driver of the winter NAO? *Weather*, **73**, 307–310.

Yadav, A., Das, S., Bakar, K. S., and Chakraborti, A. (2023). Understanding the complex dynamics of climate change in south-west Australia using machine learning. *Physica A: Statistical Mechanics and its Applications*, **627**, 129139.

Yadav, A., Das, S., and Chakraborti, A. (2024). Untangling climate's complexity: Methodological insights. *Indian Journal of Theoretical Physics*, **71**, 9–29.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 24, No. 1, 2026 (New Series), pp 109–125
http://www.ssca.org.in/journal

Stress-Strength Reliability Analysis of Power Function and Nakagami Distributions using Comparative Sampling

Surinder Kumar, Rahul Shukla and Bhupendra Meena

Department of Statistics
School of Physical and Decision Sciences
Babasaheb Bhimrao Ambedkar University, Lucknow-226025

Received: 25 March 2025; Revised: 12 July 2025; Accepted: 15 July 2025

Abstract

In this article, we examine the probability of disaster in a stress-strength model where item strength follows the power function distribution and stress follows the Nakagami distribution. We compare simple random sampling (SRS) and ranked set sampling (RSS) approaches to assess their efficiency and accuracy in this context. Our study contributes to stress-strength modeling literature by introducing a novel application of ranked set sampling with Nakagami and Power function distributions. We also performed cost function optimization in this context.

Key words: Nakagami distribution; Power function distribution; Stress-strength model; Simple random sampling; Ranked set sampling.

AMS Subject Classifications: 62K05, 05B05

The video recording of the paper made under the SSCA's Online Lecture series is available at the Youtube channel URL `https://youtu.be/MRNbU0_F9mI`.

1. Introduction

The stress-strength paradigm stands as an essential methodology in reliability analysis, represented as $P = Pr(Y > X)$, where X and Y symbolize the stress and strength variables, respectively. This model is instrumental in evaluating the probability of system survival under varying operational loads. Another formulation articulates reliability as $P = Pr(X > \theta)$, where θ represents the upper threshold of the strength distribution. This metric is valuable in scenarios where system failure is contingent upon exceeding a critical strength limit.

In this research article, we investigate the reliability dynamics of a stress-strength model by examining the probability of failure through the measure $P = Pr(Y > X)$, where the stress follows the Nakagami and strength follows the Power function distributions respectively. Here, strength is taken as Power function distribution to justify the fact that

Corresponding Author: Rahul Shukla
Email: rahul.shukla.stats@gmail.com

the strength of the items is always finite. The related probability's of disaster, that occurs when $P(X > \theta)$ is studied. Our methodology encompasses both SRS and RSS approaches to derive reliability estimators, followed by a comprehensive comparative analysis of their statistical attributes, with particular emphasis on bias and mean squared error evaluation. Through rigorous mathematical formulation, we demonstrate the superior performance of RSS over conventional SRS in reliability estimation, particularly highlighting its enhanced precision and operational efficiency.

Furthermore, we develop an optimization framework for a linear cost function based on the established parametric relationships, yielding practical implications for manufacturing processes and quality control protocols. This investigation extends the existing stress-strength modeling literature by presenting an innovative implementation of ranked set sampling methodology in conjunction with Nakagami and Power function distributions. Our analytical framework offers new perspectives on reliability estimation in complex systems, potentially advancing the field of reliability engineering and quality assessment.

In the literature, early contributions to the field include the foundational studies by Basu (1964) and Barlow and Proschan (1967), which laid the groundwork for many subsequent investigations. The 1970's saw significant advancements, with notable works by Church and Harris (1970), Enis and Geisser (1971), Downton (1973), and Tong and *et al.* (1974) expanding the theoretical framework and practical applications of reliability analysis.

Further developments emerged in the late 1970's and early 1980's, with key contributions from Kelley *et al.* (1976), Sinha and Kale (1980), Sathe and Shah (1981), and Chao (1982). These researchers explored various aspects of stress-strength models and their statistical properties.

The field continued to evolve through the 1980's and 1990's, with important works by Awad and Gharraf (1986), Constantine *et al.* (1986), and Bain (2017). These studies further refined the methodologies and expanded the scope of reliability analysis.

More recent contributions, such as those by Chaturvedi and Sharma (2007), Kumar and Vaish (2014), and Kumar *et al.* (2020) have built upon this strong foundation, introducing new perspectives and addressing contemporary challenges in reliability engineering. This diverse body of literature reflects the ongoing importance and evolution of stress-strength modeling and reliability analysis in various engineering and statistical applications.

The seminal work of Meniconi and Barry (1996) provided compelling evidence for the efficacy of Power function distribution in electrical reliability analysis. Their comparative study, which evaluated multiple probability models including Exponential, Log-normal, and Weibull distributions, demonstrated through reliability measures and hazard function analysis that the Power function distribution offers superior modeling capabilities for electrical component reliability assessment.

This research presents a detailed exploration of stress-strength reliability analysis. The theoretical foundation is established in Section 3, where expressions for the probability of failure are formulated. In Section 4, the stress-strength reliability measure $P = Pr(Y > X)$ is derived for Power function and Nakagami distributions. Section 5 employs the maximum likelihood estimation method to estimate $P = Pr(Y > X)$ for both SRS and RSS. A simulation study is conducted in Section 6 to numerically validate the theoretical results.

Section 7 provides a comprehensive discussion of the findings. In Section 8, an illustrative example is presented to optimize the cost function for RSS in comparison to SRS. The article concludes with Section 9, summarizing the key insights and contributions of the study.

2. Preliminary

The mathematical formulation of the Nakagami distribution (NAD) is defined as follows:

$$f(x,\,\xi,\,\psi) = \frac{2}{\Gamma(\xi)}\left(\frac{\xi}{\psi}\right)^{\xi} x^{2\xi-1}\exp\left(-\frac{\xi}{\psi}\,x^2\right);\quad x>0,\psi>0,\xi>0.5 \tag{1}$$

$$F(x,\,\xi,\,\psi) = \frac{\gamma\left(\xi,\,\frac{\xi}{\psi}\,x^2\right)}{\Gamma(\xi)};\quad x>0,\psi>0,\xi>0.5 \tag{2}$$

where, $f(.)$ and $F(.)$ denotes the probability density function(pdf) and cumulative distribution function(cdf) respectively and $\gamma(x,a)$ is lower incomplete gamma function. The pdf $g(.)$ and cdf $G(.)$ of Power function distribution is taken as:

$$g(y,\,\mu,\,\theta) = \frac{\mu}{\theta}\left(\frac{y}{\theta}\right)^{\mu-1};\quad 0<y<\theta,\mu>0 \tag{3}$$

$$G(y,\,\mu,\,\theta) = \left(\frac{y}{\theta}\right)^{\mu};\quad 0<y<\theta,\mu>0 \tag{4}$$

3. Disaster probability ($\boldsymbol{\alpha = P(X > \theta)}$)

Theorem 1: If X follows the Nakagami distribution (1) and Y follows the power function distribution (3), then the parameter α is given by the expression:

$$\alpha = \frac{1}{\sqrt{\xi}}\Gamma\left[\xi,\,\frac{\xi}{\beta}\right] \tag{5}$$

where $\beta = \dfrac{\psi}{\theta^2}$

Proof: As we know

$$\begin{aligned}\alpha &= P(X>\theta)\\ &= \int_{\theta}^{\infty}\frac{2}{\Gamma(\xi)}\left(\frac{\xi}{\psi}\right)^{\xi} x^{2\xi-1}\exp\left(-\frac{\xi}{\psi}\,x^2\right)dx\end{aligned} \tag{6}$$

on taking $\dfrac{\xi}{\psi}x^2 = u$ in Eq. (6), we get

$$\begin{aligned}&= \frac{1}{\sqrt{\xi}}\int_{\frac{\xi\theta^2}{\psi}}^{\infty} u^{\xi-1}e^{-u}du\\ \alpha &= \frac{1}{\sqrt{\xi}}\Gamma\left[\xi,\,\frac{\xi}{\beta}\right]\end{aligned}$$

where, $\beta = \dfrac{\psi}{\theta^2}$ and $\Gamma[x, a]$ is upper incomplete gamma function.
Hence, the theorem follows.

3.1. Numerical investigation of disaster risk probability (α)

Numerical evaluation of the probability measure $\alpha = P(X > \theta)$, derived from expression (5), exhibits systematic variations across different parametric combinations of β and ξ, as presented in Table 1. The analytical results reveal a direct correlation between the disaster probability and parameter β. In the context of stress-strength modeling, where system failure is characterized by $X > \theta$ [Alam and Roohi (2003)], this relationship provides crucial insights for system reliability optimization.

Table 1: Numerical values for probability of disaster $\alpha = P(X > \theta)$

β	ξ=0.6	ξ=1.5	ξ=2	ξ=2.6	ξ=3.5
2	0.985141	0.493692	0.520260	0.694699	1.483699
1.5	0.846234	0.414194	0.434913	0.580485	1.244489
0.2	0.037432	0.001315	0.000353	0.000096	0.000020
0.1	0.001476	0.000001	0.000000	0.000000	0.000000
0.05	0.000003	0.000000	0.000000	0.000000	0.000000

Alternatively, we may also obtain the numerical values of β for fixed ξ at different tolerance level α from Eq. (5). Further, these values are used to obtain the optimum cost for manufacturing of item at desired tolerance level.

Table 2: Values of β at different tolerance level α

	ξ=2.5				
α	0.1	0.05	0.01	0.005	0.001
β	0.570651	0.470841	0.340964	0.306055	0.248595

4. Stress-strength reliability $P = Pr(Y > X)$ for nakagami and power function distributions

Theorem 2: Let stress (X) and strength (Y) be distributed according to $f(x, \xi, \psi)$ and $g(y, \mu, \theta)$, respectively. The probability $P = Pr(Y > X)$ is formulated as

$$P = \frac{1}{\Gamma\xi}\left[\gamma\left(\xi, \frac{\xi}{\beta}\right) - \left(\frac{\beta}{\xi}\right)^{\mu/2}\gamma\left(\xi + \frac{\mu}{2}, \frac{\xi}{\beta}\right)\right] \tag{7}$$

Proof: According to stress-strength reliability

$$P = \int_{x=0}^{\theta}\int_{y=x}^{\theta} f(x, \xi, \psi)\, g(y, \mu, \theta)\, dy\, dx \tag{8}$$

Substituting $y = vx$ in Eq. (8),we get

$$P = \int_{x=0}^{\theta} \int_{y=x}^{\theta/x} f(x,\, \xi,\, \psi)\, g(vx,\, \mu,\, \theta)\, dv\, dx$$
$$= \frac{1}{\theta^{\mu}} \frac{2}{\Gamma(\xi)} \left(\frac{\xi}{\psi}\right)^{\xi} \int_0^{\sqrt{\psi/\beta}} x^{2\xi+\mu-1} \exp\left(-\frac{\xi}{\psi}\, x^2\right) \left[\left(\frac{\sqrt{\psi/\beta}}{x}\right)^{\mu} - 1\right] dx \tag{9}$$

on taking $\frac{\xi}{\psi} x^2 = u$, in Eq. (9), we get

$$P = \frac{1}{\theta^{\mu}} \frac{2}{\Gamma(\xi)} \left(\frac{\xi}{\psi}\right)^{\xi} \left[\frac{1}{2} \left(\frac{\psi}{\xi}\right)^{\xi} \int_0^{\xi/\beta} u^{\xi-1} e^{-u}\, du - \frac{1}{2} \left(\frac{\psi}{\xi}\right)^{\xi+\mu/2} \right.$$
$$\left. \int_0^{\xi/\beta} u^{\xi+\mu/2-1} e^{-u}\, du \right]$$
$$= \frac{1}{\theta^{\mu}\Gamma\xi} \left[\left(\frac{\psi}{\beta}\right)^{\mu/2} \int_0^{\xi/\beta} u^{\xi-1} e^{-u}\, du - \left(\frac{\psi}{\xi}\right)^{\mu/2} \int_0^{\xi/\beta} u^{\xi+\mu/2-1} e^{-u} du\right]$$

Putting $\theta = \left(\frac{\psi}{\beta}\right)^{1/2}$

$$P = \frac{1}{\Gamma\xi} \left[\gamma\left(\xi, \frac{\xi}{\beta}\right) - \left(\frac{\beta}{\xi}\right)^{\mu/2} \gamma\left(\xi + \frac{\mu}{2}, \frac{\xi}{\beta}\right)\right]$$

where, $\gamma(a, x) = \int_0^x t^{a-1} e^{-t} dt$ is lower incomplete gamma function.
Hence, the theorem follows.

Table 3: The Stress-strength reliability of an item for fixed value of $\xi = 2.5$

$\downarrow\mu \quad \beta\rightarrow$	0.001	0.01	0.1	1	5	10
0.1	0.299133	0.213614	0.117660	0.018746	0.000812	0.000162
1.0	0.969910	0.904847	0.699099	0.150043	0.006822	0.001373
1.5	0.994576	0.969498	0.828477	0.202057	0.009395	0.001896
2.5	0.999812	0.996653	0.940481	0.278586	0.013449	0.002726
5.0	1.000000	0.999982	0.994223	0.385308	0.019856	0.004057

5. Likelihood maximization for parameter estimation

5.1. In the case of simple random sampling (SRS)

Theorem 3: For a simple random sample x_1, x_2,...,x_n from Nakagami distribution having pdf Eq.(1), the maximum likelihood estimate (MLE) of ξ and ψ is

$$\hat{\psi}_{srs} = \frac{\sum_{i=1}^{n} x_i^2}{n} \qquad \text{and} \qquad \hat{\xi}_{srs} = \text{From Eq.(13)}$$

Proof: If we take a random sample x_1, x_2,...,x_n from the Nakagami (ξ, ψ) of size n, then the likelihood function of the Nakagami distribution NAD (ξ, ψ) is given by

$$L(x,\, \xi,\, \psi) = \frac{(2\xi^{\xi})^n}{(\Gamma\xi)^n(\psi^{\xi})^n} \prod_{i=1}^{n}(x_i)^{(2\xi-1)} \exp\left(-\frac{\xi}{\psi}\sum_{i=1}^{n} x_i^2\right) \tag{10}$$

We can formulate the likelihood function as follows

$$\log\, L = n\log\, 2 - n\log\Gamma\xi + n\xi\log\,\xi - n\xi\log\psi + (2\xi-1)\sum_{i=1}^{n}\log\, x_i - \frac{\xi}{\psi}\sum_{i=1}^{n} x_i^2 \tag{11}$$

Differentiating the log likelihood function of NAD (ξ, ψ) given in Eq. (11) with respect to ψ in the case when ξ is known and with respect to ξ in the case when ψ is known. Then equating the resulting equations equal to zero, we get

$$\hat{\psi}_{srs} = \frac{\sum_{i=1}^{n} x_i^2}{n} \tag{12}$$

and

$$\frac{\partial logL}{\partial\xi} = \log\xi - \left(\frac{\partial}{\partial\xi}\log\Gamma\xi\right) - \log\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) + \frac{1}{n}\sum_{i=1}^{n}\log x_i^2 = 0 \tag{13}$$

ML estimator of ξ can be obtained from Eq. (13) by using Newton-Raphson method because Eq. (13) does not yield a closed-form solution.

Theorem 4: For a simple random sample y_1, y_2,...,y_n from Power function distribution having pdf Eq. (3), the MLE's of θ and μ are given respectively

$$\hat{\theta}_{srs} = y_{(n)}$$

and

$$\hat{\mu}_{srs} = \frac{n}{n\log\, y_{(n)} - \sum_{i=1}^{n}\log\, y_i}$$

where, $y_{(n)}$ is n^{th} order statistics.

Proof: The log likelihood function is given as:

$$\log L = n\,(\log\mu - \mu\log\theta) + (\mu-1)\sum_{i=1}^{n}\log y_i \tag{14}$$

The first-order condition reveals that n^{th} order statistic is MLE of θ i.e.

$$\hat{\theta}_{srs} = y_{(n)} \tag{15}$$

Differentiating Eq.(14) with respect to μ and equating the resulting equation to zero, we get the MLE of μ i.e.

$$\hat{\mu}_{srs} = \frac{n}{n \log y_{(n)} - \sum_{i=1}^{n} \log y_i} \tag{16}$$

Theorem 5: If $x_1, x_2,\ldots,x_n$ is a simple random sample from Nakagami distribution and $y_1, y_2,\ldots,y_m$ is a simple random sample from Power function distribution then the MLE of stress-strength reliability $P = Pr(Y > X)$ is given by

$$\hat{P}_{(srs)} = \frac{1}{\Gamma\hat{\xi}_{srs}} \left[\gamma\left(\hat{\xi}_{srs}, \frac{\hat{\xi}_{srs}}{\hat{\beta}_{srs}}\right) - \left(\frac{\hat{\beta}_{srs}}{\hat{\xi}_{srs}}\right)^{\hat{\mu}_{srs}/2} \gamma\left(\hat{\xi}_{srs} + \frac{\hat{\mu}_{srs}}{2}, \frac{\hat{\xi}_{srs}}{\hat{\beta}_{srs}}\right) \right] \tag{17}$$

Proof: Let us take a simple random sample $x_1, x_2,\ldots,x_n$ from Nakagami distribution with parameters (ξ, ψ) and $y_1, y_2,\ldots,y_m$ is a simple random sample from Power function distribution with parameters (μ, θ). The MLE's of ψ, ξ, θ and μ can be obtained by using Theorem (3) and (4), respectively. Since ML estimation procedure holds the in-variance property. Hence, on using Theorem (3) and (4) and in-variance property of ML estimator, the ML estimate $\hat{\beta}_{srs}$ is given by

$$\hat{\beta}_{srs} = \frac{\hat{\psi}_{srs}}{\hat{\theta}^2_{srs}}$$

Using the same fact the MLE of stress-strength reliability defined in Eq. (7) is given by

$$\hat{P}_{(srs)} = \frac{1}{\Gamma\hat{\xi}_{srs}} \left[\gamma\left(\hat{\xi}_{srs}, \frac{\hat{\xi}_{srs}}{\hat{\beta}_{srs}}\right) - \left(\frac{\hat{\beta}_{srs}}{\hat{\xi}_{srs}}\right)^{\hat{\mu}_{srs}/2} \gamma\left(\hat{\xi}_{srs} + \frac{\hat{\mu}_{srs}}{2}, \frac{\hat{\xi}_{srs}}{\hat{\beta}_{srs}}\right) \right]$$

5.2. In the case of ranked set sampling (RSS)

Let us consider two ranked set samples from Nakagami distribution and Power function distribution respectively. The first sample, denoted as $x_{(ij)}$, has a size of $n_1 = r_1 m_1$, where i ranges from 1 to m_1, and j from 1 to r_1. Here, m_1 represents the set size, and r_1 the number of cycles. Similarly, the second sample, $y_{(pq)}$, has a size of $n_2 = r_2 m_2$, where k ranges from 1 to m_2, and l from 1 to r_2. In this case, m_2 is the set size, and r_2 the number of cycles. We can now express the PDF's for $X_{(ij)}$ and $Y_{(pq)}$ as follows:

$$f_i(x_{ij}) = \frac{m_1!}{(i-1)!(m_1-i)!} \left[F_X(x)\right]^{i-1} \left[1 - F_X(x)\right]^{m_1-i} f(x_{ij}) \tag{18}$$

$$g_k(y_{pq}) = \frac{m_2!}{(k-1)!(m_2-k)!} \left[F_Y(y)\right]^{k-1} \left[1 - F_Y(y)\right]^{m_2-k} g(y_{pq}) \tag{19}$$

The likelihood function is given as

$$\begin{aligned} L &= \prod_{i=1}^{r_1}\prod_{j=1}^{m_1} f_i(x_{ij}) \prod_{k=1}^{r_2}\prod_{l=1}^{m_2} g_k(y_{pq}) \\ &= \prod_{i=1}^{r_1}\prod_{j=1}^{m_1} \frac{m_1!}{(i-1)!(m_1-i)!} \left[F_X(x)\right]^{i-1} \left[1-F_X(x)\right]^{m_1-i} f(x_{ij}) \\ &\quad \prod_{k=1}^{r_2}\prod_{l=1}^{m_2} \frac{m_2!}{(k-1)!(m_2-k)!} \left[F_Y(y)\right]^{k-1} \left[1-F_Y(y)\right]^{m_2-k} g(y_{pq}) \end{aligned}$$

Let $u = \dfrac{m_1!}{(i-1)!(m_1-i)!}$ and $v = \dfrac{m_2!}{(k-1)!(m_2-k)!}$, we get

$$\begin{aligned} &= u\frac{(2)^{n_1}}{(\Gamma\xi)^{n_1 m_1}} \left(\frac{\xi}{\psi}\right)^{n_1\xi} \prod_{i=1}^{r_1}\prod_{j=1}^{m_1} \left[\gamma\left(\xi, \frac{\xi}{\psi}x_{ij}^2\right)\right]^{i-1} \left[\Gamma\left(\xi, \frac{\xi}{\psi}x_{ij}^2\right)\right]^{m_1-i} \\ &\quad x^{2\xi-1} \exp\left(-\frac{\xi}{\psi}\, x_{ij}^2\right) v\,(\theta)^{-\mu n_2 m_2} \prod_{k=1}^{r_2}\prod_{l=1}^{m_2} y^{\mu k-1} \left(\theta^\mu - y^\mu\right)^{m_2-k} \end{aligned}$$

The log likelihood function is given by

$$\begin{aligned} \log L &= \log u - n_1 m_1 \log(\Gamma\xi) + n_1 \log 2 + n_1\xi\left(\log\frac{\xi}{\psi}\right) + (2\xi-1)\log x_{ij} \\ &\quad + \sum_{i=1}^{r_1}\sum_{j=1}^{m_1} (i-1)\log\left[\gamma\left(\xi, \frac{\xi}{\psi}x_{ij}^2\right)\right] + \sum_{i=1}^{r_1}\sum_{j=1}^{m_1} (m_1-i)\log\left[\Gamma\left(\xi, \frac{\xi}{\psi}x_{ij}^2\right)\right] \\ &\quad -\frac{\xi}{\psi}\sum_{i=1}^{r_1}\sum_{j=1}^{m_1} x_{ij}^2 + \log v - \mu n_2 m_2 \log\theta + \sum_{k=1}^{r_2}\sum_{l=1}^{m_2} (\mu k-1)\log y_{pq} \\ &\quad + \sum_{k=1}^{r_2}\sum_{l=1}^{m_2} (m_2-k)\log\left(\theta^\mu - y_{pq}^\mu\right) \end{aligned} \tag{20}$$

Partially differentiating Eq.(20) with respect to ξ and ψ respectively, we get

$$\begin{aligned} \frac{\partial \log L}{\partial\xi} &= n_1(\log\xi - 1) - n_1 m_1 \frac{\partial}{\partial\xi}\log\Gamma\xi - n_1\log\psi + 2\sum_{i=1}^{r_1}\sum_{j=1}^{m_1}\log x_{ij} \\ &\quad + \sum_{i=1}^{r_1}\sum_{j=1}^{m_1} (i-1)\frac{\partial}{\partial\xi}\log\left[\gamma\left(\xi, \frac{\xi}{\psi}x_{ij}^2\right)\right] - \frac{1}{\psi}\sum_{i=1}^{r_1}\sum_{j=1}^{m_1} x_{ij}^2 \\ &\quad + \sum_{i=1}^{r_1}\sum_{j=1}^{m_1} (m_1-i)\frac{\partial}{\partial\xi}\log\left[\Gamma\left(\xi, \frac{\xi}{\psi}x_{ij}^2\right)\right] \end{aligned} \tag{21}$$

and

$$\begin{aligned} \frac{\partial \log L}{\partial\psi} &= -\frac{n_1\xi}{\psi} + \sum_{i=1}^{r_1}\sum_{j=1}^{m_1} (i-1)\frac{\partial}{\partial\xi}\log\left[\gamma\left(\xi, \frac{\xi}{\psi}x_{ij}^2\right)\right] + \frac{\xi}{\psi^2}\sum_{i=1}^{r_1}\sum_{j=1}^{m_1} x_{ij}^2 \\ &\quad + \sum_{i=1}^{r_1}\sum_{j=1}^{m_1} (m_1-i)\frac{\partial}{\partial\xi}\log\left[\Gamma\left(\xi, \frac{\xi}{\psi}x_{ij}^2\right)\right] \end{aligned} \tag{22}$$

Partially differentiating Eq.(20) with respect to μ and θ respectively, we get

$$\frac{\partial \log L}{\partial \mu} = \sum_{k=1}^{r_2}\sum_{l=1}^{m_2} k \log y_{pq} + \sum_{k=1}^{r_2}\sum_{l=1}^{m_2} (m_2 - k)\left[\frac{\theta^{\mu}\log\mu - y_{pq}^{\mu}\log y_{pq}}{\theta^{\mu} - y_{pq}^{\mu}}\right] - n_2 m_2 \log\theta \tag{23}$$

and

$$\frac{\partial \log L}{\partial \theta} = -\frac{\mu n_2 m_2}{\theta} + \sum_{k=1}^{r_2}\sum_{l=1}^{m_2} (m_2 - k)\frac{\mu\theta^{\mu-1}}{(\theta^{\mu} - y_{pq}^{\mu})} \tag{24}$$

The MLE's of ξ, ψ, μ and θ can be obtained from Eq. (21), (22), (23) and (24) by using the Newton-Raphson method, respectively, as the equations are not in closed form. We get the ML estimates as $\hat{\xi}_{rss}$, $\hat{\psi}_{rss}$, $\hat{\mu}_{rss}$ and $\hat{\theta}_{rss}$, respectively. By applying the in-variance property of maximum likelihood estimation, we can find the maximum likelihood estimate of the parameter 'β' and reliability parameter P based on RSS, denoted by $\hat{\beta}_{rsss}$ and $\hat{P}_{(rss)}$, respectively.

$$\hat{\beta}_{rss} = \frac{\hat{\psi}_{rss}}{\hat{\theta}_{rss}^2} \tag{25}$$

and

$$\hat{P}_{(rss)} = \frac{1}{\Gamma\hat{\xi}_{rss}}\left[\gamma\left(\hat{\xi}_{rss}, \frac{\hat{\xi}_{rss}}{\hat{\beta}_{rss}}\right) - \left(\frac{\hat{\beta}_{rss}}{\hat{\xi}_{rss}}\right)^{\hat{\mu}_{rss}/2}\gamma\left(\hat{\xi}_{rss} + \frac{\hat{\mu}_{rss}}{2}, \frac{\hat{\xi}_{rss}}{\hat{\beta}_{rss}}\right)\right] \tag{26}$$

6. Simulation study

In this section, the simulation studies are conducted to compare the performances using different sample datasets and different stress-strength dependent parameters. We mainly compare the performances of the ML estimates in terms of their biases and mean square errors (MSE) from the formula $\text{Bias}(\hat{P}) = E(\hat{P} - P)$ and $\text{MSE}(\hat{P}) = E(\hat{P} - P)^2$ respectively. In this particular case the stress variable is set to obey Nakagami distribution, the strength variable is set to obey Power function distribution. Based on the simple random sample of stress, the maximum likelihood estimations of ξ and ψ are obtained by using the R software through the Eq. (12) and (13) respectively. Similarly, based on the simple random sample of strength, the ML estimation of θ and μ are also obtained from Eq. (15) and (16), respectively. ML estimator of the combined parameter ($\beta = \frac{\psi}{\theta^2}$) is obtained from Eq. (25). The methodology to draw the ranked set sample from the population is given below:

1. A random subset of the population consisting of m^2 units is selected.
2. The m^2 units are then divided arbitrarily into m sets, each containing m units.
3. The units within each set are ranked based on either professional judgment or correlation with the variable of interest.

4. An individual quantile sample is constructed by taking the lowest ranked unit from the first set, the second lowest ranked unit from the second set, and continuing in this fashion.

5. To obtain a larger sample of size $n = r * m$, steps 1 through 4 can be repeated for r cycles.

The ranked set sampling method takes only one observation from each set in each cycle. In the first cycle, it chooses the lowest observation $x_{(11)r}$. In later cycles, it independently selects the second lowest $x_{(22)r}$ from a different set of m observations and the highest $x_{(mm)r}$ from the final set of m. Let $x_{(ii)k}, i = 1, 2, \ldots, m; k = 1, 2, \ldots, r$, be a ranked sample set with set size m and r cycles. For convenience, this paper will use the notation $x_{(i)r}$ in place of the full description.
Simulation steps are given below-

Step :1 We generate 1000 simple random samples of x_1, x_2,. . .,x_n, and y_1, y_2,. . .,y_m from Nakagami distribution and Power function distribution with the sample sizes of $(n_1, n_2) =$ (15, 15), (15, 20), (15, 25), (20, 20), (20, 25), (25, 25) in Case 1 and (20, 20), (20, 30), (20, 40), (30, 30), (30, 40), (40, 40) in Case 2, respectively.

Step :2 We generate 1000 ranked set samples of x_{11}, x_{22},. . .,$x_{m_1 r_1}$ and y_{11}, y_{22},. . .,$y_{m_2 r_2}$ from Nakagami distribution and from Power function distribution for the first case when the number of cycles is taken as $r_1 = r_2 = 5$ with set sizes $m_1 = m_2 = 3, 4, 5$ and for the second case when the number of cycles is taken as $r_1 = r_2 = 10$ with set sizes $m_1 = m_2 = 2, 3, 4$, respectively.

Step :3 To generate the SRS sample and RSS sample for the stress variable we consider the parametric value of Nakagami distribution as $\psi = 0.5$ and $\xi = (0.6, 1.5, 3.0)$. Similarly for the strength variable we take the parametric value of Power function distribution as $\theta = 2.23$ and $\mu = (0.5, 1.5, 3.0)$. We take the single values of the parameters ψ and θ to fix the parameter $\beta = 0.1$ which is equal to $\dfrac{\psi}{\theta^2} = \dfrac{0.5}{2.23^2} = 0.1$.

Step :4 The Biases, MSES and relative efficiency (RE) are presented in the Table 4.

Table 4: Biases, MSES and RE of P under SRS and RSS when the combined parameter $\beta = 0.1$ and $r_1 = r_2 = 5$

Case 1				SRS			RSS			
(ξ, μ)	(n_1, n_2)	(m_1, m_2)	P_{true}	$\hat{P}_{srs}$	Bias	MSE	$\hat{P}_{rss}$	Bias	MSE	RE
(0.6, 0.5)	(15,15)	(3,3)	0.5227125	0.591339	0.068626	0.010507	0.556335	0.033623	0.010297	1.0204
	(15,20)	(3,4)		0.588641	0.065928	0.009249	0.550926	0.028214	0.008555	1.0812
	(15,25)	(3,5)		0.592958	0.070245	0.008456	0.545796	0.023083	0.007778	1.0871
	(20,20)	(4,4)		0.592077	0.069364	0.008230	0.548540	0.025828	0.008066	1.0203
	(20,25)	(4,5)		0.593785	0.071073	0.007818	0.545833	0.023120	0.006213	1.2584
	(25,25)	(5,5)		0.593685	0.070973	0.007670	0.542317	0.019604	0.006451	1.1890
(1.5, 1.5)	(15,15)	(3,3)	0.8322674	0.856630	0.024362	0.005255	0.838484	0.006217	0.003951	1.3299
	(15,20)	(3,4)		0.849698	0.017430	0.004190	0.836275	0.004008	0.002382	1.7594
	(15,25)	(3,5)		0.846347	0.014080	0.003431	0.837606	0.005338	0.001767	1.9420
	(20,20)	(4,4)		0.852620	0.020353	0.003971	0.838037	0.005769	0.002269	1.7500
	(20,25)	(4,5)		0.848200	0.015933	0.003263	0.836344	0.004077	0.001616	2.0192
	(25,25)	(5,5)		0.849415	0.017147	0.003098	0.835449	0.003181	0.001590	1.9479
(3.0,3.0)	(15,15)	(3,3)	0.9646058	0.982819	0.018213	0.000861	0.970658	0.006052	0.000617	1.3948
	(15,20)	(3,4)		0.983515	0.018909	0.000686	0.968606	0.004000	0.000554	1.2388
	(15,25)	(3,5)		0.984642	0.020036	0.000622	0.968473	0.003867	0.000445	1.3987
	(20,20)	(4,4)		0.983449	0.018843	0.000700	0.968972	0.004366	0.000515	1.3595
	(20,25)	(4,5)		0.984943	0.020338	0.000637	0.968501	0.003896	0.000395	1.6099
	(25,25)	(5,5)		0.985938	0.021332	0.000648	0.967211	0.002606	0.000458	1.4141

Table 5: Biases, MSES and RE of P under SRS and RSS when the combined parameter $\beta = 0.1$ and $r_1 = r_2 = 10$

Case 2				SRS			RSS			
(ξ, μ)	(n_1, n_2)	(m_1, m_2)	P_{true}	$\hat{P}_{srs}$	Bias	MSE	$\hat{P}_{rss}$	Bias	MSE	RE
(0.6, 0.5)	(20,20)	(2,2)	0.522713	0.552325	0.029613	0.008021	0.530530	0.007817	0.005433	1.4763
	(20,30)	(2,3)		0.545121	0.022408	0.005856	0.534239	0.011527	0.003404	1.7204
	(20,40)	(2,4)		0.543142	0.020429	0.004903	0.533177	0.010465	0.002741	1.7885
	(30,30)	(3,3)		0.545193	0.022481	0.005481	0.538708	0.015995	0.003038	1.8042
	(30,40)	(3,4)		0.545384	0.022671	0.004278	0.535602	0.012890	0.002183	1.9590
	(40,40)	(4,4)		0.546516	0.023803	0.004058	0.535543	0.012830	0.001909	2.1260
(1.5, 1.5)	(20,20)	(2,2)	0.832267	0.850239	0.017972	0.003988	0.836635	0.004368	0.003431	1.1623
	(20,30)	(2,3)		0.845973	0.013705	0.002882	0.835628	0.003360	0.001993	1.4467
	(20,40)	(2,4)		0.842181	0.009914	0.002301	0.833979	0.001712	0.001265	1.8186
	(30,30)	(3,3)		0.843625	0.011358	0.002692	0.835259	0.002992	0.001631	1.6509
	(30,40)	(3,4)		0.840826	0.008558	0.002008	0.832193	-0.000074	0.001201	1.6725
	(40,40)	(4,4)		0.839773	0.007505	0.002061	0.836465	0.004197	0.000997	2.0671
(3.0,3.0)	(20,20)	(2,2)	0.964606	0.981699	0.017093	0.000835	0.969461	0.004855	0.000506	1.6491
	(20,30)	(2,3)		0.983959	0.019353	0.000638	0.967085	0.002479	0.000352	1.8115
	(20,40)	(2,4)		0.984990	0.020384	0.000569	0.967147	0.002541	0.000290	1.9611
	(30,30)	(3,3)		0.985509	0.020903	0.000650	0.966986	0.002381	0.000372	1.7469
	(30,40)	(3,4)		0.984823	0.020217	0.000569	0.966024	0.001418	0.000266	2.1379
	(40,40)	(4,4)		0.984142	0.019536	0.000535	0.967573	0.002968	0.000276	1.9361

Table (4) and (5) clearly indicates that the relative efficiency exceeds one in all instances; thus, it can be concluded that ranked set sampling demonstrates higher efficiency compared to simple random sampling for estimating stress-strength reliability.

7. Discussion

In manufacturing, when the strength of a device follows a Power function distribution, its maximum feasible operational value often has an upper threshold, denoted as θ_0. For instance, a turbine's rotational speed must not exceed its engineered safety limit to prevent mechanical failure. Suppose θ_α is the target operational value at a predefined tolerance α. If $\theta_\alpha < \theta_0$ manufacturers can derive the corresponding parameter μ_α (From Table 3) to design the item with a strength distribution defined by $(\mu_\alpha, \theta_\alpha)$ ensuring reliability. Conversely, if $\theta_\alpha > \theta_0$, adjustments to α or alternative designs become necessary. A real-world analogy is elevator systems: exceeding their maximum load capacity (θ_0) necessitates either reducing passenger limits (α) or upgrading components to meet safety standards.

8. An illustrative example

Without loss of generality, Let the ranking cost per unit be C_R and the measurement cost per unit be C_M, where typically $C_R < C_M$ since visual ordering or quick assessments are generally less expensive than precise measurements. For stress-strength models, we need to consider measurements for both stress (X) and strength (Y) components.
For a fixed budget B, consider these constraints:

- In RSS, we need to rank mx^2 units for stress and my^2 units for strength to obtain samples of sizes mx and my respectively.
- In SRS, we can directly measure $n * x$ units for stress and $n * y$ units for strength.
- The total cost must not exceed the budget B

The efficiency of RSS relative to SRS can be measured through the ratio of their respective mean squared errors (MSE) in estimating R. Let MSE$(\hat{P}_{rss})$ be the mean squared error of the RSS reliability estimator MSE$(\hat{P}_{srs})$ be the mean squared error of the SRS reliability estimator. The relative efficiency (RE) under perfect ranking is given by RE = MSE$(\hat{P}_{srs})$/MSE$(\hat{P}_{rss})$.

For stress-strength models with underlying distributions RE $\approx$ $[(mx + 1)(my + 1)]/4$.(Dell and Clutter (1972)) This shows that RSS can potentially provide greater efficiency gains for reliability estimation compared to mean estimation, as we benefit from improved estimation in both X and Y samples.

Cost functions

Let the total cost functions be:

- For RSS: $C_{rss} = (mx^2 * C_R + mx * C_M) + (my^2 * C_R + my * C_M)$
- For SRS: $C_{srs} = nx * C_M + ny * C_M$

Optimization problem

The optimization problem can be formulated as follows:
Minimize MSE($\hat{P}_{rss}$) subject to:

1. $(mx^2 * C_R + mx * C_M) + (my^2 * C_R + my * C_M) \leq B$
2. $mx, my \geq 2$ (to ensure meaningful ranking)
3. Cost ratio $r = C_M/C_R \geq 1$
4. Desired precision: MSE($\hat{P}_{rss}$) $\leq \epsilon$

For a reliability study, we consider a balanced design with ranking cost $C_R =$ ₹3, measurement cost $C_M =$ ₹15, total budget B = ₹2000, cost ratio $r = C_M/C_R = 5$, and desired reliability threshold $P(Y > X) \geq 0.99$. Using equal set sizes $mx = my = m$, the total units ranked are $2m^2$ (m^2 each for X and Y), with $2m$ total units measured (m each for X and Y). The number of cycles (k) is determined by the budget constraint through the total cost function $C_{RSS} = k(2m^2C_R + 2mC_M) \leq B$. Analysis is performed for set sizes $m = 2, 3, 4, 5$, calculating: maximum possible cycles (k) under budget, total sample sizes ($nX = nY = mk$), relative efficiency RE = $[(m + 1)(m + 1)]/4$, and total cost.

Set Size (m)	Cycle k	Sample size per component	RE	Total Cost
2	30	60	2.25	1980
3	15	45	4.00	1980
4	8	32	6.25	1824
5	5	25	9.00	1650

Detailed calculations for $m = 3$:

1. Cost per cycle = $(2 \times 3^2 \times$ ₹3) + $(2 \times 3 \times$ ₹15) = 54 + 90 = ₹144
2. Maximum cycles = floor(₹2000/₹144) = 15
3. Sample size per component = $3 \times 15 = 45$
4. RE = [(3+1)(3+1)]/4 = 16/4 = 4.0
5. Total cost = $15 \times$ ₹144 = ₹1980

Comparison with SRS Under the same budget (₹2000):

- Cost per unit = ₹15 (measurement only)
- Maximum total sample size = ₹2000/₹15 $\approx$ 133
- Sample size per component = 66

Reliability estimation

For m = 3 RSS design:

1. Variance reduction in each component:
   - $\text{Var}(\bar{X}_{rss}) = \text{Var}(\bar{X}_{srs})/4 = (4/45)/(4) = 0.0222$
   - $\text{Var}((\bar{Y}_{rss}) = \text{Var}(\bar{Y}_{srs})/4 = (4/45)/(4) = 0.0222$
2. Approximate variance of reliability estimator:
   - $\text{Var}(\hat{P}_{rss}) \approx 0.0156$ (using delta method)
   - $\text{Var}(\hat{P}_{srs}) \approx 0.0624$ (for equivalent SRS)
3. 95% Confidence interval for P using RSS: $\hat{P}_{rss} \pm 1.96\sqrt{0.0156} = \hat{P}_{rss} \pm 0.245$

Based on the calculations, $m = 3$ emerges as the optimal choice, offering 4 times the efficiency of SRS, maintaining adequate sample size (45 per component), and using the budget effectively (₹1980 of ₹2000). This choice is practically sound as ranking 9 units (3^2) at a time is manageable, with reasonable ranking error risks and sufficient sample size for normal approximations. While larger set sizes ($m = 4$, 5) offer higher theoretical efficiency, they become impractical due to increased ranking difficulty, higher error risks, smaller final sample sizes, and potential departure from asymptotic properties. Therefore, we recommend using $m = 3$ with 15 cycles to achieve reliable estimation of $P(Y > X)$, cost-effective budget utilization, manageable ranking requirements, and sufficient sample sizes for inference.

9. Conclusion

This study presents a comprehensive analysis of stress-strength reliability estimation using ranked set sampling (RSS) compared to simple random sampling (SRS). The investigation yields several significant findings. First, the maximum likelihood estimation under RSS demonstrates superior efficiency compared to SRS. This efficiency gain is particularly noteworthy in stress-strength applications where measurement costs are high. In examining finite strength cases within the stress-strength model, we derived explicit expressions for the disaster probability and analyzed its behavior under various parameter configurations. This analysis provides crucial insights for reliability engineers and quality control practitioners in assessing system failure risks. The numerical evaluations of α across different parameter values offer practical guidelines for system design and maintenance protocols. The optimization framework developed for cost-efficient implementation of RSS in stress-strength reliability estimation addresses the practical challenges of sampling design. By balancing statistical efficiency against economic constraints, we demonstrated that moderate set sizes (particularly $m = 3$) often provide the most practical solution, achieving four-fold efficiency gains while maintaining manageable ranking requirements and adequate sample sizes for inference. These findings have significant implications for reliability testing and quality control applications, particularly in scenarios where testing costs are substantial or destructive testing is required. The demonstrated efficiency gains of RSS over SRS suggest that its implementation could lead to considerable cost savings while maintaining or improving estimation precision in stress-strength reliability assessment.

Acknowledgments

We are indeed grateful to the Editors for their guidance and counsel. We are very grateful to the reviewer for valuable comments and suggestions of generously listing many useful references.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Alam, S. and Roohi (2003). On facing an exponential stress with strength having power function distribution. *Aligarh Journal of Statistics*, **23**, 57–63.

Awad, A. M. and Gharraf, M. K. (1986). Estimation of P (Y<X) in the Burr case: A comparative study. *Communications in Statistics-Simulation and Computation*, **15**, 389–403.

Bain, L. (2017). *Statistical Analysis of Reliability and Life-Testing Models: Theory and Methods*. Routledge,New York.

Barlow, R. E. and Proschan, F. (1967). Exponential life test procedures when the distribution has monotone failure rate. *Journal of the American Statistical Association*, **62**, 548–560.

Basu, A. (1964). Estimates of reliability for some distributions useful in life testing. *Technometrics*, **6**, 215–219.

Chao, A. (1982). On comparing estimators of Pr $\{Y < X\}$ in the exponential case. *IEEE Transactions on Reliability*, **31**, 389–392.

Chaturvedi, A. and Sharma, V. (2007). A family of inverse distributions and related estimation and testing procedures for the reliability function. *International Journal of Statistics and Management System*, **2**, 44–66.

Church, J. D. and Harris, B. (1970). The estimation of reliability from stress-strength relationships. *Technometrics*, **12**, 49–54.

Constantine, K., Tse, S.-K., and Karson, M. (1986). Estimation of P (Y< X) in the gamma case. *Communications in Statistics-Simulation and Computation*, **15**, 365–388.

Dell, T. and Clutter, J. (1972). Ranked set sampling theory with order statistics background. *Biometrics*, **28**, 545–555.

Downton, F. (1973). The estimation of Pr (Y< X) in the normal case. *Technometrics*, **15**, 551–558.

Enis, P. and Geisser, S. (1971). Estimation of the probability that Y< X. *Journal of the American Statistical Association*, **66**, 162–168.

Kelley, G. D., Kelley, J. A., and Schucany, W. (1976). Efficient estimation of P (Y< X) in the exponential case. *Technometrics*, **18**, 359–360.

Kumar, S., Singh, V., and Gautam, P. L. (2020). Study of disaster probability when strength follows power function distribution and stress follows odd generalised exponential Gompertz distribution(OGE-G). *Journal of Statistics Applications & Probability*, **9**, 585–594.

Kumar, S. and Vaish, M. (2014). On the estimation of R= P (Y> X) for a class of life-time distributions by transformation method. *Journal of Statistics Applications & Probability*, **3**, 369–378.

Meniconi, M. and Barry, D. (1996). The power function distribution: A useful and simple distribution to assess electrical component reliability. *Microelectronics Reliability*, **36**, 1207–1212.

Sathe, Y. and Shah, S. (1981). On estimating P (X>Y) for the exponential distribution. *Communications in Statistics-Theory and Methods*, **10**, 39–47.

Sinha, S. K. and Kale, B. K. (1980). *Life Testing and Reliability Estimation*. Wiley Eastern Limited,New Delhi.

Tong, H. and *et al.* (1974). A note on the estimation of P (Y<X) in the exponential case. *Technometrics*, **16**, 625.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 24, No. 1, 2026 (New Series), pp 127–153
http://www.ssca.org.in/journal

Zero-One-Inflated Poisson-Garima Distribution and its Applications in Biomedical Studies

Divya A.[1,2], Prasanth C. B.[3] and Muhammed Anvar P.[4]

[1]*Department of Statistics, Govt. Victoria College, Affiliated to University of Calicut, Palakkad, Kerala-678001*
[2]*Department of Statistics, St.Thomas college, Thrissur, Kerala-680001*
[3]*Department of Statistics, Sree Kerala Varma College, Thrissur, Kerala-680011*
[4]*Department of Statistical Sciences, Kannur University, Kannur, Kerala-670567*

Received: 13 December 2024; Revised: 21 July 2025; Accepted: 23 July 2025

Abstract

Count data appears in diverse field of study with a variety of patterns in certain frequencies such as excessive occurrence of zeros and ones compared to other possible values. Here a new zero-one-inflated Poisson-Garima distribution and its applications is studied. After introducing the modified Poisson-Garima distribution, its various statistical properties and estimation of parameters are discussed. The estimation methods are then illustrated with simulated samples and the proposed model was applied to two real data sets. It is demonstrated that the new model outperformed its competitors like zero-one-inflated Poisson distribution and zero-one-inflated Poisson-Lindley distribution.

Key words: Count data; EM algorithm; Garima distribution; Inflated distributions; Poisson mixture; Zero-one-inflated data.

AMS Subject Classifications: 62E10, 62E15, 62F10

1. Introduction

In any statistical data analysis, the choice of a suitable probability distribution has a significant role. In the case of discrete valued count data, the most adopted model is the standard Poisson distribution. This is due to its greater simplicity and its property of equi-dispersion. But in reality, most of the count data shows overdispersion and long tail behaviour. So, for modeling such overdispersed count data, many researchers have relied upon distributions such a Negative Binomial, Generalized Poisson etc. There are many reasons for over/under dispersion. Existence of excess/less number of zero counts is one of them. In such cases we use zero-inflated/deflated models (zero-modified models). Zero-inflated models are a mixture of two distributions of which one generates only zeros and the other generates non-negative counts from a baseline distribution. Thus, by taking mixtures of a degenerate distribution at zero and a non-negative count distribution such as Poisson,

Corresponding Authors: Divya A., Muhammed Anvar P.
Email: divy.312@gmail.com, anvarp@kannuruniv.ac.in

Negative Binomial, Geometric, Poisson-Lindley etc., we get a zero-inflated model. Many zero-inflated models are developed to address the issue of presence of an excess number of zeros. The zero-inflated Poisson (ZIP) regression model proposed by Lambert (1992) is one among them to handle count data with excess zeros.

In addition to the excessive occurrence of zeros alone, there are cases in which count data contain excess zeros and ones simultaneously. For example, virus infections occur for at most one time as after the first infection, antibodies will be generated and so there is less chance of further infection. In their unpublished manuscript Melkersson and Olsson (1999) expanded upon the modeling of count data characterized by both excess zeros and excess ones. They introduced an extension of the ZIP distribution, creating a novel distribution known as the zero-one-inflated Poisson distribution (ZOIP). Later, Zhang *et al.* (2016) conducted a comprehensive investigation of the ZOIP distribution, initially establishing five equivalent stochastic representations for the ZOIP random variable and then deduced its various essential distributional properties. The statistical inference for ZOIP models was explored by Tang *et al.* (2017). They addressed the inference problem by employing both maximum likelihood estimation and Bayesian estimation techniques. Liu *et al.* (2018) have done the estimation of ZOIP by using expectation-maximization (EM) algorithm to get maximum likelihood estimate (MLE) and Gibbs sampling to get samples from posterior distribution based on latent variables. Also they have compared the MLEs with Bayesian estimates by Monte Carlo simulation. Recently Tajuddin *et al.* (2022) introduced the zero–one-inflated Poisson–Lindley (ZOIPL) distribution as an alternative to ZOIP model. They studied the statistical properties and characteristics of this distribution. Altun *et al.* (2023) proposed a zero-inflated Poisson generalized Lindley regression model as an alternative to the zero-inflated negative-binomial regression model. Recently, more discrete distributions based on mixing of a Poisson rate parameter with new continuous models are emerging in the literature. For example, the Poisson-Garima (PG) distribution introduced by Shanker (2017) is one among them. It has been demonstrated that this model is capable of handling the over dispersion and heavy tailed-ness in a more effective way than that of previously discussed models.This model was extended to the case of modeling count data without zeros by Shanker and Shukla (2017), resulting in the zero-truncated PG distribution.

It is observed that in some situations, count data contains large values with non-negligible frequency, in addition to the presence of excessive zeros and/or ones. For example, in the monthly number of drug offenses recorded from January 1990 to December 2001, in Pittsburgh census tract 2206 (Garay *et al.* (2022)) which we analyze in a later section, the observed maximum count is 29 and the data is right skewed. The basic Poisson model, however, cannot effectively capture this pattern as for having higher frequency for large values, the rate parameter θ should be large. But when θ becomes large, it tends to a symmetric behaviour around θ in contrast to right tailed nature of the data, as we can see in the frequency pattern given in Figure 1a (See Appendix I). In Figure 1b (See Appendix I) the tail probabilities are plotted for both Poisson and PG distributions with different values of θ. For small values of θ, the tail probabilities of PG are higher than Poisson. Although the tail probabilities of the Poisson distribution become slightly larger for higher values of θ, the distribution tends to lose its right-skewness, which may limit its suitability for modeling right skewed data. This problem persists even when we take mixtures of zero/one inflated Poisson model. To model this pattern in a better way it would be beneficial to use PG distribution with mixtures of zeros and ones. In view of this, we aim to study the properties

of multiple values inflated PG model with special focus on zero inflation and simultaneous inflation of both zero and one. In this paper, a new zero-one-inflated PG distribution is introduced and studied its properties.

The paper is arranged as follows. Section 2 revisits the details of PG distribution and we introduce a modified version of PG distribution that is, zero-one-inflated PG (ZOIPG) distribution. Also its statistical properties and different methods of estimation are discussed. In the same section, a special case of ZOIPG is introduced, i.e., zero-inflated PG (ZIPG) distribution and its statistical properties and estimation details are discussed. Section 3 describes the details of simulation study. In section 4, we demonstrate the applications of the proposed models with three real data sets. Some concluding remarks are given in the last section.

2. Inflated Poisson-Garima distribution

2.1. Poisson-Garima distribution

The PG distribution was introduced by Shanker (2017) by compounding Poisson distribution with Garima distribution which was introduced by Shanker (2016). That is, when the parameter of the Poisson distribution follows a Garima distribution, we obtain PG distribution. The probability mass function (PMF) of PG distribution with parameter θ is

$$f(x) = \frac{\theta}{\theta+2}\frac{\theta x + (\theta^2+3\theta+1)}{(\theta+1)^{x+2}}; x = 0, 1, 2, \cdots ; \ \theta > 0. \tag{1}$$

The first four moments about origin of PG distribution are given by

$$\mu_1' = \frac{\theta+3}{\theta(\theta+2)},$$

$$\mu_2' = \frac{\theta^2+5\theta+8}{\theta^2(\theta+2)},$$

$$\mu_3' = \frac{\theta^3+9\theta^2+30\theta+30}{\theta^3(\theta+2)},$$

$$\mu_4' = \frac{\theta^4+17\theta^3+92\theta^2+204\theta+144}{\theta^4(\theta+2)}.$$

Shanker (2017) examined the properties of PG distribution, including its shape, moments, skewness, and kurtosis. Additionally, PG distribution exhibits an increasing hazard rate, making it a unimodal distribution. In the next section, we study ZOIPG distribution.

2.2. Zero-one-inflated Poisson-Garima distribution

Let V be a discrete random variable with non-negative integers as support having a PMF $g(.)$. A modification of the PMF g is required when excess frequencies at certain values of V are observed relative to the base distribution g. For example, a situation may arise when the expected number of occurrences of zeros and ones are larger than the baseline

distribution. A random variable X is said to follow a ZOIPG distribution with parameters p_1, p_2 and θ, if it admits the following stochastic representation.

$$X = V(1 - B_1) + B_1(1 - B_2), \tag{2}$$

where B_1 is a Bernoulli random variable with success probability p_1, $0 < p_1 < 1$, B_2 is another Bernoulli random variable with success probability p_2, $0 < p_2 < 1$, and V follows a PG distribution with rate parameter θ. Also B_1, B_2 and V are assumed to be mutually independent.
The relation between X and (B_1, B_2, V) can be understood by noticing the following expression,

$$\begin{cases} (X = 0) \Leftrightarrow (V = 0, B_1 = 0) \cup (B_1 = 1, B_2 = 1)\,, \\ (X = 1) \Leftrightarrow (V = 1, B_1 = 0) \cup (B_1 = 1, B_2 = 0)\,, \\ (X = k) \Leftrightarrow (V = k, B_1 = 0)\,. \end{cases} \tag{3}$$

Then the PMF of a ZOIPG random variable with parameters p_1, p_2 and θ, (ZOIPG(p_1, p_2, θ))is given by

$$P[X = k] = f(k) = \begin{cases} p_1p_2 + (1 - p_1)\frac{\theta}{\theta+2}\frac{(\theta^2+3\theta+1)}{(\theta+1)^2} & \text{for } k = 0 \\ p_1(1 - p_2) + (1 - p_1)\frac{\theta}{\theta+2}\frac{(\theta^2+4\theta+1)}{(\theta+1)^3} & \text{for } k = 1 \\ (1 - p_1)\frac{\theta}{\theta+2}\frac{(\theta^2+3\theta+1+\theta k)}{(\theta+1)^{k+2}} & \text{for } k \geq 2. \end{cases} \tag{4}$$

Remark: When $p_1 = 0$, this model reduces to PG distribution and when $p_2 = 1, 0 < p_1 < 1$, this model becomes ZIPG model.
Next, we discuss the properties of ZOIPG model.

Distribution function

Let X follows a ZOIPG(p_1, p_2, θ) distribution. Then its distribution function $F(x)$ is obtained as

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ p_1p_2 + K_0 & \text{for } 0 \leq x < 1 \\ p_1 + \sum_{r=0}^{\lfloor x \rfloor} K_r & \text{for } x \geq 1 \end{cases} \tag{5}$$

where $K_r = (1 - p_1)\frac{\theta}{\theta+2}\frac{(\theta^2+3\theta+1+\theta r)}{(\theta+1)^{r+2}}$, $r = 0, 1, 2, \ldots$ and $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x.

2.2.1. Moments and other related measures

The r^{th} factorial moment of a ZOIPG(p_1, p_2, θ) distribution is obtained as follows

$$\begin{aligned} \mu'_{(r)} &= E[X(X - 1)(X - 2)\ldots(X - r + 1)] \\ &= (1 - p_1)\frac{r!(\theta + r + 2)}{\theta^r(\theta + 2)}, \quad r \geq 2. \end{aligned}$$

For $r = 1$, we have $\mu'_{(1)}$ directly from the definition and we obtain it as

$$\mu'_{(1)} = p_1(1 - p_2) + (1 - p_1)\frac{\theta + 3}{\theta(\theta + 2)}. \tag{6}$$

For $r = 2$, we get

$$\mu'_{(2)} = (1 - p_1)\frac{2(\theta + 4)}{\theta^2(\theta + 2)}.$$

By using the relationship between factorial moments and moments about the origin we can get the first four moments about the origin as follows

$$\mu'_1 = p_1(1 - p_2) + (1 - p_1)\frac{\theta + 3}{\theta(\theta + 2)}, \tag{7}$$

$$\mu'_2 = p_1(1 - p_2) + (1 - p_1)\frac{\theta^2 + 5\theta + 8}{\theta^2(\theta + 2)}, \tag{8}$$

$$\mu'_3 = 5p_1(1 - p_2) + (1 - p_1)\frac{5\theta^3 + 21\theta^2 + 30\theta + 30}{\theta^3(\theta + 2)}, \tag{9}$$

$$\mu'_4 = 25p_1(1 - p_2) + (1 - p_1)\frac{25\theta^4 + 89\theta^3 + 92\theta^2 + 204\theta + 144}{\theta^4(\theta + 2)}. \tag{10}$$

2.2.2. Moment generating function

The Moment generating function (MGF) of ZOIPG(p_1, p_2, θ) can be obtained as

$$\begin{aligned} M_X(t) &= E(e^{tx}) \\ &= p_1p_2 + p_1(1 - p_2)e^t + (1 - p_1)M_Y(t) \\ &= p_1p_2 + p_1(1 - p_2)e^t + (1 - p_1)\frac{\theta^3 + (4 - e^t)\theta^2 + 2(2 - e^t)\theta + 1 - e^t}{(\theta + 1)(\theta + 2)(\theta + 1 - e^t)^2}, \end{aligned}$$

where $M_Y(t)$is the MGF of PG distribution.

2.2.3. Probability generating function

The probability generating function (PGF) of ZOIPG(p_1, p_2, θ) is obtained as

$$\begin{aligned} P_X(t) &= E(t^x) \\ &= p_1p_2 + p_1(1 - p_2)t + (1 - p_1)\frac{\theta^3 + (4 - t)\theta^2 + 2(2 - t)\theta + 1 - t}{(\theta + 1)(\theta + 2)(\theta + 1 - t)^2}. \end{aligned}$$

2.2.4. Estimation of parameters

In this section, we discuss the estimation methods for the ZOIPG distribution. The parameter vector to be estimated is denoted as $\Lambda = (p_1, p_2, \theta)'$. First we give details of method of moment (MOM) estimation and then discuss the ML estimation method. The mixture structure of ZOIPG makes it easy to handle the ML estimation method using the EM algorithm.

Method of moments estimation

Let $x_1, x_2, ...x_n$ be a observed sample of size n from ZOIPG(p_1, p_2, θ). Define

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}, \quad m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i(x_i - 1)$$

and

$$m_3 = \frac{1}{n}\sum_{i=1}^{n} x_i(x_i - 1)(x_i - 2).$$

The corresponding population moments are given by

$$E[X] = p_1(1 - p_2) + (1 - p_1)\frac{(\theta + 3)}{\theta(\theta + 2)},$$

$$E[X(X - 1)] = (1 - p_1)\frac{2(\theta + 4)}{\theta^2(\theta + 2)},$$

$$E[X(X - 1)(X - 2)] = (1 - p_1)\frac{6(\theta + 5)}{\theta^3(\theta + 2)}.$$

Equating the sample moments with the corresponding population moments and solving the equations, we get the MOMs of p_1, p_2 and θ as follows

$$\tilde{\theta} = \frac{(3m_2 - 4m_3) + \sqrt{(3m_2 + 4m_3)^2 + 12m_2m_3}}{2m_3},$$

$$\tilde{p_1} = \frac{2\tilde{\theta} + 8 - m_2\tilde{\theta}^3 - 2m_2\tilde{\theta}^2}{2(\tilde{\theta} + 4)},$$

$$\tilde{p_2} = \frac{\tilde{p_1} + (1 - \tilde{p_1})\frac{\tilde{\theta}+3}{\tilde{\theta}(\tilde{\theta}+2)} - m_1}{\tilde{p_1}}.$$

The MOM $\tilde{\theta}$ needs special attention while conducting numerical computations. We need to ensure the positivity of resulting value of the $\tilde{\theta}$ for a given sample. This is ensured by taking the positive square root of $(4m_3 + 3m_2)^2 + 12m_2m_3$. This is important because other estimators depend on this value of $\tilde{\theta}$. The details of the behaviour of $\tilde{\theta}$ is thus discussed in simulation study section. Next we describe the ML procedure.

Maximum likelihood estimation

Let $x = (x_1, x_2, ...x_n)$ be an observed sample of size n from ZOIPG(p_1, p_2, θ). Then the likelihood function is given by

$$\begin{aligned} L(p_1, p_2, \theta) &= \left[p_1 p_2 + (1-p_1)\frac{\theta(\theta^2+3\theta+1)}{(\theta+2)(\theta+1)^2}\right]^{S_0} \\ &\times \left[p_1(1-p_2) + (1-p_1\frac{\theta(\theta^2+4\theta+1)}{(\theta+2)\theta+1^3})\right]^{S_1} \\ &\times \left[\frac{(1-p_1)\theta}{(\theta+2)(\theta+1)^2}\right]^{n-S_0-S_1} \prod_{x_i\geq 2} \frac{\theta^2+3\theta+1+\theta x_i}{(\theta+1)^{x_i}} \\ &= {q_0}^{S_0}{q_1}^{S_1}\left[\frac{(1-p_0)\theta}{(\theta+2)(\theta+1)^2}\right]^{n-S_0-S_1} \prod_{x_i\geq 2} \frac{\theta^2+3\theta+1+\theta x_i}{(\theta+1)^{x_i}}, \end{aligned}$$

where S_0 and S_1 are the number of zeros and ones in the data respectively and $q_0 = P[X=0]$ and $q_1 = P[X=1]$.
Then the likelihood function becomes

$$L(q_0, q_1, \theta|x) = {q_0}^{S_0}{q_1}^{S_1}\left[\frac{(1-q_0-q_1)\theta(\theta+1)}{\theta^2+5\theta+2}\right]^{n-S_0-S_1} \prod_{x_i\geq 2} \frac{\theta^2+3\theta+1+\theta x_i}{(\theta+1)^{x_i}}.$$

Now the log likelihood function can be written as

$$\begin{aligned} \log L = (n-S_0-S_1)[\log(1-q_0-q_1) + \log\theta + \log(\theta+1) - \log(\theta^2+5\theta+2)] \\ + S_0\log q_0 + S_1\log q_1 + \sum_{x_i\geq 2}[\log(\theta^2+3\theta+1+\theta x_i) - x_i\log(\theta+1)]. \end{aligned}$$

Solving the likelihood equations $\frac{\partial \log L}{\partial q_0} = 0$ and $\frac{\partial \log L}{\partial q_1} = 0$ we get

$$\hat{q_0} = \frac{S_0}{n}$$

and

$$\hat{q_1} = \frac{S_1}{n}.$$

Then the MLE of θ, $\hat{\theta}$, is the solution of the non-linear equation

$$(n-S_0-S_1)\left[\frac{4\theta^2+4\theta+2}{\theta(\theta+1)(\theta^2+5\theta+2)}\right] + \sum_{x_i\geq 2}\left[\frac{2\theta+3+x_i}{\theta^2+3\theta+1+\theta x_i} - \frac{x_i}{\theta+1}\right] = 0,$$

which can be solved numerically by Newton-Raphson type iterative algorithm. Also note that the transformation between $(p_1, p_2, \theta)'$ and $(q_1, q_2, \theta)'$ is one to one. Thus the MLE of p_1 and p_2 are obtained as follows

$$\hat{p_1} = \frac{S_0+S_1}{n} - \frac{\hat{\theta}(\hat{\theta}^3+5\hat{\theta}^2+8\hat{\theta}+2)}{(\hat{\theta}+2)(\hat{\theta}+1)^3} \quad \text{and}$$

$$\hat{p_2} = \frac{\frac{S_0}{n} - (1-\hat{p_1})\frac{\hat{\theta}(\hat{\theta}^2+3\hat{\theta}+1)}{(\hat{\theta}+2)(\hat{\theta}+1)^2}}{\hat{p_1}}.$$

The finite sample properties of the above MLEs are studied using simulated data sets and are reported in Section 3. The special structure of the model in the form of mixture distribution makes it easier to do ML estimation by applying the EM algorithm. Next we describe the details of ML estimation using EM algorithm.

Maximum likelihood estimation with EM algorithm

For the iterative computation of MLE, we can use the EM algorithm. A ZOIPG random variable can be represented by three independent latent variables i.e., two Bernoulli variables B_1, B_2 and a PG random variable V. If we can observe these latent variables then the likelihood function of ZOIPG will be a product of three likelihood functions. Let $X = (X_1, X_2, ..., X_n)$ be a random sample from ZOIPG with corresponding latent variables $B_1 = (B_{11}, B_{12}, ..., B_{1n})$, $B_2 = (B_{21}, B_{22}, ..., B_{2n})$ and $V = (V_1, V_2, ..., V_n)$. Also we have

$$X_i = V_i(1 - B_{1i}) + B_{1i}(1 - B_{2i}) \tag{11}$$

So if we could observe B_{1i} and B_{2i} and hence V_i then by relation (11), the augmented likelihood function with augmented data (B_1, B_2, V) would be

$$\begin{aligned} L_c(p_1, p_2, \theta|X, B_1, B_2) = \prod_{i=1}^{n} p_1^{B_{1i}}(1-p_1)^{1-B_{1i}} \prod_{i=1}^{n} p_2^{B_{2i}}(1-p_2)^{1-B_{2i}} \\ \prod_{i=1}^{n}\left\{\frac{\theta(\theta X_i + \theta^2 + 3\theta + 1)}{(\theta+2)(\theta+1)^{X_i+2}}\right\}^{1-B_{1i}} \\ \triangleq L_c(p_1|X, B_1, B_2)L_c(p_2|X, B_1, B_2)L_c(\theta|X, B_1, B_2) \end{aligned} \tag{12}$$

The maximisation of $L_c(p_1, p_2, \theta|X, B_1, B_2)$ is carried out by maximizing $L_c(p_1|X, B_1, B_2)$, $L_c(p_2|X, B_1, B_2)$ and $L_c(\theta|X, B_1, B_2)$ separately. Thus the MLEs of p_1, p_2 and θ with EM algorithm can be found iteratively using two steps i.e. E step and M step respectively. In the E step, the expected values of B_{1i} and B_{2i} under the current estimates of (p_1, p_2, θ) are found. In the M step, maximization of $L_c(p_1, p_2, \theta|X, B_1, B_2)$ with the expected values of B_{1i} and B_{2i} from the E step, is done. The iteration is continued until the estimated values of p_1, p_2 and θ converge. The final iteration gives the MLEs of p_1, p_2 and θ, i.e. $\hat{p_1}_{EM}, \hat{p_2}_{EM}$ and $\hat{\theta_{EM}}$. The $(k+1)^{th}$ iteration of the EM algorithm is described as follows.

E step : Let $(p_1^{(k)}, p_2^{(k)}, \theta^{(k)})$ be the estimates obtained in the k^{th} step. Then

$$B_{1i}^{(k+1)} = E[B_{1i}|X_i, p_1^{(k)}, p_2^{(k)}, \theta^{(k)}] = \begin{cases} \frac{p_1^{(k)}p_2^{(k)}}{p_1^{(k)}p_2^{(k)} + (1-p_1^{(k)})\frac{\theta^{(k)}}{\theta^{(k)}+2}\frac{(\theta^{(k)})^2+3\theta^{(k)}+1)}{(\theta^{(k)}+1)^2}} & ; X_i = 0 \\ \frac{p_1^{(k)}(1-p_2^{(k)})}{p_1^{(k)}(1-p_2^{(k)}) + (1-p_1^{(k)})\frac{\theta^{(k)}}{\theta^{(k)}+2}\frac{(\theta^{(k)})^2+4\theta^{(k)}+1)}{(\theta^{(k)}+1)^3}} & ; X_i = 1 \\ 0 & ; X_i = 2, 3, ... \end{cases}$$

$$B_{2i}^{(k+1)} = E[B_{2i}|X_i, p_1^{(k)}, p_2^{(k)}, \theta^{(k)}] = \begin{cases} \frac{p_1^{(k)}p_2^{(k)} + (1-p_1^{(k)})p_2^{(k)} \frac{\theta^{(k)}}{\theta^{(k)}+2} \frac{{\theta^{(k)}}^2+3\theta^{(k)}+1)}{{(\theta^{(k)}+1)}^2}}{p_1^{(k)}p_2^{(k)} + (1-p_1^{(k)}) \frac{\theta^{(k)}}{\theta^{(k)}+2} \frac{{\theta^{(k)}}^2+3\theta^{(k)}+1)}{{(\theta^{(k)}+1)}^2}} & ; X_i = 0 \\ \frac{p_2^{(k)}(1-p_1^{(k)}) \frac{\theta^{(k)}}{\theta^{(k)}+2} \frac{{\theta^{(k)}}^2+4\theta^{(k)}+1)}{{(\theta^{(k)}+1)}^3}}{p_1^{(k)}(1-p_2^{(k)}) + (1-p_1^{(k)}) \frac{\theta^{(k)}}{\theta^{(k)}+2} \frac{{\theta^{(k)}}^2+4\theta^{(k)}+1)}{{(\theta^{(k)}+1)}^3}} & ; X_i = 1 \\ p_1^{(k)} & ; X_i = 2, 3, ... \end{cases}$$

M step:
Let

$$B_1^{(k+1)} = (B_{11}^{(k+1)}, B_{12}^{(k+1)}, ..., B_{1n}^{(k+1)})$$

and

$$B_2^{(k+1)} = (B_{21}^{(k+1)}, B_{22}^{(k+1)}, ..., B_{2n}^{(k+1)}).$$

Maximising $L_c(p_1|X, B_1^{(k+1)}, B_2^{(k+1)})$ and $L_c(p_2|X, B_1^{(k+1)}, B_2^{(k+1)})$ we get the MLE of p_1 and p_2 respectively as follows.

$$p_1^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} B_{1i}^{(k+1)}$$

and

$$p_2^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} B_{2i}^{(k+1)}.$$

$\theta^{(k+1)}$ is obtained by maximising $L_c(\theta|X, B_1^{(k+1)}, B_2^{(k+1)})$ and it is the solution of the equation

$$\sum_{i=1}^{n}(1 - B_{1i})\left\{\frac{1}{\theta} + \frac{X_i + 2\theta + 3}{\theta X_i + \theta^2 + 3\theta + 1} - \frac{1}{\theta + 2} - \frac{X_i + 2}{\theta + 1}\right\} = 0,$$

which can be solved numerically. Under the regularity conditions, it can be shown that the estimators are consistent and asymptotically normally distributed (See McLachlan and Krishnan (2008)).

2.3. Zero-inflated Poisson-Garima distribution

A special case of the ZOIPG distribution is the situation when $p_2 = 1$ and $0 < p_1 < 1$, that is, ZIPG. A random variable Y is said to have a ZIPG distribution with parameters p_1 and θ, (ZIPG(p_1, θ)) if the PMF of Y is specified as follows.

$$P[Y = k] = \begin{cases} p_1 + (1 - p_1)\frac{\theta}{\theta+2}\frac{(\theta^2+3\theta+1)}{(\theta+1)^2} & \text{for } k = 0 \\ (1 - p_1)\frac{\theta}{\theta+2}\frac{(\theta^2+3\theta+1+\theta k)}{(\theta+1)^{k+2}} & \text{for } k \geq 1. \end{cases} \tag{13}$$

2.3.1. Moments and other related measures

The r^{th} factorial moment of a random variable Y following ZIPG(p_1, θ) is obtained as follows

$$\begin{aligned}\mu'_{(r)} &= E[Y(Y-1)(Y-2)\dots(Y-r+1)] \\ &= (1-p_1)\frac{r!(\theta+r+2)}{\theta^r(\theta+2)}, \quad r \geq 1\end{aligned}$$

For $r = 1$, we get $\mu'_{(1)}$ as

$$\mu'_{(1)} = (1-p_1)\frac{\theta+3}{\theta(\theta+2)}.$$

For $r = 2$, we get

$$\mu'_{(2)} = (1-p_1)\frac{2(\theta+4)}{\theta^2(\theta+2)}.$$

By using the relationship between factorial moments and moments about the origin we can get the first four moments about the origin as

$$\mu'_1 = (1-p_1)\frac{\theta+3}{\theta(\theta+2)}, \qquad \mu'_2 = (1-p_1)\frac{\theta^2+5\theta+8}{\theta^2(\theta+2)},$$

$$\mu'_3 = (1-p_1)\frac{\theta^3+9\theta^2+30\theta+30}{\theta^3(\theta+2)}, \qquad \mu'_4 = (1-p_1)\frac{\theta^4+17\theta^3+92\theta^2+204\theta+144}{\theta^4(\theta+2)}.$$

Thus the mean and variance of ZIPG distribution can be obtained as

$$\mu_Y = (1-p_1)\frac{\theta+3}{\theta(\theta+2)}, \quad \sigma^2_Y = (1-p_1)\frac{p_1\theta^2+6p_1\theta+9p_1+\theta^3+6\theta^2+12\theta+7}{\theta^2(\theta+2)^2}.$$

2.3.2. Moment generating function

Let Y follows ZIPG(p_1, θ).Then its MGF can be obtained as

$$\begin{aligned}M_Y(t) &= E(e^{ty}) \\ &= p_1 + (1-p_1)\frac{\theta(\theta^2+3\theta+1)}{(\theta+2)(\theta+1)^2} + \sum_{y=1}^{\infty} e^{ty}(1-p_1)\frac{\theta(\theta y+\theta^2+3\theta+1)}{(\theta+2)(\theta+1)^{y+2}} \\ &= p_1 + (1-p_1)M_X(t) \\ &= p_1 + (1-p_1)\frac{\theta\{\theta^3+(4-e^t)\theta^2+2(2-e^t)\theta+1-e^t\}}{(\theta+1)(\theta+2)(\theta+1-e^t)^2},\end{aligned}$$

where $M_X(t)$is the MGF of PG distribution with parameter θ.

2.3.3. Probability generating function

The PGF of a random variable Y following ZIPG(p_1, θ) is obtained as

$$\begin{aligned}P_Y(t) &= E(t^y) \\ &= p_1 + (1-p_1)\frac{\theta\{\theta^3+(4-t)\theta^2+2(2-t)\theta+1-t\}}{(\theta+1)(\theta+2)(\theta+1-t)^2}.\end{aligned}$$

2.3.4. Estimation of parameters

Like in the case of ZOIPG, we describe the different methods of estimation used for ZIPG as follows. In this case the parameter vector to be estimated is given by $\Lambda = (p_1, \theta)'$.

Method of moments estimation

Let $y_1, y_2, \ldots, y_n$ be the observed sample of size n from ZIPG distribution with probability mass as given in equation (13)
Let

$$m_1 = \bar{y}, \quad m_2 = \frac{1}{n}\sum_{i=1}^{n} y_i(y_i - 1),$$

where $\bar{y}$ is the sample mean. The corresponding population moments are given by

$$E[Y] = (1 - p_1)\frac{(\theta + 3)}{\theta(\theta + 2)}, \quad E[Y(Y - 1)] = (1 - p_1)\frac{2(\theta + 4)}{\theta^2(\theta + 2)}.$$

Equating the sample moments with the corresponding population moments and solving the resulting equations, we get the MOMs of p_1 and θ as follows

$$\tilde{\theta} = \frac{(2m_1 - 3m_2) + \sqrt{(2m_1 + 3m_2)^2 + 8m_1m_2}}{2m_2}, \quad \tilde{p_1} = \frac{\tilde{\theta} + 3 - m_1\tilde{\theta}(\tilde{\theta} + 2)}{(\tilde{\theta} + 3)}.$$

As in the case of ZOIPG, the value of the MOM $\tilde{\theta}$ for a given sample should satisfy the positivity condition. To ensure the positivity of $\tilde{\theta}$, we take positive square root of $(2m_1 + 3m_2)^2 + 8m_1m_2$.

Maximum likelihood estimation

Let $y = (y_1, y_2, \ldots, y_n)$ be the observed sample from ZIPG(p_1, θ). Then the likelihood function is given by

$$\begin{aligned} L(p_1, \theta|y) &= \left[p_1 + (1 - p_1)\frac{\theta(\theta^2 + 3\theta + 1)}{(\theta + 2)(\theta + 1)^2}\right]^{S_0} \\ &\quad \times \left[\frac{(1 - p_1)\theta}{(\theta + 2)(\theta + 1)^2}\right]^{n - S_0} \prod_{y_i \geq 1} \frac{\theta^2 + 3\theta + 1 + \theta y_i}{(\theta + 1)^{y_i}} \\ &= {q_0}^{S_0}\left[\frac{(1 - p_0)\theta}{(\theta + 2)(\theta + 1)^2}\right]^{n - S_0} \prod_{y_i \geq 2} \frac{\theta^2 + 3\theta + 1 + \theta y_i}{(\theta + 1)^{y_i}}, \end{aligned}$$

where S_0 is the number of zeros in the data and $q_0 = P[Y = 0]$.
Then the likelihood function becomes

$$L(q_0, \theta|y) = {q_0}^{S_0}\left[\frac{(1 - q_0)\theta}{7\theta^2 + 6\theta + 2}\right]^{n - S_0} \prod_{y_i \geq 1} \frac{\theta^2 + 3\theta + 1 + \theta y_i}{(\theta + 1)^{y_i}}.$$

The log-likelihood is given by

$$\log L = S_0 \log q_0 + (n - S_0)\Big[\log(1-q_0) + \log\theta - \log(7\theta^2+6\theta+2)\Big] + \sum_{y_i \geq 1}\Big[\log(\theta^2+3\theta+1+\theta y_i) - y_i \log(\theta+1)\Big].$$

Solving the likelihood equation $\frac{\partial \log L}{\partial q_0} = 0$, we get $\hat{q_0} = \frac{S_0}{n}$.
The MLE of θ, $\hat{\theta}$, is the solution of the non-linear equation

$$(n - S_0)\left[\frac{2-7\theta^2}{\theta(\theta^2+6\theta+2)}\right] + \sum_{y_i \geq 1}\left[\frac{2\theta+3+y_i}{\theta^2+3\theta+1+\theta y_i} - \frac{y_i}{\theta+1}\right] = 0,$$

which can be solved numerically by Newton-Raphson iterative algorithm. Note that the transformation between (p_1, θ) and (q_1, θ) is one to one, thus the MLE of p_1, $\hat{p_1}$, is obtained as follows

$$\hat{p_1} = 1 - \frac{(1 - \frac{S_0}{n})(\hat{\theta}+1)^2(\hat{\theta}+2)}{(7\hat{\theta}^2+6\hat{\theta}+2)}.$$

Remark: The MLE can again be computed using the help of EM algorithm, we skip the details for the sake of space considerations.

3. Simulation study

To assess the finite sample performance of the above mentioned estimators, we conducted a simulation study. First, we consider the zero-one-inflated situation. Since the ZOIPG distribution can be expressed as mixture, the simulation from the model is easy. We consider different parameter combinations and simulate observations of varying sizes like $n = 100, 200, 1000$ from that particular ZOIPG. Then we apply the three estimation methods to this artificial data set. The resulting point estimates are then saved and this procedure is repeated for a large number, say, $N = 1000$. After that, we take the arithmetic mean of these 1000 saved estimates as the final estimate in each of the estimation methods. Root mean squared errors (RMSE) are then evaluated to assess the variability of the estimates. Let $\tilde{\Lambda} = (\tilde{p_1}, \tilde{p_2}, \tilde{\theta})'$, $\hat{\Lambda} = (\hat{p_1}, \hat{p_2}, \hat{\theta})'$ and $\hat{\Lambda}_{EM} = (\hat{p_1}_{EM}, \hat{p_2}_{EM}, \hat{\theta}_{EM})'$ denote respectively the MOM, MLE and EM estimators of Λ. Since the moment estimators have closed form expressions, first we computed the moment estimates. These estimates are then used as starting values to solve the non-linear system of likelihood equations in both the case of MLE and MLE using EM algorithm. The non-linear system of equations derived from equation (12) is solved numerically using iterative Newton-Raphson method implemented in **R** function **optim**. For EM method, we stop the iterative algorithm when difference of consecutive values of log-likelihood is less than a threshold, i.e., $|\log L(\hat{\theta}_k) - \log L(\hat{\theta}_{k-1})| < 0.001$. We also examined the speed of convergence of estimates obtained from EM algorithm for all parameter combinations and for different sample sizes. As an illustration, we give details of convergence for the parameter combination $p_1 = 0.8,\ p_2 = 0.7,\ \theta = 0.4$, exhibited in Figure (2)(See Appendix I). From the figure, it is clear that for large sample size, the convergence speed is high compared to the combinations with small sample size. In both cases, estimator of p_2 converges rapidly. When n is small, estimates of p_1 and θ converges after 10 iterations where as it takes just 4 iterations when n increases to 500.

The simulation results are given in Tables 1,2,3 and 4 (see Appendix I). In Tables 1 and 2, the mean estimates of p_1, p_2 and θ are given when $\theta = 0.4$ and $\theta = 0.8$ respectively. We

have chosen different combinations of values of (p_1, p_2), which may represent some practical situations. For instance, we cover the values of parameters including those estimated from fitting real data sets (discussed in the next section). For large values of θ, say, $\theta \geq 4$, we get only lesser number of distinct possible values from the range of a ZOIPG distribution $\{0, 1, 2, 3, \ldots\}$. Hence, we decided to focus on lower values of θ. Other combinations were also studied, but the results are not reported here due to the space restrictions. The complete simulation results are available up on request from the corresponding author.The simulation study uses sample sizes 100, 200 and 1000 which reflects typical data volumes covering small to moderately large samples encountered in practice. These sizes also match with those of the datasets analyzed in the data analysis section, ensuring relevant and realistic evaluation. As seen from the Tables 1 and 2, the bias of the estimates in all three methods decreases when sample size n increases. Tables 3 and 4 exhibit the RMSE of the three estimators when $\theta = 0.4$ and $\theta = 0.8$ respectively. Again from these Tables, we can see that the RMSE is decreasing when the sample size is increasing. For instance, when $\theta = 0.4, p_1 = p_2 = 0.7$, substantial reduction in RMSE is observed when there is an increase of sample size from $n = 100$ to $n = 1000$. This pattern is seen in all other parameter combinations and across the estimation methods. Considering the performance measures of ML and EM algorithm based estimation, almost similar results are obtained. But, as expected, the RMSE of moment estimates are not comparable with that of MLE and EM estimates.

Next we describe the simulation study of ZIPG model. Here, the parameter vector is $\Lambda = (p_1, \theta)'$. Similar to the case of ZOIPG, we repeated the simulation experiment $N = 1000$ times and stored the resulting estimates. The results are exhibited in Table 5 (See Appendix I). In Table 5, we report the point estimates along with their RMSE, when $\theta = 0.2$ and $\theta = 0.8$ respectively. Here also, the performance of MLEs in terms of RMSE is satisfactory. All the computations in the simulation are done through R Core Team (2022) software. The R code of the computations are available in Appendix II. As a final remark on the simulation study, we calculated computational times for different estimation methods. The average computational times are calculated separately for each estimation method across all parameter combinations over the entire $N = 1000$ repetitions. The results are given in Table 6 (See Appendix I). The MOM and MLEs take a little time to computations compared to EM. This is expected since explicit expressions for MOM are available. EM is slow, but better RMSE is obtained. Next, we illustrate the applicability of our proposed models with three real data sets.

4. Data analysis

In order to demonstrate the performance of ZOIPG and ZIPG, these distributions are fitted to three real data sets. Comparison is done with other distributions such as ZIP, ZIPL, ZOIP and ZOIPL. Parameters are computed by using ML estimation technique. For each of these distributions, values of χ^2 statistic, p-value, AIC, BIC, DI(Dispersion Index) values are calculated to compare how well the distribution fits the data.

Data set I

Zeger (1988) analyzed a data set which consists of monthly information about Polio cases reported by U.S. Centre for Disease Control over the years 1970 to 1983. This data was reanalyzed by many authors, among them, see Qi *et al.* (2019). In this data set, there

is a total of 168 observations of which 37.87 % are zeroes and 32.74 % are ones. The data have a mean of 1.33 and a variance of 3.35, which indicates overdispersion. Also it is right skewed. Figure 3a (See Appendix I) gives the barplot of the data. Since the data consists of excess number of zeroes and ones, we fitted ZOIP, ZOIPL and ZOIPG models. The results obtained are summarized in Table 7 (See Appendix I). The AIC and BIC values of ZOIPG are comparable with that of ZOIPL, with a slight improvement. Note that the value of χ^2 is minimum for ZOIPG. It is also observed that the ZOIPG model-implied DI is close to the empirical DI. The fitted probabilities for all models against empirical proportions are plotted in Figure 4 (See Appendix I). It is clear that on an average, the ZOIPG models fits well.

Data set II

A data set consisting of weekly number of syphilis cases in the United States from 2007 to 2010 is obtained from **R** package **ZIM** by Yang *et al.* (2018). The data consists of 209 observations of which 77.03 % are zeroes. Figure 3b(See Appendix I) gives the barplot of the data. Since this data consists of excessive zeros, we fitted ZIP, ZIPG and ZIPL for the data. This data have mean 0.9760 and variance 6.6773, implying a clear presence of overdispersion. The results obtained are summarized in Table 8 (See Appendix I). The values of AIC, BIC and χ^2 values are minimum for ZIPG model. Comparing to other models, the DI implied by ZIPG model is closer to empirical DI, indicating the suitability of ZIPG model to this data set. The fitted probabilities for all models against empirical probabilities are plotted in Figure 5(See Appendix I). It is obvious that on an average, the ZIPG models fits well the data.

Data set III

The monthly number of drug offenses recorded from January 1990 to December 2001, in Pittsburgh census tract 2206 (Garay *et al.* (2022)) is analyzed. The data consists of 144 observations and it includes 43.06 % zeroes and 14.58 % ones which can be seen from Figure 3c (See Appendix I). The data have mean 2.11 and variance 12.9106, leading to confirming presence of overdispersion. So we fitted ZOIP, ZOIPG and ZOIPL models for the data. The results are obtained in Table 9 (See Appendix I). A substantial reduction in AIC, BIC and χ^2 values of ZOIPG model are observed in comparison to ZOIP and ZOIPL models. Although none of the implied dispersion indices of all three models is close to empirical DI, the one implied by ZOIPG is still better among them. The fitted probabilities for all models against empirical probabilities are plotted in Figure 6(See Appendix I).

It is observed that, for all the data sets, the ZOIPG/ZIPG models consistently outperform the competing models. Specifically, these models yield the lowest values for the AIC, BIC, and χ^2 criteria. Additionally, the p-values from the goodness-of-fit tests are consistently higher for the proposed models, indicating better fit. Furthermore, the DI of the ZOIPG/ZIPG models is closer to the empirical DI across all data sets.

5. Conclusion

When count data exhibit presence of excessive number of zeros and ones, a mixture of degenerate random variables at 0 and 1 and a Poisson distribution is usually employed. However, in some situations, in addition to the presence of extra zeros and ones, large values

are also observed. In this case the zero-one-inflated Poisson model may not be well fitted. Several other Poisson mixture distributions are introduced in literature. In this paper, we have introduced a new inflated Poisson-Garima distribution and its properties. In particular, ZOIPG and ZIPG distributions are considered. The estimation of model parameters using MOM, ML method and ML using EM algorithm are then discussed. A simulation study is conducted to evaluate the finite sample performance of the above estimators. Three real data sets were analyzed to show the applicability of the proposed models. The introduced model outperformed the other competing models in terms of AIC, BIC and DI.

This class of distributions falls under the modified Poisson mixture models where the rate parameter of Poisson distribution follows a one parameter continuous random variable namely Garima distribution. One can generalize this model as a family of Poisson mixture distributions where the intensity parameter follows some non-negative distributions other than Garima distribution. Also, we can construct regression models for a response variable with abundance of occurrence of zero and one in presence of covariates. Moreover, one can develop an auto regressive count time series models based on the ZOIPG distribution. We leave these aspects for our future research study.

Acknowledgements

The authors are grateful to the editor and reviewer for their guidance and valuable comments and suggestions which led to substantial improvement of the earlier draft. Divya A. and Prasanth C. B. acknowledges the valuable support of the Department of Statistics, St.Thomas college, Thrissur. Muhammed Anvar P. wishes to acknowledge the project grant received under PM USHA Scheme G.O number G.O.(Rt)No.239/2025/HEDN dated 22.02.2025

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Altun, E., Alqifari, H., and Eliwa, M. S. (2023). A novel approach for zero-inflated count regression model: Zero-inflated Poisson generalized-Lindley linear model with applications. *AIMS Mathematics*, **8**, 23272–23290.

Garay, A. M., Medina, F. L., Jales CS, I., and Bertail, P. (2022). First-order integer valued ar processes with zero-inflated innovations. In *Nonstationary Systems: Theory and Applications: Contributions to the 13th Workshop on Nonstationary Systems and Their Applications, February 3-5, 2020, Grodek nad Dunajcem, Poland 13*, pages 19–40. Springer.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Liu, W., Tang, Y., and Xu, A. (2018). A zero-and-one inflated Poisson model and its application. *Statistics and its Interface*, **11**, 339–351.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. John Wiley & Sons.

Melkersson, M. and Olsson, C. (1999). *Is Visiting the Dentist a Good Habit?: Analyzing Count Data with Excess Zeros and Excess Ones*. University of Umeå.

Qi, X., Li, Q., and Zhu, F. (2019). Modeling time series of count with excess zeros and ones based on INAR (1) model with zero-and-one inflated Poisson innovations. *Journal of Computational and Applied Mathematics*, **346**, 572–590.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shanker, R. (2016). Garima distribution and its application to model behavioral science data. *Biometrics and Biostatistics International Journal*, **4**, 1–9.

Shanker, R. (2017). The discrete Poisson-Garima distribution. *Biometrics and Biostatistics International Journal*, **5**, 1–7.

Shanker, R. and Shukla, K. K. (2017). Zero-truncated Poisson-Garima distribution and its applications. *Biostat Biometrics Open Access Journal*, **3**, 555605.

Tajuddin, R. R. M., Ismail, N., Ibrahim, K., and Bakar, S. A. A. (2022). A new zero-one-inflated Poisson-Lindley distribution for modelling overdispersed count data. *Bulletin of the Malaysian Mathematical Sciences Society*, **45**, 21–35.

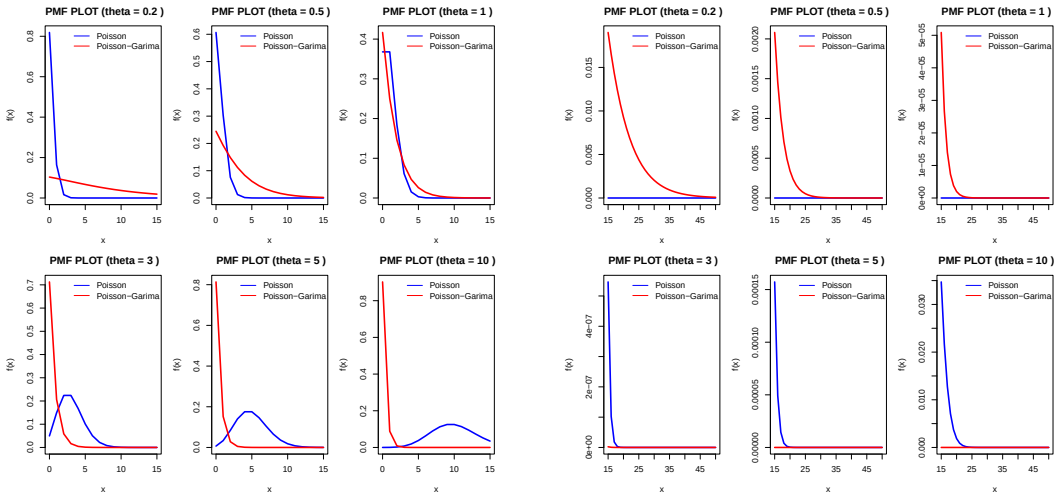
Tang, Y., Liu, W., and Xu, A. (2017). Statistical inference for zero-and-one-inflated Poisson models. *Statistical Theory and Related Fields*, **1**, 216–226.

Yang, M., Zamba, G., Cavanaugh, J., and Yang, M. M. (2018). Package 'zim'. .

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika*, **75**, 621–629.

Zhang, C., Tian, G.-L., and Ng, K.-W. (2016). Properties of the zero-and-one inflated Poisson distribution and likelihood-based inference methods. *Statistics and its Interface*, **9**, 11–32.

Appendix I



(a) **(b)**

Figure 1: Probability plots for Poisson and Poisson-Garima

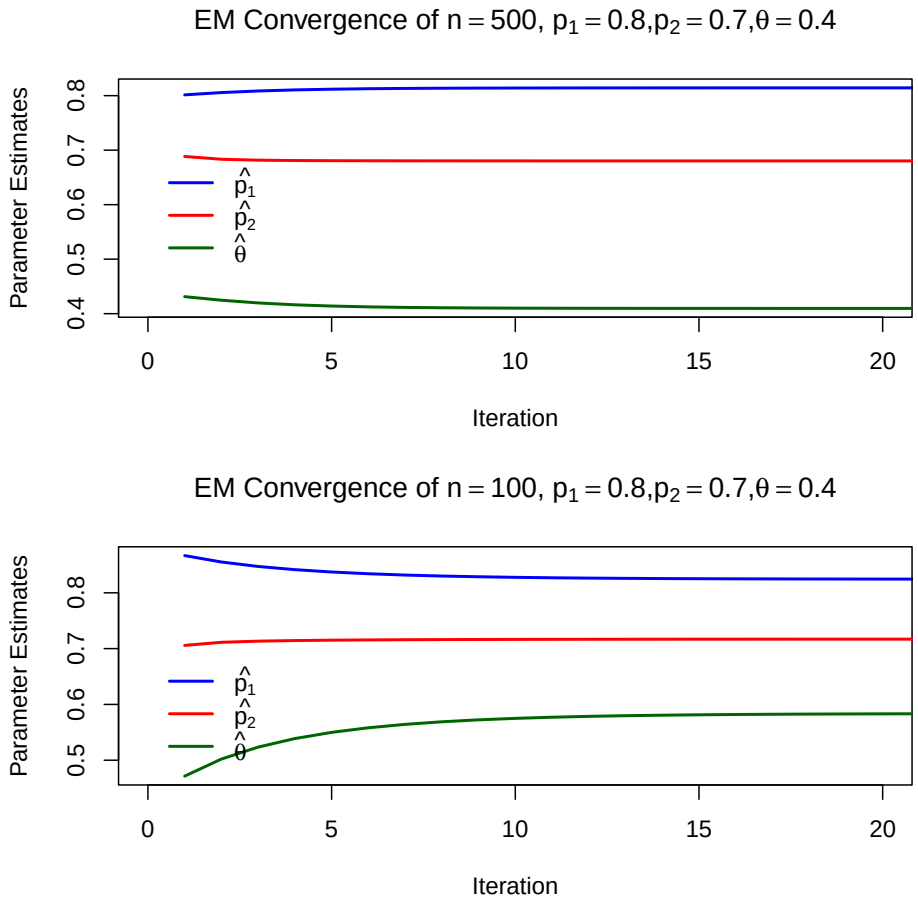
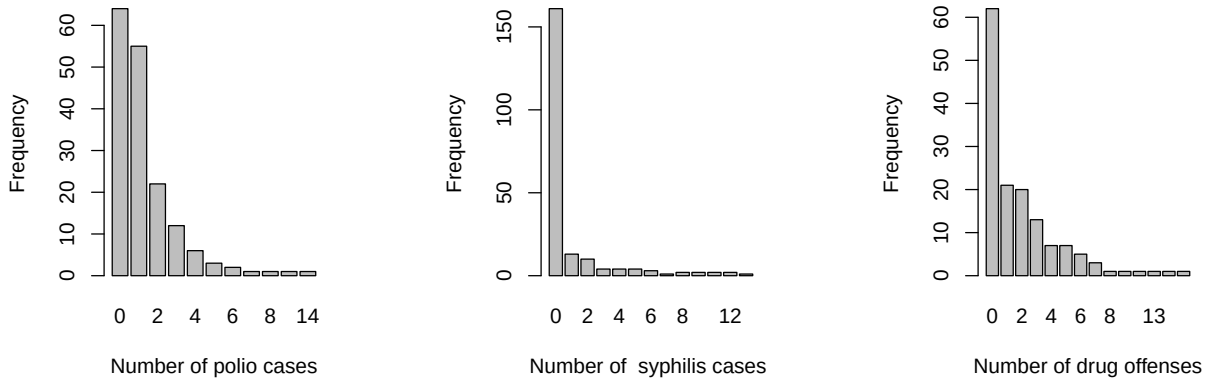


Figure 2: Convergence of EM estimates



(a) Barplot of monthly polio data

(b) Barplot of weekly number of syphilis cases

(c) Barplot of monthly number of drug offenses

Figure 3: Barplots of datasets

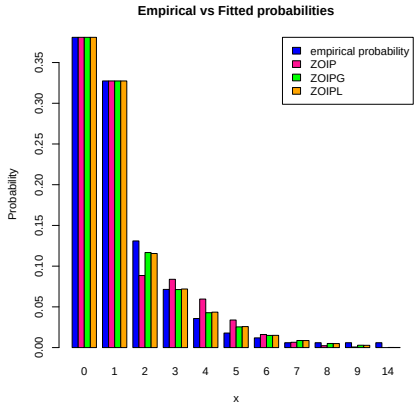


Figure 4: Bar Plot of fitted probabilities of polio data

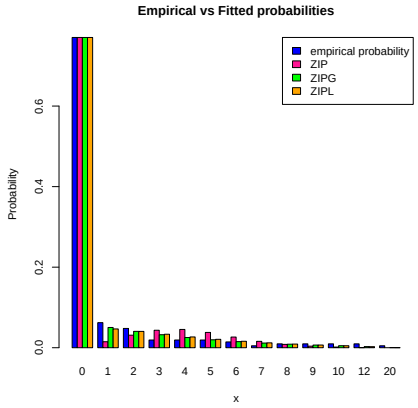


Figure 5: Bar Plot of fitted probabilities of syphilis cases data

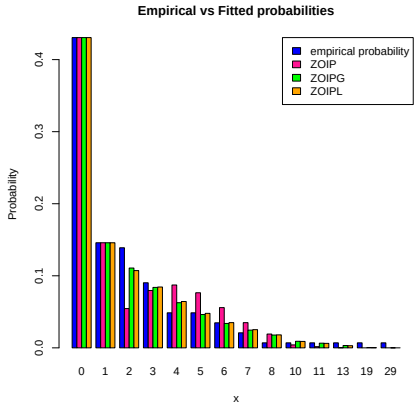


Figure 6: Bar Plot of fitted probabilities of drug data

Table 1: Parameter estimates of ZOIPG distribution with $\theta = 0.4$.

			MOM			MLE			EM		
n	p_1	p_2	$\tilde{\theta}$	$\tilde{p}_1$	$\tilde{p}_2$	$\hat{\theta}$	$\hat{p}_1$	$\hat{p}_2$	$\hat{\theta}_{EM}$	$\hat{p}_{1EM}$	$\hat{p}_{2EM}$
100			0.4457	0.3427	0.2063	0.4073	0.1225	0.4537	0.4070	0.1925	0.3537
200	0.2	0.4	0.4248	0.2977	0.2162	0.4038	0.1245	0.4471	0.4041	0.1954	0.3775
1000			0.7975	0.2825	0.3895	0.7927	0.0845	0.4767	0.7923	0.2045	0.4044
100			0.4457	0.3427	0.2069	0.4073	0.1391	0.4943	0.4100	0.186	0.4263
200	0.2	0.5	0.425	0.2977	0.2257	0.4038	0.1266	0.6027	0.4049	0.1937	0.4703
1000			0.4066	0.1814	0.3536	0.3985	0.1291	0.6746	0.3988	0.2036	0.5004
100			0.4678	0.3923	0.4080	0.4128	0.3086	0.6507	0.4141	0.4875	0.4990
200	0.5	0.5	0.4370	0.4226	0.4826	0.4053	0.3140	0.6726	0.4054	0.4970	0.4987
1000			0.4099	0.4732	0.5521	0.4002	0.3150	0.6712	0.4001	0.4983	0.4981
100			0.5170	0.5649	0.5781	0.4279	0.4284	0.6638	0.4283	0.6894	0.4977
200	0.7	0.5	0.4771	0.6165	0.5721	0.4268	0.4253	0.6625	0.4247	0.6886	0.4912
1000			0.4110	0.6823	0.5228	0.3999	0.4433	0.6707	0.3999	0.7006	0.4980
100			0.5151	0.5686	0.5660	0.4292	0.4277	0.5187	0.4537	0.6669	0.6810
200	0.7	0.7	0.4771	0.6165	0.5705	0.4268	0.4253	0.4861	0.4247	0.6886	0.7123
1000			0.4111	0.6809	0.7332	0.3977	0.4449	0.6347	0.3977	0.7017	0.6979

Table 2: Parameter estimates of ZOIPG distribution with θ=0.8

			MOM			MLE			EM		
n	p_1	p_2	$\tilde{\theta}$	$\tilde{p}_1$	$\tilde{p}_2$	$\hat{\theta}$	$\hat{p}_1$	$\hat{p}_2$	$\hat{\theta}_{EM}$	$\hat{p}_{1EM}$	$\hat{p}_{2EM}$
100			0.9528	0.4040	0.4159	0.8406	0.1817	0.3501	0.8080	0.2028	0.3188
200	0.2	0.4	0.8422	0.3602	0.3650	0.7884	0.1172	0.4025	0.7801	0.2105	0.3578
1000			0.7975	0.2825	0.3895	0.7928	0.0845	0.4767	0.7923	0.2045	0.4044
100			0.9529	0.4042	0.4149	0.8405	0.1817	0.3672	0.8203	0.1878	0.3968
200	0.2	0.5	0.8422	0.3603	0.3741	0.7884	0.1172	0.4287	0.7887	0.1988	0.4459
1000			0.7975	0.2825	0.4325	0.7928	0.0845	0.7097	0.7929	0.2037	0.5057
100			1.0290	0.4148	0.4902	0.8621	0.2029	0.5395	0.8569	0.4576	0.4507
200	0.5	0.5	0.9146	0.4174	0.4908	0.8275	0.1976	0.6362	0.8302	0.4775	0.4794
1000			0.8031	0.4823	0.5177	0.7942	0.2063	0.7169	0.7941	0.5013	0.5039
100			1.1868	0.5223	0.4732	0.9192	0.2768	0.5827	0.9167	0.6554	0.4663
200	0.7	0.5	0.9735	0.5881	0.4889	0.8449	0.2785	0.6467	0.8477	0.6844	0.4897
1000			0.8342	0.6729	0.5038	0.8048	0.2856	0.6997	0.8048	0.6992	0.4985
100			1.1870	0.5220	0.6474	0.9192	0.2768	0.2203	0.9945	0.6030	0.6447
200	0.7	0.7	0.9735	0.5881	0.7109	0.8449	0.2785	0.1494	0.8716	0.6682	0.6854
1000			0.8343	0.6729	0.7178	0.8048	0.2856	0.1000	0.8048	0.6992	0.7013

Table 3: RMSE of ZOIPG distribution with θ =0.4

			MOM			MLE			EM		
n	p_1	p_2	$\tilde{\theta}$	$\tilde{p_1}$	$\tilde{p_2}$	$\hat{\theta}$	$\hat{p_1}$	$\hat{p_2}$	$\hat{\theta}_{EM}$	$\hat{p_1}_{EM}$	$\hat{p_2}_{EM}$
100			0.1122	0.2298	0.2931	0.0592	0.1000	1.5571	0.0592	0.0900	0.2117
200	0.2	0.4	0.0755	0.2020	0.2891	0.0400	0.0889	0.7426	0.0400	0.0686	0.1609
1000			0.0985	0.1703	0.1575	0.0520	0.1183	0.2322	0.0520	0.0510	0.0922
100			0.1122	0.2298	0.3672	0.0600	0.1063	0.2924	0.0616	0.0975	0.2263
200	0.2	0.5	0.0755	0.2020	0.3610	0.0400	0.0894	0.2579	0.0412	0.0714	0.1587
1000			0.0316	0.1030	0.3226	0.0176	0.0728	0.1910	0.0173	0.0245	0.0490
100			0.1459	0.2102	0.3081	0.0800	0.2035	0.2193	0.0819	0.0889	0.0954
200	0.5	0.5	0.0985	0.1749	0.2769	0.0520	0.1926	0.1985	0.0529	0.0608	0.0608
1000			0.0424	0.0831	0.1497	0.0224	0.1860	0.1758	0.0224	0.0224	0.0245
100			0.2124	0.2175	0.2289	0.1122	0.2835	0.2035	0.1131	0.0787	0.0700
200	0.7	0.5	0.1884	0.1741	0.1887	0.0849	0.2804	0.1780	0.0837	0.0490	0.0436
1000			0.0583	0.0678	0.0819	0.0332	0.2583	0.1738	0.0332	0.0224	0.0200
100			0.2161	0.2161	0.3592	0.1145	0.2839	0.4468	0.1936	0.1517	0.1030
200	0.7	0.7	0.1884	0.1741	0.3406	0.0849	0.2804	0.4734	0.0831	0.0490	0.0458
1000			0.0557	0.0671	0.0949	0.0300	0.2563	0.4241	0.0300	0.0224	0.0200

Table 4: RMSE of estimates of ZOIPG distribution with θ=0.8

			MOM			MLE			EM		
n	p_1	p_2	$\tilde{\theta}$	$\tilde{p_1}$	$\tilde{p_2}$	$\hat{\theta}$	$\hat{p_1}$	$\hat{p_2}$	$\hat{\theta}_{EM}$	$\hat{p_1}_{EM}$	$\hat{p_2}_{EM}$
100			0.3503	0.2621	0.2490	0.2059	0.1694	0.3126	0.1487	0.1323	0.2681
200	0.2	0.4	0.2015	0.2466	0.2128	0.1109	0.1367	0.2897	0.1057	0.1054	0.2163
1000			0.0985	0.1703	0.1575	0.0520	0.1179	0.2322	0.0520	0.0510	0.0922
100			0.3503	0.2621	0.2655	0.2059	0.1694	0.3561	0.1572	0.1400	0.2933
200	0.2	0.5	0.2015	0.2466	0.2559	0.1109	0.1367	0.3342	0.1109	0.1145	0.2332
1000			0.0985	0.1703	0.2093	0.0520	0.1179	0.2903	0.0520	0.0520	0.0768
100			0.4631	0.2095	0.2441	0.2709	0.3138	0.2978	0.2445	0.1664	0.1718
200	0.5	0.5	0.2800	0.2052	0.1931	0.1616	0.3094	0.2615	0.1649	0.1136	0.1005
1000			0.1091	0.1170	0.0624	0.0616	0.2950	0.2276	0.0616	0.0424	0.0283
100			0.8141	0.2750	0.2020	0.5019	0.4382	0.2657	0.4004	0.1584	0.0269
200	0.7	0.5	0.3780	0.2102	0.1200	0.2149	0.4287	0.2347	0.2218	0.0922	0.0616
1000			0.1476	0.0911	0.0300	0.0900	0.4163	0.3783	0.0900	0.0387	0.0200
100			0.4004	0.1584	0.1269	0.5019	0.4382	0.5590	0.5085	0.2581	0.1741
200	0.7	0.7	0.2218	0.0922	0.0616	0.2149	0.4287	0.5851	0.2812	0.1449	0.0970
1000			0.0900	0.0387	0.0200	0.0900	0.4163	0.6000	0.0900	0.0387	0.0200

Table 5: Parameter estimates and their RMSE of ZIPG model

			MOM		**RMSE**		**MLE**		**RMSE**	
n	θ	p_1	$\tilde{\theta}$	$\tilde{p}_1$	$\tilde{\theta}$	$\tilde{p}_1$	$\hat{\theta}$	$\hat{p}_1$	$\hat{\theta}$	$\hat{p}_1$
100			0.2369	0.0243	0.0374	0.1758	0.2241	0.2309	0.0241	0.0308
200	0.2	0.2	0.1721	0.2775	0.0279	0.0775	0.2139	0.1566	0.0138	0.0436
1000			0.2004	0.2532	0.0004	0.0532	0.2106	0.1807	0.0105	0.0192
100			0.2096	0.4198	0.0096	0.0802	0.2249	0.4116	0.0249	0.0883
200	0.2	0.5	0.1707	0.5174	0.0293	0.0173	0.2289	0.4669	0.0289	0.0332
1000			0.1974	0.5344	0.0026	0.0346	0.2131	0.5010	0.0130	0.0010
100			0.2049	0.6575	0.0049	0.0425	0.2569	0.6772	0.0566	0.0228
200	0.2	0.7	0.1767	0.6290	0.0232	0.0707	0.2144	0.7189	0.0141	0.0190
1000			0.2044	0.7155	0.0044	0.0155	0.2171	0.6982	0.0170	0.0018
100			0.4431	0.6777	0.1212	0.0768	0.4185	0.6963	0.0883	0.0548
200	0.4	0.7	0.4198	0.6889	0.0837	0.0548	0.4085	0.6978	0.0648	0.0387
1000			0.4046	0.6979	0.0346	0.0245	0.4020	0.7000	0.0265	0.0173
100			1.0900	0.0711	0.2972	0.1288	0.8915	0.2519	0.0917	0.0520
200	0.8	0.2	0.7042	0.3856	0.0959	0.1857	0.8126	0.1244	0.0126	0.0755
1000			0.8217	0.1877	0.0217	0.0122	0.8147	0.2361	0.0148	0.0361
100			0.9431	0.4228	0.1432	0.0775	0.9374	0.5176	0.1374	0.0173
200	0.8	0.5	0.7586	0.5963	0.0412	0.0964	0.8700	0.3805	0.0700	0.1196
1000			0.8926	0.4621	0.0927	0.0374	0.7803	0.5094	0.0197	0.0094
100			1.0030	0.6309	0.2040	0.0693	1.0750	0.7159	0.2750	0.0173
200	0.8	0.7	0.6375	0.7758	0.1625	0.0755	0.8778	0.5961	0.0781	0.1039
1000			0.9161	0.6739	0.1162	0.0261	0.8206	0.7160	0.0207	0.0158

Table 6: Computation time comparison for ZOIPG and ZIPG models across different estimation methods

		ZOIPG		ZIPG
Sample Size (n)	Method	Time(in seconds)	Method	Time(in seconds)
100	MOM	0.2664	MOM	0.1637
	MLE	0.7506	MLE	1.7238
	EM	115.5400		
200	MOM	0.3100	MOM	0.2340
	MLE	0.8207	MLE	2.9412
	EM	210.2300		
1000	MOM	0.6338	MOM	0.4839
	MLE	1.2178	MLE	14.1458
	EM	619.9900		

Table 7: Fitted distributions on Dataset-I, along with their expected frequencies, estimates of parameters and model selection statistics.

Distribution of monthly number of Polio cases				
		Fitted Distributions with expected frequencies		
x	f	ZOIP	ZOIPG	ZOIPL
0	64	64	64	64
1	55	55	55	55
2	22	15	20	19
3	12	14	12	12
4	6	10	7	7
5	3	6	4	4
6	2	3	3	3
7	1	1	1	2
8	1	0	1	1
9	1	0	1	1
14	1	0	0	0
Total	168	168	168	168
Estimates	p_1	0.6241	0.2288	0.2742
	p_2	0.5753	0.3878	0.4548
	θ	2.8421	0.8712	0.9316
AIC		554.6511	**536.0119**	536.387
BIC		564.023	**545.4384**	545.7589
χ^2		5.2524	**0.4429**	0.9802
d.f		2	2	2
p-value		0.0724	**0.8014**	0.6126
DI	2.6287	1.9437	**2.2516**	2.2240

Table 8: Fitted distributions on Dataset-II, along with their expected frequencies, estimates of parameters and information statistics.

Distribution of weekly number of syphilis cases				
		Fitted Distributions with expected frequencies		
x	f	ZIP	ZIPG	ZIPL
0	161	161	161	161
1	13	3	11	10
2	10	7	9	9
3	4	9	7	7
4	4	9	5	6
5	4	8	4	5
6	3	6	3	3
7	1	3	2	3
8	2	2	2	2
9	2	1	2	1
10	2	0	2	1
12	2	0	1	1
20	1	0	0	0
Total	209	209	209	209
Estimates	p_1	0.7667	0.7095	0.7211
	θ	4.19	0.4194	0.4802
AIC		512.5916	**452.5824**	453.8479
BIC		519.2763	**459.2671**	460.5325
χ^2		46.34127	**2.0716**	7.0302
d.f		5	4	5
p-value		$< .001$	**0.7226**	0.2184
DI	6.8410	4.209	**6.1837**	5.8709

Table 9: Fitted distributions on Dataset-III, along with their expected frequencies, estimates of parameters and model selection statistics.

Distribution of monthly number of drug offenses				
		Fitted Distributions with expected frequencies.		
x	f	ZOIP	ZOIPG	ZOIPL
0	62	62	62	62
1	21	21	21	21
2	20	8	16	16
3	13	12	13	12
4	7	13	9	10
5	7	11	7	7
6	5	8	5	5
7	3	5	4	4
8	1	3	3	3
10	1	1	2	2
11	1	0	1	1
13	1	0	1	1
19	1	0	0	0
29	1	0	0	0
Total	144	144	144	144
Estimates	p_1	0.5458	0.2508	0.3026
	p_2	0.7784	0 .9913	0.9463
	θ	4.3820	0.4962	0.5491
AIC		641.4777	**559.5282**	561.2334
BIC		650.3872	**568.4377**	570.1428
χ^2		23.4321	**1.8081**	2.3470
d.f		4	4	6
p-value		0.0001	**0.7710**	0.6722
DI	6.1156	3.0198	**4.0614**	3.9262

Appendix II

R code for ZOIPG Simulation

```
library(rootSolve)
sim = 1000
n = 100
p0 = 0.3
p1 = 0.9
theta = 0.5
p = (theta + 1) / (theta + 2)

MM1 = MM2 = MM3 = c()
ML1 = ML2 = ML3 = c()
EM1 = EM2 = EM3 = c()
```

```
for (i in 1:sim) {
 x1=rexp(n,theta)
x2=rgamma(n,rate=theta,shape=2)
mix=runif(n)
data=ifelse(mix<p,x1,x2)
V=rpois(n,data)

  B1 = rbinom(n, 1, p0)
  B2 = rbinom(n, 1, p1)
  y = B1 * (1 - B2) + (1 - B1) * V

  # Method of Moments
  fm1 = mean(y)
  fm2 = mean(y * (y - 1))
  fm3 = mean(y * (y - 1) * (y - 2))
  s1 = fm3 / fm2
  MM3[i] = ((3 - 4 * s1) + sqrt((4 * s1 + 3)^2 + 12 * s1)) / (2 *
     s1)
  MM1[i] = 1 - (fm2 * (MM3[i] + 2) * MM3[i]^2) / (2 * (MM3[i] + 4))
  muhat = (MM3[i] + 3) / (MM3[i] * (MM3[i] + 2))
  p21 = MM1[i] - fm1 + (1 - MM1[i]) * muhat
  MM2[i] = p21 / MM1[i]
  if (MM1[i] < 0 || MM1[i] > 1) MM1[i] = 0.5
  if (MM2[i] < 0 || MM2[i] > 1) MM2[i] = 0.1

  # MLE Estimation
  s0 = sum(y == 0)
  s1 = sum(y == 1)
  y1 = y[y >= 2]

  loglt = function(x) {
    a1 = n - s0 - s1
    a2 = (4 * x^2 + 4 * x + 2) / (x * (x + 1) * (x^2 + 5 * x + 2))
    a3 = sum((2 * x + 3 + y1) / (x^2 + 3 * x + 1 + x * y1))
    a4 = sum(y1 / (x + 1))
    F = a1 * a2 + a3 - a4
    return(F)
  }
  ML3[i] = uniroot(loglt, c(0, 100))$root

  loglp = function(x) {
    p00 = x[1]; p10 = x[2]
    c1 = (ML3[i] * (ML3[i]^2 + 3 * ML3[i] + 1)) / ((ML3[i] + 2) *
       (ML3[i] + 1)^2)
    c2 = (ML3[i] * (ML3[i]^2 + 4 * ML3[i] + 1)) / ((ML3[i] + 2) *
       (ML3[i] + 1)^3)
    F1 = c1 * (1 - p00) + p00 * p10 - s0 / n
    F2 = c2 * (1 - p00) + p00 * (1 - p10) - s1 / n
    return(c(F1, F2))
  }
```

```
ML1[i] = multiroot(loglp, c(0.5, 0.5))$root[1]
ML2[i] = multiroot(loglp, c(0.5, 0.5))$root[2]

if (ML1[i] < 0 || ML1[i] > 1) ML1[i] = 0.5
if (ML2[i] < 0 || ML2[i] > 1) ML2[i] = 0.1

# EM Algorithm
EMfit = function(par) {
  EM = function(para) {
    EMp0 = para[1]; EMp1 = para[2]; EMtheta = para[3]
    a1 = (EMp0 * EMp1) / (EMp0 * EMp1 + (1 - EMp0) * (EMtheta *
       (EMtheta^2 + 3 * EMtheta + 1)) / ((EMtheta + 2) * (EMtheta
       + 1)^2))
    a2 = (EMp0 * (1 - EMp1)) / (EMp0 * (1 - EMp1) + (1 - EMp0) *
       (EMtheta * (EMtheta^2 + 4 * EMtheta + 1)) / ((EMtheta + 2)
       * (EMtheta + 1)^3))
    B1em = a1 * (y == 0) + a2 * (y == 1)
    b1 = (EMp0 * EMp1 + (1 - EMp0) * EMp1 * EMtheta * (EMtheta^2
       + 3 * EMtheta + 1) / ((EMtheta + 2) * (EMtheta + 1)^2)) /
         (EMp0 * EMp1 + (1 - EMp0) * EMtheta * (EMtheta^2 + 3 *
            EMtheta + 1) / ((EMtheta + 2) * (EMtheta + 1)^2))
    b2 = ((1 - EMp0) * EMp1 * EMtheta * (EMtheta^2 + 4 * EMtheta
       + 1) / ((EMtheta + 2) * (EMtheta + 1)^3)) /
         (EMp0 * (1 - EMp1) + (1 - EMp0) * EMtheta * (EMtheta^2 +
            4 * EMtheta + 1) / ((EMtheta + 2) * (EMtheta + 1)^3))
    b3 = EMp1
    B2em = b1 * (y == 0) + b2 * (y == 1) + b3 * (y >= 2)

    logtheta = function(th) {
      -sum((1 - B1em) * log(th * (th * y + th^2 + 3 * th + 1) /
         ((th + 2) * (th + 1)^(y + 2))))
    }
    thetaEM = optimize(logtheta, c(0, 100))$minimum
    para[1] = mean(B1em)
    para[2] = mean(B2em)
    para[3] = thetaEM
    loglik = function(theta) {
      p01 = theta[1]; p11 = theta[2]; th = theta[3]
      if (p01 < 0 || p01 > 1) p01 = 0.5
      if (p11 < 0 || p11 > 1) p11 = 0.5
      logL = sum(B1em * log(p01) + (1 - B1em) * (log(1 - p01) + 2
         * log(th) + log(y + th + 2) - (y + 3) * log(th + 1)) +
                  B2em * log(p11) + (1 - B2em) * log(1 - p11))
      return(logL)
    }
    list(pa = para, ll = loglik(para))
  }
  iter = 0; para.old = par; like.old = EM(par)$ll
  repeat {
    iter = iter + 1
```

```
      para.new = EM(para.old)$pa
      loglik.new = EM(para.new)$ll
      if (abs(loglik.new - like.old) < 0.001) break
      para.old = para.new; like.old = loglik.new
    }
    list(para = para.new, loglik = loglik.new)
  }
  mm = c(MM1[i], MM2[i], MM3[i])
  EM1[i] = EMfit(mm)$para[1]
  EM2[i] = EMfit(mm)$para[2]
  EM3[i] = EMfit(mm)$para[3]
}

True = c(theta, p0, p1)
MM = c(mean(MM3), mean(MM1), mean(MM2))
MLE = c(mean(ML3), mean(ML1), mean(ML2))
EM = c(mean(EM3), mean(EM1), mean(EM2))

MSEmm = c(mean((MM3 - theta)^2), mean((MM1 - p0)^2), mean((MM2 -
   p1)^2))
MSEml = c(mean((ML3 - theta)^2), mean((ML1 - p0)^2), mean((ML2 -
   p1)^2))
MSEem = c(mean((EM3 - theta)^2), mean((EM1 - p0)^2), mean((EM2 -
   p1)^2))

cbind(True, MM, MSEmm, MLE, MSEml, EM, MSEem)
```

More R codes are available at https://github.com/Divya-1203/zoipg/blob/main/zoipggithub.R

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 24, No. 1, 2026 (New Series), pp 155–176
http://www.ssca.org.in/journal

Semi-supervised Feature Selection using Maximum Mutual Information and Minimum Correlation through Augmented Learning

Arghya Kusum Das[1], Saptarsi Goswami[2], Amlan Chakrabarti[3] and Basabi Chakraborty[4]

[1]*Department of Computer Science and Engineering, Techno International New Town, Kolkata 700156, India*
[2]*Department of Computer Science, Bangabasi Morning College, Kolkata, 700012, India*
[3]*A.K. Choudhury School of Information Technology, University of Calcutta, Kolkata 700009, India*
[4]*Faculty of Software and Information Science, Prefectural University, Takizawa City, Iwate, Japan*
[4]*School of Computer Science, Madanapalle Institute of Technology and Science, AP, India*

Received: 18 February 2025; Revised: 26 July 2025; Accepted: 29 July 2025

Abstract

Feature selection is a critical pre-processing step in machine learning. For supervised problems, class labels are used to identify important features. However, labeling or annotating the data is labor-intensive and hence costly. Consequently, there is an abundance of unlabeled data and limited labeled data. Therefore, semi-supervised learning is very pertinent in this case. The problem of feature selection is equally relevant to semi-supervised learning. In this research work, a novel semi-supervised method of feature selection, **R**epeated **S**ampled **S**emi-supervised **F**eature **S**election(RSSFS), is proposed with a time complexity of O(nd), which is significantly better than other existing algorithms. This follows an unbiased and neutral strategy in producing a final feature set that has less redundancy and frequently occurring feature components, thereby inducing better stability in the feature set. In the first step, a decision tree classifier is used for labeling the unlabeled portion of the data and augmenting the training set. Repeated sampling is done from the pseudo-labeled portion, to generate multiple augmented training sets. The top k features are selected based on mutual information from each augmented training set. While choosing the features, it is ensured that the features are not redundant by applying a correlation coefficient based threshold. A voting-based approach is used to combine these multiple features into the final feature set. The proposed algorithm is compared with a) a benchmark, using the full feature set, and b) top k % supervised feature selection on the labeled portion. Comparing these three methods across 18 datasets, it was found that RSSFS outperformed the supervised method based on F1 scores by 2.79% and the benchmark by 0.36%. Thus, the proposed algorithm would prove impactful in applications where there is a plethora of unlabeled data compared to labeled data.

Corresponding Author: Arghya Kusum Das
Email: arghya.das@gmail.com

Key words: Semi-supervised learning; Feature selection; Stability; Mutual information; Correlation.

AMS Subject Classifications: 68T01 The video recording of the paper made under the SSCA's Online Lecture series is available at the Youtube channel URL `https://youtu.be/OWQ_FuYt55E`.

1. Introduction

The curse of dimensionality [*c.f.* Bellman and Kalaba (1959)] has been a very important problem in data mining/machine learning applications as it requires high processing power and storage capacity which causes the model to overfit and increases the error rate of the learning algorithm. A Feature X_i is said to be an irrelevant feature if the output Y is not conditionally dependent on X_i [*c.f.* John *et al.* (1994)]. The features that are highly associated with each other are called redundant features [*c.f.* Kumar and Minz (2014)]. In high-dimensional datasets, feature reduction or selection is used as a dimensionality reduction method and is employed in machine learning and data mining. Here, a subset of the features is selected for the learning algorithm that leaves out the irrelevant and redundant features.

Feature selection can also be categorized into three types based on the presence of class label information: supervised feature selection, unsupervised feature selection and semi-supervised feature selection. In supervised feature selection,feature importance is calculated based on the degree of association between the class and the feature [*c.f.* Benabdeslem and Hindawi (2011)], [*c.f.* Song *et al.* (2016)] and [*c.f.*Chang *et al.* (2014)]. However, one challenge with this approach is the costly and potentially unreliable process of labeling the data using external knowledge [*c.f.* Kalakech *et al.* (2011)].

Traditionally, the performance of supervised approaches has been compared to unsupervised approaches [*c.f.* Yarowsky (1995)]. With the abundance of unlabeled data and the cost associated with labeling, the importance of semi-supervised learning has increased. Though there has been a lot of theoretical work in this domain [*c.f.* Zhao and Liu (2007)], the exploration of semi-supervised feature selection has been relatively limited. However, the introduction of semi-supervised feature selection has expanded the range of applications [*c.f.* Rosenberg *et al.* (2005)], [*c.f.* Dópido *et al.* (2013)], [*c.f.* Wu *et al.* (2012)] and [*c.f.* Kok *et al.* (2021)]. Contemporary research work focused on this domain covers a wide range of applications, from detecting liver diseases [*c.f.* Tran *et al.* (2019)] to semiconductor materials and manufacturing [*c.f.* Liu *et al.* (2022)]. Typically, in semi-supervised learning setups, a classifier trained on the labeled data is used to assign pseudo-labels to the unlabeled portion of the dataset. A dilemma exists regarding whether to fully utilize the pseudo-labeled data, as this may introduce bias if the classifier assigns incorrect labels.

In this paper, a simple semi-supervised feature selection method is proposed. In different iterations, various samples of the pseudo-labeled data are selected, creating multiple augmented training sets. On each of these augmented training sets, a simple mutual information based feature selection is applied, with a threshold for feature redundancy. Hence, in n iterations, n such feature sets are retrieved. However, the selection of the final feature set from the n feature sets each having k features is determined using a voting mechanism based

on the most repetitive and non-redundant features. The feature set produced in this manner outperforms all features set and also the partially labeled dataset after applying supervised selection in terms of both F1-Score and feature stability.

The contributions of the paper are as follows:

i. A novel and simple sampling-based feature selection technique for semi-supervised learning is proposed.

ii. An intuitive voting mechanism is employed to derive the final feature subset by aggregating multiple intermediate subsets, ensuring the selection of the most relevant and non-redundant features.

iii. With only 20% of the features, the proposed algorithm outperforms the supervised feature selection by 2.79% and the benchmark by 0.36% based on the F1 Score evaluated across 18 benchmark datasets.

iv. The proposed algorithm exhibits a time complexity of O(nd), which is significantly lower than that of existing semi-supervised models.

v. As the secondary focus of the proposed algorithm was on stability by similarity of the feature sets, the feature set of the proposed algorithm achieved a mean similarity of 39.4% with benchmark, and 42.64% with the supervised model feature sets.

The remainder of this work is organized into the following sections. In the literature survey *i.e.* Section 2, the related work in semi-supervised learning is discussed. The preliminaries Section provides an overview of fundamental concepts related to feature selection and semi-supervised learning. In Section 4, the proposed methodology is explained through a flow diagram followed by the algorithm used. The simulation experiment Section details the experimental setup and environmental considerations. In the result and discussion Section, the experimental results are presented, along with an analysis. The proposed work is summarized in the conclusions Section, which also outlines future scope.

2. Literature survey

Semi-supervised feature selection methods can be classified based on different perspectives. The viewpoint discussed here is based on the fundamental nomenclature of feature selection methods, which categorizes semi-supervised feature selection methods into several categories depending on how they interact with the learning algorithm. The first way of classification is similar to regular feature selection categories, namely a) filter b) wrapper c) hybrid d) embedded. More specific categorizations from the perspective of semi-supervised feature selection are i) Co-training ii) Self-Training iii) Entropy Based methods *etc.*

Co-Training: This is a method of semi-supervised learning. In the first step, the entire feature set is divided into two subsets, let us call them f_1 and f_2. They are mutually exclusive and exhaustive. Hence $f_1 \cup f_2 = f$, and, $f_1 \cap f_2 = \phi$. Next, two classifiers C1 and C2 are trained on the f1 and f2 view of the labeled data. Now the algorithm proceeds iteratively. In each iteration, an unlabeled portion of the data, on which C1 is most confident, is added to the training set of C2 and vice versa. These training sets are often called augmented training sets, let us denote them by T_1 and T_2. A simpler way to extend this to feature selection is to select top k features separately from T_1 and T_2 based on measures like feature

selection and take a union from them.

Self-Training: Similar to Co-Training, Self-Training is also a kind of semi-supervised learning. Initially, a particular classifier is selected. Then iteratively predicted points from the unlabeled set are added to the training set based on confidence of the prediction. Similar to co-training, such a training set is called the augmented training set. Feature selection now can be done in the usual manner from the augmented training set. The major developments in semi-supervised learning are summarized in Table 1 shown next.

In this section, the different models of semi-supervised learning that have evolved are summarized, along with their significance or drawbacks, discussed in Table 1. This study has helped to draw the motivation for this work, as discussed next.

The time complexity of existing methods like Laplacian Score(semi-supervised) [*c.f.* He *et al.* (2005)] is O(n^2d), while Semi-Supervised Feature Selection via Spectral Analysis(SSFS) [*c.f.* Cai *et al.* (2007)] is O(n^2d+d^3) and Semi-supervised Feature Selection using Max-margin Criterion (SSMMC) [*c.f.* Wu *et al.* (2013)] is O(nd^2+d^3) where n is the number of samples and d is the number of features. Hence, improving the time complexity was the foremost motivation and hence the algorithm was proposed with a complexity of O(nd).

This work was inspired by the self-training and co-training methods of wrapper-based approaches. This research is different in the sense that it does not only add the most confident points based on a confidence measure of prediction but, through sampling, adds a portion of the predicted samples to the partially labeled dataset.This is done to ensure that there is no bias against the confident points. It operates by taking the predicted labels, which, in turn, depend on the classifier's performance. With this augmented training set, feature selection is implemented and compared with the supervised mode, which is discussed in the results and discussion section to justify this approach. The stability of the selected feature subset is also studied to assess the quality of the feature selection process adopted.

3. Preliminaries

In this Section, the factors/criteria for the selection of features (mutual information and correlation coefficient), the evaluation method of feature sets based on stability, and the confidence of prediction using confidence score are discussed.

3.1. Mutual information

Mutual Information(MI) of a feature is the degree of dependence between the feature and the class variable. Mutual information of a feature, denoted as F, concerning the class variable C, is the difference between the entropy of the class $H(C)$ and the conditional entropy of the class variable given the value of the feature variable $H(C|F)$. The following equations are used to calculate the feature-to-class relationship:

$$I(F;C) = H(C) - H(C|F) \tag{1}$$

$$I(F;C) = H(C) + H(F) - H(C.F) \tag{2}$$

Table 1: Summary of semi-supervised methods based on the taxonomy of feature selection methods

Related Works		
Methodology	**Method**	**Conclusion/Remarks**
Graph theory and cluster assumption [*c.f.* Zhao and Liu (2007)] Type-Filter.	• It constructs a neighborhood graph and transforms each feature vector into a cluster indicator, which is evaluated based on separability and consistency. • Laplacian score is associated with normalized mutual information.	• The method overlooks the correlation between features and evaluates the features one by one.
Laplacian score [*c.f.* Cheng *et al.* (2011)] and [*c.f.* Zhao *et al.* (2008)] Type-Filter.	• It constructs a within-class graph and a between-class graph. • The features are estimated through their degree of preserving the graph structures.	• The method disregards the correlation among features and evaluates the features one by one.
Constraint score [*c.f.* Kalakech *et al.* (2011)] and [*c.f.* Benabdeslem and Hindawi (2011)] Type-Filter.	• The metric uses some supervision information in the form of pairwise constraints. • It constructs two graphs using pairwise constraints and unlabeled data. • Finally, it evaluates the features based on their locality and constraint-preserving ability.	• It depends on the subsets of pairwise constraints created by the user. • The constraints can be redundant or incoherent. The method evaluates the features individually, ignoring the correlation among features.
Fisher criterion [*c.f.* Chen *et al.* (2010)], [*c.f.* Lv *et al.* (2013)], [*c.f.* Liu *et al.* (2013)] and [*c.f.* Liu *et al.* (2010)] Type-Filter.	• This metric utilizes the Fisher criterion and considers the local structure of both labeled and unlabeled data. • The features are estimated based on their discriminant and locality-preserving abilities.	• The method disregards the correlation among features and evaluates the features one by one.

Methodology	Method	Conclusion/Remarks
Sparse-based filter methods [*c.f.* Han *et al.* (2014)] Type-Filter.	• It combines two supervised and unsupervised scatter matrices. • It preserves the discriminant information from labeled data and the local geometric structure from both labeled and unlabeled data. • It adds a l2 norm to the objective function, making it suitable for feature selection. • The method utilizes an iterative algorithm to solve the objective function.	• The features are estimated jointly while considering the correlation among the features. The objective function is non-smooth and difficult to solve.
Single Learner [*c.f.* Ren *et al.* (2008)] Type-Wrapper.	• The initial labeled training set is augmented with predicted unlabeled data. • Data is randomly selected from unlabeled set to create new training sets. • Next, it adds the most frequently selected feature to the feature subset during each iteration	• The method ignores confidence measures for unlabeled data. • On adding mislabeled data, it may degrade performance. • The method neglects the discriminative power of feature combinations. • It takes high computational time.
Ensemble learning [*c.f.* Bellal *et al.* (2012)], [*c.f.* Han *et al.* (2011)] and [*c.f.* Barkia *et al.* (2011)] Type-Wrapper.	• The method uses ensemble learning with self-training or co-training to predict the labels of unlabeled data.	• It considers the reliance on features. • It uses a confidence measure to select predicted unlabeled data. • But if the confidence measures are inaccurate, this may lead to mislabeling of data.

Methodology	Method	Conclusion/Remarks
		• It enhances generalization ability using an ensemble classifier. • It takes high computational time.
Sparse-based Embedded methods [*c.f.* Song *et al.* (2016)], [*c.f.* Ma *et al.* (2012)], [*c.f.* Shi *et al.* (2014)] and [*c.f.* Ma *et al.* (2011)].	• The method directly learns the classifiers during feature selection. • It constructs a graph using both labeled and unlabeled data.	• This combines the strengths of joint feature selection and semi-supervised learning.
SVM-based [*c.f.* Yang and Wang (2007)], [*c.f.* Xu *et al.* (2010)] and [*c.f.* Dai *et al.* (2013)] Type-Embedded.	• Select the features by maximizing the margin between different classes and at the same time exploiting the local structure of both labeled and unlabeled data.	• The objective function is difficult to solve.

3.2. Correlation coefficient

Correlation coefficient is a statistical measure of the strength of the relationship between two variables. The correlation coefficient, denoted as (ρ), between two variables x and y is defined as follows:

$$\rho(x, y) = \frac{cov(x, y)}{\sqrt{var(x)var(y)}} \tag{3}$$

where $cov(x, y)$ refers to the covariance between x and y , and $var(x)$ is the variance of x.

$$cov(x, y) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{4}$$

where $\bar{x}$ and $\bar{y}$ are the means of x and y respectively.

$$var(x) = \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{5}$$

3.3. Stability of feature sets

The stability of a feature selection algorithm results in a persistent feature subset when new training samples are added or removed [*c.f.* Xin *et al.* (2015)]. Stability can be categorized into three groups: stability by rank, stability by weight, and stability by

similarity [*c.f.* Khaire and Dhanalakshmi (2022)] and [*c.f.* Chelvan and Perumal (2016)]. In this work, we consider stability by similarity between feature sets. In this case, similarity is estimated by using the ratio of the intersection to the union of two selected feature subsets or the amount of overlap between the overall subset of selected features [*c.f.* Khaire and Dhanalakshmi (2022)] and [*c.f.* Yu *et al.* (2011)].

Let X and Y be two different feature sets such that

$X = \{x_1, x_2, x_3, x_4, x_5\}$ and $Y = \{y_1, y_2, y_3, y_4, y_5\}$ where $x_1 \ldots\ldots x_5$ and $y_1 \ldots\ldots y_5$ are the individual feature components such that $x_1 = y_2$ and $x_3 = y_5$

$$Similarity = \frac{|X \cap Y|}{|X \cup Y|} \tag{6}$$

$|X \cup Y| = 8, |X \cap Y| = 2, Similarity = \frac{2}{8} = 0.25$

3.4. Confidence score

The confidence score is used in predictions made by models. For binary classification or a yes-no answer, it predicts 0/1 based on a score known as the confidence score. A confidence score is a number between 0 and 1, representing the likelihood that the output of a machine learning model is correct and will satisfy a user's request. One way to interpret it, is by considering a 'yes' or '1' if the value is > 0.5, where 0.5 serves as the minimum confidence score or threshold. Increasing the threshold will result in lower recall and improved precision.

4. Proposed methodology

The working of the proposed method is explained in Section 4.1 followed by the algorithm in Section 4.2 and finally the complexity analysis in Section 4.3.

4.1. Working of the algorithm

The step-by-step working of the algorithm is mentioned below:
Step 1 (Divide Dataset): To simulate the semi-supervised setup, a part of the dataset is unlabeled and the labeled portion is divided into train and test as shown in Figure 1.

Figure 1: Splitting the dataset

Step 2 (Train using classifier and add label to unlabeled): The selected classifier is trained on the training dataset and the unlabeled portion is now labeled using the classifier as shown in Figure 2.

Step 3 (Construct different sample sets from the pseudo label and form augmented training set) : This sample of pseudo labeled data is added to the training

Figure 2: Labeling the unlabeled to pseudo label

set each time to form an augmented training set as shown in Figure 3.

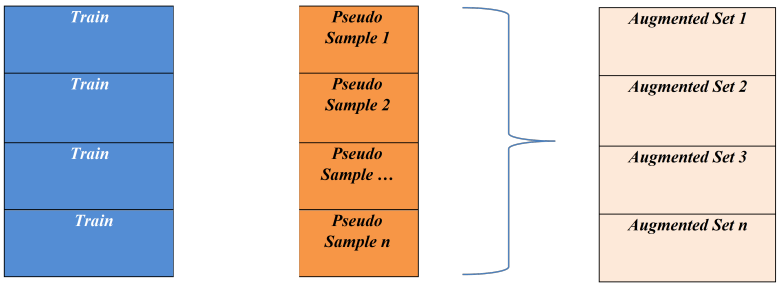


Figure 3: Construction of Augmented set

Step 4 (Select *K* features from each augmented set based on MI and Redundancy Level): Using the set, feature selection is performed based on MI and redundancy threshold to form n feature sets into the Master Feature Set(MFS) as shown in Figure 4.

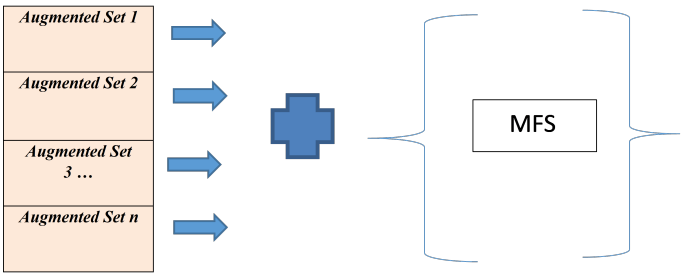


Figure 4: Construction of Master Feature Set

Step 5 (Select *K* features into FFS based on frequency count and Redundancy Level): Once these n feature sets are obtained in MFS, voting has been done to select the final k features into Final Feature Set(FFS) as shown in Figure 5. Here, the redundant features are eliminated.

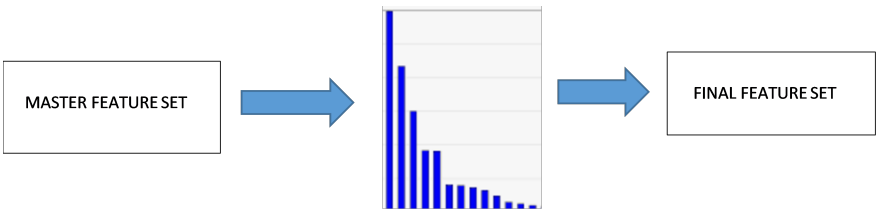


Figure 5: Retrieval of the Final Feature Set

4.2. Algorithm RSSFS (Repeated Sampled Semi-Supervised Feature Selection)

Input (D_L: Labeled Dataset,D_U: Unlabeled Dataset, C: Classifier, α: Redundancy Threshold, I: No of Iterations, p:Randomness, K: No of Features (can be an absolute value or a percentage))
Output: *FFS* // Final Feature set
Initialization :
$MFS \rightarrow \phi$,$FFS \rightarrow \phi$// MFS and FFS are set to empty set
Step 1: Divide D_L into Train + Test
Step 2: Train C on Train and predict the labels for D_U and Assign to *D_Pseudo*.
Step 3:Construct *D_Augment*
For iteration (i) 1 to I
Step 3a: p % of observations selected from *D_Pseudo* and Assigned to $D_PseudoSample_i$
Step 3b: $D_PseudoSample_i$ are added to Train to form $D_Augment_i$
Step4: Select K features from $D_Augment_i$ to form Master Feature Set(*MFS*)
Step 4a: Set K to 0
For iteration (i)= 1 to I
Step 4b: The First feature is added to MFS_i based on maximum mutual information with the target variable
Step 4c: The next feature is added which has the next highest mutual information and does not cross the redundancy Threshold α, with any of the selected features in MFS_i.
Step 4d: Increment K, Continue up to K features
Step 5: Compose the final feature set *FFS* from the *MFS*
Step 5a: Out of the $I * k$ features in *MFS*, count the frequency of the features
Step 5b: The first feature is added to *FFS*, based on the highest frequency
Step 5c: Set K =1
Step 5d: The next feature is added which has the next highest frequency count and does not cross the redundancy Threshold α, with any of the selected features in *FFS*
Step 5e: Increment K, Continue up to K features
Return *FFS*

4.3. Case study

Let us assume that we have a dataset with thirty features, denoted as f_1 to f_{30}, for binary classification. T represents the redundancy threshold, and our goal is to select the top 20 % of non-redundant augmented features, which starts as an empty set. So, by the end of step five, after six iterations, the features selected by RSSFS, *i.e. ssl_set*, and the supervised set, *i.e. sl_set*, are mentioned using their column indices which are as follows:

ssl_set = [[1,3, 7, 11, 13,21], [1,3, 10, 12, 20,22], [1,3, 9, 11, 23,25], [1,3, 7, 19, 17,30], [1,3, 7, 11, 20,29], [1,5, 7, 11, 23,30]]
sl_set = [[3.7,11,17,21,28], [3.7,11,17,21,28], [3.7,11,17,23,19], [3.7,11,18,25,29], [3.7,9,16,27,22], [3.6,10,12,24,26]]

The most frequent features from *ssl_set* and *sl_set* are enlisted in the *freq_ssl_set* and *freq_sl_set* respectively as shown next.

freq_ssl_set = [1,3,7,11,23, 30]
freq_sl_set = [3.7,11,17,21,28]

Find the similarity between the frequent feature sets of semi-supervised and supervised.

Similar features = [3,7,11]
$|Similar features| = 3$
Total features = [1,3,7,11,17,21,23,28,30]
$|Total features| = 9$

$$Similarity = \frac{|Similar features|}{|Total features|} = 3/9 = 33.33\% \tag{7}$$

4.4. Computational complexity

With n being the number of samples and d being the number of attributes, time complexity of feature selection is O(nd). Finding the frequent feature set from the feature sets has a complexity of O(nd). The complexity of evaluating the precision and recall of the frequent feature sets is O(n). Hence, the overall complexity is of the quadratic order as evident in equation 9 shown next.

$$T(n) = O(nd) + O(nd) + O(n) \tag{8}$$

$$T(n) = O(nd) \tag{9}$$

5. Simulation experiment

In this Section, the working environment is described in four subsections which comprises of the dataset description, the experimental setup, the parameter sensitivity analysis and the performance metrics used.

5.1. Dataset description

The datasets used are available in public online repositories of the University of California, Irvine [*c.f.* Lichman *et al.* (2013)]. The proposed algorithm is compared with the benchmark having all the features and the supervised feature selection having the top k% features. These three methods were conducted on eighteen datasets. Among these, six datasets have binary classes, while twelve datasets have multi-class labels, with the isolet dataset having a maximum of twenty six classes. The dataset with the highest number of features used is six hundred and seventeen in the isolet dataset. The dataset with the maximum number of patterns used is 42,158 in the Dry Bean dataset. Details of the datasets are provided in Table 2.

While working, it has been observed that some of the datasets were similar to each other in terms of input size, where the input size of a dataset is the product of the number of records and the number of features/attributes in the dataset. These datasets are then categorized into the same group. Ultimately, this categorization resulted in the creation of three distinct categories based on input size. The first category *i.e.* Group A has an input size up to 50000, whereas the input size for the second group, B, is from 50001-300000, and for the third group C, it ranges from 300001-1000000. The main reason behind this approach is to study and analyze the performance of the three methods across these three different categories having varying input size.

Table 2: Details of the datasets used

Sl No.	Dataset	Records	Features	InputSize	Category
1	lung_cancer	32	57	1824	Group-A/1 (0-50000)
2	wine	178	13	2314	
3	lymphography	147	19	2793	
4	cleave	297	13	3861	
5	sonar	208	60	12480	
6	Vehicle	846	18	15228	
7	wbdc	569	30	17070	
8	ctg	2126	34	72284	Group-B/2 (50001- 300000)
9	Colon	62	2000	124000	
10	Arythmia	452	279	126108	
11	mfeat-karhunen	2000	64	128000	
12	texture	5500	40	220000	
13	Spambase	4601	57	262257	
14	digits	1593	256	407808	Group-C/3 (300001-1000000)
15	Dry Bean	42158	16	674528	
16	ECG	4998	140	699720	
17	isolet	1559	617	961903	
18	Madelon	2000	500	1000000	

5.2. Experimental setup

i. The processor used is a Core i5, with 8GB of memory.

ii. The operating system used in this context was Windows 10, 64-bit, and the computing environment Python 3.6 was used for the experiments. Various Python libraries were also utilized.

iii. The classifier used is the decision tree classifier.

iv. This experiment was conducted having different proportions of unlabeled, training, and test data. Initially, the dataset was divided into a ratio of 50% unlabeled, 30% training, and 20% test data. The dataset was trained using a random 30% of the samples. This knowledge was then used to produce pseudo-labels for the remaining 50% of unlabeled data.

v. A separate test was performed where the proposed algorithm was compared against Supervised Feature Selection (SFS) which was retrieved from the labeled samples only having an equal number of features compared to RSSFS, and the Benchmark having all the features.

vi. While performing feature selection for all the three methods the top k% non-redundant features were selected. The value of k is chosen to be twenty in this case, as ten would be very insignificant and thirty would be much more significant.

vii. The final feature set for RSSFS and SFS is retrieved by selecting the most frequent features from n feature sets retrieved in n iterations with different seed values. The value of n is chosen to be one hundred in this case, as it would provide an average of the results on different random samples.

viii. While selecting features it is ensured that the correlation between the features in the feature subset is less than 0.67.

ix. Additionally, the goal was to assess the category-wise performance, which is the average accuracy of these three methods that is calculated for all the datasets within a particular category.

x. For each dataset, feature selection for RSSFS is performed using the mixed training set, which combines original and pseudo labels, to select the top $k\%$ of original non-redundant features. From this feature set, metrics are calculated for the semi-supervised set. Similarly, using the original training set, feature selection for SFS is conducted to select the top $k\%$ of original non-redundant features, and classification metrics are calculated for the supervised set. This process is repeated for one hundred iterations. Next, from the two feature sets, each composed of one fundred feature sets, each of size k, the frequency of each feature in both sets is determined. Based on the frequency of the features, two feature sets are finally composed, each containing the k most frequent features.

5.3. Parameter sensitivity analysis

In the adopted strategy RSSFS, the parameters chosen are redundancy threshold α, number of iterations I and randomness parameter p, and fine-tuning these parameters may affect the proposed algorithm to a certain extent as discussed next.

i. **Redundancy threshold** α: The redundancy threshold used is 0.67. This is because if the value is increased to 0.75 or 0.72, sometimes in an iteration a fixed number of k features cannot be retrieved, as it goes down to less than k features say *k-1* or *k-2* features which results in a different number of feature components in different iterations. Hence the strategy is sensitive to the redundancy threshold.

ii. **Number of iterations** $\boldsymbol{I}$: The number of iterations used uniformly for all datasets was one hundred. However, for certain datasets, it was tested with one hundred and twenty but the metrics calculated such as precision, recall and F1-score did not vary much. Hence, the method is not sensitive to a much higher value of I.

iii. **Randomness parameter** $\boldsymbol{p}$: The randomness parameter p used here is 60% which suggests that for a distribution of one hundred samples with 50% unlabeled samples only 60% *i.e.* thirty pseudo samples are added to the earlier training set. If the randomness is further increased, then there is a chance of picking the majority of the pseudo samples. Also if the randomness is decreased then again there is a chance of picking less number of augmented samples into the training set. In both ways, the earlier training set augmented may be affected, where either the pseudo samples may dominate, or, the training set from label data may have more preference. Hence the method is sensitive to randomness parameter.

5.4. Performance metrics

In the case of RSSFS and SFS, on computing the mutual information for all original and extracted features, the top 20% of features were selected. To validate the proposed methodology, three perspectives or views were utilized, as described below:

i. The classification performance metrics used for comparing the performance of the semi-supervised feature set against the supervised feature set included precision, recall and F1 scores.

ii. For visualizing graphically, sensitivity and specificity were also considered from where the AUC(Area Under the Curve) was computed.

iii. To assess the stability of the feature sets, stability by similarity was calculated. This involved determining the percentage of common feature components in both the sets.

iv. As the proposed algorithm was compared against the benchmark and Supervised Feature Selection(SFS), a t-test was conducted to analyze the statistical significance.

6. Results and discussion

In this Section, the results of the research work conducted are provided, followed by a comprehensive discussion. The result analysis is mainly divided into four Subsections namely comparison test, t-test, analysis of similarity of feature sets and AUC analysis.

6.1. Comparison test: performance of RSSFS against Benchmark and SFS

In each iteration, the performance is evaluated thrice, first using the entire feature set that sets the benchmark, secondly, using a semi-supervised training set *i.e.* the proposed algorithm (RSSFS), and third using a portion of the train set with labels which is the supervised mode of feature selection. After one hundred iterations, the average precision, recall and F1 Score are recorded. Here, the proposed algorithm with 20% features is compared against two methods, firstly the benchmark having 100 % features, and secondly, the Supervised Feature Selection (SFS) with 20% features.

The results of the performance of these methods are shown next in Table 3 for all the 18 datasets used individually. Also, the average scores of the performance metrics used *i.e.* precision, recall and F1 Score for each of the three categories A, B and C have been computed to draw a conclusion. From Table 3, the observations are mentioned as follows :

- In Group A, the proposed algorithm outperforms supervised feature selection by 1.68%, 1.58%, and 1.65% in precision, recall, and F1 Score, respectively. Also, it lags behind the benchmark by a marginal amount *i.e.* 0.44%, 0.54% and 0.34% in precision, recall and F1 Score respectively.
- In Group B, the proposed algorithm outperforms supervised feature selection by 5.18%, 4.38%, and 4.85% in precision, recall, and F1 Score, respectively. Also, it exceeds the benchmark by 1.7%, 2.15% and 1.92% in precision, recall and F1 Score respectively.
- In Group C, the proposed algorithm outperforms supervised feature selection by 1.52% in precision, 2.26% in recall, and 1.89% in F1 Score. Also, it lags behind the benchmark by 0.85%,0.17% and 0.51% in precision, recall and F1 Score respectively.
- In each of the three categories, the proposed algorithm has outperformed the supervised feature selection, and is at par with the benchmark.

Table 3: Comparison of the performance of the proposed algorithm against Benchmark and Supervised Feature Selection

		Benchmark			RSSFS			Supervised Feature Selection		
Sl No.	Dataset	Avg.Pr.	Avg.Re.	F1	Avg.Pr.	Avg.Re.	F1	Avg.Pr.	Avg.R.	F1
1	lung_cancer	56.42	38.57	45.82	65.07	52.85	58.33	64.81	51.42	57.35
2	wine	97.56	97.22	97.39	90.67	89.58	90.13	90.43	88.19	89.3
3	lymphography	72.66	74.85	73.74	77.27	74	75.6	75.41	72.66	74.01
4	cleave	57.21	59.5	58.34	52.94	55.16	54.03	50.93	54.66	52.73
5	Sonar	79.48	78.57	79.03	72.19	72.02	72.11	71.9	71.42	71.66
6	Vehicle	68.08	68.52	68.3	69.94	69.67	69.81	65.32	66.29	65.81
7	wbdc	94.35	94.3	94.33	94.58	94.52	94.55	92.15	92.1	92.13
Group-A Average		**75.11**	**73.08**	**73.85**	**74.67**	**72.54**	**73.51**	**72.99**	**70.96**	**71.86**
8	ctg	91.03	90.61	90.82	90.08	90.07	90.08	86.7	86.9	86.8
9	Colon	77.12	74.61	75.85	77.11	75.38	76.24	71.35	68.46	69.88
10	Arythmia	71.13	72.97	72.04	68.33	71.78	70.02	50.26	59.88	54.65
11	mfeat-karhunen	92.56	92.37	92.47	91.89	91.62	91.76	88.67	88	88.34
12	texture	78.75	78.78	78.77	93.04	92.99	93.02	92.74	92.69	92.72
13	Spambase	92.54	92.51	92.53	92.91	92.9	92.91	92.53	92.5	92.52
Group-B Average		**83.86**	**83.64**	**83.75**	**85.56**	**85.79**	**85.67**	**80.38**	**81.41**	**80.82**
14	digits	86.31	86.01	86.16	86.8	87.28	87.04	81.92	81.27	81.6
15	Dry Bean	90.71	90.68	90.7	90.15	90.11	90.13	89.77	89.73	89.75
16	ECG	98.61	98.62	98.62	98.67	98.67	98.67	98.42	98.42	98.42
17	isolet	85.88	84.58	85.23	84.1	85.41	84.75	84.51	83.05	83.78
18	Madelon	69.77	69.65	69.71	67.32	67.22	67.27	64.82	64.91	64.87
Group-C Average		**86.26**	**85.91**	**86.08**	**85.41**	**85.74**	**85.57**	**83.89**	**83.48**	**83.68**

Table 4: Comparative summary of the performance of the proposed algorithm against Benchmark and Supervised Feature Selection

		Benchmark			RSSFS			Supervised Feature Selection		
Sl No.	Dataset	Avg.Pr.	Avg.Re.	F1	Avg.Pr.	Avg.Re.	F1	Avg.Pr.	Avg.R.	F1
1-7	Group-A	75.11	73.08	73.85	74.67	72.54	73.51	72.99	70.96	71.86
8-13	Group-B	83.86	83.64	83.75	85.56	85.79	85.67	80.38	81.41	80.82
14-18	Group-C	86.26	85.91	86.08	85.41	85.74	85.57	83.89	83.48	83.68
	Avg.	**81.74**	**80.87**	**81.22**	**81.88**	**81.35**	**81.58**	**79.08**	**78.61**	**78.78**

To further summarize the performance of the proposed algorithm against the Benchmark and Supervised Feature Selection(SFS) a study is done as shown next in Table 4.

Analyzing Table 4, the observations are mentioned as follows:

- Across eighteen datasets, the proposed algorithm, RSSFS outperforms supervised feature selection (SFS) by 2.79% in precision, 2.74% in recall, and 2.79% in F1 Score.
- Also, the proposed algorithm, RSSFS excels over the benchmark by 0.13%,0.48% and 0.36% in precision, recall, and F1 Score, respectively.

To compare the behavior of F1 Scores across different input sizes categorized into distinct classes, please refer to Figure 6. Analyzing Figure 6, it is evident that the proposed algorithm has dominated over the benchmark and SFS in Group-B and Group-C. For an alternative perspective, the digit dataset was used to capture multiple readings of F1-scores through a box plot as displayed in Figure 7 next.

On analyzing the box plot on the digit dataset as shown in Figure 7 above, the observations are mentioned as follows:

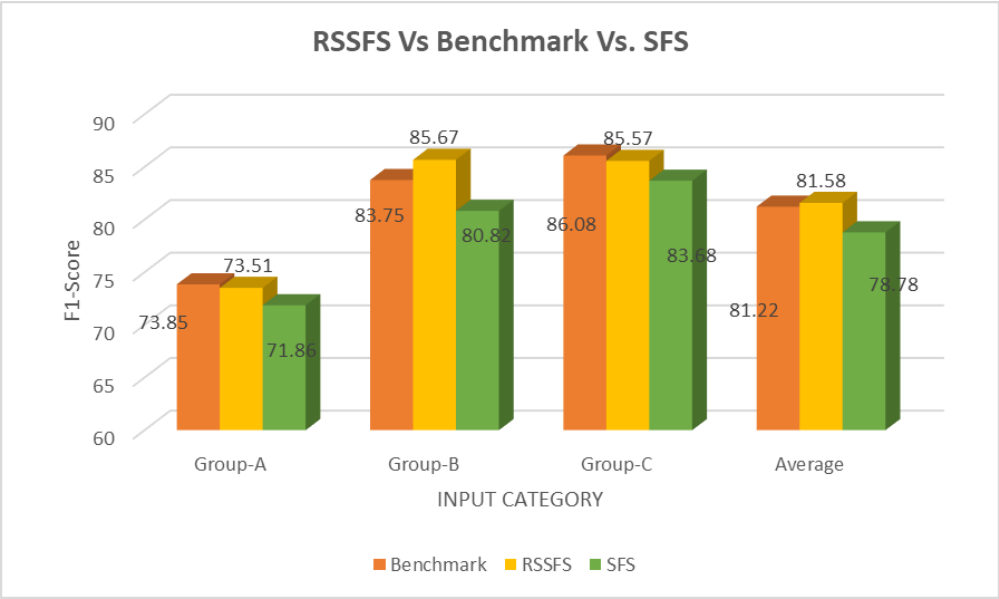


Figure 6: Comparison of performance of RSSFS against benchmark and SFS

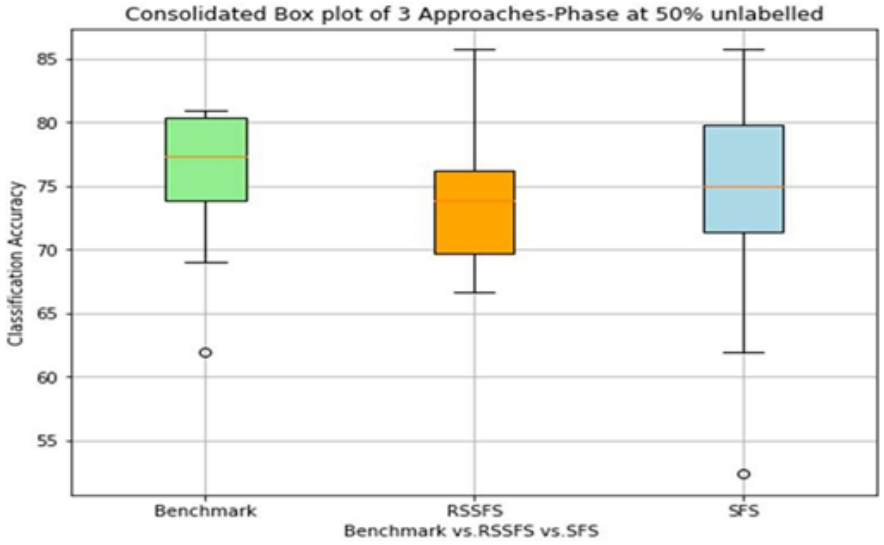


Figure 7: Box plot of F1 scores of the most frequent feature set for digit dataset

- The benchmark and SFS have performed better than the proposed algorithm concerning accuracy based on F1 Scores.
- The Interquartile Range(IQR) for the proposed algorithm is around 6%, whereas that of the benchmark and SFS is around 7% and 9% respectively, which suggests a lesser dispersion for the proposed algorithm.

6.2. Test of statistical significance through t-test

Since the number of samples was eighteen that is than thirty, a t-test was used instead of a z-test. The t-test was applied to compare the performance of our proposed algorithm against Supervised Feature Selection (SFS). The resulting p-value was found to be less than

Table 5: Comparison of similarity of frequent feature sets between proposed algorithm and SFS, and, proposed algorithm and benchmark

Sl No.	Dataset	Similarity between RSSFS and SFS (%)	Similarity between RSSFS and Benchmark (%)
1	lung_cancer	41.17	33.33
2	wine	50	50
3	lymphography	33.33	60
4	cleave	50	50
5	Sonar	33.33	20
6	Vehicle	100	60
7	wbdc	33.33	20
	Group-A Average	**48.74**	**41.9**
8	ctg	55.55	75
9	Colon	66.07	43.9
10	Arythmia	45.45	12
11	mfeat-karhunen	85.71	85.67
12	texture	60	30
13	Spambase	84.61	71.42
	Group-B Average	**66.23**	**53**
14	digits	33.33	42.46
15	Dry Bean	60	100
16	ECG	60	40
17	isolet	39.35	32.62
18	Madelon	23.45	11.73
	Group-C Average	**43.23**	**45.36**
	Mean	**42.64**	**39.4**
	Standard Deviation	15.07	26.91

0.05, indicating that the 2.79% improvement achieved by our proposed algorithm over SFS is statistically significant. In contrast, the 0.36% reduction in performance compared to using all features was not statistically significant.

6.3. Analysis of stability using the similarity of feature sets

The extent of similarity between the feature components found in the feature sets of the proposed algorithm and those of supervised feature selection and the benchmark, is a crucial aspect of the study, thus serving as an indicator of stability. Therefore, a detailed comparison is made between the most frequent semi-supervised feature set along with the most frequent supervised feature set retrieved after n operations, and the full feature set considered for the benchmark is presented in Table 5 next.

From Table 5, it can be observed that the mean similarity between the benchmark and RSSFS is 39.4%, and the mean similarity between SFS and RSSFS is 42.64%. The similarity between the most frequent feature sets from the two methods, namely the proposed algorithm and the supervised mode of feature selection along with the full feature set of the benchmark, are also evaluated as they provide insight into the stability of the feature sets.

This is pictorially represented in Figure 8 as shown next.

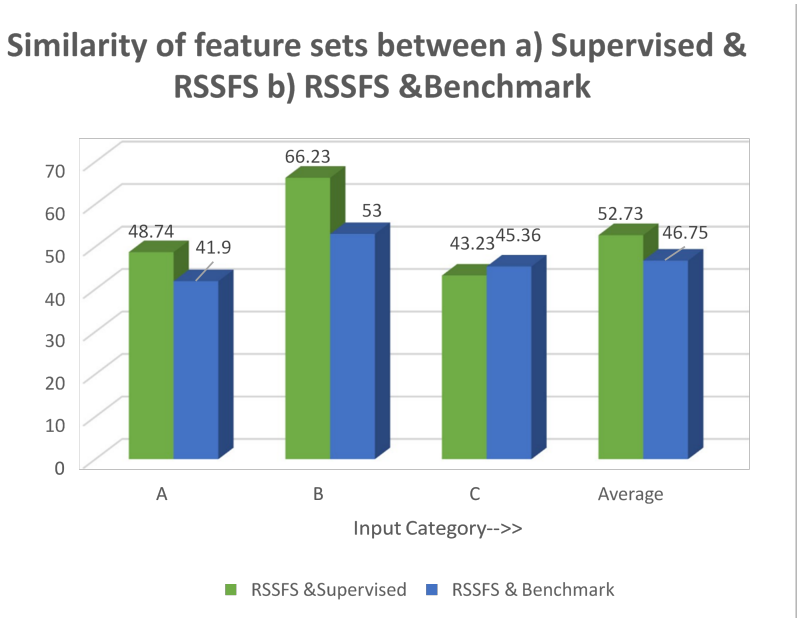


Figure 8: Similarity of the feature sets of the proposed algorithm against benchmark and SFS in all categories

From Figure 8, it can be seen that the similarity of the feature components of the proposed algorithm with the benchmark and that of the supervised feature selection is in the range of 40-60% approximately in all the three categories.

6.4. Analysis of the Area Under the Curve (AUC) from the RoC curve

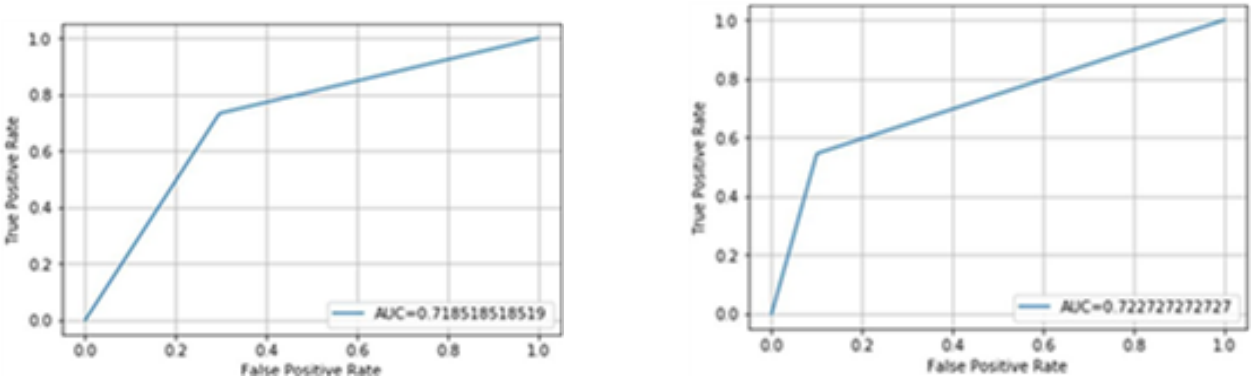
Another way to justify the proposed algorithm is by considering metrics such as sensitivity and specificity. In this analysis, the AUC was calculated from the RoC [*c.f.* Fawcett (2006)]. The performance of the proposed algorithm is compared against Supervised Feature Selection (SFS) using the sonar dataset as shown next in Figure 9 and the isolet dataset as shown in Figure 10. The classifier used in these tests is decision tree.

On analyzing both the figures 9 and 10, the observations are mentioned as follows:

i. For the sonar dataset, the proposed algorithm outperforms SFS, exceeding it by 0.42% in terms of AUC, as shown in Figures 9 (a) and (b).
ii. For the isolet dataset, the proposed algorithm again outperforms SFS significantly, exceeding it by 6.67% in terms of AUC, as shown in Figures 10 (a) and (b). Hence for both the categories A/1 and category C/3 the proposed algorithm excels over SFS.

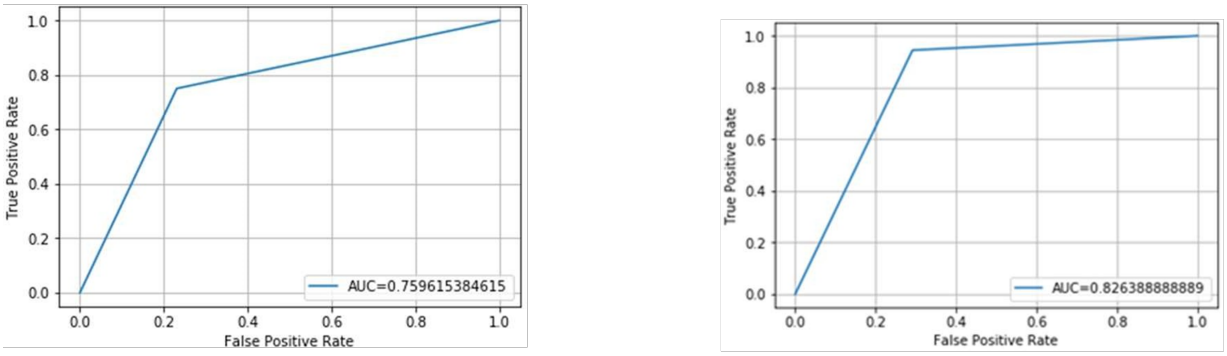
7. Conclusions

The proposed work aimed to implement feature subset selection in partially labeled datasets, with the majority portion being unlabeled. This methodology was evaluated on 18 datasets and compared against all features and supervised feature selection . For comparison,



(a) SFS **(b) RSSFS**

Figure 9: sonar dataset–RoC curve and the AUC



(a) SFS **(b) RSSFS**

Figure 10: isolet dataset–RoC curve and the AUC

conventional performance metrics and similarity measures were used. On analyzing across 18 datasets,the major observations are summarized as follows:

i. 20% of selected features of the proposed algorithm outperformed the 20% features of the supervised set and the benchmark having the full feature set, by 2.79% and 0.36% respectively based on the F1 Score.

ii. The mean similarity of the feature components of the feature sets among benchmark and RSSFS is 39.4%, while the mean similarity among the feature sets of SFS and RSSFS is 42.64%.

In datasets where the number of labeled data is much less, maybe 5-10% at most, with 90-95% having unlabeled data, it is challenging for the proposed algorithm, and here, the performance metrics may be affected. Also, in datasets where there is a clear imbalance in the number of classes /label representations, then the proposed method is prone to failure as it may learn in a biased environment given the dataset.

While the proposed work has been validated on representative datasets, future research will focus on improving the model's performance in scenarios where labeled data is extremely limited. Also in datasets, where there is a class imbalance problem, this presents an intriguing opportunity to develop a feasible solution.

Apart from predicting the labels, this work can also be applied in those domains where the availability of data on relevant features is scarce. In conclusion, it is anticipated that the significant gap between clustering and classification will greatly diminish. In essence, both methods can be viewed as special cases of semi-supervised learning, where either only labeled data or only unlabeled data is available.

Acknowledgements

We are indeed grateful to the Editors for their guidance and counsel. We are very grateful to the reviewer for the valuable comments and suggestions of generously listing many useful references.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Barkia, H., Elghazel, H., and Aussem, A. (2011). Semi-supervised feature importance evaluation with ensemble learning. In *2011 IEEE 11th International Conference on Data Mining*, 31–40.

Bellal, F., Elghazel, H., and Aussem, A. (2012). A semi-supervised feature ranking method with ensemble learning. *Pattern Recognition Letters*, **33**, 1426–1433.

Bellman, R. and Kalaba, R. (1959). A mathematical theory of adaptive control processes. *Proceedings of the National Academy of Sciences*, **45**, 1288–1290.

Benabdeslem, K. and Hindawi, M. (2011). Constrained laplacian score for semi-supervised feature selection. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part I 11*, 204–218.

Cai, D., He, X., and Han, J. (2007). Semi-supervised feature selection via spectral analysis. In *Proceedings of the 2007 SIAM International Conference on Data Mining (SDM)*, 199–207.

Chang, X., Shen, H., Wang, S., Liu, J., and Li, X. (2014). Semi-supervised feature analysis for multimedia annotation by mining label correlation. In *Advances in Knowledge Discovery and Data Mining: 18th Pacific-Asia Conference, PAKDD 2014, Tainan, Taiwan, May 13-16, 2014. Proceedings, Part II 18*, 74–85.

Chelvan, P. M. and Perumal, K. (2016). A study on selection stability measures for various feature selection algorithms. In *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 1–4. IEEE.

Chen, L., Huang, R., and Huang, W. (2010). Graph-based semi-supervised weighted band selection for classification of hyperspectral data. In *2010 International Conference on Audio, Language and Image Processing*, 1123–1126.

Cheng, H., Deng, W., Fu, C., Wang, Y., and Qin, Z. (2011). Graph-based semi-supervised feature selection with application to automatic spam image identification. In *Computer Science for Environmental Engineering and EcoInformatics: International Workshop, CSEEE 2011, Kunming, China, July 29-31, 2011, Proceedings, Part II*, 259–264.

Dai, K., Yu, H.-Y., Li, Q., et al. (2013). A semisupervised feature selection with support vector machine. *Journal of Applied Mathematics*, **2013(1)**, 416320.

Dópido, I., Li, J., Marpu, P. R., Plaza, A., Dias, J. M. B., and Benediktsson, J. A. (2013). Semisupervised self-learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, **51**, 4032–4044.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, **27**, 861–874.

Han, Y., Park, K., and Lee, Y.-K. (2011). Confident wrapper-type semi-supervised feature selection using an ensemble classifier. In *2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*, 4581–4586.

Han, Y., Yang, Y., Yan, Y., Ma, Z., Sebe, N., and Zhou, X. (2014). Semisupervised feature selection via spline regression for video semantic recognition. *IEEE Transactions on Neural Networks and Learning Systems*, **26**, 252–264.

He, X., Cai, D., and Niyogi, P. (2005). Laplacian score for feature selection. In *Advances in Neural Information Processing Systems (NeurIPS)*, **18**, 507–514.

John, G. H., Kohavi, R., and Pfleger, K. (1994). Irrelevant features and the subset selection problem. In Cohen, W. W. and Hirsh, H., editors, *Machine Learning Proceedings 1994*, 121–129. Morgan Kaufmann, San Francisco (CA).

Kalakech, M., Biela, P., Macaire, L., and Hamad, D. (2011). Constraint scores for semi-supervised feature selection: A comparative study. *Pattern Recognition Letters*, **32**, 656–665.

Khaire, U. M. and Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review. *Journal of King Saud University-Computer and Information Sciences*, **34**, 1060–1073.

Kok, T. T., Krempl, G., and Schnack, H. G. (2021). Implementation of and experimental software for active selection of classification features. *Software Impacts*, **9**, 100103.

Kumar, V. and Minz, S. (2014). Feature selection. *SmartCR*, **4**, 211–229.

Lichman, M. et al. (2013). UCI machine learning repository.

Liu, D.-Y., Xu, L.-M., Lin, X.-M., Wei, X., Yu, W.-J., Wang, Y., and Wei, Z.-M. (2022). Machine learning for semiconductors. *Chip*, **1**, 100033.

Liu, Y., Nie, F., Wu, J., and Chen, L. (2010). Semi-supervised feature selection based on label propagation and subset selection. In *2010 International Conference on Computer and Information Application*, 293–296.

Liu, Y., Nie, F., Wu, J., and Chen, L. (2013). Efficient semi-supervised feature selection with noise insensitive trace ratio criterion. *Neurocomputing*, **105**, 12–18.

Lv, S., Jiang, H., Zhao, L., Wang, D., and Fan, M. (2013). Manifold based fisher method for semi-supervised feature selection. In *2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 664–668.

Ma, Z., Nie, F., Yang, Y., Uijlings, J. R., Sebe, N., and Hauptmann, A. G. (2012). Discriminating joint feature analysis for multimedia data understanding. *IEEE Transactions on Multimedia*, **14**, 1662–1672.

Ma, Z., Yang, Y., Nie, F., Uijlings, J., and Sebe, N. (2011). Exploiting the entire feature space with sparsity for automatic image annotation. In *Proceedings of the 19th ACM International Conference on Multimedia*, 283–292.

Ren, J., Qiu, Z., Fan, W., Cheng, H., and Yu, P. S. (2008). Forward semi-supervised feature selection. In *Advances in Knowledge Discovery and Data Mining: 12th Pacific-Asia Conference, PAKDD 2008 Osaka, Japan, May 20-23, 2008 Proceedings 12*, 970–976.

Rosenberg, C., Hebert, M., and Schneiderman, H. (2005). Semi-supervised self-training of object detection models. In *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)*, **1**, 29–36.

Shi, C., Ruan, Q., and An, G. (2014). Sparse feature selection based on graph Laplacian for web image annotation. *Image and Vision Computing*, **32**, 189–201.

Song, X., Zhang, J., Han, Y., and Jiang, J. (2016). Semi-supervised feature selection via hierarchical regression for web image classification. *Multimedia Systems*, **22**, 41–49.

Tran, T. N., Vu, D. M., Tran, M. T., and Le, B. D. (2019). The combination of fuzzy min–max neural network and semi-supervised learning in solving liver disease diagnosis support problem. *Arabian Journal for Science and Engineering*, **44**, 2933–2944.

Wu, Q., Wang, J., and Zhang, J. (2013). Semi-supervised feature selection using max-margin criterion. *IEEE Transactions on Neural Networks and Learning Systems*, **24**, 1279–1291.

Wu, Z., Wu, J., Cao, J., and Tao, D. (2012). HySAD: A semi-supervised hybrid shilling attack detector for trustworthy product recommendation. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 985–993.

Xin, B., Hu, L., Wang, Y., and Gao, W. (2015). Stable feature selection from brain sMRI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, **29**.

Xu, Z., King, I., Lyu, M. R.-T., and Jin, R. (2010). Discriminative semi-supervised feature selection via manifold regularization. *IEEE Transactions on Neural networks*, **21**, 1033–1047.

Yang, L. and Wang, L. (2007). Simultaneous feature selection and classification via semi-supervised models. In *Third International Conference on Natural Computation (ICNC 2007)*, **1**, 646–650.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, 189–196.

Yu, L., Han, Y., and Berens, M. E. (2011). Stable gene selection from microarray data via sample weighting. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **9**, 262–272.

Zhao, J., Lu, K., and He, X. (2008). Locality sensitive semi-supervised feature selection. *Neurocomputing*, **71**, 1842–1849.

Zhao, Z. and Liu, H. (2007). Semi-supervised feature selection via spectral analysis. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, 641–646.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 24, No. 1, 2026 (New Series), pp 177–186
http://www.ssca.org.in/journal

R-optimal Mixture Designs for Special Cubic Model

Mahesh Kumar Panda

Department of Statistics
Ravenshaw University, Cuttack-753003

Received: 20 June 2025; Revised: 21 August 2025; Accepted: 26 August 2025

Abstract

R-optimality has been proposed in the literature as an alternative to the widely used D-optimality criterion, particularly when the goal is to construct rectangular confidence regions. In this study, R-optimal designs are investigated for the special cubic model in mixture experiments. The optimality of the proposed designs is verified using the equivalence theorem.

Key words: R-optimality; D-optimality; Special cubic model; Equivalence theorem; Mixture experiment.

AMS Subject Classifications: 62K05

The video recording of the paper made under the SSCA's Online Lecture series is available at the Youtube channel URL `https://youtu.be/rxRh9Z9uJ80`.

1. Introduction

The mixture experiments have a lot of applications in different industrial fields like the food industry, biological engineering, pharmaceutical industry, and building materials since many industrial products can be viewed as mixtures of several mixture components [*ref.* Moldes *et al.* (2007), Zaitri *et al.* (2014), and Liu *et al.* (2016), *etc.*]. Mixture experiments are a special type of design of experiments where the response of interest is only a function of the proportion of the ingredients present in the mixture.

Let x_i represents the proportion of the i^{th} component present in a q component mixture then

$$0 \leq x_i \leq 1, i = 1, 2, \ldots, q, \text{ and } \sum_{i=1}^{q} x_i = 1.$$

Due to the above restriction, the experimental region becomes a $(q-1)$-dimensional simplex given by

$$\mathrm{T} = \left\{ \boldsymbol{x} = (x_1, x_2, \ldots, x_q)' \in R^q \mid \sum_{i=1}^{q} x_i = 1, 0 \leq x_i \leq 1, i = 1, 2, \ldots, q \right\}. \tag{1}$$

Corresponding Author: Mahesh Kumar Panda
Email: mahesh2123ster@gmail.com

Consider a regression model of the form

$$\eta(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\beta} \tag{2}$$

where $\eta(\boldsymbol{x})$ denotes the expected response, $\boldsymbol{x}$ is the vector of ingredient proportions, and $\boldsymbol{f}(\boldsymbol{x})$ is the model expansion of $\boldsymbol{x}$, and $\boldsymbol{\beta}$ is the vector of unknown parameters. Here we assume that the errors are independently and identically distributed with mean 0 and constant variance σ^2.

Different model forms have been proposed in the literature to analyze mixture data *e.g.*, Scheffè's canonical polynomial models, Becker's mixture models, Darroch and Waller's additive quadratic and cubic mixture models, log contrast models, *etc.* Among these models, Scheffè's quadratic polynomial models are the most widely used polynomial models to analyze mixture data. However, when the true underlying response surface exhibits considerable complexity or curvature, a cubic model typically yields a substantially better fit compared to a quadratic model. Cubic models are capable of capturing nonlinear blending effects and synergistic/antagonistic interactions among three components, phenomena that are commonly observed in complex mixtures such as those found in food, pharmaceutical, and polymer systems [*ref.* Cornell (2002)]. Further, the special cubic models (SCMs) reduce model complexity while retaining the ability to capture key higher-order interactions, thereby offering greater efficiency and a lower risk of overfitting compared to full cubic models.

The construction of optimal design based on a certain optimality criterion aims to make the predicted response nearer to the average response over a certain region of interest. In the last few years, many authors have contributed to the domain of optimal design for mixture experiments [*ref.* Aggrawal *et al.* (2011), Pal and Mandal (2012), Mandal and Pal (2017), Panda (2024b)]. The pioneering work on the construction of optimal design for cubic mixture models was due to Kiefer (1961). He obtained D-optimal designs for the full cubic model, cubic model without a 3-way effect, and special cubic model (SCM) for the three-component mixture. Uranisi (1964) proved that a $\{q, 3\}$ simplex centroid design that assigns equal weight to each support point is D-optimal for a special cubic mixture model. Farrell *et al.* (1967) derived the D-optimal designs for the general cubic polynomial model with two and three mixture components respectively. Lim (1990) obtained the D-optimal designs for the same model, when $4 \leq q \leq 10$. Mikaeili (1989) obtained the D-optimal designs for the cubic model without a 3-way effect. Mikaeili (1993) investigated the D-optimal designs for the full cubic model on the set T. Panda and Sahoo (2022) obtained saturated A-optimal designs for full cubic model, cubic model without a 3-way effect, and special cubic model in three mixture components. Zhu and Hao (2024) investigated the A-optimal designs for the special cubic mixture model. Recently, Panda (2024a) found saturated A-optimal designs for the cubic model without a 3-way effect.

The D-optimal design is one of the most widely used classical optimal design criteria. Its construction is relatively straightforward when the number of unknown parameters, denoted by p, is small (*e.g.*, 2 or 3). However, as the number of parameters increases, the computation of D-optimal designs becomes increasingly complex. To address this limitation, Dette (1997) introduced the concept of R-optimal design, which aims to minimize the volume of the p-dimensional rectangular confidence region based on a Bonferroni t-interval. Subsequently, Panda (2021) derived saturated R-optimal designs for all three forms of the cubic model, when $q = 3$. Hao *et al.* (2021) extended the investigation of R-optimal designs

to the second-order Scheffé model for q number of mixture components. Building on this line of research, the present study derives the R-optimal design for the SCM when the mixture consists of q components.

The article is structured as follows. Section 2 discusses simplex centroid design. Section 3 explains the R-optimal design and the corresponding equivalence theorem. In Section 4, we obtain the R-optimal designs for the special cubic model where the mixture comprises q number of mixture ingredients. Section 5 obtains R-efficiencies of the corresponding A- and D-optimal designs. The article ends with some discussion in Section 6.

2. Simplex-centroid design

The simplex centroid design is a mixture design that consists of $2^q - 1$ number of support points, *i.e.*,

C_1^q number of pure components
C_2^q number of binary mixtures
C_3^q number of ternary mixtures
...
C_q^q number of q-nary mixtures.

Our focus is on the SCM with a simplex-centroid design consisting of pure components, binary mixtures, and ternary mixtures. This design may be called a $\{q, 3\}$ simplex centroid design. For example, the $\{3, 3\}$ simplex-centroid design consists of three pure components $(1, 0, 0), (0, 1, 0), (0, 0, 1)$; three binary blend points $(1/2, 1/2, 0), (1/2, 0, 1/2), (0, 1/2, 1/2)$; and one ternary mixture $(1/3, 1/3, 1/3)$.

To derive the R-optimal designs for a SCM, the search is restricted to the $\{q, 3\}$ simplex-centroid design. Notably, this design provides the precise set of mixture proportions necessary for the estimation of all model terms, including linear, binary (second-order), and ternary (third-order) interactions. Moreover, the number of support points in the $\{q, 3\}$ simplex-centroid design precisely matches the number of model terms, ensuring identifiability. The design points of this said design are evenly spread over the mixture space T. This balance ensures good coverage and reduces model bias.

3. R-optimal design and equivalence theorem

Let us consider a continuous design of the following form:

$$\xi = \left\{ \begin{array}{ccc} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_s \\ w_1 & \dots & w_s \end{array} \right\}$$

where $\boldsymbol{x}_i \in \mathrm{T}, 0 < w_i < 1$ for $i = 1, 2, \dots s$, and $\sum_{i=1}^{s} w_i = 1.w_i$ is the weight assigned to each of the points $\boldsymbol{x}_i$. For any continuous design $\xi \in \Delta$, the non-singular information matrix is defined as

$$\boldsymbol{M}(\xi) = \sum_{i=1}^{s} w_i \boldsymbol{f}\left(\boldsymbol{x_i}\right) \boldsymbol{f}'\left(\boldsymbol{x_i}\right) \tag{3}$$

where $\boldsymbol{f}(\boldsymbol{x})$ is of the form given in Equation (2). Let Δ denotes the collection of all continuous designs.

Definition 1: A design $\xi^* \in \Delta$ with a non-singular information matrix $\boldsymbol{M}(\xi^*)$ is said to be an R-optimal design for the model Equation (2) if it minimizes following function

$$\phi(\xi) = \prod_{i=1}^{p} \boldsymbol{e_i'} \boldsymbol{M^{-1}}(\xi) \boldsymbol{e}_i \tag{4}$$

over all the designs $\xi \in \Delta$, where $\boldsymbol{e_i}$ denotes the i$^{\text{th}}$ unit vector in R^p. Here p is the number of parameters associated with the model Equation (2). The R-optimal criterion corresponds to minimizing the product of the diagonal elements of the inverse information matrix, thereby reducing the joint confidence region associated with Bonferroni t-interval. The necessary and sufficient condition of the R-optimality can be verified using the corresponding equivalence theorem proposed by Dette (1997) which is as follows:

Theorem 1: Let us define the quadratic form

$$\Psi(\boldsymbol{x}, \xi) = \boldsymbol{f'}(\boldsymbol{x}) \boldsymbol{M^{-1}}(\xi) \left(\sum_{i=1}^{p} \frac{\boldsymbol{e}_i \boldsymbol{e_i'}}{\boldsymbol{e_i'} \boldsymbol{M^{-1}}(\xi) \boldsymbol{e_i}} \right) \boldsymbol{M^{-1}}(\xi) \boldsymbol{f'}(\boldsymbol{x}). \tag{5}$$

A design $\xi^* \in \Delta$ is R-optimal if and only if

$$\text{Sup}_{\boldsymbol{x} \in \mathbf{T}} \, \Psi(\boldsymbol{x}, \xi^*) = p$$

with equality holds at the support points of ξ^*.

4. R-optimal designs for special cubic model

In this section, we derive R-optimal designs for the SCM with mixture experiment. The special cubic model for the mixture experiment can be expressed as

$$\begin{aligned} \eta(\boldsymbol{x}) &= \sum_{i=1}^{q} \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \sum_{i<j<k} \beta_{ijk} x_i x_j x_k \\ &= \boldsymbol{f'}(\boldsymbol{x}) \boldsymbol{\beta} \end{aligned} \tag{6}$$

where $\boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{\beta}$ are column vectors of dimension $\frac{q\left(q^2+5\right)}{6} \times 1$ and are defined as

$$\begin{aligned} &\boldsymbol{f}(\boldsymbol{x}) = (x_1, x_2, \ldots, x_q, x_1x_2, x_1x_3, \ldots, x_{q-1}x_q, x_1x_2x_3, x_1x_2x_4, \ldots, x_{q-2}x_{q-1}x_q)'; \\ &\boldsymbol{\beta} = \left(\beta_1, \beta_2, \ldots, \beta_q, \beta_{12}, \beta_{13}, \ldots, \beta_{(q-1)q}, \beta_{123}, \beta_{124}, \ldots, \beta_{(q-1)(q-2)q}\right)'. \end{aligned}$$

Theorem 2: Under $\{q, 3\}$ simplex centroid designs, the allocations corresponding to R-optimality criterion for model Equation (6) are attached weights α_1 to the vertices, weights α_2 to the binary mixtures, weights α_3 to the ternary mixtures where weights α_1, α_2, and α_3

satisfy the following system of equations:

$$\left.\begin{aligned}
&\frac{1}{6\alpha_1\left(2\alpha_1+\alpha_2\right)\left(27\alpha_1\alpha_2+16\alpha_1\alpha_3+\alpha_2\alpha_3\right)}\Bigg(q-81\alpha_1\alpha_2\left(4\alpha_1+\alpha_2+q\alpha_2\right)\\
&\quad\times\left(192\alpha_1^2+2(32+q(q+21))\alpha_1\alpha_2+(5+q^2)\alpha_2^2\right)\alpha_3\\
&\quad+6\alpha_1(2\alpha_1+\alpha_2)(27\alpha_1\alpha_2+16\alpha_1\alpha_3+\alpha_2\alpha_3)\lambda\Bigg)=0\\
&\frac{1}{6}q(q-1)\left(-\frac{2\alpha_1\left(81\alpha_1\alpha_2+16(q+1)\alpha_1\alpha_3+(8q-13)\right)}{\alpha_2(2\alpha_1+\alpha_2)(27\alpha_1\alpha_2+16\alpha_1\alpha_3+\alpha_2\alpha_3)}+\lambda\right)=0\\
&\frac{1}{6}q(q-1)(q-2)\left(-\frac{27\alpha_1\alpha_2}{\alpha_3(27\alpha_1\alpha_2+16\alpha_1\alpha_3+\alpha_2\alpha_3)}+\lambda\right)=0\\
&-1+\frac{1}{6}q\left(6\alpha_1+(q-1)\left(3\alpha_2+(q-2)\alpha_3\right)\right)=0
\end{aligned}\right\}\tag{7}$$

Proof: A comprehensive proof of this theorem is presented in Panda (2025) and can be obtained from the author upon request. □

4.1. Establishment of equivalence theorem

In this section, we demonstrate that the R-optimal allocations derived in Theorem 2 satisfy the necessary and sufficient conditions of the equivalence theorem for R-optimal design. The verification of the equivalence theorem (Theorem 1), however, presents two primary challenges:

1. The complexity of the system of equations in (7) makes the derivation of a general closed-form solution for the optimal weights analytically challenging.
2. Establishing the equivalence theorem for a general value of q is difficult, primarily because of the complex structure of the matrix $\boldsymbol{M^{-1}}(\xi)$.

Although the explicit form of $\boldsymbol{M^{-1}}(\xi)$ is not included in this paper, its structure related mathematical details are discussed in Panda (2025). These materials are available from the author upon request. To address part (i), the optimal weights are computed numerically. The numerical value of optimal weights α_1, α_2, and α_3 for the design ξ^*, along with the maximum value of $\Psi(\boldsymbol{x},\xi)$ for various values of q in the range $3\leq q\leq 18$, are displayed in Table 1.

Table 1: R-optimal allocations for the special cubic model and corresponding values of $\max_{\boldsymbol{x}\in\mathbf{T}} \Psi(\boldsymbol{x},\xi)$

(1)	(2)	(3)	(4)	(5)
q	α_1	α_2	α_3	$\max_{\boldsymbol{x}\in\mathbf{T}} \Psi(\boldsymbol{x},\xi)$
3	0.1796	0.1217	0.0960	7
4	0.0995	0.0675	0.0492	14
5	0.0611	0.0413	0.0281	25
6	0.0406	0.0271	0.0175	41
7	0.0286	0.0189	0.0115	63
8	0.0210	0.0137	0.0080	92
9	0.0160	0.0103	0.0058	129
10	0.0126	0.0080	0.0043	175
11	0.0101	0.0063	0.0033	231
12	0.0083	0.0051	0.0026	298
13	0.0070	0.0043	0.0020	377
14	0.0058	0.0035	0.0017	469
15	0.0049	0.0029	0.0014	575
16	0.0044	0.0026	0.0011	696
17	0.0037	0.0022	0.0009	833
18	0.0033	0.0019	0.0008	987

Subsequently, to handle the part (ii), we use appropriate MATLAB code by following the steps of Algorithm 1 and demonstrate numerically that $\mathrm{Max}_{\boldsymbol{x}\in\mathrm{T}} \Psi(\boldsymbol{x},\xi^*) = p$. We also examine that equality holds only at the support points of the design ξ^*. This confirms that the design ξ^* is the R -optimal design in the simplex region T.

Algorithm 1: Algorithm to demonstrate Equivalence theorem

Input:

Step 1: Set the value of q and the values of $\alpha_1, \alpha_2, \alpha_3$ using columns (2), (3), and (4) of Table 1 respectively.

Step 2: Input the column vector

$$\boldsymbol{f}(\boldsymbol{x}) = (x_1, x_2, \ldots, x_q, x_1x_2, x_1x_3, \ldots, x_{q-1}x_q, x_1x_2x_3, x_1x_2x_4, \ldots, x_{q-2}x_{q-1}x_q)'$$

Computation:

Step 3: Obtain the matrix $\boldsymbol{M}^{-1}(\xi^*)$ [*ref.* Panda (2025)].

Step 4: Find the functional form $\Psi(\boldsymbol{x},\xi)$

where

$$\Psi(\boldsymbol{x},\xi) = \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{M}^{-1}(\xi^*)\left(\sum_{i=1}^{p} \frac{\boldsymbol{e_i}\boldsymbol{e_i'}}{\boldsymbol{e_i'}\boldsymbol{M}^{-1}(\xi^*)\,\boldsymbol{e_i}}\right)\boldsymbol{M}^{-1}(\xi^*)\,\boldsymbol{f}(\boldsymbol{x})$$

Step 5: Evaluate $\Psi(\boldsymbol{x}, \xi)$ at the support points of design ξ^*.

Step 6: Find $\text{Max}_{\boldsymbol{x}\in T}\, \Psi\,(\boldsymbol{x}, \xi^*)$.

5. Efficiency of A- and D-optimal designs

The R-efficiency denoted by $\Delta_R(\xi)$ of a design ξ relative to an R-optimal design $\xi^{(0)}$ is given by

$$\Delta_R(\xi) = \left(\frac{\phi\left(\xi^{(0)}\right)}{\phi(\xi)} \right)^{1/p}.$$

The value of $\Delta_R(\xi)$ closer to 1 indicates that the design has a high R-efficiency value. The R-efficiency values of D-optimal design [*ref.* Uranisi (1964)] and A-optimal design [*ref.* Zhu and Hao (2024)] of the SCM for different values of $q\,(3 \leq q \leq 18)$ have been computed and shown in Table 2. Further, the R-efficiency values of the corresponding D-optimal and A-optimal designs against various values of q are displayed in Figure 1.

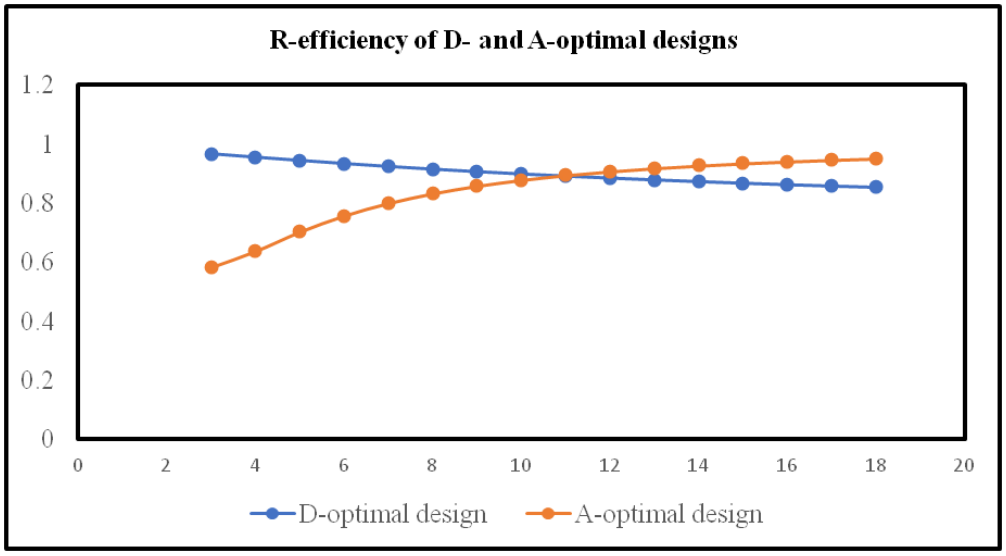


Figure 1: R-efficiency of D- and A-optimal designs for the special cubic model

Next, we consider a three-component example to establish the fact that the R-optimal design can improve the ability of parameter estimation over other designs.

Example 1: **The R-efficiency with three components example**

Cornell (*pg.* 57, 2002) cited a three-component mixture experiment that aims at the effect of the proportion of three artificial sweeteners such as glycerine, saccharin, and an enhancer on the intensity of sweetness. The amount of each of the sweeteners was fixed to 250 ml. To analyze the problem, we use the special cubic model. The R-optimal design for the said model is

$$\xi_R = \begin{pmatrix} \boldsymbol{x} \leftrightarrow (1,0,0) & \boldsymbol{x} \leftrightarrow (1/2,1/2,0) & \boldsymbol{x} \leftrightarrow (1/3,1/3,1/3) \\ 0.1796 & 0.1217 & 0.9996 \end{pmatrix}$$

where $\boldsymbol{x} \leftrightarrow (1,0,0)$ refers to the design point $(1,0,0)$ and its permutations $(0,1,0)$, and $(0,0,1)$; $\boldsymbol{x} \leftrightarrow (1/2,1/2,0)$ refers to the design point $(1/2,1/2,0)$ and its permutations $(1/2,0,1/2)$, and $(0,1/2,1/2)$; $x \leftrightarrow (1/3,1/3,1/3)$ means the design point itself. Here value of $\phi(\xi_R) = 1.06234 \times 10^{13}$. Let us consider the 10 support points that have been discussed in the same problem. Allocate a weight of $\frac{1}{10}$ to each of the support points. Let us denote the design by ξ_1.

$$\xi_1 = \begin{pmatrix} \boldsymbol{x} \leftrightarrow (1,0,0) & \boldsymbol{x} \leftrightarrow (1/2,1/2,0) & \boldsymbol{x} \leftrightarrow (1/3,1/3,1/3) & \boldsymbol{x} \leftrightarrow (2/3,1/6,1/6) \\ \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} \end{pmatrix}.$$

For the design ξ_1, the value of $\phi(\xi_1) = 1.1151 \times 10^{14}$. Consider another design ξ_2 as discussed in Cornell (2002) *i.e.*,

$$\xi_2 = \begin{pmatrix} \boldsymbol{x} \leftrightarrow (1,0,0) & \boldsymbol{x} \leftrightarrow (b,1-b,0) & \boldsymbol{x} \leftrightarrow (1/3,1/3,1/3) \\ \frac{1}{10} & \frac{1}{10} & \frac{1}{10} \end{pmatrix}$$

where $b = \frac{1-5^{-1/2}}{2}$. For this design, $\phi(\xi_2) = 9.88373 \times 10^{13}$. Similarly consider the following $\{3,3\}$ simplex-lattice design denoted by ξ_3 where equal weight has been allocated to each of the support points.

$$\xi_3 = \begin{pmatrix} \boldsymbol{x} \leftrightarrow (1,0,0) & \boldsymbol{x} \leftrightarrow (1/3,2/3,0) & \boldsymbol{x} \leftrightarrow (1/3,1/3,1/3) \\ \frac{1}{10} & \frac{1}{10} & \frac{1}{10} \end{pmatrix}$$

The R-efficiencies of the designs ξ_1, ξ_2, and ξ_3 are calculated as $0.7147, 0.7271$, and 0.7793 respectively. Thus, it can be depicted that ξ_R can improve the parameter estimation.

Table 2: R-efficiencies of various optimal designs for the special cubic mixture model

q	D-optimal design	A-optimal design
3	0.9661	0.5816
4	0.9549	0.6364
5	0.9441	0.7022
6	0.9336	0.7569
7	0.9238	0.7996
8	0.9148	0.8325
9	0.9064	0.8581
10	0.8986	0.8783
11	0.8915	0.8944
12	0.8849	0.9074
13	0.8788	0.9181
14	0.8730	0.9270
15	0.8677	0.9345
16	0.8627	0.9408
17	0.8581	0.9462
18	0.8537	0.9508

6. Discussion

The construction of R-optimal designs involves more challenges in comparison to the D-optimal design as the weights associated with the support points in the case of the former are different. Also, the weights vary as the value of q changes. This article finds R-optimal designs for the special cubic model with a mixture experiment based on q mixture components, where $3 \leq q \leq 18$. For $q = 3$, we also observe that the derived R-optimal design for the special cubic model is similar to the result derived by Panda(2021). From Example 1 as well as Table 2, we can see that the R-optimal designs have higher efficiency as compared to other designs. Based on Figure 1, it can be depicted that the R-efficiency value of the D-optimal design decreases with q whereas it increases in the case of the A-optimal design. The performance of the D-optimal design is better than the A-optimal design with q, when $3 \leq q \leq 10$ whereas the performance of the latter is better with q for $11 \leq q \leq 18$. Still, the performance of the R-optimal design is consistently better with different values of q.

Acknowledgment

The author gratefully acknowledges the referee's insightful comments, which significantly contributed to improving the clarity and quality of the manuscript.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Aggrawal, M., Singh, P., and Panda, M. K. (2011). A-optimal designs for an additive cubic model. *Statistics & Probability Letters*, **81**, 259–266.

Cornell, J. A. (2002). *Experiments With Mixtures*. 3rd edition, Wiley, New York.

Dette, H. (1997). Designing experiments with respect to 'standardized' optimality criteria. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59**, 97–110.

Farrell, R., Kiefer, J., and Walbran, A. (1967). Optimum multivariate designs. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. **5**, 113–139, University of California Press.

Hao, H., Zhu, X., Zhang, X., and Zhang, C. (2021). R-optimal design of the second-order scheffé mixture model. *Statistics & Probability Letters*, **173**, 109069.

Kiefer, J. (1961). Optimum designs in regression problems, ii. *The Annals of Mathematical Statistics*, **32**, 298–325.

Lim, Y. B. (1990). D-optimal design for cubic polynomial regression on the q-simplex. *Journal of Statistical Planning and Inference*, **25**, 141–152.

Liu, W., Luo, T., Li, Y., Song, Y., Zhu, Y., Liu, Y., Zhao, J., Wu, Z., and Xu, X. (2016). Experimental study on the mechanical properties of sediments containing ch4 and co2 hydrate mixtures. *Journal of Natural Gas Science and Engineering*, **32**, 20–27.

Mandal, N. K. and Pal, M. (2017). Optimum mixture designs in some constrained experimental regions. *Communications in Statistics—Theory and Methods*, **46**, 4240–4249.

Mikaeili, F. (1989). D-optimum design for cubic without 3-way effect on the simplex. *Journal of Statistical Planning and Inference*, **21**, 107–115.

Mikaeili, F. (1993). D-optimum design for full cubic on q-simplex. *Journal of Statistical Planning and Inference*, **35**, 121–130.

Moldes, A., Cendon, Y., and Barral, M. T. (2007). Evaluation of municipal solid waste compost as a plant growing media component, by applying mixture design. *Bioresource Technology*, **98**, 3069–3075.

Pal, M. and Mandal, N. K. (2012). D-optimum designs for optimum mixture in a quadratic log contrast model. *Communications in Statistics—Simulation and Computation*, **41**, 253–264.

Panda, M. K. (2021). R-optimal designs for canonical polynomial models with mixture experiments. *Calcutta Statistical Association Bulletin*, **73**, 146–161.

Panda, M. K. (2024a). A-optimal design for cubic model without a 3-way effect for mixture experiment. *Hacettepe Journal of Mathematics and Statistics*, **53**, 1759–1773.

Panda, M. K. (2024b). Model robust optimal designs for kronecker model for mixture experiments. *International Journal of Statistical Sciences*, **24**, 31–48.

Panda, M. K. (2025). R-optimal mixture designs for special cubic model. Unpublished Technical Report.

Panda, M. K. and Sahoo, R. P. (2022). A-optimal designs for cubic polynomial models with mixture experiments in three components. *Statistics and Applications*, **20**, 41–55.

Uranisi, J. (1964). Optimum design for the special cubic regression on the q-simplex. *Mathematical Reports Kyushu University*, **1**.

Zaitri, R., Bederina, M., Bouziani, T., Makhloufi, Z., and Hadjoudja, M. (2014). Development of high-performance concrete based on the addition of grinded dune sand and limestone rock using the mixture design modelling approach. *Construction and Building Materials*, **60**, 8–16.

Zhu, X. and Hao, H. (2024). A-optimal design for the special cubic mixture model. *Communications in Statistics—Theory and Methods*, **53**, 1081–1090.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 24, No. 1, 2026 (New Series), pp 187–220
http://www.ssca.org.in/journal

Poisson–Transmuted Geometric Convolution for Overdispersed Count Data

Anupama Nandi[1], Partha Jyoti Hazarika[1], Aniket Biswas[2], Morad Alizadeh[3], Hadi Saboori[4] and Mohamed S. Eliwa[5,6]

[1]*Department of Statistics, Dibrugarh University, India*
[2]*Department of Statistics, St. Xavier's University, Kolkata, India*
[3]*Department of Statistics, Persian Gulf university, Bushehr, 75169, Iran*
[4]*Department of Statistics, Faculty of Sciences, University of Zabol, Zabol, Iran*
[5]*Department of Statistics and Operations Research, College of Science, Qassim University, Saudi Arabia*
[6]*Department of Mathematics, Faculty of Science, Mansoura University, Mansoura 35516, Egypt*

Received: 10 May 2025; Revised: 01 September 2025; Accepted: 05 September 2025

Abstract

A novel overdispersed count distribution is obtained by convolving independently distributed Poisson and transmuted geometric random variables. This distribution extends the Poisson-geometric, geometric, and transmuted geometric distributions. Essential statistical properties are analyzed. Maximum likelihood estimators of the unknown constants are derived using numerical optimization techniques and the EM algorithm. Extensive simulation studies evaluate the estimators performance under different conditions. Additionally, a flexible regression model built on the proposed distribution is formulated. Real-world applications in modeling overdispersed count data, both with and without covariates, demonstrate the model's relevance. The proposed model demonstrates desirable statistical properties and surpasses its closest competitors in empirical applications.

Key words: Transmuted geometric distribution; EM algorithm; Computer simulation; Prediction model; Failure analysis; Statistics and numerical data.

MSC 2010: 60E05, 62E15

1. Introduction

Overdispersion is a common phenomenon in modeling applications, occurring more frequently than underdispersion or equidispersion. The literature presents numerous count models specifically designed to handle overdispersed data. Given the sustained research interest in this area, the development of a simple yet effective model remains essential. Various distributions and regression models for overdispersed count data have been extensively studied [Moghimbeigi, *et al.* (2008), Rodrigues-Motta, *et al.* (2013), Hassanzadeh

Corresponding Author: Anupama Nandi
Email: anupamanandi0702@gmail.com

and Kazemi (2016), Wongrin and Bodhisuwan (2017), Wang, *et al.* (2017), Sarvi, *et al.* (2019), Moqaddasi Amiri, *et al.* (2019), Tapak, *et al.* (2020), Ghahramani and White (2020), Tüzen, *et al.* (2020), Altun (2020), and Bar-Lev and Ridder (2021)]. Recent developments in overdispersed count regression models include the finite mixtures of mean-parameterized Conway–Maxwell–Poisson (COM-Poisson) regression introduced by Zhan and Young (2024) and the marginalized zero-inflated Bell regression model proposed by Amani, *et al.* (2025).

Although widely used, the Poisson distribution $Poi(\theta)$ is constrained by its equidispersion property. To overcome this limitation, several alternative distributions have been introduced, including the hyper-Poisson by Bardwell and Crow (1964), the double-Poisson by Efron (1986), the weighted Poisson by Del Castillo and Pérez-Casany (1998), the weighted generalized Poisson by Chakraborty (2010), the widely adopted COM-Poisson by Sellers and Shmueli (2010), the Mittag-Leffler function distribution by Chakraborty and Ong (2017), and the Tilted Beta-Binomial Distribution by Hahn (2022).

In addition to these distributions, various modifications of the geometric $Geo(\theta)$ distribution have been studied for modeling overdispersed count data. Notable contributions include the works of Chakraborty and Gupta (2015), Jain and Consul (1971), Philippou, *et al.* (1983), Tripathi, *et al.* (1987), Makcutek (2008), Gómez-Déniz (2010), and Nekoukhou, *et al.* (2012), among others. A key improvement of the geometric distribution is the transmuted geometric distribution (Chakraborty and Bhati, 2016) $TGD(q, \alpha)$ which is developed using the quadratic rank transmutation technique. For a random variable Y following the $TGD(q, \alpha)$ distribution, the probability mass function (pmf) is

$$P(Y = y) = \alpha q^{2y}(1 - q^2) + (1 - \alpha)q^y(1 - q), \;\; y \geq 0. \tag{1}$$

Here, $q \in (0, 1)$ and $\alpha \in (-1, 1)$. It exhibits underdispersion for $\alpha \in (-1, 0)$ and overdispersion for blue $\alpha \in (0, 1)$.

Despite the existence of numerous discrete distributions, there remains a need for new models that can effectively capture overdispersed count data while retaining simplicity and interpretability. The convolution method serves as a powerful yet straightforward technique for constructing new probability distributions. By convolving two known distributions, a new distribution with distinct characteristics can be derived, which is applicable to diverse data-modeling scenarios. Bourguignon and Weiß (2017) introduced a count distribution, termed BerG, by convolving a Bernoulli and a geometric variable. The Bernoulli distribution is underdispersed, with a variance lower than its mean, whereas the geometric distribution is overdispersed, with a variance greater than its mean. The BerG model effectively accommodates underdispersed, overdispersed, and equidispersed data. The Bernoulli-Poisson (BerPoi) distribution of Bourguignon, *et al.* (2022), and the Poisson-geometric (PoiG) distribution of Nandi, *et al.* (2024b) are some recent advancements in flexible count data models based on a similar approach. Later, Nandi, *et al.* (2024a) developed an overdispersed regression model, termed the PoiG regression model, based on the mean parameterization of the PoiG distribution.

The PoiG distribution is derived by convolving the Poisson and geometric distributions. Let $X_1 \sim Poi(\theta)$, where $\theta > 0$ is the parameter of the Poisson distribution and $X_2 \sim G(q)$, where $0 < q < 1$ is the parameter of the geometric distribution. If X_1 and X_2

are independent, then their sum, $X = X_1 + X_2$, follows the pmf

$$P(X = x) = \frac{q^x(1-q)}{\Gamma(x+1)} \exp\left(\frac{\theta(1-q)}{q}\right)\Gamma\left(x+1, \frac{\theta}{q}\right), \qquad x \geq 0. \tag{2}$$

The PoiG regression model of Nandi, *et al.* (2024a) is an overdispersed regression model developed by considering the mean parameterization of the PoiG distribution.

This study introduces the Poisson transmuted geometric (PoiTG) distribution using the above mentioned technique. Compared to the COM-Poisson distribution, the PoiTG distribution is more interpretable as it does not require a complex normalizing constant in its pmf. Additionally, its compact expressions for the mean and variance make it suitable for regression modeling. This distribution extends the Poisson, transmuted geometric, and PoiG distributions, offering a flexible approach to count data modeling.

The paper is structured as follows: Section 2 introduces the PoiTG distribution. Section 3 delves into its key statistical properties, such as recurrence relations, generating functions, moments, skewness, kurtosis, dispersion index, reliability function, hazard rate function, mean residual life function and Shannon entropy. Section 4 covers parameter estimation methods. Section 5 evaluates the performance of maximum likelihood estimators through simulation studies. Section 6 demonstrates the practical applicability of the PoiTG distribution by analyzing real-world data sets. Section 7 proposes a regression model based on the PoiTG distribution and presents an empirical application. Finally, the concluding remarks and potential future research directions are provided in the last section.

2. The PoiTG distribution

Let Y_1 follows the Poisson distribution with parameter $\theta > 0$, that is, $Poi(\theta)$ and Y_2 follows the transmuted geometric distribution with parameters $q \in (0,1)$ and $\alpha \in (0,1)$ (Chakraborty and Bhati, 2016) independent of Y_1, that is, $TGD(q,\alpha)$. Chakraborty and Bhati (2016) originally considered $-1 < \alpha < 1$. However, we restrict α to $0 < \alpha < 1$ to better illustrate the overdispersion in the proposed convolution-based model. This restriction also proves useful in Section 4.2 for implementing the EM algorithm. Since both Y_1 and Y_2 can take values from the set, denoted as N_0, the convolution $Y = Y_1 + Y_2$ also has support on N_0 and

$$\Pr(Y = y) = (1-\alpha)(1-q)q^y e^{-\theta} \sum_{i=0}^{y} \frac{1}{i!}\left(\frac{\theta}{q}\right)^i + \alpha(1-q^2)q^{2y}e^{-\theta}\sum_{i=0}^{y}\frac{1}{i!}\left(\frac{\theta}{q^2}\right)^i. \tag{3}$$

The distribution in (3) is referred to as the PoiTG distribution and we denote it as $Y \sim PoiTG(\theta, q, \alpha)$. Note that the pmf of Y can also be written as

$$p_Y(y) = (1-\alpha)w_1(y,\theta,q) + \alpha w_2(y,\theta,q), \qquad y \geq 0\ , \tag{4}$$

where,

$$\begin{aligned} w_k(y,\theta,q) &= (1-q^k)q^{ky}e^{-\theta}\sum_{i=0}^{y}\frac{1}{i!}\left(\frac{\theta}{q^k}\right)^i \\ &= \frac{(1-q^k)q^{ky}}{\Gamma(y+1)}\exp\left(\frac{\theta}{q^k}-\theta\right)\Gamma\left(y+1,\frac{\theta}{q^k}\right), \qquad k = 1,2\ . \end{aligned}$$

The incomplete gamma function (Abramowitz and Stegun, 1964) in (4) is defined as $\Gamma(k,x) = \int_x^\infty t^{(k-1)}e^{-t}dt$ and it can also be defined as

$$\Gamma(k,x) = (k-1)!\sum_{n=0}^{k-1}\frac{e^{-x}x^n}{n!},$$

which holds for non-negative values of k and any real value of x. The incomplete gamma function in $w_k(y,\theta,q)$ can be rewritten as

$$\Gamma\left(y+1,\frac{\theta}{q^k}\right) = y!\sum_{i=0}^{y}\frac{1}{i!}\exp\left(-\frac{\theta}{q^k}\right)\left(\frac{\theta}{q^k}\right)^i,$$

The nature of the pmf in (3) of $Y \sim PoiTG(\theta,q,\alpha)$ is demonstrated by Figure 1 for different combinations of θ, q, and α. These plots clearly depict that the distribution of $PoiTG(\theta,q,\alpha)$ is unimodal and has, at most, one exponential tail. The cumulative distribution function (cdf) of the proposed distribution is

$$F_Y(y) = \frac{\Gamma(y+1,\theta)}{\Gamma(y+1)} - (1-\alpha)W_1(y,\theta,q) - \alpha W_2(y,\theta,q), \qquad y \geq 0, \tag{5}$$

where,

$$W_k(y,\theta,q) = \frac{q^{k(y+1)}}{\Gamma(y+1)}\exp\left(\frac{\theta(1-q^k)}{q^k}\right)\Gamma\left(y+1,\frac{\theta}{q^k}\right), \qquad k = 1,2\ .$$

The nature of the cdf in (5) of $PoiTG(\theta,q,\alpha)$ is demonstrated by Figure 2 for different combinations of θ, q, and α.

Remark 1: As $\theta \to 0$, the $PoiTG(\theta,q,\alpha)$ distribution exhibits behavior similar to the $TGD(q,\alpha)$. Similarly, as $\alpha \to 0$, it behaves akin to the $PoiG(\theta,1-q)$ distribution. Furthermore, when both θ and $\alpha \to 0$, it resembles the $G(1-q)$ distribution.

3. Statistical properties

3.1. Recurrence relation

The probability recurrence relation facilitates the determination of a term based on its preceding term(s). This approach is particularly useful for computing probability masses at various support points. Note that,

$$p_Y(y) = \alpha(1-q^2)q^{2y}e^{-\theta}s''_y + (1-\alpha)(1-q)q^y e^{-\theta}s'_y$$

where,

$$s'_y = \sum_{i=0}^{y}\frac{1}{i!}\left(\frac{\theta}{q}\right)^i,$$

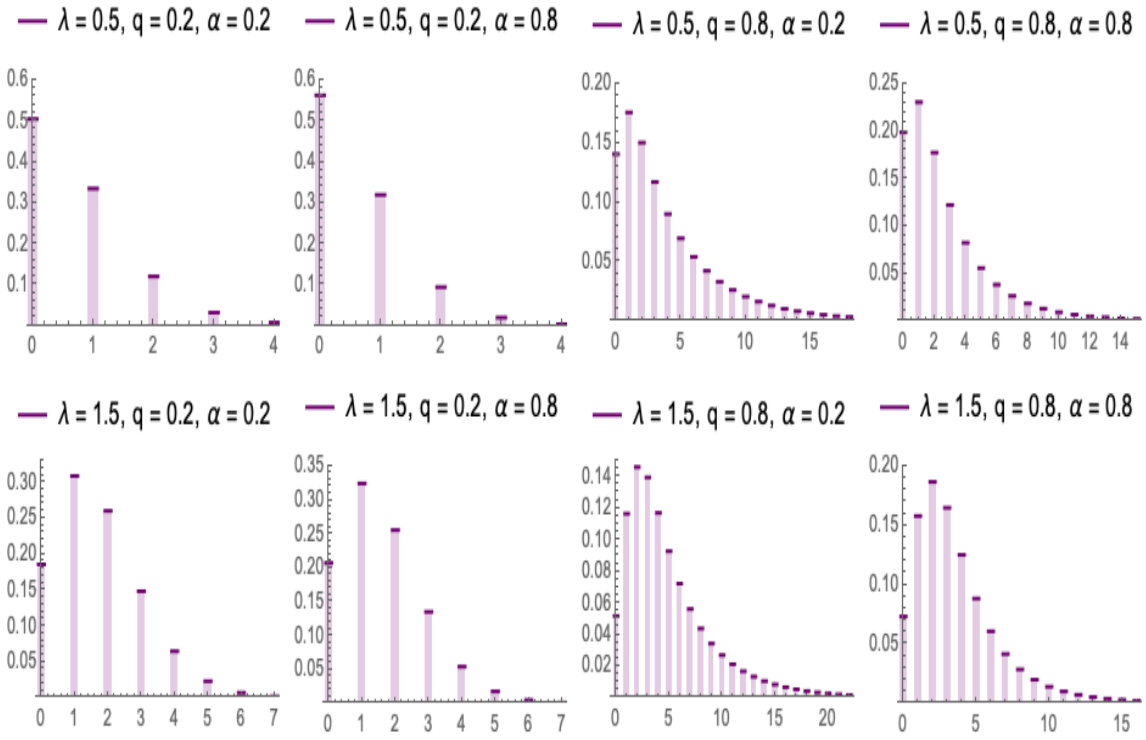


Figure 1: The pmf of $PoiTG(\theta, q, \alpha)$ for multiple choices of θ, q and α

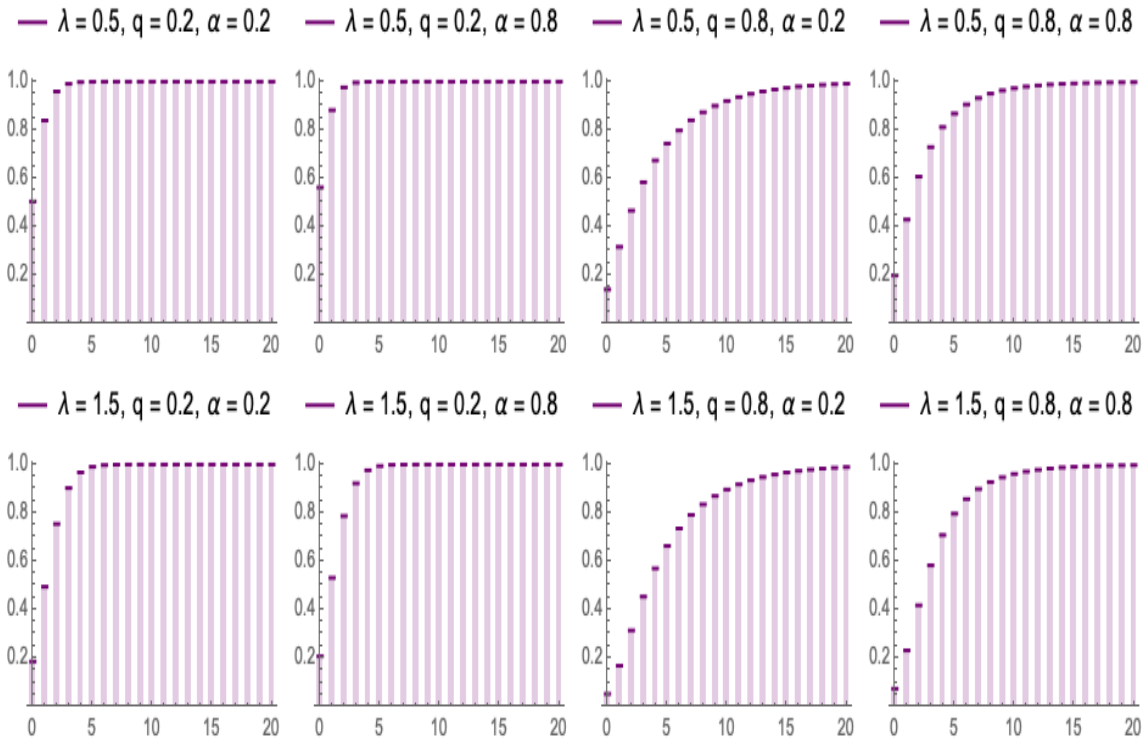


Figure 2: The cdf of $PoiTG(\theta, q, \alpha)$ for multiple choices of θ, q and α

$$s''_y = \sum_{i=0}^{y} \frac{1}{i!} \left(\frac{\theta}{q^2}\right)^i,$$

$$s'_{y+1} = s'_y + \frac{1}{\Gamma(y+2)} \left(\frac{\theta}{q}\right)^{y+1},$$

and

$$s''_{y+1} = s''_y + \frac{1}{\Gamma(y+2)} \left(\frac{\theta}{q^2}\right)^{y+1}.$$

Thus,

$$p_Y(y+1) = \alpha(1-q^2)q^{2(y+1)}e^{-\theta}s''_{y+1} + (1-\alpha)(1-q)q^{y+1}e^{-\theta}s'_{y+1}. \tag{6}$$

Equation 6 is the recurrence relation of the PoiTG distribution. It is easy to verify that

$$\begin{aligned}\frac{p_Y(y+1)}{p_Y(y)} &= q\frac{s'_{y+1} + \dfrac{\alpha(1+q)}{1-\alpha}q^{y+1}s''_{y+1}}{s'_y + \dfrac{\alpha(1+q)}{1-\alpha}q^y s''_y} \\ &= q\left(1 + \frac{\alpha(1+q)q^y s''_y}{(1-\alpha)s'_y}\right)^{-1}. \end{aligned} \tag{7}$$

3.2. Generating functions

We use G to denote the probability generating function (pgf) and to be specific in its use, the corresponding random variable is mentioned in the subscript of the function G. For $Y_1 \sim Poi(\theta)$ and $Y_2 \sim TGD(\alpha, q)$,

$$G_{Y_1}(s) = e^{\theta(s-1)} \quad \text{and} \quad G_{Y_2}(s) = \frac{(1-q)(1+\alpha q(1-s)-q^2 s)}{(1-qs)(1-q^2 s)}, \ |q^2 s| < 1.$$

Since Y_1 and Y_2 are independent, $G_Y(s) = G_{Y_1}(s)G_{Y_2}(s)$. Thus the pgf of $PoiTG(\theta, q, \alpha)$ is

$$G_Y(s) = \frac{e^{\theta(s-1)}(1-q)(1+\alpha q(1-s)-q^2 s)}{(1-qs)(1-q^2 s)}. \tag{8}$$

In similar way, we derive the other generating functions including moment generating function ($M_Y(t)$), characteristic function ($\phi_Y(t)$) and cumulant generating function ($K_Y(t)$). These are provided below as

$$M_Y(t) = \frac{e^{\theta(e^t-1)}(1-q)(1+\alpha q(1-e^t)-q^2 e^t)}{(1-qe^t)(1-q^2 e^t)}, \tag{9}$$

$$\phi_Y(t) = \frac{e^{\theta(e^{it}-1)}(1-q)(1+\alpha q(1-e^{it})-q^2 e^{it})}{(1-qe^{it})(1-q^2 e^{it})}, \tag{10}$$

and

$$K_Y(t) = \log\left[\frac{(1-q)(1+\alpha q(1-e^t)-q^2 e^t)}{(1-qe^t)(1-q^2 e^t)}\right] + \theta(e^t - 1). \tag{11}$$

3.3. Moments, skewness and kurtosis

The r^{th} order raw moment can be computed by taking the r^{th} derivative of the moment generating function specified in (9) with respect to t and evaluating it at $t = 0$.

$$E(Y^r) = M_Y^{(r)}(0) = \frac{d^r}{dt^r}[M_Y(t)]_{t=0}.$$

Let μ'_r denote the r^{th} order raw moment, that is $\mu'_r = E(Y^r)$. The closed form expressions of the first, second, third and fourth moments are

$$\mu'_1 = \theta + \frac{q(1-\alpha)+q^2}{1-q^2}, \tag{12}$$

$$\begin{aligned}\mu'_2 = \frac{1}{(1-q^2)^2}&[\theta(1+\theta)+q(1-\alpha)(1+2\theta)+q^2(3-2\alpha-2\theta^2)\\ &+q^3(1-\alpha)(3-2\theta)+q^4(1-\theta+\theta^2)],\end{aligned} \tag{13}$$

$$\begin{aligned}\mu'_3 = \frac{1}{(1-q^2)^3}&[\theta(\theta^2+3\theta+1)+q(1-\alpha)(3\theta^2+6\theta+1)\\ &-q^2(3\theta^3+6\theta^2-9\theta+6\alpha(\theta+1)-7)-2q^3(1-\alpha)(3\theta^2-8)\\ &+q^4(3\theta^3+3\theta^2-9\theta+6\alpha(\theta-2)+16)+q^5(1-\alpha(3\theta^2-6\theta+7)\\ &-q^6(\theta^3+\theta-1)],\end{aligned} \tag{14}$$

and

$$\begin{aligned}\mu'_4 = \frac{1}{(1-q^2)^4}&[\theta(\theta^3+6\theta^2+7\theta+1)+q(1-\alpha)(4\theta^3+18\theta^2+14\theta+1)\\ &-q^2(4\theta^4+20\theta^3-2\theta^2-46\theta+2\alpha(7+18\theta+6\theta^2)-15)\\ &-q^3(1-\alpha)(12\theta^3+30\theta^2-54\theta-61)\\ &+q^4(6\theta^4+24\theta^3-24\theta^2+8\alpha(3\theta^2-13)+115)\\ &+q^5(1-\alpha)(12\theta^3+30\theta^2-54\theta+115)\\ &-q^6(4\theta^4+12\theta^3-14\theta^2+46\theta+2\alpha(25-18\theta+6\theta^2)-61)\\ &-q^7(1-\alpha)(-15+14\theta-6\theta^2+4\theta^3)+q^8(\theta^4+2\theta^3+\theta^2-\theta+1)].\end{aligned} \tag{15}$$

By utilizing the above formulas for the raw moments, we can derive the explicit formulae for the first, second, four central moments of the random variable Y. The r^{th} order central moment, denoted by μ_r, is defined as $\mu_r = E(Y-\mu'_1)^r$. The explicit expressions for the central moments of the PoiTG distribution are

$$\mu_2 = \frac{\theta(1-q^2)^2+q(1-\alpha+q(2+q(1-\alpha)-\alpha^2))}{(1-q^2)^2}, \tag{16}$$

$$\begin{aligned}\mu_3 = \frac{1}{(1-q^2)^3}&(\theta+q(1-\alpha)-q^2(3\alpha^2+3\theta-4)-2q^3(\alpha^3+2\alpha-3)\\ &-q^4(3\alpha^2-3\theta-4)+q^5(1-\alpha)-q^6\theta),\end{aligned} \tag{17}$$

and

$$\begin{aligned}\mu_4 = \frac{1}{(1-q^2)^4}&((1+q^8)\theta(3\theta+1)+(q+q^7)(1-\alpha)(6\theta+1)-(q^2+q^6)\\ &(\alpha^2(6\theta+4)+6\alpha+12\theta^2-8\theta-11)+(q^3+q^5)(1-\alpha)\\ &(6\alpha^2+12\alpha-6\theta+35)+q^4(3\alpha^4-4\alpha^2(3\theta-7)+12\alpha-2(9\theta^2-9\theta+25))).\end{aligned} \tag{18}$$

Here, explicit expressions for skewness and kurtosis are not provided because of the computational complexity and the involvement of large equations. Instead, 3-D surface plots for these measures are presented in Figure 3 and 4, respectively.

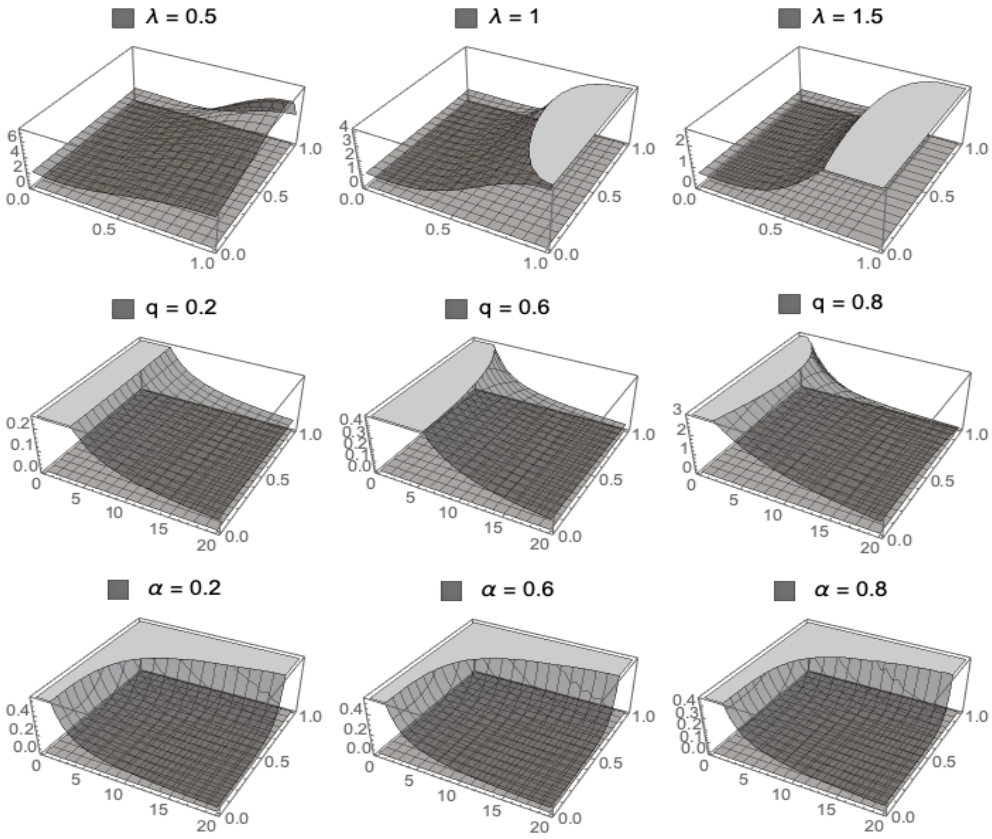


Figure 3: Skewness of $PoiTG(\theta, q, \alpha)$ for different combinations of θ, q and α

Figure 3 represents three rows of surface plots. Each row corresponds to a different combination of fixed parameters: θ, q, and α. The first row shows the surface plot of skewness with respect to q and α, while keeping θ constant. The second row depicts the surface plot of skewness in relation to θ and α, with q held constant. The third row exhibits the surface plot of skewness with respect to θ and q, while maintaining α at a fixed value. Notably, all the surface plots indicate positive skewness for any combination of θ, q, and α.

Similarly, in Figure 4, the same parameter combinations are used for each row as in Figure 3. A horizontal surface is drawn at a value of 3 in each graph, which never intersects the kurtosis surface, indicates the leptokurtic nature of the PoiTG distribution. The second row shows that the parameter q significantly influences kurtosis, with higher q values leading to greater leptokurticity. In contrast, the first and third rows suggest that for large θ and α, the distribution exhibits no substantial variation in peakedness.

3.4. Index of dispersion

The index of dispersion (Hoel, 1943), denoted by ID_Y, provides a measure of the level of dispersion of a distribution. It assesses whether a distribution is appropriate for

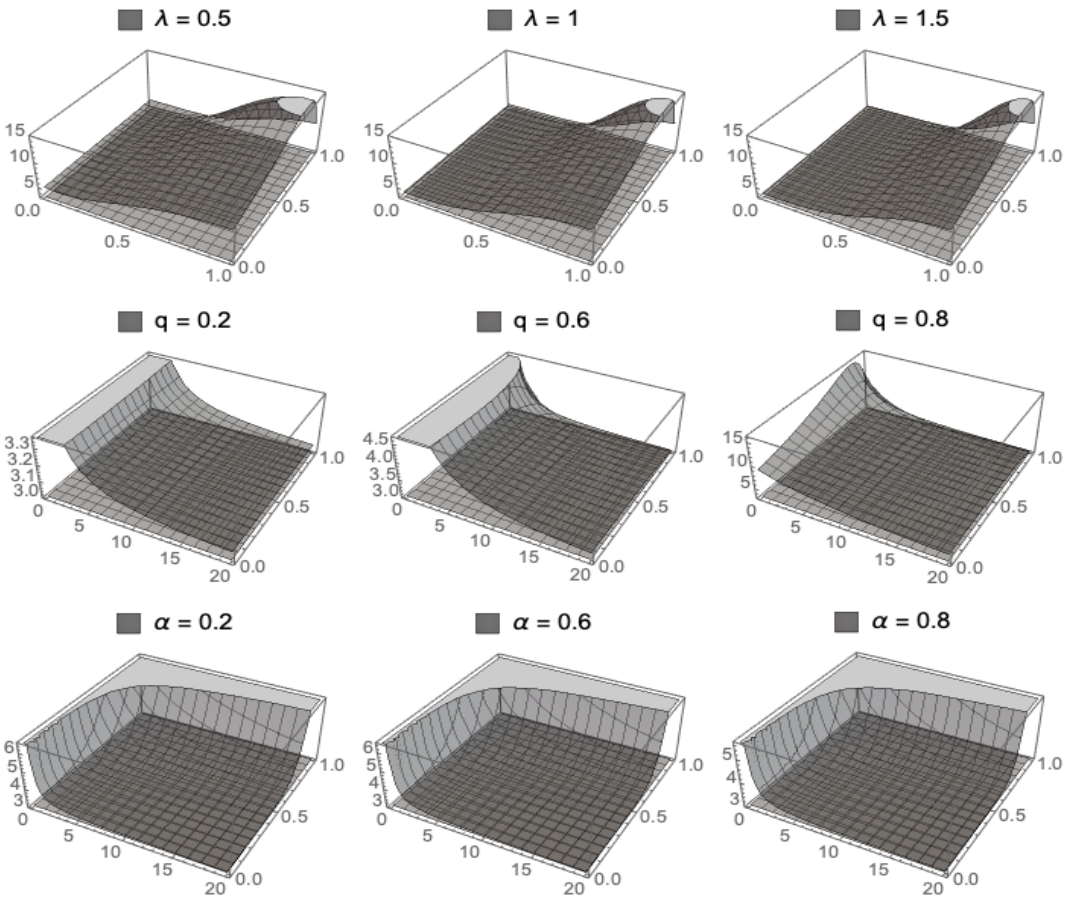


Figure 4: Kurtosis of $PoiTG(\theta, q, \alpha)$ for different combinations of θ, q and α

modeling an overdispersed, underdispersed, or equidispersed dataset. If ID_Y is greater than one, the distribution of Y can accommodate overdispersion, and if ID_Y is less than one, the distribution can accommodate underdispersion. The distribution is said to be equidispersed when $ID_Y = 1$. The dispersion index of $PoiTG(\theta, q, \alpha)$ is given by

$$ID_Y = \frac{\theta(1-q^2)^2 + q(1-\alpha+q(2+q(1-\alpha)-\alpha^2))}{(1-q^2)(\theta(1-q^2)+q(1-\alpha)+q^2)}.$$

We investigate the nature of ID_Y in Figure 5 for different combinations of θ, q, and α. The plots in the first row exhibits the fact that ID_Y is slightly influenced by θ. As the value of θ increases, the corresponding distribution becomes less overdispersed. The values of ID_Y for $\theta = 1.5$ suggest a lower level of overdispersion compared to the case where $\theta = 0.5$ and $\theta = 1$. The plots in the second row shows that ID_Y is greatly influenced by the choice of q values. As the value of q increases, the corresponding distribution becomes more overdispersed. The values of ID_Y for $q = 0.6$ and $q = 0.8$ suggest a higher level of overdispersion compared to the case where $q = 0.2$. The plots in the third row shows that ID_Y is also substantially influenced by the choice of α values. As the value of α increases, the corresponding distribution becomes less overdispersed.

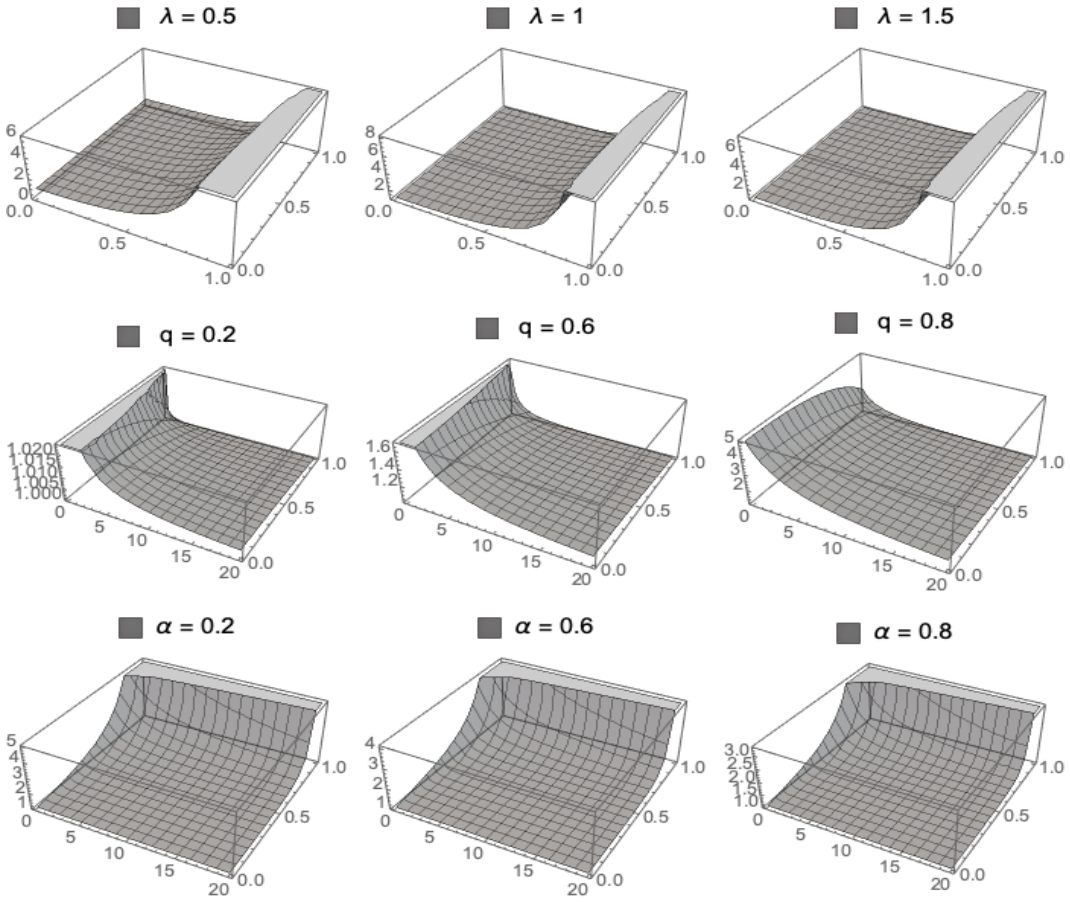


Figure 5: Index of dispersion of $PoiTG(\theta, q, \alpha)$ for multiple combinations of θ, q and α

3.5. Reliability properties

The survival function(sf) is important in reliability analysis and is used to quantify the probability that a system will survive beyond a certain number of events. It is also known as the reliability function because it provides information about the reliability of a system. The sf of $Y \sim PoiTG(\theta, q, \alpha)$ is

$$S_Y(y) = 1 - \frac{\Gamma(y, \theta)}{\Gamma y} + (1 - \alpha)W_1(y - 1, \theta, q) + \alpha W_2(y - 1, \theta, q). \tag{19}$$

In the above expression,

$$W_k(y - 1, \theta, q) = \frac{q^{ky}}{\Gamma y} \exp\left(\frac{\theta(1 - q^k)}{q^k}\right)\Gamma\left(y, \frac{\theta}{q^k}\right), \qquad k = 1, 2\,.$$

The hazard rate function (hrf) of $Y \sim PoiG(\theta, q, \alpha)$ is

$$h_Y(y) = \frac{(1 - q)q^y\left(\alpha(1 + q)e^{\frac{\theta}{q^2}}q^y\Gamma\left(y + 1, \frac{\theta}{q^2}\right) + (1 - \alpha)e^{\frac{\theta}{q}}\Gamma\left(y + 1, \frac{\theta}{q}\right)\right)}{ye^{\theta}(\Gamma y - \Gamma(y, \theta)) + q^y y\left(\alpha e^{\frac{\theta}{q^2}}q^y\Gamma\left(y, \frac{\theta}{q^2}\right) + (1 - \alpha)e^{\frac{\theta}{q}}\Gamma\left(y, \frac{\theta}{q}\right)\right)}. \tag{20}$$

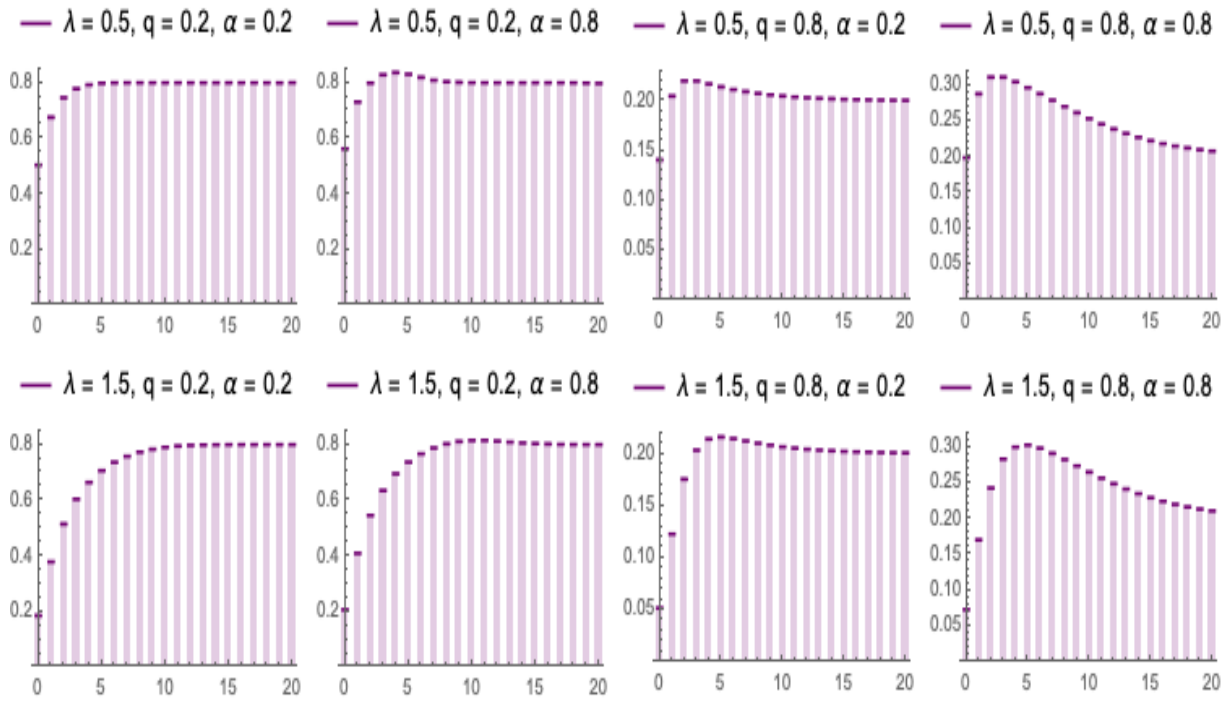


Figure 6: The hrf of $PoiTG(\theta, q, \alpha)$ for multiple choices of θ, q and α

Figure 6 presents the hrf for different parameter settings. In the PoiTG distribution, the behavior of the hrf varies based on parameter values. For small θ, the distribution exhibits a constant failure rate. As θ increases, the failure rate initially rises to a certain point before stabilizing at a constant rate. Conversely, as α increases, the failure rate decreases after a certain time and then remains constant. The mean residual life of Y is

$$\mu_Y(y) = \frac{\sum_{y=t}^{\infty}\left(1 - \frac{\Gamma(y,\theta)}{\Gamma y} + (1-\alpha)W_1(y-1,\theta,q) + \alpha W_2(y-1,\theta,q)\right)}{1 - \frac{\Gamma(t-1,\theta)}{(\Gamma t - 1)} + (1-\alpha)W_1(t-2,\theta,q) + \alpha W_2(t-2,\theta,q)}. \tag{21}$$

In the derivation of the expression in (21), $\bar{F}(y)$ denotes $1-F(y-1)$, and similarly $\bar{F}(t-1) = 1-F(t-2)$. The functions $W_k(y-1,\theta,q)$ and $W_k(t-2,\theta,q)$ hold their meaning as in (19).

3.6. Shannon entropy

Shannon entropy measures the randomness in a distribution. A higher entropy implies the distribution is more spread out across many possible values, whereas a lower entropy implies the distribution is more concentrated at a few values. Let $p_Y(y)$ be the pmf of random variable $Y \sim PoiTG(\theta, q, \alpha)$ given in (4). The Shannon entropy (Shannon, 1948) of Y can be written as

$$S(\theta, q, \alpha) = -\sum_{y=0}^{\infty}((1-\alpha)w_1(y,\theta,q) + \alpha w_2(y,\theta,q))\log((1-\alpha)w_1(y,\theta,q) + \alpha w_2(y,\theta,q)). \tag{22}$$

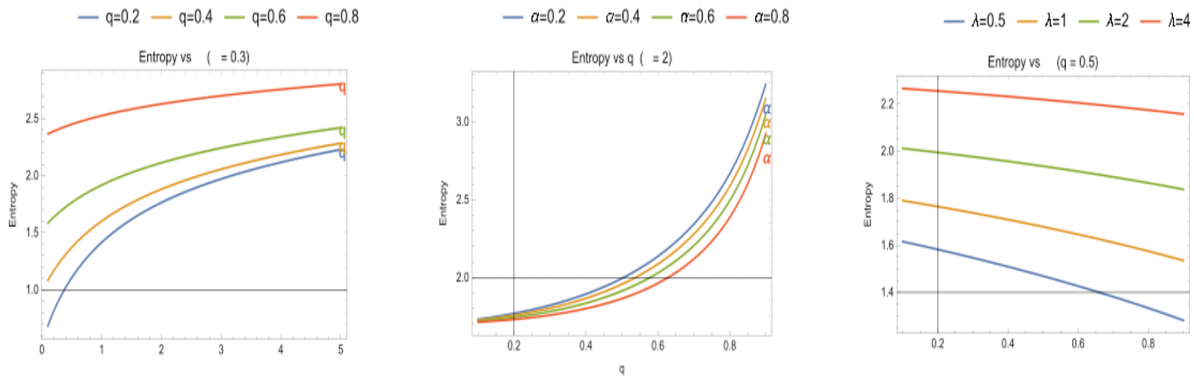


Figure 7: Shannon entropy of PoiTG(λ, q, α)

Figure (7) depicts the Shannon entropy for different parameter choices with the maximum value of y set to 50. The results show that entropy increases with larger values of λ and q, whereas it decreases as α increases. The rising entropy curves for λ and q show the effect of overdispersion increases randomness whereas the falling entropy curves for α illustrate how underdispersion reduces randomness.

4. Parameter estimation

Let $\mathbf{Y}$ be a sample consisting of n observations selected from the $PoiTG(\theta, q, \alpha)$, and a realization on $\mathbf{Y}$ is $(y_1, y_2, \ldots, y_n)$. Here we obtain the maximum likelihood estimates (MLE) of the parameters by two procedures, by direct numerical optimization and by the expectation-maximization (EM) algorithm. Additionally, we derive asymptotic confidence limits for the parameters using the large sample distribution of the MLE.

4.1. Maximum likelihood estimator

Using the pmf of Y in (4), the log-likelihood function of the PoiTG distribution can easily be written as

$$l(\theta, q, \alpha; \boldsymbol{y}) = \sum_{i=1}^{n} \log((1-\alpha)w_{i1} + \alpha w_{i2}). \tag{23}$$

In the above expression,

$$w_{i1} = (1-q)e^{\theta(1-q)/q}\frac{q^{y_i}}{y_i!}\Gamma\left(y_i+1, \frac{\theta}{q}\right), \text{ and}$$
$$w_{i2} = (1-q^2)e^{\theta(1-q^2)/q^2}\frac{q^{2y_i}}{y_i!}\Gamma\left(y_i+1, \frac{\theta}{q^2}\right),$$

for each $i = 1, 2, ..., n$. We list some useful notations below to be used in writing the score functions. For each $i = 1, 2, ..., n$,

$$u_{ij} = \frac{e^{-\theta} q^{j(y_i-1)}}{y_i!} \left(\frac{\theta}{q^j}\right)^{y_i} (1-q^j),$$
$$v_{ij} = \frac{e^{\theta(1-q^j)/q^j} q^{jy_i}}{y_i!} \Gamma\left(y_i+1, \frac{\theta}{q^j}\right), \text{ and}$$
$$q_{jk} = \frac{1-q^j}{q^k}, \qquad j = 1,2 \text{ and } k = 0,1,2,3.$$

The score functions are obtained by differentiating (23) with respect to the parameters, these are given as

$$s_1(\theta, q, \alpha; \boldsymbol{y}) = \sum_{i=1}^{n} \left(\frac{\alpha(u_{i1} - u_{i2} + q^2 q_{22}^2 v_{i2} - q q_{11}^2 v_{i1}) + u_{i1} + q q_{11}^2 v_{i1}}{\alpha q_{20} v_{i2} + (1-\alpha) q_{10} v_{i1}} \right), \tag{24}$$

$$s_2(\theta, q, \alpha; \boldsymbol{y}) \tag{25}$$
$$= \sum_{i=1}^{n} \left(\frac{2\alpha(\frac{\theta u_{i2}}{q} - q v_{i2} - \theta q_{23} v_{i2} + q_{21} v_{i2} y_i) + (1-\alpha)(\frac{\theta u_{i1}}{q} - v_{i1} - \theta q_{12} v_{i1} + q_{11} v_{i1} y_i)}{\alpha q_{20} v_{i2} + (1-\alpha) q_{10} v_{i1}} \right), \tag{26}$$

and

$$s_3(\theta, q, \alpha; \boldsymbol{y}) = \sum_{i=1}^{n} \left(\frac{q_{20} v_{i2} - q_{10} v_{i1}}{\alpha q_{20} v_{i2} + (1-\alpha) q_{10} v_{i1}} \right). \tag{27}$$

Due to the structural complexity of equations (24), (26), and (27), obtaining explicit expressions for the MLE is challenging. Therefore, we employ numerical optimization methods to maximize the log-likelihood function with respect to the parameters. Specifically, we use the *constrOptim* function in *R*, which adequately implements the Nelder-Mead method. Let $\hat{\theta}_{ML}$, $\hat{q}_{ML}$, and $\hat{\alpha}_{ML}$ denote the MLEs of θ, q, and α, respectively.

To obtain the information matrix, the second order partial derivatives of the function given in (23) with respect to θ, q, and α must be computed. However, deriving exact expressions for all second-order partial derivatives is analytically intractable.

The recently introduced $PoiG(\theta, q)$ distribution (Nandi, *et al.*, 2024b) is a special case of the $PoiTG(\theta, q, \alpha)$ distribution when the parameter α is set to zero. Now, we express the pmf of $PoiTG(\theta, q, \alpha)$ as a mixture of the pmf of $PoiG(\theta, q)$ provided in (2). Note that, $Y \sim PoiTG(\theta, q, \alpha)$ can be expressed as

$$Y = \begin{cases} X_1 \text{ with probability } (1-\alpha), \\ X_2 \text{ with probability } \alpha, \end{cases}$$

where $X_1 \sim PoiG(\theta, q)$ and $X_2 \sim PoiG(\theta, q^2)$ are independent. Thus, the pmf is $p_Y(y) = (1-\alpha)p_{X_1}(y) + \alpha p_{X_2}(y)$.Using this formulation, we have derived the Fisher's information matrix and constructed confidence intervals for all parameters of $PoiTG(\theta, q, \alpha)$ (see Appendix I).

4.2. Expectation maximization algorithm

The expectation-maximization (EM) algorithm is a useful iterative procedure to compute MLEs of parameters of a mixture distribution. First, the mixture distribution is represented in such a way that the mixing coefficient is the parameter of a latent Bernoulli random variable. As pointed out earlier, $Y = (1-\alpha)X_1 + \alpha X_2$, where $X_1 \sim PoiG(\theta, q)$ and $X_2 \sim PoiG(\theta, q^2)$. We consider Z to be another random variable such that, Z is independent of X_1 and X_2, and Y is expressed as the following linear combination of X_1 and X_2.

$$Y = (1-Z)X_1 + ZX_2. \tag{28}$$

Clearly, a reasonable assumption is $Z \sim Bernoulli(\alpha), 0 < \alpha < 1$. The estimates obtained by this method are consistent and unique (Dempster, *et al.* (1977), Redner and Walker (1984)). The EM algorithm iterates through two steps: the Expectation (E-step) and the Maximization (M-step). We have n iid copies of Y. However, in the incomplete-data framework of EM algorithm, we find that observations of Z are not available. Thus, the hypothetical complete dataset is $(Y_i, Z_i) : i = 1, 2, ..., n$. Under the formulation, E-step of each EM iteration requires the expectation of $(Z|Y; \Psi^{(k)})$, where $\Psi^{(k)} = (\theta^{(k)}, q^{(k)}, \alpha^{(k)})$ is the estimate of $\Psi = (\theta, q, \alpha)$ in the k^{th} iteration. Since,

$$Z_i|Y_i, \Psi^{(k)} \sim Bernoulli(\alpha_i^{(k)}),$$

$$\alpha_i^{(k+1)} = \frac{\alpha^{(k)}(1-q^{(k)^2})2^{2y_i}\, e^{\frac{\theta^{(k)}(1-q^{(k)^2})}{q^{(k)^2}}}\, \Gamma\left(y_i+1, \frac{\theta^{(k)}}{q^{(k)^2}}\right)}{\alpha^{(k)}(1-q^{(k)^2})2^{2y_i}\, e^{\frac{\theta^{(k)}(1-q^{(k)^2})}{q^{(k)^2}}}\, \Gamma\left(y_i+1, \frac{\theta^{(k)}}{q^{(k)^2}}\right) + (1-\alpha^{(k)})(1-q^k)q^{y_i}\, e^{\frac{\theta^{(k)}(1-q^{(k)})}{q^{(k)}}}\, \Gamma\left(y_i+1, \frac{\theta^{(k)}}{q^{(k)}}\right)}. \tag{29}$$

Now, it is clear that ,

$$E[Z_i|Y_i, \Psi^{(k)}] = \alpha_i^{k+1} \text{and } V[Z_i|Y_i, \Psi^{(k)}] = \alpha_i^{k+1}(1-\alpha_i^{k+1}). \tag{30}$$

We estimate α for the $(k+1)^{th}$ iteration by

$$\alpha^{k+1} = \frac{1}{n}\sum_{i=1}^{n} \alpha_i^{k+1}. \tag{31}$$

Then, we proceed to M-step, where we maximize the log-likelihood of observed $(y_1, y_2, ..., y_n)$ on $(Y_1, Y_2, ..., Y_n)$ with respect to θ and q for given $\alpha^{(k+1)}$ and obtain $\theta^{(k+1)}$ and $q^{(k+1)}$. That is,

$$(\theta^{(k+1)}, q^{(k+1)}) = \arg \max_{\theta, q}\, l(\theta, q, \alpha^{(k+1)}; \vec{y}).$$

We continue iterating the successive E-step and M-step until $|\theta^{(k+1)} - \theta^{(k)}| < \epsilon$, $|q^{(k+1)} - q^{(k)}| < \epsilon$, and $|\alpha^{(k+1)} - \alpha^{(k)}| < \epsilon$, simultaneously. We set $\alpha_0 = 0.5$ and $\epsilon = 0.0001$ in the implementation of EM algorithm.

5. Simulation study

The initial aim is to extract a sample of size n from $Y \sim PoiTG(\theta, q, \alpha)$. We obtain two samples, each of size n, namely $\mathbf{y_1}$ from $Y_1 \sim Poi(\theta)$ and $\mathbf{y_2}$ from $Y_2 \sim TGD(q, \alpha)$. Since Y is the sum of Y_1 and Y_2, a sample from the PoiTG distribution can be obtained as $\mathbf{y} = \mathbf{y_1} + \mathbf{y_2}$.

The main objective is to assess the performance of the MLEs obtained using a general optimization approach and the estimates obtained using the EM algorithm. We generate a sample of size n from the PoiTG distribution for 8 different combinations of the parameters θ, q, and α. Additionally, we consider six different values of n, such as 30, 50, 100, 250, 500 and 1000. The MLEs obtained through the EM algorithm are denoted as $\hat{\theta}_{EM}$, $\hat{q}_{EM}$, and $\hat{\alpha}_{EM}$. For each sample of size n, we obtain the MLEs using both approaches mentioned in Section 4.1 and Section 4.2. Subsequently, we calculate the bias and squared error of these estimates.

To evaluate the performance of the estimators, we replicate the experiment 10,000 times for each value of n and each fixed combination of parameter values. We then compute the average bias and average mean squared error (MSE). Table 5 (see Appendix I) displays the values of the corresponding average biases for $\hat{\theta}_{ML}$, $\hat{q}_{ML}$, $\hat{\alpha}_{ML}$, $\hat{\theta}_{EM}$, $\hat{q}_{EM}$, and $\hat{\alpha}_{EM}$ in the non-shaded regions. The shaded regions represent the average MSEs of these estimates.

From Table 5, it can be observed that the average biases and the average MSEs of $\hat{\theta}_{ML}$, $\hat{\theta}_{EM}$, $\hat{q}_{ML}$, and $\hat{q}_{EM}$ decreases with the increase in sample size n. In fact, as the sample size increases from 30 to 1000, a consistent pattern is observed, where both the average bias and the MSE decrease with larger n. It is also observed that for particular combination of n, θ, and greater value of q the average biases and mean squared errors of $\hat{\theta}_{ML}$ and $\hat{\theta}_{EM}$ decrease as α increases. However, for the same combinations, the average mean squared error of $\hat{q}_{ML}$ and $\hat{q}_{EM}$ exhibits an opposite trend. When keeping n, q, and α fixed, the average mean squared errors of both the estimators $\hat{\theta}_{ML}(\hat{\theta}_{EM})$ and $\hat{q}_{ML}$ $(\hat{q}_{EM})$ increase with θ. Similarly, for fixed n, θ, and α, the average mean squared errors of $\hat{\theta}_{ML}$ and $\hat{\theta}_{EM}$ increase with q, while the same for $\hat{q}_{ML}$ and $\hat{q}_{EM}$ decrease with increasing q. On the other hand, the biases and MSEs of $\hat{\alpha}_{ML}$ and $\hat{\alpha}_{EM}$ do not exhibit any clear trend. Additionally, it is noteworthy that for any combination of parameters, the average biases and average mean squared errors of the maximum likelihood estimators obtained using the EM algorithm ($\hat{\theta}_{EM}$, $\hat{q}_{EM}$, and $\hat{\alpha}_{EM}$) are smaller compared to their counterparts for the optimization approach ($\hat{\theta}_{ML}$, $\hat{q}_{ML}$, and $\hat{\alpha}_{ML}$).

6. Goodness-of-Fit evaluation in data analysis

This section highlights the flexibility of the PoiTG distribution in effectively modeling diverse datasets across various fields. The goodness of fit is assessed using the Chi-square (χ^2) test along with its associated P-value. Additionally, the fit of the PoiTG distribution is compared with that of other competing distributions listed below.

Table 1: Competing models

Distribution	Abbreviation	Author(s)
Discrete log-logistic	DsLogL	Para and Jan (2016a)
Discrete inverse Weibull	DsIW	Jazi, *et al.* (2010)
Discrete Inverted Nadarajah-Haghighi	DsINH	Singh, *et al.* (2022)
Discrete Burr type II	DsBX-II	Para and Jan (2016b)
Discrete Belal	DsBL	Altun, *et al.* (2022)
Geometric	Geo	-
Discrete Rayleigh	DsR	Roy (2004)
Discrete inverse Rayleigh	DsIR	Hussain and Ahmad (2014)
Poisson	Poi	-
Discrete Burr-Hatke	DsBH	El-Morshedy, *et al.* (2020)
Discrete Pareto	DsPa	Krishna and Pundir (2009)
Poisson Geometric	PoiG	Nandi, *et al.* (2024b)
Transmuted Geometric	TGD	Chakraborty and Bhati (2016)

6.1. Dataset I

The data set is sourced from (https://www.worldometers.info/coronavirus/country/south-korea/) and consists of daily COVID-19 fatalities in South Korea from February 15 to December 12, 2020. Figure 8 provides a primary graphical representation of these data using nonparametric methods. In Figure 8, it is observed that the frequency distribution is unimodal, has a mode at zero, exhibits a heavier tail compared to the Poisson distribution, and contains a small proportion of outliers. Furthermore, Table 2 presents the observed frequency (OF), expected frequency (EF), maximum likelihood estimators (MLEs) of the parameters, log-likelihood ($-L$), and Chi-square test values along with their corresponding p values for each competing distribution applied to Data set I. In particular, Table 2 shows that the χ^2 test fails to reject the hypothesis of a good fit only for the PoiTG distribution.

Among all models tested, the PoiTG distribution exhibits superior performance as the optimal fit for this data set. This conclusion is supported by its lowest χ^2 value compared to other distributions. The estimated pmfs for Data set I are illustrated in Figure 9.

6.2. Dataset II

The data set in question captures the number of computer malfunctions over 128 successive weeks of operation, as documented by Hand, *et al.* (1995). Non-parametric graphs are recommended for the first visual analysis of this data, with the findings illustrated in Figure 10. From Figure 10, we note that the frequency distribution is multimodal, has maximum frequency at 2, has a heavy tail and a moderate proportion of outliers. Table 3 outlines the OF, EF, MLEs for the parameters, $-L$, and Chi-square test results, along with their respective P-values, for each competing distribution within dataset II.

From Table 3 we note that the χ^2 test does not reject the hypothesis of a good fit for the PoiTG distribution. Among all the tested models, the PoiTG distribution and others distributions such as PoiG and TGD perform in similar manner for this dataset. The estimated pmfs for Dataset II are illustrated in Figure 11.

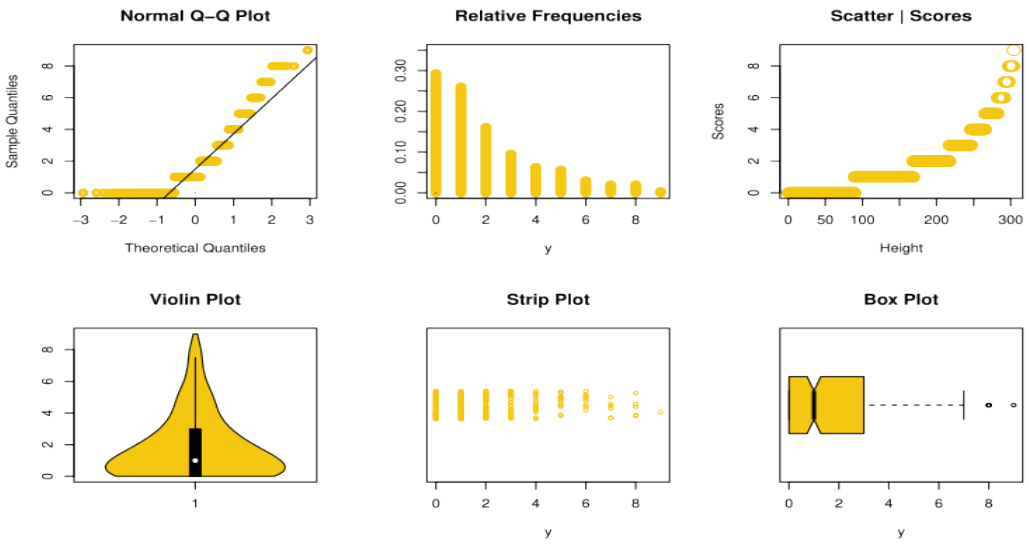


Figure 8: Plots based on non-parametric methods for dataset I

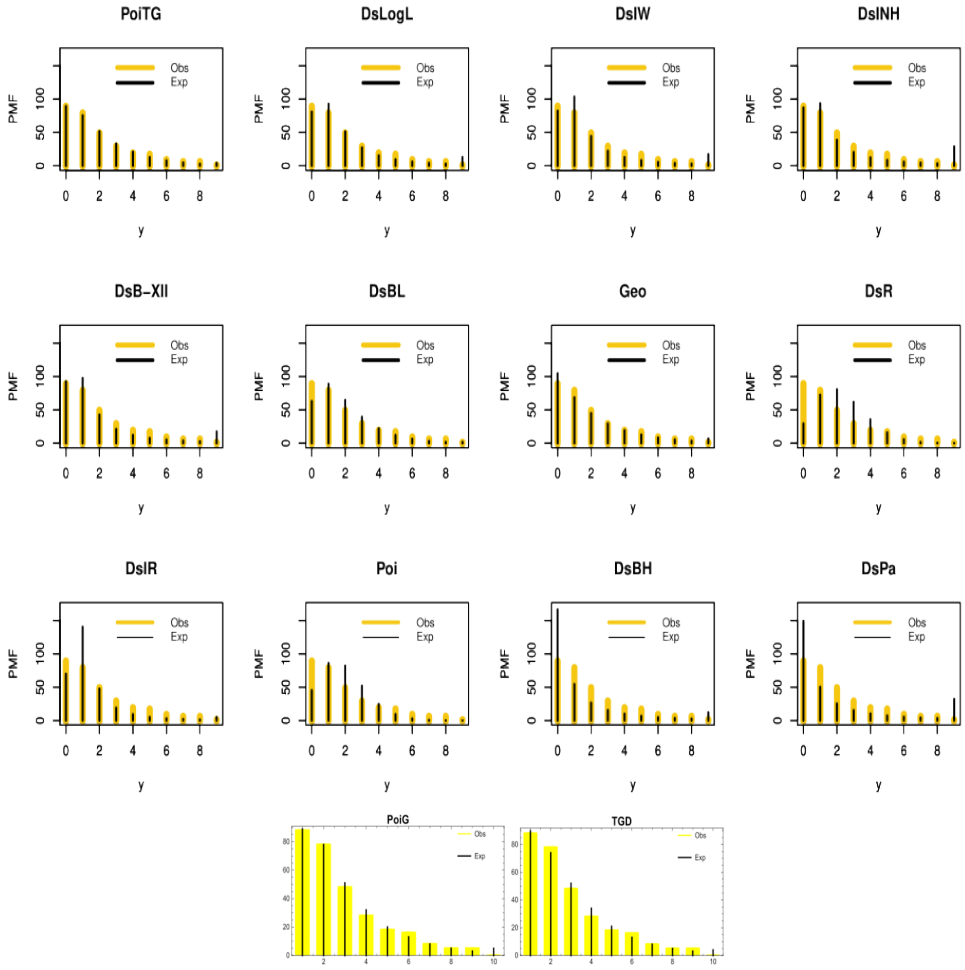


Figure 9: The estimated probability mass functions for dataset I

Table 2: The MLEs and goodness-of fit measures of all competing distributions for dataset I

No. ECB	OF	Expected frequency (EF)												
		PoiTG	TGD	PoiG	DsLogL	DsIW	DsINH	DsB-XII	DsBL	Geo	DsR	DsIR	Poi	DsBH	DsPa
0	89	89.423	89.772	89.129	80.931	82.351	87.190	92.887	63.089	104.780	29.499	69.890	45.408	166.600	149.356
1	79	75.642	74.420	78.023	92.775	103.702	93.636	97.788	89.133	68.665	72.407	140.613	86.336	54.598	50.500
2	49	51.726	51.965	51.414	51.431	44.390	38.648	42.676	64.868	44.998	80.783	47.685	82.074	26.667	25.478
3	29	33.332	33.767	32.241	27.336	22.317	20.378	21.174	39.578	29.489	61.938	19.125	52.014	15.540	15.379
4	19	20.893	21.168	20.082	15.559	12.926	12.508	12.172	22.307	19.325	35.679	9.333	24.725	10.017	10.312
5	17	12.898	13.015	12.500	9.518	8.248	8.442	7.754	12.037	12.664	15.984	5.198	9.402	6.885	7.387
6	9	7.893	7.914	7.780	6.193	5.636	6.076	5.308	6.328	8.299	5.664	3.162	2.979	4.953	5.533
7	6	4.805	4.781	4.842	4.247	4.052	4.581	3.831	3.273	5.439	1.603	2.098	0.809	3.682	4.408
8	6	2.917	2.878	3.014	3.061	3.031	3.576	2.879	1.674	3.564	0.364	1.429	0.192	2.808	3.466
9	1	4.471	4.316	5.008	12.949	17.347	28.965	17.531	1.713	6.777	0.079	5.467	0.061	12.25	32.181
Total	304	304	304	304	304	304	304	304	304	304	304	304	304	304	304
$-L$		565.047	565.087	565.258	577.011	586.855	596.877	587.652	575.338	568.083	638.905	606.870	621.098	620.466	633.531
MLE_α		0.275	−0.441	–	1.716	0.271	6.241	0.591	0.707	0.655	0.903	0.229	1.901	0.904	0.377
MLE_q		0.604	0.598	0.377	1.878	1.411	0.139	2.466	–	–	–	–	–	–	–
MLE_θ		0.115	–	0.2529	–	–	–	–	–	–	–	–	–	–	–
χ^2		2.771	2.9967	3.59	25.019	41.868	51.008	44.784	28.203	11.767	184.989	92.204	115.896	109.333	128.631
d.f		4	5	6	6	6	6	6	6	7	5	6	4	6	7
P-value		0.597	0.7005	0.7319	≤ 0.001	≤ 0.001	≤ 0.001	≤ 0.001	≤ 0.001	0.109	≤ 0.001	≤ 0.001	≤ 0.001	≤ 0.001	≤ 0.001

Table 3: The MLEs and goodness-of fit measures for all competing distributions in dataset II

X	OF	PoiTG	TGD	PoiG	DsLogL	DsIW	DsINH	DsB-XII	DsBL	Geo	DsR	DsIR	Poi	DsBH	DsPa
0	15	13.754	14.871	13.758	11.012	9.689	12.401	19.886	9.7251	25.527	3.636	3.237	2.310	65.853	47.806
1	19	21.773	20.205	21.772	23.309	33.087	30.161	36.588	19.441	20.434	10.302	47.809	9.272	21.915	19.190
2	23	21.140	19.678	21.137	23.010	23.104	19.935	19.052	20.731	16.360	15.306	34.021	18.613	10.931	10.760
3	14	17.242	16.958	17.238	17.906	14.481	12.782	10.690	18.511	13.096	18.041	16.649	24.912	6.538	7.021
4	15	13.267	13.740	13.265	12.868	9.539	8.721	6.832	15.191	10.491	18.440	8.773	25.014	4.342	5.001
5	10	10.045	10.741	10.044	9.105	6.637	6.291	4.762	11.873	8.385	16.918	5.079	20.085	3.088	3.774
6	8	7.577	8.211	7.577	6.506	4.840	4.732	3.520	8.993	6.720	14.172	3.174	13.440	2.304	2.967
7	4	5.712	6.184	5.712	4.738	3.653	3.693	2.724	6.666	5.381	10.944	2.106	7.713	1.782	2.404
8	6	4.305	4.612	4.307	3.525	2.849	2.950	2.171	4.865	4.312	7.839	1.466	3.870	1.417	1.994
9	2	3.245	3.415	3.246	2.677	2.273	2.412	1.774	3.510	3.456	5.228	1.059	1.736	1.151	1.686
10	3	2.446	2.517	2.447	2.073	1.848	2.013	1.482	2.510	2.753	3.256	0.789	0.690	0.953	1.447
11	3	1.843	1.848	1.844	1.633	1.533	1.691	1.251	1.783	2.205	1.897	0.604	0.253	0.800	1.258
12	2	1.389	1.354	1.390	1.307	1.284	1.460	1.080	1.260	1.766	1.036	0.472	0.080	0.680	1.106
+13	4	4.254	3.660	4.257	8.331	13.183	18.758	16.188	2.941	7.114	0.985	2.762	0.0062	6.246	21.586
Total	128	128	128	128	128	128	128	128	128	128	128	128	128	128	128
$-l$		311.638	311.038	311.663	319.854	330.446	331.931	342.581	318.545	320.703	347.148	356.525	384.974	379.346	369.766
MLE_α		< 0.001	–	−0.787	3.337	0.076	1.244	0.785	0.831	0.801	0.972	0.025	4.016	0.971	0.509
MLE_q		0.753	0.728	–	1.960	1.235	1.634	3.309	–	–	–	–	–	–	–
MLE_θ		0.829	0.727	0.829	–	–	–	–	–	–	–	–	–	–	–
χ^2		1.540	1.580	1.539	6.018	18.709	20.199	40.183	6.361	11.651	49.897	50.533	88.995	143.174	94.971
d.f		6	7	7	7	6	6	5	8	9	8	5	6	5	6
P-value		0.9567	0.9793	0.9809	0.537	0.005	0.003	0	0.607	0.234	0	0	0	0	0

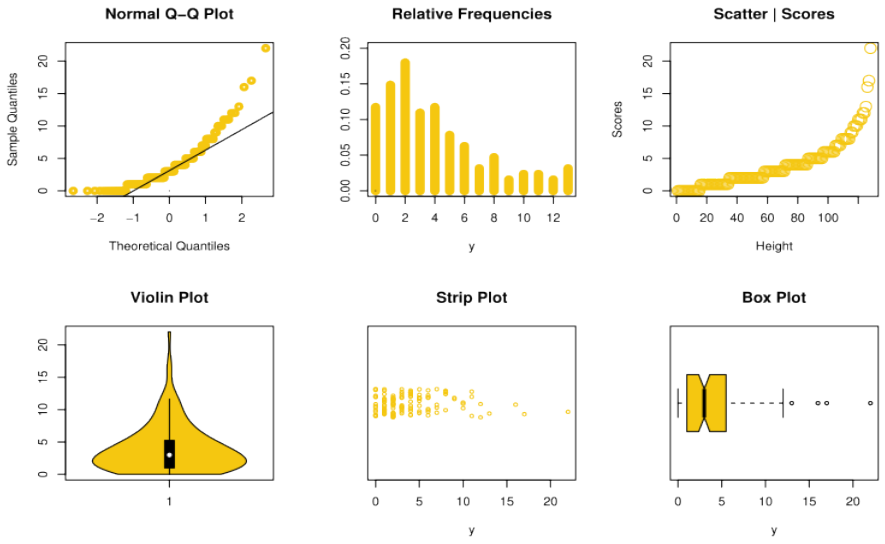


Figure 10: Plots based on non-parametric methods for dataset II

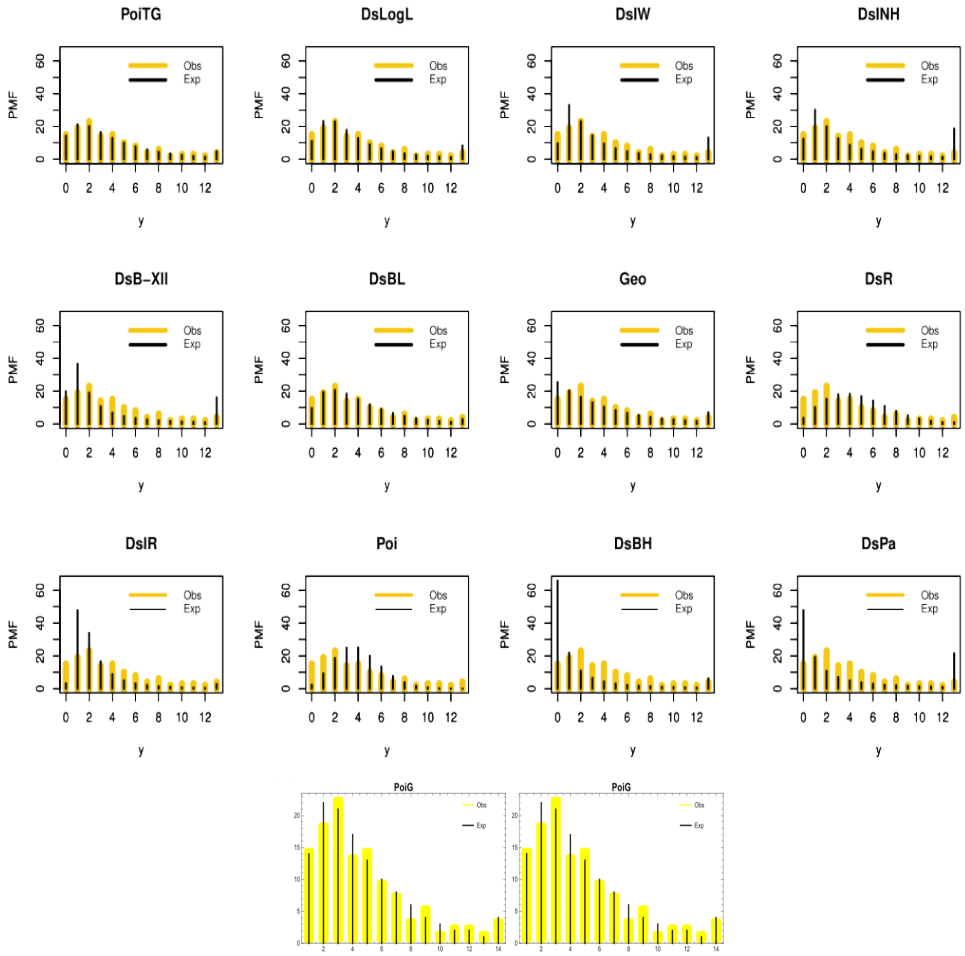


Figure 11: The estimated probability mass functions for dataset II

Remark 2: The PoiTG distribution provides a noticeably better fit compared to most of the competing models considered for Datasets I & II. It is noted, however, that its performance is equivalent to that of PoiG and TGD. Despite this equivalence, the PoiTG distribution remains particularly useful, as it unifies different hazard rate shapes within a single framework and demonstrates greater flexibility when extended to regression settings, as shown in Section 7. This makes it a strong candidate for wider application in future case studies.

7. Regression models: theory and practical application

In this section, we reparameterize the PoiTG model in terms of its mean (μ) and the two parameters q and α for modeling overdispersed count responses. Using this reparametrization, we develop a new count regression model within the framework of generalized linear models (GLM). The $PoiTG(\theta, q, \alpha)$ model is re-parametrized by considering $\theta = \mu - g(q, \alpha)$, where $g(q, \alpha) = [q(1-\alpha) + q^2]/[1-q^2]$ with the restriction that $\mu > g(q, \alpha)$ since $\theta > 0$. The reparameterized PoiTG distribution is is denoted as $PoiTG(\mu, q, \alpha)$ with $\mu > g(q, \alpha)$, $0 < q < 1$ and $0 < \alpha < 1$. Under this reparameterization, the pmf of the PoiTG distribution in (4) can be rewritten as

$$p_Y(y) = (1-\alpha)\, w_1(y, \mu, q, \alpha) + \alpha\, w_2(y, \mu, q, \alpha), \qquad y = 0, 1, 2, \ldots\,. \tag{32}$$

In the above expression,

$$\begin{aligned} w_k(y, \mu, q, \alpha) &= (1-q^k) q^{ky} \exp[-(\mu - g(q,\alpha))] \sum_{i=0}^{y} \frac{1}{i!} \left(\frac{\mu - g(q,\alpha)}{q^k}\right)^i \\ &= \frac{(1-q^k) q^{ky}}{\Gamma(y+1)} \exp\left[(\mu - g(q,\alpha)) \left(\frac{1}{q^k} - 1\right)\right] \Gamma\left(y+1, \frac{\mu - g(q,\alpha)}{q^k}\right),\ k = 1, 2\,. \end{aligned}$$

Clearly, the mean is $E(Y) = \mu$ and the variance is

$$V(Y) = \frac{\mu(1-q^2)^2 + q^2[(1+q)^2 - 2q\alpha - \alpha^2]}{(1-q^2)^2}.$$

Let y be a random sample of n from the $PoiTG(\mu, q, \alpha)$ distribution, as defined in (32). The PoiTG_{GLM} is formulated using the log link function *i.e.*, which establishes the relationship between the p-dimensional covariates and the mean response.

$$\eta_i = g(\mu_i) = \log(\mu_i) = \sum_{j=0}^{p} \gamma_j x_{ij} = \boldsymbol{x_i'\gamma}, \quad i = 1, 2, \ldots n, \tag{33}$$

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_p)'$ and $\boldsymbol{x_i'} = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$ are the vector of unknown regression coefficients and covariates augmented with a 1 to account for the intercept. Let $Y_i|\boldsymbol{x_i'}$ be a response for a given set of covariates, $\boldsymbol{x_i'}$. By substituting $\mu = \mu_i = \exp(\boldsymbol{x_i'\gamma})$ into (32), the pmf of $Y_i|\boldsymbol{x_i'} \sim PoiTG_{GLM}(\mu, q, \alpha)$ can be expressed as

$$p_Y(y_i|\boldsymbol{x_i'}) = (1-\alpha)\, w_1(y_i, \exp(\boldsymbol{x_i'\gamma}), q, \alpha) + \alpha\, w_2(y_i, \exp(\boldsymbol{x_i'\gamma}), q, \alpha), \quad i = 1, 2, \ldots n\,. \tag{34}$$

In the above expression,

$$w_k(y_i, \exp(\boldsymbol{x_i'\gamma}), q, \alpha) = \frac{(1-q^k)q^{ky_i}}{\Gamma(y_i+1)} \exp\left[(\exp(\boldsymbol{x_i'\gamma}) - g(q,\alpha))\left(\frac{1}{q^k}-1\right)\right] \Gamma\left(y_i+1, \frac{\exp(\boldsymbol{x_i'\gamma}) - g(q,\alpha)}{q^k}\right), \ k = 1, 2 \ .$$

The log-likelihood function of the PoiTG regression model is given by

$$l(\boldsymbol{\delta}) = \sum_{i=1}^{n} \log\left[(1-\alpha)\, w_1(y_i, \exp(\boldsymbol{x_i'\gamma}), q, \alpha) + \alpha\, w_2(y_i, \exp(\boldsymbol{x_i'\gamma}), q, \alpha)\right], \tag{35}$$

where $\boldsymbol{\delta} = (\boldsymbol{\gamma},\, q,\, \alpha)'$. The unknown parameters, $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, ..., \gamma_p)'$, q and α are estimated by maximizing (34) using the *constrOptim* function in R, which efficiently implements the Nelder-Mead method. The asymptotic confidence intervals for the regression coefficients $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, ..., \gamma_p)'$, q and α can be derived from asymptotic theory as $n \to \infty$. Based on this theory,

$$\sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \to N_{p+3}(0,\, I^{-1}(\hat{\boldsymbol{\delta}})) \text{ as } n \to \infty,$$

where $I(\hat{\boldsymbol{\delta}})$ is the observed information matrix of $\hat{\boldsymbol{\delta}}$.

7.1. Bootstrap standard error and confidence interval

Calculating asymptotic confidence intervals for each regression coefficient and the parameters q and α is laborious. Bootstrapping provides an alternative approach to obtain precise standard errors (SE) and confidence intervals (CI) for the estimates. To determine these for the PoiTG regression model, we adopt the following nonparametric bootstrap routine.

First, we randomly select a sample with replacement B times from the set of row indices $\{1, 2, ..., n\}$ of the given dataset. For the j^{th} repetition, let $\{i_1^{(j)}, i_2^{(j)}, ..., i_n^{(j)}\}$ denote the randomly selected indices. The j^{th} bootstrap dataset is

$$D^{(j)} = \begin{pmatrix} y_{i_1^{(j)}} & x_{1i_1^{(j)}} & \dots & x_{pi_1^{(j)}} \\ y_{i_2^{(j)}} & x_{1i_2^{(j)}} & \dots & x_{pi_2^{(j)}} \\ \vdots & \vdots & & \vdots \\ y_{i_n^{(j)}} & x_{1i_n^{(j)}} & \dots & x_{pi_n^{(j)}} \end{pmatrix}, \ j = 1, 2, ..., B.$$

Based on $D^{(j)}$, we compute the MLEs of $\gamma_1, \gamma_2, ..., \gamma_p$ as $\hat{\gamma_1}^{(j)}, \hat{\gamma_2}^{(j)}, ..., \hat{\gamma_p}^{(j)}$ for each $j = 1, 2, ..., B$. Consequently, for any given $k \in 0, 1, ..., p$, we obtain a set of MLEs of γ_k, represented as $\boldsymbol{\gamma_k^*} = (\hat{\gamma_k}^{(1)}, \hat{\gamma_k}^{(2)}, ..., \hat{\gamma_k}^{(B)})$.

To obtain the ordinary bootstrap estimate of the SE of $\hat{\gamma_k}$, we compute the standard deviation of the array $\boldsymbol{\gamma_k^*}$ as

$$SE(\hat{\gamma_k}) = \sqrt{\frac{1}{B}\sum_{j=1}^{B}(\gamma_k^{(j)} - \bar{\gamma_k})^2},$$

where, $\bar{\gamma_k}$ is the mean of the bootstrap estimates. To construct $100(1-\nu)\%$ bootstrap CI for γ_k, first we obtain the empirical cumulative distribution function:

$$\hat{F}_B(t) = \frac{1}{B}\sum_{j=1}^{B} I(\sqrt{n}(\hat{\gamma}_k^{(j)} - \hat{\gamma_k}) \leq t).$$

Let, $t_{\nu/2} = \hat{F_B^{-1}}(\nu/2)$ and $t_{1-\nu/2} = \hat{F_B^{-1}}(1-\nu/2)$. Using these threshold points, the $100(1-\nu)\%$ bootstrap CI for γ_k is given by

$$CI(\hat{\gamma_k}) = \left[\hat{\gamma_k} - \frac{t_{1-\nu/2}}{\sqrt{n}},\, \hat{\gamma_k} + \frac{t_{\nu/2}}{\sqrt{n}}\right].$$

Typically, B is chosen to be large, on the order of thousands. A similar procedure can be applied to estimate the SEs and CIs for q and α. The corresponding formulas for q and α are as follows:

$$SE(\hat{q}) = \sqrt{\frac{1}{B}\sum_{j=1}^{B}(q^{(j)} - \bar{q})^2}\ ,\ SE(\hat{\alpha}) = \sqrt{\frac{1}{B}\sum_{j=1}^{B}(\alpha^{(j)} - \bar{\alpha})^2},$$

$$CI(\hat{q}) = \left[\hat{q} - \frac{t_{1-\nu/2}}{\sqrt{n}},\, \hat{q} + \frac{t_{\nu/2}}{\sqrt{n}}\right] \text{ and } CI(\hat{\alpha}) = \left[\hat{\alpha} - \frac{t_{1-\nu/2}}{\sqrt{n}},\, \hat{\alpha} + \frac{t_{\nu/2}}{\sqrt{n}}\right].$$

7.2. Applied regression fitting

In this section, we evaluate the practicality of the PoiTG regression model (PoiTG_{GLM}) using a real-world dataset. DeHart*, et al.* (2008) conducted a study in which individuals classified as "moderate to heavy drinkers" (defined as consuming at least 12 alcoholic drinks per week for women and 15 for men) were asked to maintain a daily record of their alcohol intake over a 30-day period. Participants also completed various rating scales related to daily life events and self-esteem.

One of the key hypotheses explored in the study was that negative events, particularly those involving romantic relationships, could be associated with the amount of alcohol consumed, especially among people with low self-esteem. The primary focus of our analysis is to model the variable *numall*, which represents the number of alcoholic beverages (or "drinks") consumed in a single day. Furthermore, we examine how *number* is influenced by factors such as daily desire to drink (*desired*), age (*age*), negative romantic relationship events (*nevgent*), and state self-esteem (*state*). The predictive model is formulated as follows:

$$g(\mu_i) = \log(\mu_i) = \gamma_0 + \gamma_1\, negevent_i + \gamma_2\, state_i + \gamma_3\, age_i + \gamma_4\, desired_i$$

for $i = 1, ..., 618$.

We compare the fitted values of the PoiTG regression model (PoiTG_{GLM}) with those of established regression models for equidispersed count data, such as the Poisson regression model (Poi_{GLM}), and for overdispersed count data, including the PoiG regression model

Table 4: Analysis of the Alcohol consumption dataset

	Estimators(SE)				
	Poi_{GLM}	NB_{GLM}	CMP_{GLM}	PoiG_{GLM}	PoiTG_{GLM}
γ_0	-0.02541(0.3198)	-0.2559(0.4662)	-0.0486(0.2225)	0.4054(0.5121)	0.7128(0.16209)
	(-0.6522, 0.6014)	(-1.1696 ,0.6578)	(-0.4847 ,0.3875)	(-0.5983,1.4091)	(0.3951,1.0305)
γ_1	-0.2726(0.0709)	-0.2420(0.1002)	-0.1452(0.0509)	-2.0305(0.12040)	0.6884(0.0347)
	(-0.4116,-0.1336)	(-0.4384 ,-0.0456)	(-0.2450,-0.0454)	(-2.2665,-1.7945)	(0.6204,0.7564)
γ_2	-0.0514(0.0570)	-0.0600(0.0843)	-0.0670(0.0389)	-0.6384(0.0749)	-0.1148(0.0244)
	(-0.1631,0.0603)	(-0.2252,0.1052)	(-0.1432,0.0092)	(-0.7851,-0.4917)	(-0.1627,-0.0669)
γ_3	-0.0018(0.0058)	0.0039(0.0083)	-0.0058(0.0040)	-0.1126(0.0138)	0.2720(0.0029)
	(-0.0132,0.0096)	(-0.0124,0.0202)	(-0.0136,0.0020)	(-0.1396,-0.0856)	(0.2663,0.2777)
γ_4	0.2755(0.01636)	0.2871(0.0232)	0.1342(0.0139)	0.0133(0.0502)	-0.0959(0.0077)
	(0.2434,0.3076)	(0.2416,0.3326)	(0.1070,0.1614)	(-0.0851,0.1117)	(-0.1109,-0.0809)
ϕ	—	2.4720(0.3050)	-1.2003(0.1392)	0.4611(0.0190)	—
	—	(1.8742,3.0698)	(-1.4731,-0.9275)	(0.4239,0.4983)	—
q	—	—	—	—	0.0148(0.0071)
	—	—	—	—	(0.0009,0.0287)
α	—	—	—	—	0.2167(0.0334)
					(0.1512,0.2822)
$-\text{l}$	1317.750	1217.189	1223.979	1227.197	**1197.856**
AIC	2645.500	2446.378	2459.958	2466.394	**2409.712**
BIC	2667.632	2472.937	2486.517	2492.953	**2440.697**

The bootstrap SEs and CIs of the MLEs are presented in parentheses next to and below the estimates.

(PoiG_{GLM}), the COM-Poisson regression model (CMP_{GLM}), and the NB regression model (NB_{GLM}). Model selection is based on the estimated negative log-likelihood (-LL), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). Table 4 presents the estimated parameters, -LL, AIC, and BIC values for the fitted models using the Alcohol consumption dataset. The results demonstrate that the PoiTG regression model offers the highest modeling accuracy, as it attains the lowest -LL, AIC, and BIC values. The estimated parameters of the PoiTG regression model further implies that the predictive model can be written as

$$\mu_i = \exp(0.7128 + 0.6884\, negevent_i - 0.1148\, state_i + 0.2720\, age_i - 0.0959\, desired_i).$$

Figure 12 includes scatter plot, Q-Q plot and histogram of the randomized quantile residuals of PoiTG regression model fitted to Alcohol consumption dataset.

The question arises whether there is a statistically significant difference between the PoiTG regression model and its closest competitor, the NB regression model, in fitting this data set. To answer this question, we use the generalized likelihood ratio test based on LLs, whose test statistic is $\text{LL}(\text{NB}_{GLM})$-$\text{LL}(\text{PoiTG}_{GLM})$=19.333, which results in a significance level of 0.00001. To perform this test, we compare the obtained statistics with $\chi^2(1)$ (Chi-square distribution with one degree of freedom). Therefore, there is a significant difference

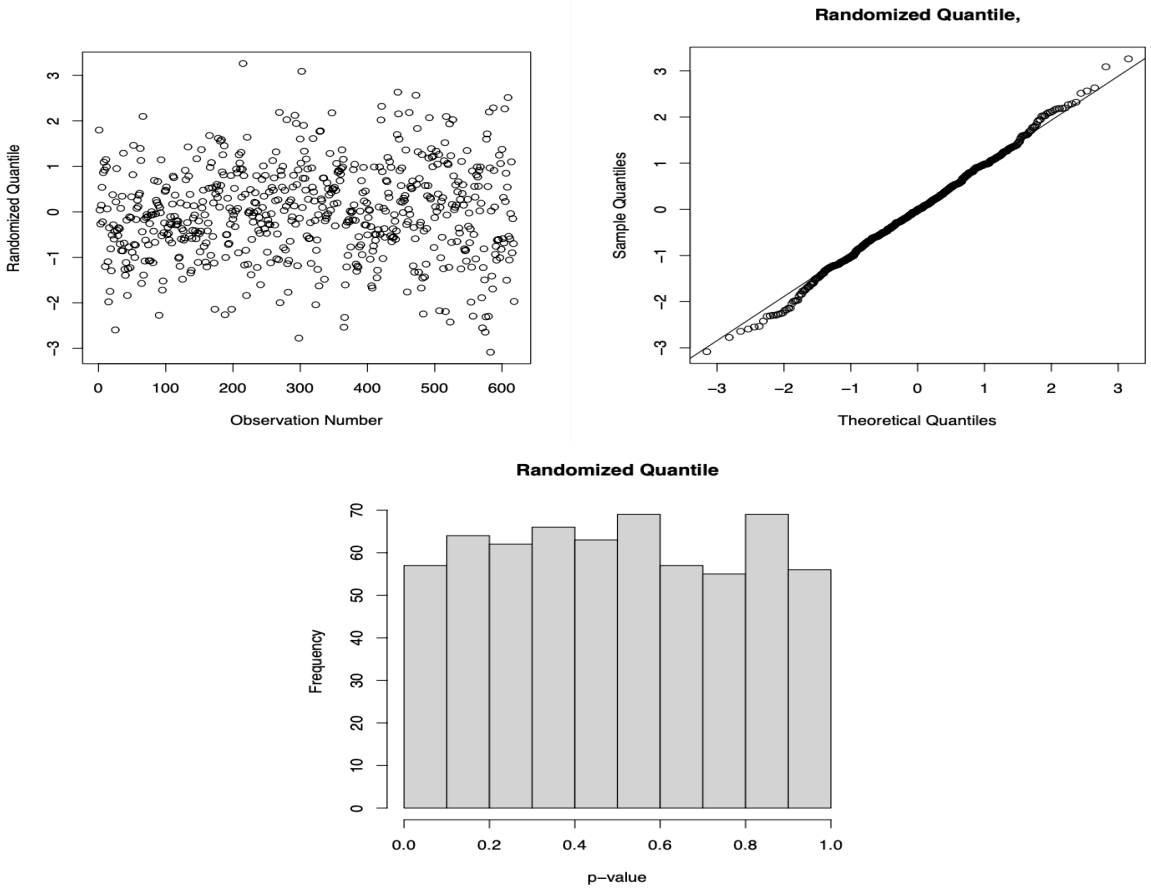


Figure 12: The randomized quantile residuals and corresponding Q-Q plot and histogram

between these two models (for more details, see Self and Liang (1987)). Thus, among the used models, the PoiTG regression model is the best model to fit this data set.

To check the efficiency of the fitted models, in addition to using the previous criteria, we also use the Pearson chi-square statistic χ^2 Tests.

$$\chi^2 = \sum_{i=1}^{c} \frac{(o_i - e_i)^2}{e_i},$$

where, o_i is observed frequency, e_i is the expected frequency of the ith category, and c is the total number of categories. To perform this test, we sorted the data from 0 to 7 and more than 7. Then, the observed and expected frequency of the data set are calculated. The value of the test statistic is 7.87 and the P-value corresponding to the test is 0.2475, which supports a good fit of the PoiTG model to the data.

8. Discussion

This article introduces a novel discrete distribution and thoroughly studies its important properties. This distribution possesses a closed form of the mean, which eventually facilitates the corresponding generalized linear model. However, the usp of the present work is the real-life applications which should motivate the practitioners to adopt it as a viable alternative to the COM-Poisson and other overdispersed count data models. The interpretability of the proposed model is also worth adoption in practice. The parameters θ, q, and α control the distribution's tail behavior, while q and α account for the overdispersion in the dataset. Their combined influence provides flexibility in shaping the distribution. In the case of large θ, the distribution maintains a bell-shaped mass distribution. In Section 3.5, we observed that the hazard rate function of the PoiTG distribution can capture near-constant, increasing, and decreasing behaviors under different parameter settings. One possible extension of this work is to suitably modify the distribution so as to capture bathtub-shaped hazard behaviors, thereby enriching its applicability to a wider range of reliability scenarios. This work is complete in its objectives. However, further research is warranted on exploring the applicability of this new distribution for count time series data. Another important direction of future research may be to study the inferential aspects of the proposed model under the Bayesian perspective.

Acknowledgements

The authors gratefully acknowledge the editor and the learned reviewers for the insightful comments and constructive suggestions, which significantly improved the clarity and presentation of the manuscript.

Conflict of interest

The authors declare that there is no conflict of interest. Additionally, no funding was received for the preparation of this manuscript.

References

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. US Government printing office.

Altun, E. (2020). A new generalization of geometric distribution with properties and applications. *Communications in Statistics-Simulation and Computation*, **49**, 793–807.

Altun, E., El-Morshedy, M., and Eliwa, M. S. (2022). A study on discrete Bilal distribution with properties and applications on integervalued autoregressive process. *REVSTAT-Statistical Journal*, **20**, 501–528.

Amani, K. M., Kouakou, K. J. G., and Hili, O. (2025). Marginalized zero-inflated bell regression models for overdispersed count data. *Journal of Statistical Theory and Practice*, **19**, 17.

Bar-Lev, S. K. and Ridder, A. (2021). Exponential dispersion models for overdispersed zero-inflated count data. *Communications in Statistics-Simulation and Computation*, **52**, 3286–3304.

Bardwell, G. E. and Crow, E. L. (1964). A two parameter family of hyper-Poisson distributions. *Journal of the American Statistical Association*, **59**, 133–141.

Bourguignon, M., Gallardo, D. I., and Medeiros, R. M. R. (2022). A simple and useful regression model for underdispersed count data based on Bernoulli–Poisson convolution. *Statistical Papers*, **63**, 821–848.

Bourguignon, M. and Weiß, C. H. (2017). An INAR (1) process for modeling count time series with equidispersion, underdispersion and overdispersion. *Test*, **26**, 847–868.

Chakraborty, S. (2010). On some distributional properties of the family of weighted generalized poisson distribution. *Communications in Statistics—Theory and Methods*, **39**, 2767–2788.

Chakraborty, S. and Bhati, D. (2016). Transmuted geometric distribution with applications in modeling and regression analysis of count data. *SORT-Statistics and Operations Research Transactions*, **40**, 153–176.

Chakraborty, S. and Gupta, R. D. (2015). Exponentiated geometric distribution: another generalization of geometric distribution. *Communications in Statistics-Theory and Methods*, **44**, 1143–1157.

Chakraborty, S. and Ong, S. H. (2017). Mittag-leffler function distribution-a new generalization of hyper-Poisson distribution. *Journal of Statistical Distributions and Applications*, **4**, 1–17.

DeHart, T., Tennen, H., Armeli, S., Todd, M., and Affleck, G. (2008). Drinking to regulate negative romantic relationship interactions: The moderating role of self-esteem. *Journal of Experimental Social Psychology*, **44**, 527–538.

Del Castillo, J. and Pérez-Casany, M. (1998). Weighted poisson distributions for overdispersion and underdispersion situations. *Annals of the Institute of Statistical Mathematics*, **50**, 567–585.

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 (Methodological)*, **39**, 1–22.

Efron, B. (1986). Double exponential-families and their use in generalized linear-regression. *Journal of the American Statistical Association*, **81**, 709–721.

El-Morshedy, M., Eliwa, M. S., and Altun, E. (2020). Discrete Burr-Hatke distribution with properties, estimation methods and regression model. *IEEE Access*, **8**, 74359–74370.

Ghahramani, M. and White, S. S. (2020). Time series regression for zero-inflated and overdispersed count data: A functional response model approach. *Journal of Statistical Theory and Practice*, **14**, 29.

Gómez-Déniz, E. (2010). Another generalization of the geometric distribution. *Test*, **19**, 399–415.

Hahn, E. D. (2022). The tilted beta-binomial distribution in overdispersed data: Maximum likelihood and Bayesian estimation. *Journal of Statistical Theory and Practice*, **16**, 43.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K., and Ostrowski, E. (1995). A handbook of small data sets. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **158**, 339.

Hassanzadeh, F. and Kazemi, I. (2016). Analysis of over-dispersed count data with extra zeros using the Poisson log-skew-normal distribution. *Journal of Statistical Computation and Simulation*, **86**, 2644–2662.

Hoel, P. G. (1943). On indices of dispersion. *The Annals of Mathematical Statistics*, **14**, 155–162.

Hussain, T. and Ahmad, M. (2014). Discrete inverse Rayleigh distribution. *Pakistan Journal of Statistics*, **30**, 203–2022.

Jain, G. C. and Consul, P. C. (1971). A generalized negative binomial distribution. *SIAM Journal on Applied Mathematics*, **21**, 501–513.

Jazi, M. A., Lai, C. D., and Alamatsaz, M. H. (2010). A discrete inverse Weibull distribution and estimation of its parameters. *Statistical Methodology*, **7**, 121–132.

Krishna, H. and Pundir, P. S. (2009). Discrete Burr and discrete Pareto distributions. *Statistical Methodology*, **6**, 177–188.

Makcutek, J. (2008). A generalization of the geometric distribution and its application in quantitative linguistics. *Romanian Reports in Physics*, **60**, 501–509.

Moghimbeigi, A., Eshraghian, M. R., Mohammad, K., and Mcardle, B. (2008). Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. *Journal of Applied Statistics*, **35**, 1193–1202.

Moqaddasi Amiri, M., Tapak, L., and Faradmal, J. (2019). A mixed-effects least square support vector regression model for three-level count data. *Journal of Statistical Computation and Simulation*, **89**, 2801–2812.

Nandi, A., Biswas, A., Hazarika, P. J., and Das, J. (2024a). A new regression model for over-dispersed count responses based on Poisson and geometric convolution. *Journal of the Indian Society for Probability and Statistics*, **Online first**, 1–18.

Nandi, A., Chakraborty, S., and Biswas, A. (2024b). A new over-dispersed count model based on poisson-geometric convolution. *Communications in Statistics - Simulation and Computation*, **Online first**, 1–26.

Nekoukhou, V., Alamatsaz, M. H., and Bidram, H. (2012). A discrete analogue of the generalized exponential distribution. *Communications in Statistics - Theory and Methods*, **41**, 2000–2013.

Para, B. A. and Jan, T. R. (2016a). Discrete version of log-logistic distribution and its applications in genetics. *International Journal of Modern Mathematical Sciences*, **14**, 407–422.

Para, B. A. and Jan, T. R. (2016b). On discrete three-parameter Burr type XII and discrete Lomax distributions and their applications to model count data from medical science. *Biometrics and Biostatistics International Journal*, **4**, 1–15.

Philippou, A. N., Georghiou, C., and Philippou, G. N. (1983). A generalized geometric distribution and some of its properties. *Statistics and Probability Letters*, **1**, 171–175.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, **26**, 195–239.

Rodrigues-Motta, M., Pinheiro, H. P., Martins, E. G., Araújo, M. S., and dos Reis, S. F. (2013). Multivariate models for correlated count data. *Journal of Applied Statistics*, **40**, 1586–1596.

Roy, D. (2004). Discrete rayleigh distribution. *IEEE Transactions on Reliability*, **53**, 255–260.

Sarvi, F., Moghimbeigi, A., and Mahjub, H. (2019). GEE-based zero-inflated generalized Poisson model for clustered over or under-dispersed count data. *Journal of Statistical Computation and Simulation*, **89**, 2711–2732.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610.

Sellers, K. F. and Shmueli, G. (2010). A flexible regression model for count data. *The Annals of Applied Statistics*, **4**, 943–961.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379–423.

Singh, B., Singh, R. P., Nayal, A. S., and Tyagi, A. (2022). Discrete inverted Nadarajah-Haghighi distribution: Properties and classical estimation with application to complete and censored data. *Statistics, Optimization & Information Computing*, **10**, 1293–1313.

Tapak, L., Hamidi, O., Amini, P., and Verbeke, G. (2020). Random effect exponentiated-exponential geometric model for clustered/longitudinal zero-inflated count data. *Journal of Applied Statistics*, **47**, 2272–2288.

Tripathi, R. C., Gupta, R. C., and White, T. J. (1987). Some generalizations of the geometric distribution. *Sankhya Ser. B*, **49**, 218–223.

Tüzen, F., Erbaş, S., and Olmuş, H. (2020). A simulation study for count data models under varying degrees of outliers and zeros. *Communications in Statistics-Simulation and Computation*, **49**, 1078–1088.

Wang, S., Cadigan, N. G., and Benoît, H. P. (2017). Inference about regression parameters using highly stratified survey count data with over-dispersion and repeated measurements. *Journal of Applied Statistics*, **44**, 1013–1030.

Wongrin, W. and Bodhisuwan, W. (2017). Generalized Poisson–Lindley linear model for count data. *Journal of Applied Statistics*, **44**, 2659–2671.

Zhan, D. and Young, D. S. (2024). Finite mixtures of mean-parameterized Conway–Maxwell–Poisson regressions. *Journal of Statistical Theory and Practice*, **65**, 1469–1492.

Appendix I

The log-likelihood functions of the $PoiG(\theta, q)$ distribution is

$$l'(\theta, q; \boldsymbol{y}) = n\log(1-q) + n\overline{y}\log q + \frac{n\theta(1-q)}{q} + \sum_{i=0}^{n}\log\left(\frac{\Gamma\left(y_i+1, \frac{\theta}{q}\right)}{\Gamma(y_i+1)}\right). \tag{36}$$

The score functions are

$$s_1'(\theta, q; \boldsymbol{y}) = \frac{\partial}{\partial\theta} l'(\theta, q; \boldsymbol{y}) = \frac{n(1-q)}{q} - \sum_{i=1}^{n} \alpha_1(y_i)\beta^{y_i},$$

$$\text{and} \quad s_2'(\theta, q; \boldsymbol{y}) = \frac{\partial}{\partial q} l'(\theta, q; \boldsymbol{y}) = -\frac{n}{1-q} - \frac{n(\theta - \bar{y})}{q} - \frac{n\theta(1-q)}{q^2} + \sum_{i=1}^{n} \theta\alpha_2(y_i)\beta^{y_i}.$$

In the above expressions, $\beta = \dfrac{\theta}{q}$ and

$$\alpha_j(y_i) = \frac{e^{-\beta}}{\Gamma(y_i + 1, \beta)} \frac{1}{q^j} \text{ for } i = 1, 2, ..., n, j = 1, 2.$$

The second order partial derivatives of the log-likelihood function of the $PoiG(\theta, q)$ distribution (Nandi, *et al.* (2024b)) are

$$\frac{\partial^2 l'(\theta, q; \boldsymbol{y})}{\partial\theta^2} = \sum_{i=1}^{n} \left[(\beta^{y_i} - y_i\beta^{y_i - 1})\alpha_2(y_i) - \beta^{2y_i}\alpha_1(y_i)^2 \right],$$

$$\frac{\partial^2 l'(\theta, q; \boldsymbol{y})}{\partial\theta\partial q} = -\frac{n}{q^2} - \sum_{i=1}^{n} \left[\theta(\beta^{y_i} - y_i\beta^{y_i - 1})\alpha_3(y_i) - \beta^{y_i}\alpha_2(y_i) - \frac{\theta}{q}\beta^{2y_i}\alpha_1(y_i)^2 \right],$$

$$\text{and} \quad \frac{\partial^2 l'(\theta, q; \boldsymbol{y})}{\partial q^2} = \frac{2n\theta - n\bar{y}q}{q^3} - \frac{n}{(1-q)^2} + \sum_{i=1}^{n} \left[((\theta^2 - 2\theta q)\beta^{y_i} - \theta^2 y_i\beta^{y_i - 1})\alpha_4(y_i) - \theta^2\beta^{2y_i}\alpha_2(y_i)^2 \right].$$

The above listed second order partial derivatives of the log-likelihood functions for the PoiG distribution (Nandi, *et al.*, 2024b) can be utilized to derive the expression for the 2^{nd} partial derivatives of the log-likelihood function for PoiTG distribution,

$$l(\theta, q, \alpha; \boldsymbol{y}) = \sum_{i=1}^{n} \log L(\theta, q, \alpha; y_i) = \sum_{i=1}^{n} \log((1-\alpha)P(\theta, q; y_i) + \alpha P(\theta, q_*; y_i)).$$

Here, we use q^* to denote q^2. For each $i = 1, 2, ..., n$, let us introduce the following notations for further use.

$$A_{i1} = (1-\alpha)L'(\theta, q; y_i)\frac{\partial l'(\theta, q; y_i)}{\partial\theta} + \alpha L'(\theta, q_*; y_i)\frac{\partial l'(\theta, q_*; y_i)}{\partial\theta},$$

$$A_{i2} = (1-\alpha)L'(\theta, q; y_i)\left(\left(\frac{\partial l'(\theta, q; y_i)}{\partial\theta}\right)^2 + \frac{\partial^2 l'(\theta, q; y_i)}{\partial\theta^2}\right),$$

$$A_{i3} = \alpha L'(\theta, q_*; y_i)\left(\left(\frac{\partial l'(\theta, q_*; y_i)}{\partial\theta}\right)^2 + \frac{\partial^2 l'(\theta, q_*; y_i)}{\partial\theta^2}\right),$$

$$B_{i1} = (1-\alpha)L'(\theta, q; y_i)\frac{\partial l'(\theta, q; y_i)}{\partial q} + 2\alpha q L'(\theta, q_*; y_i)\frac{\partial l'(\theta, q_*; y_i)}{\partial q^*},$$

$$B_{i2} = (1-\alpha)L'(\theta, q; y_i)\left(\left(\frac{\partial l'(\theta, q; y_i)}{\partial q}\right)^2 + \frac{\partial^2 l'(\theta, q; y_i)}{\partial q^2}\right),$$

$$B_{i3} = 2\alpha L'(\theta, q_*; y_i)\left(\frac{\partial l'(\theta, q_*; y_i)}{\partial q_*} + 2q^2\left(\frac{\partial l'(\theta, q_*; y_i)}{\partial q_*}\right)^2 + 2q^2\frac{\partial^2 l'(\theta, q_*; y_i)}{\partial q_*^2}\right),$$

$$C_{i1} = (1-\alpha)L'(\theta, q; y_i)\left(\frac{\partial l'(\theta, q; y_i}{\partial \theta}\frac{\partial l'(\theta, q; y_i)}{\partial q} + \frac{\partial^2 l'(\theta, q; y_i)}{\partial \theta \partial q}\right),$$

$$C_{i2} = 2\alpha q L'(\theta, q_*; y_i)\left(\frac{\partial l'(\theta, q_*; y_i)}{\partial \theta}\frac{\partial l'(\theta, q_*; y_i)}{\partial q_*} + \frac{\partial^2 l'(\theta, q_*; y_i)}{\partial \theta \partial q_*}\right),$$

and $$C_{i3} = 2qL'(\theta, q_*; y_i)\frac{\partial l'(\theta, q_*; y_i)}{\partial q_*} - L'(\theta, q; y_i)\frac{\partial l'(\theta, q; y_i)}{\partial q}.$$

The second order partial derivatives of the log-likelihood function for $PoiTG(\theta, q, \alpha)$ are

$$\frac{\partial^2 l(\theta, q, \alpha; \boldsymbol{y})}{\partial \theta^2} = \sum_{i=1}^{n}\frac{L(\theta, q, \alpha; y_i)(A_{i2} + A_{i3}) - A_{i1}^2}{L(\theta, q, \alpha; y_i)^2}, \tag{37}$$

$$\frac{\partial^2 l(\theta, q, \alpha; \boldsymbol{y})}{\partial q^2} = \sum_{i=1}^{n}\frac{L(\theta, q, \alpha; y_i)(B_{i2} + B_{i3}) - B_{i1}^2}{L(\theta, q, \alpha; y_i)^2}, \tag{38}$$

$$\frac{\partial^2 l(\theta, q, \alpha; \boldsymbol{y})}{\partial \theta \partial q} = \sum_{i=1}^{n}\frac{L(\theta, q, \alpha y_i)(C_{i1} + C_{i2}) - A_{i1}B_{i1}}{L(\theta, q, \alpha; y_i)^2}, \tag{39}$$

$$\frac{\partial^2 l(\theta, q, \alpha; \boldsymbol{y})}{\partial \alpha^2} = -\sum_{i=1}^{n}\frac{(L'(\theta, q_*; y_i) - L'(\theta, q; y_i))^2}{L(\theta, q, \alpha;)^2}, \tag{40}$$

$$\frac{\partial^2 l(\theta, q, \alpha; \boldsymbol{y})}{\partial \theta \partial \alpha} = \sum_{i=1}^{n}\frac{\left(L'(\theta, q_*; y_i)\frac{\partial l'(\theta, q_*; y_i)}{\partial \theta} - L'(\theta, q; y_i)\frac{\partial l'(\theta, q; y_i)}{\partial \theta}\right)}{L(\theta, q, \alpha; y_i)^2}, \tag{41}$$

and $$\frac{\partial^2 l(\theta, q, \alpha; \boldsymbol{y})}{\partial q \partial \alpha} = \sum_{i=1}^{n}\frac{L(\theta, q, \alpha; y_i)C_{i3}}{L(\theta, q, \alpha; y_i)^2} - \sum_{i=1}^{n}\frac{(L'(\theta, q_*; y_i) - L'(\theta, q; y_i))\, B_{i1}}{L(\theta, q, \alpha; y_i)^2}. \tag{42}$$

The resulting fisher's information matrix for θ, q, and α is

$$I_Y(\theta, q, \alpha) = \begin{pmatrix} -E\left(\frac{\partial^2 l(\theta, q, \alpha; \boldsymbol{y})}{\partial \theta^2}\right) & -E\left(\frac{\partial^2 l(\theta, q, \alpha; \boldsymbol{y})}{\partial \theta \partial q}\right) & -E\left(\frac{\partial^2 l(\theta, q, \alpha; \boldsymbol{y})}{\partial \theta \partial \alpha}\right) \\ -E\left(\frac{\partial^2 l(\theta, q, \alpha; \boldsymbol{y})}{\partial \theta \partial q}\right) & -E\left(\frac{\partial^2 l(\theta, q, \alpha; \boldsymbol{y})}{\partial q^2}\right) & -E\left(\frac{\partial^2 l(\theta, q, \alpha; \boldsymbol{y})}{\partial q \partial \alpha}\right) \\ -E\left(\frac{\partial^2 l(\theta, q, \alpha; \boldsymbol{y})}{\partial \theta \partial \alpha}\right) & -E\left(\frac{\partial^2 l(\theta, q, \alpha; \boldsymbol{y})}{\partial q \partial \alpha}\right) & -E\left(\frac{\partial^2 l(\theta, q, \alpha; \boldsymbol{y})}{\partial \alpha^2}\right) \end{pmatrix}.$$

This can be approximated by

$$\widehat{I}_Y(\theta,q,\alpha) \approx \begin{pmatrix} -\frac{\partial^2 l(\theta,q,\alpha;\boldsymbol{y})}{\partial\theta^2} & -\frac{\partial^2 l(\theta,q,\alpha;\boldsymbol{y})}{\partial\theta\partial q} & -\frac{\partial^2 l(\theta,q,\alpha;\boldsymbol{y})}{\partial\theta\partial\alpha} \\ -\frac{\partial^2 l(\theta,q,\alpha;\boldsymbol{y})}{\partial\theta\partial q} & -\frac{\partial^2 l(\theta,q,\alpha;\boldsymbol{y})}{\partial q^2} & -\frac{\partial^2 l(\theta,q,\alpha;\boldsymbol{y})}{\partial q\partial\alpha} \\ -\frac{\partial^2 l(\theta,q,\alpha;\boldsymbol{y})}{\partial\theta\partial\alpha} & -\frac{\partial^2 l(\theta,q,\alpha;\boldsymbol{y})}{\partial q\partial\alpha} & -\frac{\partial^2 l(\theta,q,\alpha;\boldsymbol{y})}{\partial\alpha^2} \end{pmatrix}_{(\theta,q,\alpha)=(\hat{\theta}_{ML},\hat{q}_{ML},\hat{\alpha}_{ML})}.$$

The MLEs $\hat{\theta}_{ML}$, $\hat{q}_{ML}$, and $\hat{\alpha}_{ML}$ are consistent and asymptotically normal with the mean vector (θ, q, α) and the dispersion matrix $\hat{I}_Y^{-1} = [d_{ij}]_{3\times 3}$. Let z_ν denote the $(1-\nu)$-th quantile of the standard normal distribution. The $(1-\nu)\times 100\%$ confidence interval for the parameters θ, q and α are given by

$$\left(\hat{\theta}_{ML} - z_{\nu/2}\sqrt{d_{11}}\ ,\ \hat{\theta}_{ML} + z_{\nu/2}\sqrt{d_{11}}\right), \left(\hat{q}_{ML} - z_{\nu/2}\sqrt{d_{22}}\ ,\ \hat{q}_{ML} + z_{\nu/2}\sqrt{d_{22}}\right)$$

and

$$\left(\hat{\alpha}_{ML} - z_{\nu/2}\sqrt{d_{33}}\ ,\ \hat{\alpha}_{ML} + z_{\nu/2}\sqrt{d_{33}}\right), \text{ respectively.}$$

Table 5: Simulated bias and mean squared error of the estimators

Bias (Unshaded region) and MSE (Shaded region)									
θ	q	α	n	$\hat{\theta}_{ML}$	$\hat{q}_{ML}$	$\hat{\alpha}_{ML}$	$\hat{\theta}_{EM}$	$\hat{q}_{EM}$	$\hat{\alpha}_{EM}$
0.5	0.2	0.2	30	-0.1791	0.1034	-0.0222	-0.1141	0.1105	0.2673
				0.0780	0.0240	0.0726	0.0611	0.0270	0.0737
			50	-0.1324	0.0847	0.0021	-0.0705	0.0881	0.2753
				0.0529	0.0191	0.0893	0.0385	0.0202	0.0772
			100	-0.0808	0.0641	0.0357	-0.0258	0.0624	0.2836
				0.0296	0.0139	0.1092	0.0204	0.0130	0.0808
			250	-0.0359	0.0465	0.0798	0.0103	0.0376	0.2886
				0.0140	0.0103	0.1353	0.0101	0.0077	0.0834
			500	-0.0143	0.0370	0.1037	0.0273	0.0247	0.2905
				0.0085	0.0082	0.1496	0.0069	0.0052	0.0845
			1000	-0.0003	0.0327	0.1266	0.0376	0.0175	0.2916
				0.0059	0.0071	0.1611	0.0054	0.0037	0.0851
0.5	0.2	0.8	30	-0.2459	0.0732	-0.6155	-0.1894	0.0789	-0.3301
				0.0955	0.0176	0.4509	0.0733	0.0198	0.1107
			50	-0.2033	0.0524	-0.6031	-0.1480	0.0548	-0.3232
				0.0697	0.0139	0.4501	0.0487	0.0145	0.1055
			100	-0.1618	0.0350	-0.5610	-0.1131	0.0308	-0.3155

				0.0444	0.0101	0.4265	0.0279	0.0089	0.0999
			250	-0.1082	0.0110	-0.5106	-0.0695	-0.0013	-0.3095
				0.0211	0.0076	0.3947	0.0120	0.0052	0.0959
			500	-0.0828	0.0005	-0.4653	-0.0501	-0.0198	-0.3073
				0.0129	0.0069	0.3675	0.0070	0.0043	0.0945
			1000	-0.0609	-0.0096	-0.4260	-0.0331	-0.0359	-0.3057
				0.0078	0.0064	0.3425	0.0040	0.0044	0.0935
0.5	0.8	0.2	30	0.1161	-0.0112	0.0396	0.1821	0.0055	0.2266
				0.2452	0.0033	0.0735	0.2834	0.0026	0.0616
			50	0.0583	-0.0025	0.0524	0.1268	0.0133	0.232
				0.1462	0.0018	0.0796	0.1685	0.0013	0.0612
			100	0.0134	0.0033	0.0628	0.0863	0.0187	0.2424
				0.0748	0.0012	0.0823	0.0845	0.0004	0.0616
			250	-0.0075	0.0050	0.0513	0.0678	0.0219	0.2479
				0.0313	0.0008	0.0721	0.0346	0.0007	0.0622
			500	-0.0070	0.0049	0.0469	0.0648	0.0226	0.2496
				0.0175	0.0006	0.0609	0.0194	0.0006	0.0624
			1000	-0.0104	0.0032	0.0257	0.0584	0.0231	0.2495
				0.0095	0.0004	0.0411	0.0108	0.0006	0.0623
0.5	0.8	0.8	30	0.0894	-0.0992	-0.5173	0.1509	-0.0840	-0.3562
				0.1982	0.0187	0.3517	0.2322	0.0148	0.1386
			50	0.0374	-0.0809	-0.4789	0.0972	-0.0689	-0.3494
				0.1162	0.0114	0.3228	0.1314	0.0081	0.1302
			100	0.0045	-0.0655	-0.4264	0.0619	-0.0592	-0.3414
				0.0569	0.0072	0.2816	0.0614	0.0048	0.1199
			250	0.0014	-0.0516	-0.3507	0.0499	-0.0548	-0.3389
				0.0250	0.0048	0.2184	0.0258	0.0035	0.1156
			500	0.0009	-0.0435	-0.2985	0.0417	-0.0534	-0.3395
				0.0124	0.0036	0.1733	0.0127	0.0031	0.1155
			1000	0.0081	-0.0358	-0.2422	0.0404	-0.0530	-0.3403
				0.0068	0.0027	0.1255	0.0072	0.0029	0.1159
1.5	0.2	0.2	30	-0.3463	0.1663	-0.0455	-0.2485	0.1748	0.2729
				0.2543	0.0456	0.0686	0.1904	0.0506	0.0772
			50	-0.2760	0.1476	-0.0160	-0.1866	0.1532	0.2824
				0.1621	0.0366	0.0840	0.1144	0.0394	0.0809
			100	-0.1969	0.1227	0.0231	-0.1193	0.1220	0.2891
				0.0901	0.0278	0.1038	0.0589	0.0274	0.0838
			250	-0.1208	0.0932	0.0590	-0.0554	0.0851	0.2929
				0.0412	0.0191	0.1263	0.0252	0.0161	0.0858
			500	-0.0809	0.0762	0.0854	-0.0223	0.0626	0.2944
				0.0233	0.0149	0.1410	0.0138	0.0108	0.0867

			1000	-0.0459	0.0605	0.1089	0.0080	0.0412	0.2955
				0.0134	0.0118	0.1547	0.0087	0.0072	0.0873
1.5	0.2	0.8	30	-0.4406	0.1534	-0.6475	-0.3446	0.1611	-0.3263
				0.3136	0.0407	0.4862	0.2331	0.0450	0.1090
			50	-0.3688	0.1326	-0.6151	-0.2831	0.1368	-0.3171
				0.2141	0.0322	0.4623	0.1518	0.0343	0.1015
			100	-0.2953	0.1057	-0.5853	-0.2207	0.1057	-0.3105
				0.1335	0.0235	0.4427	0.0890	0.0232	0.0966
			250	-0.2164	0.0748	-0.5409	-0.1544	0.0651	-0.3066
				0.0701	0.0157	0.4167	0.0430	0.0127	0.0940
			500	-0.1711	0.0554	-0.5070	-0.1169	0.0386	-0.3050
				0.0438	0.0121	0.3942	0.0253	0.0080	0.0930
			1000	-0.1326	0.0380	-0.4703	-0.0846	0.0128	-0.3038
				0.0268	0.0097	0.3699	0.0143	0.0052	0.0923
1.5	0.8	0.2	30	0.1059	-0.0143	0.0282	0.1997	0.0014	0.2211
				0.4456	0.0044	0.0739	0.4877	0.0035	0.0629
			50	0.0477	-0.0044	0.0398	0.1439	0.0108	0.2279
				0.2499	0.0021	0.0786	0.2696	0.0016	0.0622
			100	0.0121	0.0029	0.0584	0.1106	0.0173	0.2415
				0.1283	0.0013	0.0847	0.1361	0.0009	0.0633
			250	-0.0128	0.0050	0.0500	0.0888	0.0210	0.2511
				0.0529	0.0008	0.0748	0.0559	0.0007	0.0642
			500	-0.0111	0.0046	0.0423	0.0877	0.0217	0.2535
				0.0293	0.0006	0.0623	0.0319	0.0006	0.0645
			1000	-0.0129	0.0034	0.0276	0.0821	0.0224	0.2541
				0.0163	0.0004	0.0453	0.0187	0.0006	0.0646
1.5	0.8	0.8	30	0.0615	-0.1017	-0.5276	0.1475	-0.0875	-0.3514
				0.3552	0.0210	0.3636	0.3868	0.0175	0.1350
			50	0.0292	-0.0844	-0.4853	0.1106	-0.0731	-0.3420
				0.2117	0.0136	0.3322	0.2293	0.0103	0.1255
			100	0.0009	-0.0664	-0.4291	0.0763	-0.0607	-0.3333
				0.1049	0.0077	0.2878	0.1086	0.0054	0.1147
			250	-0.0105	-0.0515	-0.3545	0.0537	-0.0547	-0.3309
				0.0426	0.0049	0.2269	0.0418	0.0036	0.1105
			500	-0.0067	-0.0420	-0.2935	0.0450	-0.0527	-0.3315
				0.0229	0.0035	0.1736	0.0215	0.0031	0.1103
			1000	-0.0005	-0.0340	-0.2342	0.0398	-0.0521	-0.3331
				0.0117	0.0025	0.1242	0.0111	0.0028	0.1111

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 24, No. 1, 2026 (New Series), pp 221–236
http://www.ssca.org.in/journal

Nonparametric Estimation and Analysis of Conditional Dynamic Failure Extropy in Bivariate Systems

Lekshmi Krishnan C. U. and E. I. Abdul Sathar

Department of Statistics, University of Kerala, Thiruvananthapuram-695 581, India.

Received: 05 February 2025; Revised: 28 July 2025; Accepted: 25 September 2025

Abstract

This study introduces Conditional Dynamic Failure Extropy ($CDFE_X$), a novel measure for quantifying uncertainty in bivariate systems. $CDFE_X$ captures interdependence between components via the joint distribution**.** We develop nonparametric estimators for $CDFE_X$ and establish their asymptotic properties under mild conditions. Through simulations and real-world datasets, we demonstrate the robustness of the proposed estimators. Our results show that $CDFE_X$ outperforms both univariate Dynamic Failure Extropy (DFE_X) and Conditional Dynamic Cumulative Past Entropy, highlighting its potential to enhance reliability analysis in complex systems.

Key words: Conditional dynamic failure extropy; Bivariate reversed hazard rate; Bivariate mean inactivity time.

AMS Subject Classifications: 94A17, 62G07

1. Introduction

Extropy complements Shannon's entropy (Shannon, 1948), serving to quantify uncertainty in probability distributions. Differential extropy (Lad, 2015) extends this concept to continuous variables, enabling analysis of time-to-event data and lifetimes in continuous domains. In fields such as information theory, statistics, and reliability, it offers a valuable alternative for measuring uncertainty.

Extropy has evolved into a dynamic concept with applications in order statistics and record values (Qiu, 2017), and has been extended to residual lifetime analysis (Qiu and Jia, 2018). Estimators and goodness-of-fit tests Qiu and Jia (2017) have been developed, and bounds based on variational distance have also been explored (Yang *et al.*, 2019). Its adaptability has enabled broad use across reliability, risk assessment, and information theory, including in mixed systems (Qiu *et al.*, 2018). Continued research by Raqab and Qiu (2019), Noughabi and Jarrahiferiz (2019), and Krishnan *et al.* (2020) emphasizes the versatility and ongoing relevance of extropy in modern statistical analysis.

To extend extropy to multivariate systems, recent studies have introduced bivari-

Corresponding Author: E. I. Abdul Sathar
Email: sathare@gmail.com

ate survival extropy (Krishnan and Sathar, 2022) and shift-dependent bivariate weighted survival extropy (Sathar and Krishnan, 2023), which capture interdependent behavior in two-component systems. These measures have found applications in reliability engineering, survival analysis, and forensic science, particularly for analyzing time-since-failure scenarios. By quantifying uncertainty in multi-component settings, such tools support decision-making in maintenance and system upgrades.

The remainder of the paper is structured as follows: Section 2 introduces conditional dynamic failure extropy and explores its main properties, including monotonicity and characterization results. Section 3 discusses nonparametric estimation using empirical and kernel methods, investigates their asymptotic behavior, and presents simulation and data-based illustrations.

2. Conditional dynamic failure extropy

In bivariate systems, marginal distributions alone are insufficient to capture joint dependence unless the variables are independent. When conditional distributions are known, both marginal and conditional components are essential for understanding the joint behavior. The following definition provides a conditional extension of bivariate failure extropy, based on the approach of Kayal (2021).

Definition 1: Let $X = (X_1, X_2)$ be a bivariate random vector with joint distribution function F. Define $Y_i = (X_i \mid X_1 < t_1, X_2 < t_2)$, $i = 1, 2$, representing the conditional distribution of X_i given both components fail within intervals $(0, t_1)$ and $(0, t_2)$, respectively. Then the distribution functions of Y_1 and Y_2 are

$$P(Y_1 \leq y_1) = \frac{F(y_1, t_2)}{F(t_1, t_2)}, \quad 0 < y_1 < t_1,$$

$$P(Y_2 \leq y_2) = \frac{F(t_1, y_2)}{F(t_1, t_2)}, \quad 0 < y_2 < t_2.$$

These conditional distributions capture the failure behavior of each component, given that both components have failed within the specified time window. The corresponding conditional dynamic failure extropy measures are

$$J_1^F(X : t_1, t_2) = -\frac{1}{2} \int_0^{t_1} \left[\frac{F(x_1, t_2)}{F(t_1, t_2)} \right]^2 dx_1, \tag{1}$$

$$J_2^F(X : t_1, t_2) = -\frac{1}{2} \int_0^{t_2} \left[\frac{F(t_1, x_2)}{F(t_1, t_2)} \right]^2 dx_2. \tag{2}$$

$CDFE_X$ quantifies the dispersion of uncertainty in a component's conditional lifetime, given joint system failure before thresholds t_1 and t_2. A value closer to zero indicates lower concentration (i.e., a broader spread of failure times), whereas a more negative value reflects greater concentration, meaning failures are more tightly clustered within the conditional window. Thus, $CDFE_X$ captures residual uncertainty under joint failure constraints, supporting applications in reliability modeling, risk assessment, and preventive maintenance planning.

Table 1: Expressions of conditional dynamic failure extropy ($CDFE_X$) for some well-known bivariate distributions

Bivariate Distribution $F(t_1,t_2)$	**Conditional Dynamic Failure Extropy** $J_i^F(X:t_1,t_2)$
$F(t_1,t_2)=t_i^{1+\theta\log t_j}\cdot t_j,\quad 0\le t_i,t_j\le 1$	$J_i^F=\dfrac{-t_i}{2\,(2\theta\log t_j+3)}$
$F(t_1,t_2)=\exp\left(2-\dfrac{1}{t_i}-\dfrac{1}{t_j}\right)$	$J_i^F=\dfrac{-t_i^2}{4},\quad i,j=1,2,\ i\neq j$
$F(t_1,t_2)=\dfrac{t_it_j}{bd},\quad 0\le t_i\le b,\ 0\le t_j\le d$	$J_i^F=\dfrac{-t_i}{6},\quad i,j=1,2,\ i\neq j$
$F(t_1,t_2)=\left(\dfrac{t_i}{b_i}\right)^{c_i}\left(\dfrac{t_j}{b_j}\right)^{c_j+\theta\log\left(\frac{t_i}{b_i}\right)}$ $\theta\le 0,\ 0\le t_i\le b_i,\ 0\le t_j\le b_j$	$J_i^F=\dfrac{-t_i}{2\left[2c_i+2\theta\log\left(\dfrac{t_j}{b_j}\right)+1\right]}$
$F(x_1,x_2)=\left(1+e^{-x_1}+e^{-x_2}\right)^{-1},$ $-\infty<x_1,x_2<\infty$	$J_i^F(X:t_1,t_2)=\dfrac{-1}{2(1+e^{t_j})^2}\,e^{2t_j}\left(1+e^{-t_i}+e^{-t_j}\right)^2$ $\times\left[\dfrac{-1}{2+e^{-t_j}}+\dfrac{1}{1+e^{t_i}(1+e^{t_j})}\right.$ $\left.-\log(1+2e^{t_j})+\log\left(e^{t_i}e^{t_j}(1+e^{t_i})\right)\right],$ for $i,j=1,2,\ i\neq j$

Example 1: To illustrate the physical meaning of $CDFE_X$, we analyze failure times (in months) for a two-motor parallel system (Table 2). Assuming a bivariate normal distribution fitted to the data, we compute the conditional dynamic failure extropies $J_1^F(t_1,t_2)$ and $J_2^F(t_1,t_2)$. Figure 1 presents a 3D overlay J_1^F for Motor A (black) and J_2^F for Motor B (white). The plot shows that $|J_2^F|$ is consistently greater than $|J_1^F|$, indicating that, conditional on joint failure before (t_1,t_2), Motor B's failure times are more concentrated, while Motor A shows greater variability. This supports the interpretation of $CDFE_X$ as a measure of conditional uncertainty, offering valuable guidance for reliability planning and maintenance prioritization.

Table 2: Failure times (in months) for motors A and B in 11 parallel systems

System	1	2	3	4	5	6	7	8	9	10	11
Motor A	3.4	2.8	2.93	5.2	4.93	4.63	8.16	7.83	7.33	6.9	8.33
Motor B	2.16	4.93	6.73	4.03	4.1	5.0	5.2	5.73	7.33	7.13	7.06

The following theorem establishes an upper bound for $CDFE_X$.

Theorem 1: Let $X=(X_1,X_2)$ be a bivariate random vector with joint distribution function F, and let $t_1,t_2>0$. Consider the event $A=\{X_1<t_1,X_2<t_2\}$, representing the joint failure of both components before times t_1 and t_2. Then, the conditional dynamic failure extropy of each component satisfies

$$J_i^F(X;t_1,t_2)\le m_i^X(t_1,t_2),\quad i=1,2,$$

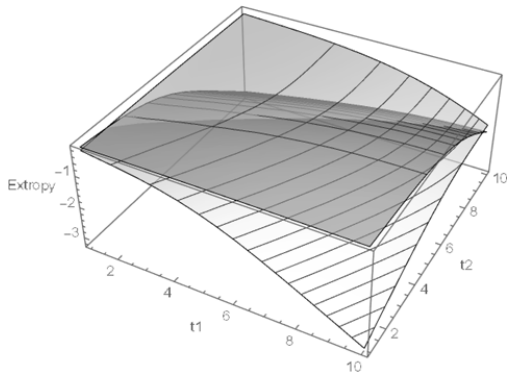


Figure 1: Overlay of conditional dynamic failure extropies J_1^F (Motor A) and J_2^F (Motor B) based on bivariate normal distribution

where the upper bounds $m_i^X(t_1, t_2)$ denote the conditional expected inactivity times

$$m_1^X(t_1, t_2) = \int_0^{t_1} \frac{F(x_1, t_2)}{F(t_1, t_2)}\, dx_1, \quad m_2^X(t_1, t_2) = \int_0^{t_2} \frac{F(t_1, x_2)}{F(t_1, t_2)}\, dx_2.$$

These correspond to the conditional mean waiting times for each component, given joint failure before t_1 and t_2 (Nair and Gopalakrishnan (2008)).

Proof: Since $0 < F(t_1, t_2) \leq 1$, we have

$$\left[\frac{F(x_i, t_j)}{F(t_1, t_2)}\right]^2 \leq \frac{F(x_i, t_j)}{F(t_1, t_2)}, \quad i, j = 1, 2,$$

for $i \neq j$. Integrating both sides over $[0, t_i]$ gives the desired inequality. □

The following example illustrates Theorem 1.

Example 2: Let the bivariate random variable $X = (X_1, X_2)$ have joint distribution function $F(t_1, t_2) = t_1 t_2$, where $0 \leq t_1, t_2 \leq 1$. Then, for $i = 1, 2$, the expected conditional waiting time is $m_i^X(t_1, t_2) = \frac{t_i}{2}$, and the conditional dynamic failure extropy is $J_i^F(X : t_1, t_2) = -\frac{t_i}{6}$. Therefore, $J_i^F(X : t_1, t_2) \leq m_i^X(t_1, t_2)$, verifying the inequality in Theorem 1.

The following theorem establishes an upper bound for conditional dynamic failure extropy ($CDFE_X$), relating it to the expected conditional waiting time and cumulative past entropy. Such bounds are instrumental in reliability modeling, offering insights into the reduction of uncertainty achieved through conditioning.

Theorem 2: Let $X = (X_1, X_2)$ be a non-negative bivariate random vector. Then, for $i = 1, 2$,

$$J_i^F(X : t_1, t_2) \leq \frac{1}{2}\left[\bar{\varepsilon}_i^*(X : t_1, t_2) - m_i^X(t_1, t_2)\right], \tag{3}$$

where $\bar{\varepsilon}_i^*(X : t_1, t_2)$ denotes the conditional dynamic cumulative past entropy (CDCPE), given by

$$\bar{\varepsilon}_1^*(X : t_1, t_2) = -\int_0^{t_1} \frac{F(x_1, t_2)}{F(t_1, t_2)} \log \frac{F(x_1, t_2)}{F(t_1, t_2)}\, dx_1, \tag{4}$$

$$\bar{\varepsilon}_2^*(X : t_1, t_2) = -\int_0^{t_2} \frac{F(t_1, x_2)}{F(t_1, t_2)} \log \frac{F(t_1, x_2)}{F(t_1, t_2)}\, dx_2, \tag{5}$$

and $m_i^X(t_1, t_2)$ is the conditional mean waiting time of component X_i, given joint failure before (t_1, t_2).

Proof: Using the inequality $-\log x \geq 1 - x$, we obtain

$$\bar{\varepsilon}_1^*(X : t_1, t_2) \geq \int_0^{t_1} \left[\frac{F(x_1, t_2)}{F(t_1, t_2)} - \left(\frac{F(x_1, t_2)}{F(t_1, t_2)} \right)^2 \right] dx_1,$$

which leads directly to the desired bound in (3). The result for $i = 2$ follows similarly. □

The following theorem examines the behavior of $CDFE_X$ under strictly monotonic, differentiable transformations, as often encountered in reliability and survival analysis through scaling or shifts.

Theorem 3: Let $V = (V_1, V_2)$ with $V_i = \phi_i(X_i)$, where each ϕ_i is non-negative, strictly monotone, differentiable, and absolutely continuous. Then,

$$J_1^F(V : t_1, t_2) = \begin{cases} -\frac{1}{2} \int_0^{\phi_1^{-1}(t_1)} \left[\frac{F(x_1, \phi_2^{-1}(t_2))}{F(\phi_1^{-1}(t_1), \phi_2^{-1}(t_2))} \right]^2 \phi_1'(x_1) dx_1, & \text{if each } \phi_i \text{ is strictly increasing} \\ -\frac{1}{2} \int_{\phi_1^{-1}(t_1)}^{\infty} \left[\frac{\bar{F}(x_1, \phi_2^{-1}(t_2))}{\bar{F}(\phi_1^{-1}(t_1), \phi_2^{-1}(t_2))} \right]^2 \phi_1'(x_1) dx_1 & \text{if each } \phi_i \text{ is strictly decreasing.} \end{cases} \tag{6}$$

The expression for $J_2^F(V : t_1, t_2)$ follows analogously by symmetry.

In particular, for affine transformations of the form $\phi_i(X_i) = c_i X_i + d_i$ with $c_i > 0$, $d_i \geq 0$, we obtain the scaled identity

$$J_1^F(V : t_1, t_2) = c_1\, J_1^F \left(X : \tfrac{t_1 - d_1}{c_1}, \tfrac{t_2 - d_2}{c_2} \right).$$

The connection between $CDFE_X$ and the reversed hazard rate improves interpretability by relating the uncertainty in conditional failure times to the instantaneous recovery rate, as shown in the following theorem.

Theorem 4: Let $J_i^F(X : t_1, t_2)$ denote the conditional dynamic failure extropy, and let the reversed hazard rate be defined by $\bar{h}_i(t_1, t_2) = \frac{\partial}{\partial t_i} \log F(t_1, t_2)$ for $i = 1, 2$. Then,

$$\bar{h}_i(t_1, t_2) = \frac{2 \frac{\partial}{\partial t_i} J_i^F(X : t_1, t_2) + 1}{-4 J_i^F(X : t_1, t_2)}. \tag{7}$$

Proof: From Equations (1) and (2), consider the case $i = 1$

$$\frac{\partial}{\partial t_1} J_1^F(X : t_1, t_2) = \frac{-1}{2} \left[1 - 2\bar{h}_1(t_1, t_2) \int_0^{t_1} \frac{F(x_1, t_2)}{F(t_1, t_2)}\, dx_1 \right]. \tag{8}$$

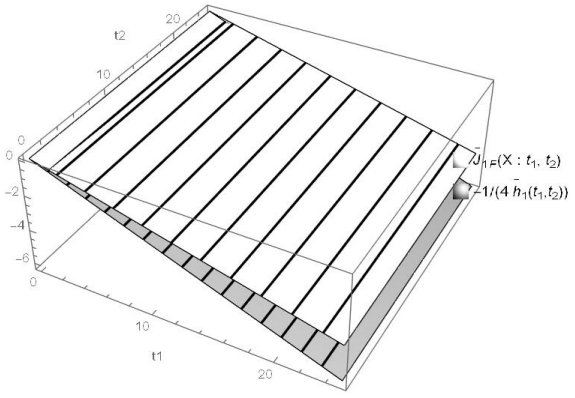


Figure 2: $J_i^F(X : t_1, t_2)$ **with RHS of the inequality** (9)

The result follows directly by solving the above equation for $\bar{h}_1(t_1, t_2)$. The case for $i = 2$ is analogous. □

Nonparametric classification criteria can be developed using the monotonicity properties of conditional dynamic failure extropy ($CDFE_X$). This classification helps characterize the reliability behavior of the system components over time.

Definition 2: A distribution function F is said to be increasing (decreasing) in conditional dynamic failure extropy ($CDFE_X$) if the measure $J_i^F(X : t_1, t_2)$ increases (decreases) monotonically with respect to t_i, for $i = 1, 2$.

Example 3: If X has the distribution function $F(t_1, t_2) = \exp\left(2 - \frac{1}{t_1} - \frac{1}{t_2}\right)$ for $0 \leq t_1, t_2 \leq 1$, then $J_i^F(X : t_1, t_2) = \frac{-t_i^2}{4}$, $i, j = 1, 2,\ i \neq j$, which is decreasing in t_i. Thus, the distribution is decreasing in $CDFE_X$ ($DCDFE_X$).

A necessary and sufficient condition for a distribution to exhibit increasing ($ICDFE_X$) or decreasing ($DCDFE_X$) conditional dynamic failure extropy can be expressed using the reversed hazard rate, as presented in the following theorem.

Theorem 5: A distribution function F is said to be $ICDFE_X$ ($DCDFE_X$) if and only if, for all $t_i > 0$, $i = 1, 2$,

$$J_i^F(X : t_1, t_2) \leq (\geq) \frac{-1}{4\bar{h}_i(t_1, t_2)}, \tag{9}$$

where $\bar{h}_i(t_1, t_2)$ denotes the i-th component of the bivariate reversed hazard rate.

The following example illustrates Theorem 5.

Example 4: If X follows a bivariate uniform distribution with joint distribution function $F(t_1, t_2) = \frac{t_1 t_2}{bd}$, $0 \leq t_1 \leq b$, $0 \leq t_2 \leq d$, then $J_i^F(X : t_1, t_2) = \frac{-t_i}{6}$ and $\bar{h}_i(t_1, t_2) = \frac{1}{t_i}$. Thus, the inequality in (9) holds, indicating that F is $DCDFE_X$.

Figure 2 illustrates the relationship between $J_1^F(X : t_1, t_2)$ and the right-hand side

(RHS) of inequality (9) over the range $t_1, t_2 \in [0, 25]$. As shown, $J_1^F(X : t_1, t_2)$ decreases with increasing t_1 and consistently remains above the RHS of the inequality. This confirms the validity of inequality (9) and demonstrates the decreasing trend of $CDFE_X$ with respect to t_1.

Theorem 6: Let $V = (V_1, V_2)$ be a non-negative bivariate random vector such that $V_i = a_i X_i + b_i$, where $a_i > 0$ and $b_i > 0$, for $i = 1, 2$, and $X = (X_1, X_2)$. Then, $J_i^F(V : t_1, t_2)$ is increasing in t_i if and only if $J_i^F(X : t_1, t_2)$ is increasing in t_i.

The following theorem shows that conditional dynamic failure extropy ($CDFE_X$) can uniquely determine the underlying distribution, making it a useful tool for reliability analysis from lifetime data.

Theorem 7: Let X be a non-negative bivariate random vector with absolutely continuous distribution function F and bivariate reversed hazard rate components $\bar{h}_i(t_1, t_2)$, $i = 1, 2$. Then, $J_i^F(X : t_1, t_2)$ uniquely determines the joint distribution function F.

Proof: Let X and V be two non-negative random vectors such that

$$J_i^F(X : t_1, t_2) = J_i^F(V : t_1, t_2), \quad i = 1, 2.$$

Differentiating both sides with respect to t_i yields $\bar{h}_i^X(t_1, t_2) = \bar{h}_i^V(t_1, t_2)$. Since the reversed hazard rate uniquely determines the distribution (see Nair and Gopalakrishnan (2008)), it follows that $F_X = F_V$. □

The following theorem characterizes distributions where $CDFE_X$ is linearly proportional to the conditional mean inactivity time, leading to a specific power-form joint distribution.

Theorem 8: Let $X = (X_1, X_2)$ be a bivariate random vector supported on $[0, a_i]$, $i = 1, 2$, with finite a_i. Then, for all $t_i \in [0, a_i]$, $i = 1, 2$, we have

$$J_i^F(X : t_1, t_2) = c_i \, m_i(t_1, t_2), \quad \text{for } i = 1, 2, \tag{10}$$

if and only if

$$F(t_1, t_2) = \left(\frac{t_1}{a_1}\right)^{\frac{-g_1}{1+g_1}} \left(\frac{t_2}{a_2}\right)^{\frac{-g_2}{1+g_2}}, \tag{11}$$

where $g_i = \frac{2c_i+1}{2c_i}$ and $c_i \in (-1, 0)$, $i = 1, 2$.

Proof: The "if" part follows directly. For the "only if" part, differentiating (10) with respect to t_1 (for $i = 1$), we get

$$\frac{\partial}{\partial t_1} J_1^F(X : t_1, t_2) = c_1 \frac{\partial}{\partial t_1} m_1(t_1, t_2). \tag{12}$$

From this, we obtain

$$\frac{\partial}{\partial t_1} m_1(t_1, t_2) = 1 + g_1, \tag{13}$$

which leads to the form of $F(t_1, t_2)$ in (11). □

Theorem 8 leads to specific distributional forms for suitable choices of c_i. One such case characterizes the bivariate uniform distribution.

Corollary 8.1: A bivariate random vector X follows a uniform distribution on $[0, b] \times [0, d]$ with

$$F(t_1, t_2) = \frac{t_1 t_2}{bd}, \quad 0 \leq t_1 \leq b, \quad 0 \leq t_2 \leq d,$$

if and only if $J_i^F(X : t_1, t_2) = -m_i(t_1, t_2)$ for $i = 1, 2$.

Another class of distributions characterized via J_i^F is given below.

Theorem 9: Let X be a bivariate random vector supported on $(-\infty, a_1) \times (-\infty, a_2)$ with absolutely continuous distribution. Then

$$J_i^F(X : t_1, t_2) = -\frac{1}{4[k_i + k_3(t_j - a_j)]}, \quad i \neq j$$

holds if and only if

$$F(t_1, t_2) = \exp\left(k_1(t_1 - a_1) + k_2(t_2 - a_2) + k_3(t_1 - a_1)(t_2 - a_2)\right).$$

Remark 1: Many real-world systems involve more than two interdependent components. While the bivariate Conditional Dynamic Failure Extropy ($\mathrm{CDFE_X}$) captures pairwise uncertainty, extending this measure to multivariate settings is essential for analyzing complex systems with higher-dimensional dependencies.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be an n-dimensional non-negative random vector with joint distribution function $F(\mathbf{t}) = F(t_1, t_2, \ldots, t_n)$. Define the failure threshold vector $\mathbf{t} = (t_1, t_2, \ldots, t_n)$ and the joint failure event $\mathcal{A} = \{X_i < t_i, \forall i = 1, 2, \ldots, n\}$.

Then, the Multivariate Conditional Dynamic Failure Extropy ($CDFE_X$) for component X_i under $\mathcal{A}$ is defined as

$$J_i^F(\mathbf{X} : \mathbf{t}) = -\frac{1}{2}\int_0^{t_i} \left[\frac{F(t_1, \ldots, t_{i-1}, x_i, t_{i+1}, \ldots, t_n)}{F(t_1, \ldots, t_n)}\right]^2 dx_i, \quad i = 1, 2, \ldots, n. \tag{14}$$

Note: All theoretical results, properties, and estimators presented in this paper are derived and validated for the bivariate case ($n = 2$). The above multivariate extension is proposed to illustrate its potential applicability, but it is not theoretically or empirically explored in this work.

3. Non parametric estimation

In this section, we introduce a nonparametric framework for estimating Conditional Dynamic Failure Extropy ($CDFE_X$), suitable when the underlying distribution is unknown. We propose both empirical and kernel-based estimators that allow flexible modeling from complete data. The empirical estimator, based on the empirical distribution function, provides a practical and straightforward approach, especially effective for large samples. We also examine the asymptotic behavior of these estimators and validate their performance through simulations and real-world applications.

Definition 3: Let (X_{1i}, X_{2i}), $i = 1, 2, \ldots, n$, be a random sample from a population with joint distribution function F. Based on expressions (1) and (2), the empirical estimators of conditional dynamic failure extropy are defined as

$$\hat{J}_1^F(X : t_1, t_2) = -\frac{1}{2}\int_0^{t_1} \left[\frac{\hat{F}(x_1, t_2)}{\hat{F}(t_1, t_2)}\right]^2 dx_1, \tag{15}$$

$$\hat{J}_2^F(X : t_1, t_2) = -\frac{1}{2}\int_0^{t_2} \left[\frac{\hat{F}(t_1, x_2)}{\hat{F}(t_1, t_2)}\right]^2 dx_2, \tag{16}$$

where $\hat{F}(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} I(X_{1i} \le t_1, X_{2i} \le t_2)$ is the bivariate empirical distribution function, and

$$I(X_{1i} \le t_1, X_{2i} \le t_2) = \begin{cases} 1, & \text{if } X_{1i} \le t_1 \text{ and } X_{2i} \le t_2, \\ 0, & \text{otherwise.} \end{cases}$$

To improve estimation accuracy, we propose a kernel-based nonparametric estimator for $CDFE_X$. Kernel methods provide smooth approximations of the underlying distribution, reducing the variance inherent in empirical estimators.

Definition 4: Let (X_{1i}, X_{2i}), $i = 1, 2, \ldots, n$, be a sample from a population with joint distribution function F. The kernel estimator of $CDFE_X$ is defined as

$$\tilde{J}_1^F(X : t_1, t_2) = -\frac{1}{2}\int_0^{t_1} \left[\frac{\widetilde{F}(x_1, t_2)}{\widetilde{F}(t_1, t_2)}\right]^2 dx_1, \tag{17}$$

$$\tilde{J}_2^F(X : t_1, t_2) = -\frac{1}{2}\int_0^{t_2} \left[\frac{\widetilde{F}(t_1, x_2)}{\widetilde{F}(t_1, t_2)}\right]^2 dx_2, \tag{18}$$

where

$$\widetilde{F}(t_1, t_2) = \frac{1}{na_n^2}\sum_{j=1}^{n} K_1\left(\frac{t_1 - X_{1j}}{a_n}\right) K_2\left(\frac{t_2 - X_{2j}}{a_n}\right),$$

and $K_i(z) = a_n \int_0^z k_i(v)\, dv$, $i = 1, 2$, with $k_i(v)$ being known kernel density functions. The bandwidth sequence $\{a_n\}$ satisfies $a_n \to 0$ and $na_n \to \infty$ as $n \to \infty$.

To simplify notation, define

$$\hat{A}(t_1, t_2) = \int_0^{t_1} \left[\widetilde{F}(x_1, t_2)\right]^2 dx_1, \quad \hat{B}(t_1, t_2) = \left[\widetilde{F}(t_1, t_2)\right]^2,$$

$$A(t_1, t_2) = \int_0^{t_1} [F(x_1, t_2)]^2 dx_1, \quad B(t_1, t_2) = [F(t_1, t_2)]^2.$$

We employ kernel functions such as the Epanechnikov and Quartic kernels, known for their optimality in minimizing mean squared error.

To ensure consistency of the kernel-based estimator, we impose regularity conditions on the kernel function $k(t)$. These conditions ensure that the estimated density is well-defined and converges to the true distribution as sample size increases. The kernel $k(t)$ must satisfy

$$\int k(t)\,dt = 1, \quad \int t\,k(t)\,dt = 0, \quad \int tt^T\,k(t)\,dt = I,$$

where I is the identity matrix, ensuring proper normalization and moment conditions.

The following lemma, from Jin and Shao (1999), provides the asymptotic properties of the kernel estimator $\widetilde{F}$.

Lemma 1: Let x be a continuity point of F. Then, as $n \to \infty$

1. $\mathbb{E}[\widetilde{F}(x)] \to F(x)$,
2. $n\,\mathrm{Var}[\widetilde{F}(x)] \to F(x)[1 - F(x)]$.

The following theorem establishes the consistency of the kernel-based estimator for Conditional Dynamic Failure Extropy ($CDFE_X$), ensuring convergence in probability to the true value as the sample size increases.

Theorem 10: Let $\tilde{J}_i^F(X : t_1, t_2)$, $i = 1, 2$, be the kernel-based nonparametric estimator of the conditional dynamic failure extropy $J_i^F(X : t_1, t_2)$. Then $\tilde{J}_i^F(X : t_1, t_2)$ is a consistent estimator of $J_i^F(X : t_1, t_2)$.

Proof:

$$\int_0^{t_1} \widetilde{F}(x_1, t_2)^2\,dx_1 = \int_0^{t_1} F(x_1, t_2)^2\,dx_1 + 2\int_0^{t_1} F(x_1, t_2)\left[\widetilde{F}(x_1, t_2) - F(x_1, t_2)\right]dx_1 + O\left(\left[\widetilde{F}(x_1, t_2) - F(x_1, t_2)\right]^2\right).$$

Using Taylor expansion, we write

The bias and variance of $\hat{A}(t_1, t_2) = \int_0^{t_1} \widetilde{F}(x_1, t_2)^2 dx_1$ are

$$\mathrm{Bias}[\hat{A}(t_1, t_2)] \simeq 2\int_0^{t_1} F(x_1, t_2)\mathrm{Bias}[\widetilde{F}(x_1, t_2)]dx_1,$$
$$\mathrm{Var}[\hat{A}(t_1, t_2)] \simeq \frac{4}{n}\int_0^{t_1} F^3(x_1, t_2)[1 - F(x_1, t_2)]dx_1.$$

Similarly, for $\hat{B}(t_1, t_2) = \widetilde{F}(t_1, t_2)^2$

$$\mathrm{Bias}[\hat{B}(t_1, t_2)] \simeq 2F(t_1, t_2)\mathrm{Bias}[\widetilde{F}(t_1, t_2)],$$
$$\mathrm{Var}[\hat{B}(t_1, t_2)] \simeq \frac{4}{n}F^3(t_1, t_2)[1 - F(t_1, t_2)].$$

Therefore, we have the convergence in probability

$$\tilde{J}_1^F(X:t_1,t_2) \xrightarrow{P} J_1^F(X:t_1,t_2) \quad \text{as } n \to \infty.$$

A similar argument holds for $i = 2$. Hence, $\tilde{J}_i^F(X:t_1,t_2)$ is consistent. □

The following theorem derives the asymptotic bias and variance of the proposed estimator to evaluate its efficiency and understand the bias–variance trade-off.

Theorem 11: The asymptotic bias and variance of $\tilde{J}_1^F(X:t_1,t_2)$ are given by

$$\text{Bias}(\tilde{J}_1^F(X:t_1,t_2)) \simeq 0,$$

and

$$\begin{aligned}\text{Var}(\tilde{J}_1^F(X:t_1,t_2)) = {} & \frac{1}{nB^2(t_1,t_2)} \int_0^{t_1} F^3(x_1,t_2)[1-F(x_1,t_2)]dx_1 \\ & + \frac{A^2(t_1,t_2)}{nB^4(t_1,t_2)} F^3(t_1,t_2)[1-F(t_1,t_2)].\end{aligned}$$

Proof: Using the approximation

$$\text{Bias}\left(\frac{\hat{A}}{\hat{B}}\right) \simeq \frac{1}{B}\left[\text{Bias}(\hat{A}) - \frac{A}{B}\text{Bias}(\hat{B})\right],$$

and

$$\text{Var}\left(\frac{\hat{A}}{\hat{B}}\right) \simeq \frac{1}{B^2}\left[\text{Var}(\hat{A}) + \frac{A^2}{B^2}\text{Var}(\hat{B})\right],$$

where $A = A(t_1,t_2)$ and $B = B(t_1,t_2)$. Substituting known expressions for bias and variance completes the proof. □

This section develops a flexible nonparametric estimation framework for conditional dynamic failure extropy, applicable to reliability, survival, and risk analysis. The following subsections further explore the implementation and evaluation of these methods using simulated and real-world datasets.

3.1. Simulation study

This subsection evaluates the performance of the empirical and kernel estimators defined in Definitions 3 and 4, using simulation to estimate the bias and mean squared error (MSE) of $CDFE_X$. Simulations were implemented in Mathematica on synthetic datasets of sizes $n = 200$, 500, and 800, generated from the joint distribution

$$F(t_1,t_2) = \left(\frac{t_1}{b_1}\right)^{c_1}\left(\frac{t_2}{b_2}\right)^{c_2} + \theta \log\left(\frac{t_1}{b_1}\right), \quad \theta \le 0,$$

with parameters $b_1 = b_2 = 0.01$, $c_1 = c_2 = 0.02$, and $\theta = -0.1$. Both Epanechnikov and Quartic kernels were employed to estimate $CDFE_X$. For each n, 1,000 iterations were run.

Table 3: Bias and Mean squared errors (MSEs) of the conditional DFE_X function using empirical estimator, Epanechnikov kernel, and quartic kernel for simulated data

Sample Size	(t_1, t_2)	$CDFE_X$ (Empirical)		$CDFE_X$ (Epanechnikov Kernel)		$CDFE_X$ (Quartic Kernel)	
		Bias	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**
200	(0.17, 0.17)	(0.0858, 0.1029)	(0.0073, 0.0105)	(0.0467, 0.0466)	(0.0021, 0.0021)	(0.0536, 0.0573)	(0.0035, 0.0037)
	(0.2, 0.7)	(0.2858, 0.1929)	(0.0173, 0.0305)	(0.0492, 0.0949)	(0.0024, 0.0080)	(0.0592, 0.1049)	(0.0047, 0.0153)
	(0.45, 0.45)	(0.46892, 0.1830)	(0.0891, 0.0483)	(0.1173, 0.0971)	(0.0137, 0.0137)	(0.1571, 0.1071)	(0.0424, 0.0432)
	(0.5, 0.5)	(0.5575, 0.2329)	(0.0941, 0.0778)	(0.1293, 0.1391)	(0.0167, 0.0177)	(0.3215, 0.1991)	(0.0364, 0.0747)
500	(0.17, 0.17)	(0.0849, 0.1018)	(0.0053, 0.0096)	(0.0423, 0.0421)	(0.0009, 0.0009)	(0.0521, 0.0543)	(0.0029, 0.0031)
	(0.2, 0.7)	(0.2851, 0.1922)	(0.0116, 0.0274)	(0.0415, 0.0243)	(0.0011, 0.0052)	(0.0524, 0.0949)	(0.0024, 0.0097)
	(0.45, 0.45)	(0.4681, 0.1797)	(0.0825, 0.0424)	(0.1152, 0.0768)	(0.0122, 0.0124)	(0.1542, 0.1025)	(0.0404, 0.0418)
	(0.5, 0.5)	(0.5524, 0.2285)	(0.0912, 0.0734)	(0.1263, 0.1324)	(0.0126, 0.0145)	(0.2860, 0.1928)	(0.0327, 0.0722)
800	(0.17, 0.17)	(0.0795, 0.0938)	(0.0013, 0.0057)	(0.0315, 0.0247)	(0.0003, 0.0002)	(0.0482, 0.0473)	(0.0023, 0.0016)
	(0.2, 0.7)	(0.2783, 0.1839)	(0.0082, 0.0139)	(0.0361, 0.0172)	(0.0008, 0.0039)	(0.0472, 0.0928)	(0.0015, 0.0081)
	(0.45, 0.45)	(0.4597, 0.1647)	(0.0791, 0.0397)	(0.1145, 0.0752)	(0.0097, 0.0089)	(0.1497, 0.1007)	(0.0375, 0.0373)
	(0.5, 0.5)	(0.5491, 0.2232)	(0.0885, 0.0719)	(0.1246, 0.1312)	(0.0101, 0.0086)	(0.2784, 0.1911)	(0.0287, 0.0699)

In every iteration, empirical and kernel estimates of $CDFE_X$ were computed, and their absolute bias and MSE were evaluated. Table 3 summarizes the simulation results across sample sizes. The simulation results show that kernel estimators consistently yield lower MSE than empirical estimators, especially with larger sample sizes. This demonstrates the reliability and reduced uncertainty of kernel-based methods for estimating $CDFE_X$, validating their applicability in complex reliability analysis. Overall, the study confirms the effectiveness of the proposed nonparametric estimators for practical use.

Table 4: Fit results for both sets of data

	Statistic	P-Value
Anderson-Darling	1.15342	0.28493
Cramér-von Mises	0.0891055	0.640949
Kolmogorov-Smirnov	0.11474	0.656902
Kuiper	0.656902	0.394967
Pearson χ^2	10.3158	0.243557
Watson U^2	0.0840573	0.383813

	Statistic	P-Value
Anderson-Darling	0.486959	0.75891
Cramér-von Mises	0.0558492	0.84009
Kolmogorov-Smirnov	0.102625	0.780647
Pearson χ^2	12.2105	0.142055

Table 5: Conditional DFE_X function bias and MSEs using empirical and Epanechnikov kernel estimators for a real dataset

Measure	**(1.7,1)**	**(2,3)**	**(2,4)**
Empirical Bias	(0.85, 0.5)	(0.12548, 0.63589)	(0.12512, 0.93445)
Empirical MSE	(0.7225, 0.25)	(0.01574, 0.40436)	(0.01565, 0.92869)
Kernel Bias	(0.00068, 0.00037)	(0.00067, 0.00042)	(0.00061, 0.00110)
Kernel MSE	$(4.73354 \times 10^{-7}, 1.42776 \times 10^{-7})$	$(4.5577 \times 10^{-7}, 1.83099 \times 10^{-7})$	$(3.7551 \times 10^{-7}, 1.23225 \times 10^{-6})$
Measure	**(3,3)**	**(3,4)**	
Empirical Bias	(0.12539, 0.63667)	(0.54465, 0.91255)	
Empirical MSE	(0.01572, 0.40536)	(0.29665, 0.9668)	
Kernel Bias	(0.00172, 0.00602)	(0.00260, 0.01090)	
Kernel MSE	$(2.99221 \times 10^{-6}, 0.00003)$	$(6.78643 \times 10^{-6}, 0.00011)$	

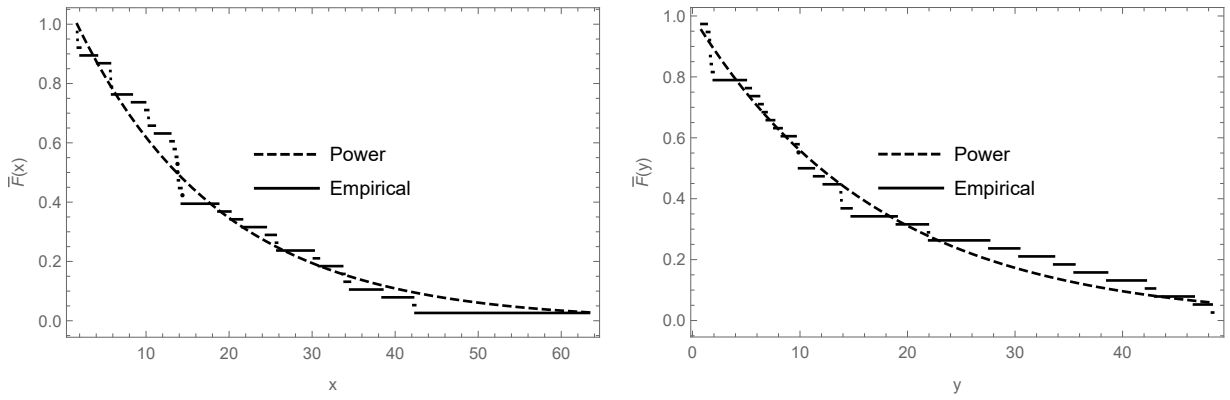


Figure 3: Graphs depicting empirical and fitted distributions for both datasets.

3.2. Application to a real dataset

In the following example, we apply real data to evaluate the empirical and kernel estimators defined in Definitions 3 and 4.

Example 5: We analyze a dataset from the National Eye Institute, involving 38 diabetic patients at risk of blindness, to illustrate the use of $CDFE_X$ estimators. One eye of each patient received laser photocoagulation, and times (in months) until blindness in both treated (X_2) and untreated (X_1) eyes were recorded Kilany and El-Qareb (2023). Each marginal distribution was fitted using a power distribution and assessed via the Kolmogorov–Smirnov test. The fit was further confirmed by comparing empirical and fitted curves in Figure 4, with detailed test statistics in Table 4. We computed $CDFE_X$ using both empirical and kernel methods, and assessed performance using 100,000 bootstrap samples (size 38 each) across (t_1, t_2) values. Bias and MSE of both estimators are summarized in Table 5. Results demonstrate that kernel estimators provide more accurate and stable estimates of $CDFE_X$ with reduced uncertainty. This highlights the practical utility of these methods in clinical reliability studies, particularly in evaluating treatment effects on failure times.

Example 6 demonstrates the superior performance of the proposed kernel estimators for Conditional Dynamic Failure Extropy ($CDFE_X$) over the univariate kernel estimator of Dynamic Failure Extropy by Nair and Sathar (2020).

Example 6: Consider the dataset provided in Kilany and El-Qareb (2023), which records the vision loss times for diabetic retinopathy patients, denoted by X_1 and X_2. Based on the univariate Dynamic Failure Extropy ($\mathrm{DFE_X}$) values reported by Nair and Sathar (2020), estimated as -1.1962 for X_1 and -1.06739 for X_2, we apply equations (17) and (18) to compute the Conditional $\mathrm{DFE_X}$ ($\mathrm{CDFE_X}$) values. For the point $(t_1, t_2) = (1, 2)$, the resulting $\mathrm{CDFE_X}$ values are obtained as $(-0.000828392,\ -0.00198842)$. For $(t_1, t_2) = (2, 3)$, the corresponding univariate $\mathrm{DFE_X}$ values are -1.1999 and -0.728436, with $CDFE_X$ estimates $(-0.000830955, -0.00135683)$. These results show that $CDFE_X$ yields consistently lower uncertainty than the univariate $\mathrm{DFE_X}$, highlighting the advantage of incorporating joint information. While univariate analysis may overstate uncertainty in component failure times, conditional extropy captures interdependence, improving predictability in past lifetime studies. Hence, bivariate failure extropy measures are preferred in such settings.

3.3. Comparison with conditional dynamic cumulative past entropy

This subsection compares the performance of the proposed Bivariate Conditional Dynamic Failure Extropy ($CDFE_X$) estimators with the Conditional Dynamic Cumulative Past Entropy ($CDCPE$), focusing on their ability to capture uncertainty in joint failure distributions. The kernel-based estimators for $CDCPE$ are given by

$$\widetilde{\bar{\varepsilon}}_1^*(X : t_1, t_2) = -\int_0^{t_1} \frac{\widetilde{F}(x_1, t_2)}{\widetilde{F}(t_1, t_2)} \log\left(\frac{\widetilde{F}(x_1, t_2)}{\widetilde{F}(t_1, t_2)}\right) dx_1, \tag{19}$$

$$\widetilde{\bar{\varepsilon}}_2^*(X : t_1, t_2) = -\int_0^{t_2} \frac{\widetilde{F}(t_1, x_2)}{\widetilde{F}(t_1, t_2)} \log\left(\frac{\widetilde{F}(t_1, x_2)}{\widetilde{F}(t_1, t_2)}\right) dx_2. \tag{20}$$

Using the real-world dataset from Example 5, we estimated $CDFE_X$ via both empirical and kernel methods, alongside the $CDCPE$ values. Bias and mean squared error (MSE) were used to evaluate estimator performance, with results shown in Tables 5 and 6. The results show that bivariate $CDFE_X$ consistently achieves lower bias and MSE compared to $CDCPE$, particularly under strong variable dependencies. This highlights $CDFE_X$ as a more robust and informative measure for analyzing joint failure behavior. Bivariate $CDFE_X$ outperforms $CDCPE$ in modeling uncertainty in two-component systems, offering improved accuracy and reliability for real-world reliability analysis.

Table 6: Bias and mean squared errors (MSEs) of the conditional dynamic $CDCPE$ function with Epanechnikov kernel for real data sets

(t_1, t_2)	$CDCPE$ (Kernel)	
	Bias	MSE
(1.7,1)	(0.17196, 0.09877)	(0.02957, 0.00975)
(2,3)	(0.12924, 0.42271)	(0.01671, 0.17869)
(2,4)	(0.12901, 0.54580)	(0.01665, 0.29791)
(3,3)	(0.00651, 0.42287)	(0.00007, 0.17883)
(3,4)	(0.00597, 0.54574)	(0.00007, 0.29784)

4. Conclusions

This paper proposed Conditional Dynamic Failure Extropy ($CDFE_X$) as a novel measure to quantify uncertainty in bivariate systems. We established its theoretical properties and developed nonparametric estimators using empirical and kernel methods. Simulation studies and real data analysis demonstrated the superior performance of $CDFE_X$ over existing measures such as univariate Dynamic Failure Extropy and Conditional Dynamic Cumulative Past Entropy, particularly in capturing joint failure dependencies. Kernel estimators consistently yielded lower bias and MSE than empirical counterparts. Overall, $CDFE_X$ offers a robust and informative framework for reliability analysis in systems with dependent components. Future work may extend this framework to higher dimensions and broader applications.

Acknowledgements

We sincerely thank the reviewers and the editor for their valuable feedback and guidance, which greatly improved the quality of our manuscript.

Conflict of interest

The authors do not have any financial or non-financial conflicts of interest to declare for the research work included in this article.

References

Jin, Z. and Shao, Y. (1999). On kernel estimation of a multivariate distribution function. *Statistics and Probability Letters*, **41**, 163–168.

Kayal, S. (2021). Failure extropy, dynamic failure extropy and their weighted versions. *Stochastics and Quality Control*, **36**, 59–71.

Kilany, N. and El-Qareb, F. (2023). Modelling bivariate failure time data via bivariate extended Chen distribution. *Stochastic Environmental Research and Risk Assessment*, **37**, 1–9.

Krishnan, A. S., Sunoj, S. M., and Nair, N. U. (2020). Some reliability properties of extropy for residual and past lifetime random variables. *Journal of the Korean Statistical Society*, **49**, 457–474.

Krishnan, L. and Sathar, A. (2022). On bivariate dynamic survival extropy and its estimation. *Journal of the Indian Society for Probability and Statistics*, **23**, 425–449.

Lad, Frank; Sanfilippo, G. A. G. (2015). Extropy: Complementary dual of entropy. *Statistical Science*, **30**, 40–58.

Nair, N. and Gopalakrishnan, A. (2008). Some characterizations based on bivariate reversed mean residual life. *ProbStat Forum*, **01**, 1–14.

Nair, R. D. and Sathar, E. I. A. (2020). On dynamic failure extropy. *Journal of the Indian Society for Probability and Statistics*, **21**, 287–313.

Noughabi, H. A. and Jarrahiferiz, J. (2019). On the estimation of extropy. *Journal of Nonparametric Statistics*, **31**, 88–99.

Qiu, G. (2017). The extropy of order statistics and record values. *Statistics & Probability Letters*, **120**, 52–60.

Qiu, G. and Jia, K. (2017). Extropy estimators with applications in testing uniformity. *Journal of Nonparametric Statistics*, **30**, 182–196.

Qiu, G. and Jia, K. (2018). The residual extropy of order statistics. *Statistics & Probability Letters*, **133**, 15–22.

Qiu, G., Wang, L., and Wang, X. (2018). On extropy properties of mixed systems. *Probability in the Engineering and Informational Sciences*, **33**, 471–486.

Raqab, M. Z. and Qiu, G. (2019). On extropy properties of ranked set sampling. *Statistics*, **53**, 210–226.

Sathar, A. and Krishnan, C. (2023). Bivariate extension of weighted dynamic survival extropy and its quantile approach. *International Journal of Reliability, Quality and Safety Engineering*, **30**.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379–423.

Yang, J., Xia, W., and Hu, T. (2019). Bounds on extropy with variational distance constraint. *Probability in the Engineering and Informational Sciences*, **33**, 186–204.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 24, No. 1, 2026 (New Series), pp 237–248
http://www.ssca.org.in/journal

ROC Curve for Binary Classification using X Lindley Distribution

Sandhya Singh[1] **and Saebugari Balaswamy**[2]

[1]*Department of Community Medicine, Bundelkhand Medical College, Madhya Pradesh*
[2]*Department of Statistics, Indira Gandhi National Tribal University, Madhya Pradesh*

Received: 15 May 2025; Revised: 29 September 2025; Accepted: 10 October 2025

Abstract

A lot of ROC models have been derived in the area of classification to classify the subjects using many distributional assumptions based upon the nature of the data. Though, there are many models in the literature, still such models are of need current situation to address the different needs of the nature of data due to the skewness or non normal behavior of the data in reality. Therefore, this paper addresses such an ROC model which consists of the X Lindley (XL) distribution and this distribution has more flexibility than other one-parameter distributions. An attempt has been made to develop an ROC model in classification when the healthy population is at the higher side than the abnormal/diseased population. Further, simulations as well as real data set have been used to find out area under the curve. The simulations are done with various parameter values of the distribution, which will represent the better, moderate and worst case of classification in ROC analysis. The X Lindley distribution can be used quite effectively in analyzing the data in classification and which is easy to make computations even for a non statistician. Further, the properties of the XL ROC curve are verified mathematically and the likelihood ratio test is also proposed and the relation is established with the slope of the ROC curve.

Key words: ROC curve; X Lindley distribution; AUC; Likelihood ratio test; Slope of the ROC curve; Optimal threshold.

AMS Subject Classifications: 62P10

1. Introduction

Various statistical methods have been developed to identify class labels, which is a primary goal in classification tasks. A typical challenge in classification involves assigning an individual to one of several predefined groups, such as healthy (normal) or diseased (abnormal). To address these challenges, tools like the Receiver Operating Characteristic (ROC) curve are commonly used. The ROC curve originated in the 1940s when electrical and radar engineers developed it for detecting enemy objects during World War II. Since then, ROC analysis has found widespread application across diverse fields, including medicine, radiol-

Corresponding Author: Saebugari Balaswamy
Email: balaswamy.stats@gmail.com

ogy, biometrics, natural hazard forecasting, and meteorology, and it has become increasingly important in machine learning and data mining. An ROC curve plots sensitivity against 1-specificity for a diagnostic test. Sensitivity refers to the probability of correctly identifying individuals with a disease, while specificity measures the probability of correctly identifying healthy individuals. Additionally, the area under the ROC curve (AUC) is used as a summary measure, reflecting the binary classifier's ability to differentiate between classes.

If the outcome S of a medical test is measured on a continuous scale, a specific threshold t can be set to classify individuals. For any diagnostic test, a person with a score $s \leq t$ can be classified as healthy (H), while those with a score above t are classified as diseased (D). Based on this classification, a 2×2 contingency table, also known as a confusion matrix, can be constructed. This matrix includes four possible outcomes: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

The probabilistic definitions are given below,

1. The probability that an individual from D is correctly classified.
   True positive rate, $TPR = P(S > t|D)$ Sensitivity

2. The probability that an individual from H is misclassified.
   False positive Rate, $FPR = P(S > t|H)$

3. The probability that an individual from is correctly classified.
   True negative rate, $TNR = P(S \leq t|H)$ Specificity

4. The probability that an individual from D is misclassified.
   False Negative Rate, $FNR = P(S \leq t|D)$ 1-Sensitivity.

Over time, a wide variety of ROC (Receiver Operating Characteristic) models have been introduced in the literature, particularly those relying on bi-distributional assumptions. The following are notable examples: Bilognormal (Dorfman and Alf, 1968); Binormal (Egan, 1975); Bibeta (Zou *et al.*, 1997); Bigamma (Hussain, 2012); Biexponential (Tang and Balakrishnan, 2011); Hybrid ROC Curve (Balaswamy *et al.*, 2015). Additionally, methods for estimating confidence intervals for ROC curves have been enhanced by applying distributions such as: Generalized Half Normal and Weibull (Balaswamy *et al.*, 2015). Further, innovative applications and extensions are developed in the recent past, which includes the analysis of the relationship between the ratio β and the area under the ROC curve (AUC) using combined Half Normal and Rayleigh distributions (Balaswamy and Vardhan, 2015); AUC estimation for non-normal data (Balaswamy and Vardhan, 2022); Estimation of AUC with the Bi-Generalised Exponential ROC curve (Dashina and Vishnu Vardhan, 2022); ROC curve area estimation within gamma mixture models (Arunima and Vishnu Vardhan, 2022); Multiclass ROC statistics and non-normal data applications (Arunima and Vishnu Vardhan, 2023); Comparative AUC estimation using generalized exponential distributions (Balaswamy and Vardhan, 2023). Despite the variety of ROC models based on different distributions, new models are required to address complex, real-world data scenarios. Motivated by the limitations in existing approaches and the need for increased flexibility, the present work introduces an ROC model founded on the X Lindley distribution. This choice is justified by ease of application and greater flexibility compared to other single-parameter distributions.

This new model aims to better accommodate diverse data characteristics encountered in practical classification problems where standard assumptions may not hold.

Therefore, the next section deals with the methodology, where the ROC model is developed using the X Lindley distribution.

2. Methodology

In this section, a mixture of two known distributions (Exponential and Lindley) used to give new distribution called X Lindley distribution. Let X be a random variable with the mixture Distribution viz, X Lindley distribution. The probability density function and the distribution functions of X Lindley distribution are as follows.

$$f_{XL}(x,\theta) = \frac{\theta^2(2+\theta+x)}{(1+\theta)^2}e^{-x\theta} \quad , x,\theta > 0 \tag{1}$$

$$F_{XL}(x,\theta) = 1 - \left[1 + \frac{x\theta}{(1+\theta)^2}\right] e^{-x\theta} \quad , x,\theta > 0 \tag{2}$$

Let the test scores (S) of normal (0) and abnormal (1) populations follows a X Lindley distribution. As the ROC curve is the tradeoff between Sensitivity and 1-Specificity, the FPR (1-Specificity) is defined as

$$FPR = x(t) = P(S > t|0) = \left[1 + \frac{t\theta_0}{(1+\theta_0)^2}\right] e^{-t\theta_0} \tag{3}$$

Where, t denotes threshold of a diagnostic test and θ_0 is the parameter of XL distribution from healthy (normal) population.

Applying log on both sides in (3), we get

$$\log(x(t)) = \log\left[1 + \frac{t\theta_0}{(1+\theta_0)^2}\right] - t(\theta_0)$$

$$\log(x(t)) = \frac{t\theta_0}{(1+\theta_0)^2} - t\theta_0$$

Further, the threshold can be obtained from the above equation as

$$t = \frac{(1+\theta_0)^2}{[\theta_0 - (1+\theta_0)^2\theta_0]} \log(x(t)) \tag{4}$$

Another intrinsic measure Sensitivity is defined as

$$TPR = Y(t) = P(S > t|0) = \left[1 + \frac{t\theta_1}{(1+\theta_1)^2}\right] e^{-t\theta_1} \tag{5}$$

where θ_1 is the parameter of XL distribution from diseased (abnormal) population and it is assumed to be lesser than the value of θ_0. The ROC curve can be obtained by substituting the above threshold value t in the y(t), we get

$$ROC(t) = \left[1 + \frac{\theta_1}{(1+\theta_1)^2} \frac{(1+\theta_0)^2}{[\theta_0 - (1+\theta_0)^2\theta_0]} \log(x(t))\right] e^{\frac{-\theta_1(1+\theta_0)^2}{\theta_0-(1+\theta_0)^2\theta_0} \log(x(t))} \tag{6}$$

On further simplification, the final expression for TPR is given by

$$ROC(t) = \left[1 + \frac{\theta_1}{\theta_0}\left(\frac{1}{\left(\frac{1+\theta_1}{1+\theta_0}\right)^2 - (1+\theta_1)^2}\log(x(t))\right)\right] e^{-\frac{\theta_1}{\theta_0}\left[\frac{1}{\frac{1}{(1+\theta_0)^2}-1}\right]\log(x(t))} \tag{7}$$

The above expression is defined as the ROC curve for X Lindley distribution and is named as XL ROC curve in binary classification. Along with the above two intrinsic measures namely Sensitivity and Specificity; Area under the Curve (AUC) is also an important measure of ROC curve and plays a prominent role in assessing the performance of a test. AUC takes values between 0 and 1. A perfect diagnostic test is one with an area equal to 1 and a test with an area 0 is perfectly inaccurate. The AUC of an ROC curve can be interpreted as the average of sensitivity for all possible values of specificity and vice versa. AUC of an ROC curve can be obtained by integrating the ROC expression over the range [0, 1] *i.e.*,

$$AUC = \int_0^1 ROC(t)dt$$

Further the area under the curve for XL ROC curve is given by

$$AUC = \int_0^1 \left[1 + \frac{\theta_1}{\theta_0}\left(\frac{1}{\left(\frac{1+\theta_1}{1+\theta_0}\right)^2 - (1+\theta_1)^2}\right)\log(x(t))\right] e^{-\frac{\theta_1}{\theta_0}\left[\frac{1}{\frac{1}{(1+\theta_0)^2}-1}\right]\log(x(t))} dx(t) \tag{8}$$

For the purpose of simplification, let us consider the above AUC in (8) as,

$$AUC = \int_0^1 \{1 + a\log(x(t))\}\, e^{-b\log(x(t))} dx(t) \tag{9}$$

Where a and b are defined as follows

$$a = \frac{\theta_1}{\theta_0}\left(\frac{1}{\left(\frac{1+\theta_1}{1+\theta_0}\right)^2 - (1+\theta_1)^2}\right)$$

$$b = \frac{\theta_1}{\theta_0}\left[\frac{1}{\frac{1}{(1+\theta_0)^2}-1}\right]$$

Further, the AUC in (9) can be simplified as

$$AUC = \lim_{\epsilon\to 0}\int_\epsilon^1 \{1 + a\log(x(t))\}e^{-b\log(x(t))}dx(t)$$

$$AUC = \lim_{\epsilon\to 0}\int_\epsilon^1 \{1 + a\log(x(t))\}\frac{1}{x^b}dx(t)$$

On further simplification, we get

$$AUC = \lim_{\epsilon\to 0}\left\{\left(\frac{x^{1-b}}{1-b}\right)_\epsilon^1 + a\log(x)\int_\epsilon^1 \frac{1}{x^b}dx(t) - a\int_\epsilon^1 \frac{1}{x}\frac{x^{1-b}}{1-b}dx(t)\right\}$$

$$AUC = \lim_{\epsilon \to 0} \left(\frac{1-\epsilon^{(1-b)}}{1-b} - \frac{\epsilon^{(1-b)} a \log(\epsilon)}{1-b} - \frac{a}{(1-b)^2}\left(1-\epsilon^{1-b}\right) \right)$$

$$AUC = \frac{1}{1-b} - \frac{a}{(1-b)^2}$$

$$AUC = \frac{1-b-a}{(1-b)^2} \tag{10}$$

The above expression (10) is the final Area under the XL ROC curve, which provides the accuracy of the test in classification. The next section deals with the optimal threshold and the Youden's Index, which are crucial in identifying the optimal threshold to classify the subjects with greater accuracy.

2.1. Optimal threshold and Youden's index

Once, the ROC Curve and AUC are estimated, another important aspect in classification is to obtain the optimal threshold, which is unique and provides a better percentage of correct classification. This criterion can be met in two ways, first one depends on the condition that the pdf's of two populations are equal and second way is to make use of the widely accepted measure, namely Youden's Index (J). To obtain the optimal threshold for the ROC Curve, let us assume that both population densities are equal,

$$f_x(x, \theta_0) = f_x(x, \theta_1)$$

On further simplification, we have

$$\frac{\theta_0^2(2+\theta_0+t)e^{-t\theta_0}}{(1+\theta_0)^2} = \frac{\theta_1^2(2+\theta_1+t)e^{-t\theta_1}}{(1+\theta_1)^2}$$

The final expression for the optimal threshold is given by,

$$opt\ t = \frac{2}{(\theta_1-\theta_0)} \log\left(\frac{\theta_1}{\theta_0}\right) + \frac{2}{(\theta_1-\theta_0)} \log\left(\frac{1+\theta_0}{1+\theta_1}\right) + 1 \tag{11}$$

This is the optimal threshold for obtaining the best threshold or cutoff which will classify the populations with higher accuracy and lesser misclassification rate. Here, description on the use of Youden's Index in validating the obtained optimal threshold is given. In ROC context, the Youden's Index is defined as

$$J = \max\{Sensitivity(t) + Specificity(t) - 1\} \tag{12}$$

over all cut points t, $-\infty < t < \infty$. The index (J) ranges between 0 and 1 with a value of 1 indicating perfect diagnostic effectiveness and 0 indicating an ineffective test. With respect to the ROC curve, J is the maximum vertical distance between the curve and the diagonal (chance line) and acts as a global measure of the optimum diagnostic ability. Further, the Youden's index is obtained as

$$J = \max\left\{ \left[1 + \frac{t\theta_1}{(1+\theta_1)^2}\right] e^{-t\theta_1} - \left[1 + \frac{t\theta_0}{(1+\theta_0)^2}\right] e^{-t\theta_0} \right\} \tag{13}$$

Further, ROC curve should possess the properties of monotonic transformation, invariance with respect to some increasing transformation. That is, the ROC curve is invariant to any monotonic (*e.g.*, linear, logarithmic) transformation of the test results (Krzanowski and Hand, 2009). Here, a brief outline about these basic properties of an ROC curve is discussed.

Properties of the ROC

i. $Y = h(x)$ is the mathematical model of the ROC curve, where y denotes the true positive rate and x denotes the false positive rate. The curve is a monotonic increasing function in the positive quadrant, lying between $y = 0 \ \ at \ \ x = 0$ and $y = 1 \ \ at \ \ x = 1$. Let us consider two false positives values P_1 and P_2 such that $P_1 < P_2$

$$\begin{aligned} -b\log P_1 &< -b\log P_2 \\ e^{-b\log P_1} &< e^{-b\log P_2} \\ a\log(P_1)e^{-b\log P_1} &\leq a\log(P_2)e^{-b\log P_2} \\ e^{-b*\log P_1} + [a*\log P_1]e^{-b\log P_1} &\leq e^{-b*\log P_2}[a*\log P_2]e^{-b\log P_2} \\ e^{-b*\log P_1} + [1 + a*\log P_1] &\leq e^{-b*\log P_2}[1 + a*\log P_2] \\ XLROC(P_1) &\leq XLROC(P_2) \end{aligned}$$

Hence the XL ROC curve is monotonically increasing.

ii. The ROC curve is unaltered if the classification scores undergo a strictly increasing transformation.

Let S denote the set of scores with $S \subset R$ and $h(\cdot)$ is strictly increasing function. Let $a, \ b \in S$ and $a < b$, then by using the strictly increasing function, we can write $h(a) < h(b)$.

The transformed random variables U and V from the respective healthy and diseased classes are

$$P(U \leq t) = P[h(U) \leq h(t)]$$
$$P(V \leq t) = P[h(V) \leq h(t)]$$

Let us consider the points $(x^*(t), y^*(t))$ on the ROC Curve for the transformed scores

$$x^*(t) = P\{h(U) > h(t)|H\} = 1 - P[h(U) \leq h(t)] = 1 - P(U \leq t) = x(t)$$

$$y^*(t) = P\{h(V) > h(t)|D\} = 1 - P[h(V) \leq h(t)] = 1 - P(V \leq t) = y(t)$$

Thus, the XLROC Curve is invariant to transformation.

iii. The slope of the XL ROC curve at threshold value t is given by

$$\frac{dy}{dx} = \frac{P(S > t|1)}{P(S > t|0)}$$

The derivative of the roc curve at a given pair of coordinates equals the likelihood ratio. Let us parameterize x and y in terms of t and derivative can be written as

$$\frac{dy}{dx} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}} = \frac{g(t)}{f(t)}$$

Therefore,

$$\frac{dy}{dx} = \left\{\frac{\theta_1^2(2 + \theta_1 + t)e^{-t\theta_1}}{(1 + \theta_1)^2}\right\}\left\{\frac{(1 + \theta_0)^2}{\theta_0^2(2 + \theta_0 + t)e^{-t\theta_0}}\right\}$$

$$\frac{dy}{dx} = \left(\frac{\theta_1}{1+\theta_1}\right)^2 \left(\frac{1+\theta_0}{\theta_0}\right)^2 \left(\frac{2+\theta_1+t}{2+\theta_0+t}\right) e^{(\theta_0-\theta_1)t} \tag{14}$$

Here $\theta_0,\ \theta_1$ and $t \geq 0$. Therefore, $\frac{dy}{dx} \geq 0$. The behavior of the curve and the typical forms of ROC curve are totally dependent on the slope given in (14). This is the ratio of distribution of diseased scores to healthy scores of the two probability densities at the value of t. This is referred to as likelihood ratio of XL ROC curve, which is provided in the next section. In the next section, the well known Likelihood Ratio test is described and integrated in the context of ROC curve analysis. It is also a measure, which helps in understanding the shape of the ROC curve.

2.2. Likelihood ratio test

In order to establish an interesting relationship between the likelihood ratio test and the slope of the ROC curve, we have considered this concept of likelihood ratio test in classification as follows. In general, the likelihood ratio test has the hypothesis as

$$H_0 : \theta = \theta_0$$

Versus

$$H_1 : \theta = \theta_1$$

To test the hypothesis Likelihood Ratio Test is given by

$$\lambda = \frac{supL(x_1, x_2, \ldots, x_n; \theta_0)}{supL(x_1, x_2, \ldots, x_n; \theta_1)}$$

The likelihood function of XL distribution is

$$L = \left(\frac{\theta}{1+\theta}\right)^{2n} e^{-\theta \sum_{i=1}^{n} t_i} \prod_{i=1}^{n} (2+\theta+t_i)$$

Therefore, the likelihood ratio test for testing the hypothesis is

$$\lambda = \frac{\left(\frac{\theta_0}{1+\theta_0}\right)^{2n} e^{-\theta_0 \sum_{i=1}^{n} t_i} \prod_{i=1}^{n} (2+\theta_0+t_i)}{\left(\frac{\theta_1}{1+\theta_1}\right)^{2n} e^{-\theta_1 \sum_{i=1}^{n} t_i} \prod_{i=1}^{n} (2+\theta_1+t_i)}$$

$$\lambda = \left(\frac{\theta_0}{1+\theta_0}\right)^{2n} \left(\frac{1+\theta_1}{\theta_1}\right)^{2n} e^{(-\theta_0+\theta_1) \sum_{i=1}^{n} t_i} \prod_{i=1}^{n} \frac{(2+\theta_0+t_i)}{(2+\theta_1+t_i)}$$

This is the likelihood ratio test for testing the defined hypothesis, where $\theta_1 > \theta_0$. But, in case of XL ROC curve, we have $\theta_0 > \theta_1$. Therefore, λ can be rewritten as

$$\lambda = \left(\frac{\theta_1}{1+\theta_1}\right)^{2n} \left(\frac{1+\theta_0}{\theta_0}\right)^{2n} e^{-(\theta_1+\theta_0) \sum_{i=1}^{n} t_i} \prod_{i=1}^{n} \frac{(2+\theta_1+t_i)}{(2+\theta_0+t_i)} \tag{15}$$

For $n = 1$, the slope of the XL ROC curve (14) is equals the likelihood ratio (15) in the context of classification. Therefore, the functional relationship between slope of the ROC

curve and the likelihood ratio is established to study the behavior of the ROC curve and its parameters.

The next section deals with the results and discussion to explain the proposed ROC curve in classification.

3. Results and discussion

This section presents the results obtained from the following experiments. Several tests were performed to demonstrate the characteristics and behavior of the XL ROC curve. Three different simulations were carried out using varying distribution parameters to illustrate better, moderate and worst cases classification scenarios in ROC analysis.

Table 1: Results of XLROC curve for simulated data at particular values of parameters

case	θ_0	θ_1	b	a	AUC	J	Opt t
Better	3.5	0.2	-0.06011	-0.04174	0.980441	0.32582	1.933603
Better	3	0.5	-0.17778	-0.07901	0.906016	0.673449	1.648744
Better	2.5	0.8	-0.34844	-0.10754	0.800741	0.125075	1.558186
Better	2	1.1	-0.61875	-0.14031	0.671305	0.253085	1.535916
Moderate	1.5	0.5	-0.39683	-0.17637	0.806302	0.482189	2.175573
Moderate	2.25	0.9	-0.44183	-0.12239	0.752437	0.162233	1.562207
Moderate	3.5	1.5	-0.45083	-0.07213	0.723527	0.263832	1.259511
Moderate	1.001	1	-1.33156	-0.33289	0.490134	0.000232	1.999251
Worst	0.2	3.5	-57.2727	-2.82828	0.017994	-0.41979	1.933603
Worst	0.5	3	-10.8	-0.675	0.089594	-0.05391	1.648744
Worst	0.8	2.5	-4.52009	-0.36899	0.193266	-0.1352	1.558186
Worst	1.1	2	-2.35137	-0.26126	0.321646	-0.05129	1.535916

In Table 1 (better case), as the scale parameter of healthy (normal) population decreases with opposite phenomenon in the diseased (abnormal) population, then it is found that area under the curve is decreasing and value of optimum threshold is also decreasing with the case better curve (Figure 1 - Better case). This means that as the parameters of the distributions moves away from the center to produce the largest discrimination between the two populations, the area under the curve increases with better classification rate.

Further, the moderate case of ROC curve is considered with different parameter values of the considered distribution and the results with the corresponding XL ROC curves are plotted (Figure 1 - Moderate case). Here, the area under the curve and its corresponding Youden's index are found to be good with moderate classification rate. Also, the when the AUC is about 50% (parameter values are equal), then the ROC curve is at the chance line.

Finally, the worst case of ROC curve is considered, where the normal and abnormal population parameter values are inverted and the corresponding AUC and the ROC curve are plotted (Figure 1 - Worst case). This is the case, where an abnormal population value tends to be higher than the normal population parameter values. This particular mathematical model of ROC curve is considered with the condition that the parameter values of normal population are to be higher side that the abnormal population ($\theta_0 \geq \theta_1$). So far, many

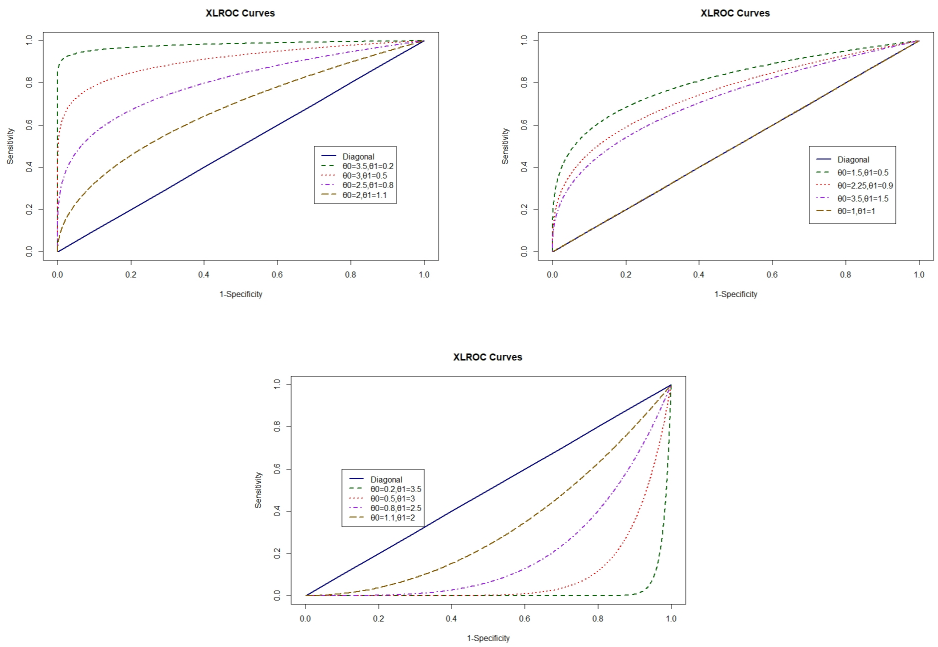


Figure 1: XL ROC curve using different simulations

models in ROC literature considered the other way, where the normal population values are at lower side than the abnormal population for example bi-normal, bi-exponential ROC curve.

3.1. Real dataset - APACHE IV

A real dataset APACHE IV (Balaswamy and Vardhan, 2023) is used to check the performance of the proposed ROC curve and the other existing ROC models like Lindley ROC (LROC) and Exponential ROC (EROC) curves (Balaswamy and Vardhan, 2021) and their corresponding AUC along with the optimal threshold values are given in Table 2.

Table 2: Comparison of different ROC curves using APACHE IV dataset

Method	θ_0	θ_1	AUC	J	Opt t (FPR, TPR)	KS test (H)	KS test (D)
XL ROC	0.2178	0.0912	0.7200	0.4400	12 (0.2022, 0.6422)	D = 0.1641 p-value = 0.0670	D = 0.1653 p-value = 0.1705
LROC	0.2464	0.1152	0.6945	0.3891	10 (0.2531, 0.6422)	D = 0.1682 p-value = 0.0477	D = 0.1413 p-value = 0.3301
EROC	0.1367	0.0607	0.6449	0.2899	10 (0.2547, 0.5446)	D = 0.0786 p-value = 0.8088	D = 0.0794 p-value = 0.9390

From Table 2, it is observed that the optimal threshold value is calculated and is reported as 12 (FPR = 0.2022 & TPR = 0.6422) and its corresponding Area under the curve for the APACHE IV dataset is 0.7200 with the youden's index value as 0.4400. This means that if the APACHE IV score of an individual who admitted to ICU is more than the optimal threshold 12, then the condition of the individual is considered as severe and belongs to the abnormal (diseased) group. The Kolmogorov-Smirnov test values are also given in the table 2 to check the validation of the data with the considered X Lindley distribution and the result shows that the data follows XL distribution. The corresponding XL ROC curve for the considered dataset is also plotted and shown in Figure 2. Further, the proposed ROC

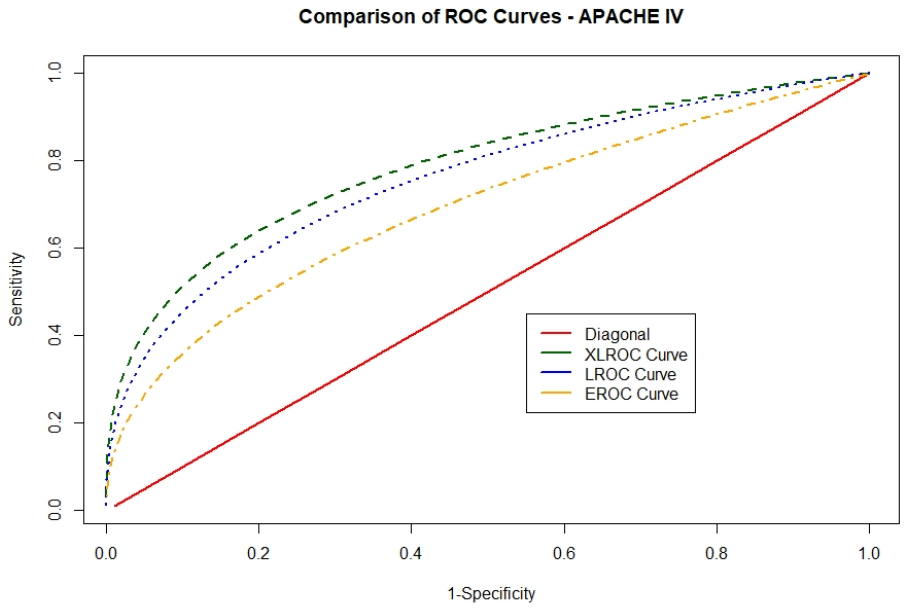


Figure 2: Comparison of ROC curves for APACHE IV dataset

curve methodology, XL ROC (AUC = 0.7200) is compared with the existing ROC models LROC (AUC = 0.6945) and EROC (AUC = 0.6449), which means that the XL ROC curve is performing better than the other models when the data follows the distribution assumed. The comparative plot is also drawn in the Figure 2, which explains that the proposed method is better than the other two models considered here.

Further, on performing the non-parametric ROC curve analysis, it is observed that the XL ROC curve methodology is found to be better with higher AUC values than the AUC of nonparametric ROC curve analysis both in simulations and real dataset. It is observed that the simulation results are also giving a higher accuracy than the non-parametric method ($At\ \theta_0 = 3.5\ \&\ \theta_1 = 0.5;\ AUC\ of\ XL\ ROC = 0.9804\ \&\ AUC\ of\ Non$-Parametric method $= 0.7435;\ At\ \theta_0 = 3\ \&\ \theta_1 = 0.5;\ AUC\ of\ XL\ ROC = 0.9060\ \&\ AUC\ of\ Non$-Parametric method $= 0.7193$).

On comparing the proposed methodology with non-parametric ROC curve, the XL ROC curve is found to be better than the non-parametric ROC method (AUC of XL ROC = 0.7200 with optimal threshold 12 and its corresponding FPR as 0.2022 & TPR as 0.6422 & AUC of Non-Parametric method = 0.6787 with optimal threshold 10 and its corresponding FPR as 0.2424, TPR as 0.6000). Therefore, the XL ROC methodology is recommended as compared to the Non-Parametric ROC method when the data follows an assumed distribution.

4. Conclusion

A Receiver Operating Characteristic (ROC) model based on the X Lindley (XL) distribution has been developed for classification tasks due to its greater flexibility compared to other single parameter distributions. The XL distribution is particularly effective for data analysis in classification contexts, offering computational simplicity that is accessible even

to non-statisticians. The mathematical properties of the XL-based ROC curve have been rigorously verified, and a likelihood ratio test has been introduced for classification purposes. Extensive simulation studies demonstrate the model's performance under various scenarios, revealing better, moderate, and poorer ROC curve shapes as the scale parameters of the healthy and diseased populations vary. This model is especially relevant when parameters of the normal (healthy) group are higher than those of the abnormal (diseased) group in classification problems.

Acknowledgements

The authors are grateful to the Editor and reviewers for their guidance, valuable comments and suggestions to improve the paper.

Conflict of interest

The authors do not have any conflict of interest to declare for the research work included in this article.

References

Arunima, S. K. and Vardhan, V. R. (2023). Estimation of area under the multiclass ROC for non-normal data, *Statistics and Applications*, **21**, 113-121.

Arunima, S. K. and Vardhan, V. R. (2022). Estimation of area under the ROC curve in the framework of gamma mixtures. *Communications in Statistics Case Studies, Data Analysis and Applications*, **8**, 714-727.

Balaswamy, S. and Vardhan, V. R. (2023). AUC estimation and ROC model comparison in the perspective of generalized exponential distribution. *International Journal of Statistics and Reliability Engineering*, **10**, 346-351.

Balaswamy, S. and Vardhan, V. R. (2015). Confidence interval estimation of an ROC curve - an application of generalised half normal and weibull distribution. *Journal of Probability and Statistics*, **ID 934362**, 1-8.

Balaswamy, S. and Vardhan, V. R. (2021). Estimation of the area under the ROC curve in the framework of lindley centered distributions. *Statistics and Applications*, **19**, 231-240.

Balaswamy, S. and Vardhan, V. R. (2022). Estimation of the area under the ROC curve for non normal data. *Communication in Statistics - Case Studies, Data Analysis and Applications*, **8**, 393-406.

Balaswamy, S. and Vardhan, V. R. (2015). Interface between the ratio β with area under the ROC curve and kullback-leibler divergence under the combination of half normal and rayleigh distributions. *American Journal of Biostatistics*, **5**, 69-77.

Balaswamy, S., Vardhan, V. R., and Sarma, K. V. S. (2015). The Hybrid ROC (HROC) curve and its divergence measures for binary classification. *International Journal of Statistics in Medical Research*, **4**, 94-102.

Dashina, P. and Vardhan, V. R. (2022). Estimation of AUC of bi-generalised exponential ROC curve and its asymptotic results. *Advances and Applications in Statistics*, **79**, 105-119.

Dorfman, D. D. and Alf, Jr. E. (1968). Maximum likelihood estimation of parameters of signal detection theory - a direct solution. *Psychometrika*, **33**, 117-124.

Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis.* Academic Press, New York.

Hussain, E. (2012). The bi-gamma ROC curve in a straightforward manner. *Journal of Basic and Applied Sciences*, **8**, 309-314.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC Curves for Continuous data*, Monographs on Statistics and Applied Probability. CRC Press, New York.

Tang, L. L. and Balakrishnan, N. (2011). A random-sum wilcoxon statistic and its application to analysis of ROC and LROC data. *Journal of Statistical Planning and Inference*, **141**, 335-344.

Zou, K. H., Hall W. J., and Shapiro, D. E. (1997). Smooth nonparametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, **16**, 2143-2156.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 24, No. 1, 2026 (New Series), pp 249–274
http://www.ssca.org.in/journal

Reliability Analysis of a Phased Mission System under Degradation using Wiener Process and Copulas

Satya Rani and Preeti Wanti Srivastava

Department of Operational Research, University of Delhi, Delhi 110007, India

Received: 20 September 2024; Revised: 06 October 2025; Accepted: 30 October 2025

Abstract

In this paper, reliability of a Phased Mission System (PMS) under internal and external degradation has been analysed. The degradation path of a component in a phase is taken as a linear combination of the internal degradation and a proportion of common external degradation that influences each component and modelled using Wiener process. The PMS is modelled using cumulative exposure model with components' dependency and that of phases modelled using different copulas: Gumbel-Hougaard, Clayton, and Frank. A numerical illustration is presented using an aircraft flight PMS with comparative study amongst different copulas carried out. Sensitivity analyses are also undertaken to determine which parameters are sensitive to small deviations from their respective true value so that extra care may be taken by an engineer while setting the true value of each of such parameters.

Key words: Phased mission systems; Copulas; Degradation; Wiener Process; Sensitivity analyses.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Various environmental and operational conditions act on real-world systems leading to their degradation, and usually culminate into the system's failure that may be costly and sometimes catastrophic. Such systems are often required to complete their tasks in several non-overlapping phases or stages in which the structure of systems change phase-wise and are addressed in the literature as phased mission systems (PMSs).

Some examples of PMSs are:

- Aircraft flight PMS (Somani *et al.* (1992) and Xing (2007)), see also Srivastava and Rani (2024),
- a satellite launching PMS, (Huang *et al.* (2019)),

Corresponding Author: Preeti Wanti Srivastava
Email: preetisrivastava.saxena@gmail.com

- a boiling water reactor PMS, (Alam *et al.* (2006)),
- Mars orbiter mission system, (Xing (2007)).
- Spaceflight tracking, telemetry and command (TT&C) system, (Yang and Wu (2014)).
- Marine transportation mission of a ship, (Eruguz *et al.* (2017)).
- A distributed computer system, (Shrestha *et al.* (2010), Wang and Trivedi (2007), Levitin *et al.* (2018)).

The changes in system configuration from phase to phase, change in failure criteria and stresses due to differing environmental conditions acting on the PMS, and phase dependencies render reliability analyses of PMSs very challenging.

Two classes of approaches to the reliability evaluation of PMSs existing in the literature are Analytical Methods and Simulation Methods.

Analytical Methods include State-Space Based Approach, Combinatorial, Modular, and Recursive Methods.

State-space models based on Markov Chains (Alam and Al-Saggaf (1986), Kim and Park (1994)), Semi-Markov process (Wu and Hillston (2015)), Bayes Network (Bing and Xiao-yue (2012),Li *et al.* (2020)), and Petri-nets (Mural *et al.* (1999), Bondavalli *et al.* (2004)) have been used to describe the system behavior of the PMS. See also Li *et al.* (2021). However, they cannot be used in large-scale PMS due to the exponential growth of the systems' state numbers.

Combinatorial methods are based on the assumption that all the components are independent, meaning no dependency exists within one phase. Four different combinatorial methods used in the literature are the Mini-component Technique, Boolean Algebraic – based binary decision diagrams (BDD), multi-valued decision diagrams (MDD), and universal generating function-based method (Xing and Amari (2008)).

However, the assumption of components' independence limits the models' applicability because dependence within and across the phases occurs in some PMSs.

Phase modular techniques, introduced by Meshkat *et al.* (2003) and Ou and Dugan (2004), combine BDD-based solutions for static modules with Markov chain-based solutions for dynamic modules. These methods are specifically designed for analyzing non-repairable, binary-state PMSs.

While the phase modular approach offers greater modeling capability than purely combinatorial solutions and higher computational efficiency than state-space-based methods, it has notable limitations. It requires a costly path enumeration of the system BDD and the calculation of joint phase module probabilities for all system-level modules in the current path. The complicated state mapping between static and dynamic phase modules adds to the computational burden. As a result, the computational overhead of existing phase modular techniques remains significant in many practical applications See also Shrestha *et al.* (2010) and Shrestha *et al.* (2009).

A **Recursive method** based on conditional probability and the branch and bound has been proposed by Levitin *et al.* (2018) to evaluate the reliability of a PMS.

However, it may not be appropriate in large-scale PMS with complex dynamic behavior (Liu *et al.* (2020)).

Simulation methods are highly versatile and flexible for modeling a system and its components' behavior.

However, they only yield approximate reliability results, which may be inadequate or unsuitable for mission-critical systems See, for example, Yang and Wu (2014) and Hu *et al.* (2021).

More recently, Mura (2021) has presented a review of stochastic modeling approaches, emphasizing phase-dependent behavior and the quantitative analysis of PMSs' dependability attributes.

Copula-based approach for reliability analyses of PMSs wherein the dependency amongst components in a phase, and dependency across the phases are modelled using copulas have been studied by Srivastava *et al.* (2022),Srivastava and Rani (2024).

Dependency among degradation mechanisms or processes has received less attention in reliability modelling of a PMS, and assume Markovian state-space model wherein estimation of lifetime depends only upon the current state. In the literature, reliability of PMS has been estimated by Si *et al.* (2015) using condition-based monitoring and the degradation data. The numerical approach used by the authors is time consuming if the number of mission phases is substantially large.

The present paper deals with reliability analysis of PMS subject to internal and external degradation using Wiener degradation model with dependencies amongst components in a phase as well as across the phases modelled using copulas. Prior to the present work, Li *et al.* (2021) have used state –space model, *viz.*, semi-Markov model for PMSs under internal degradation and external shocks and, therefore, suffer from state-space explosion problem.

Section 2 presents the methodology for evaluating the reliability of a PMS under degradation, and in Section 3, the proposed method is explained using an aircraft flight PMS. The concluding remarks have been made in the last section.

2. Model formulation

2.1. Assumptions

- The components in a phase are dependent.
- The phases are dependent with structure of the PMS varying across the phases.
- Degradation of each component follows Wiener Process (WP) $\{W(t)\ ;\ t \geq 0\}$ with some positive drift ($\mu > 0$) and diffusion parameter $\sigma^2 > 0$.
- The degradation process is independent of the mission process.

A phased mission system (PMS) may consist of multiple degrading components.

Degradation of a component in a phase is internal as well as external; the former being unexplained and intrinsic to each component and the latter that is shared by all the components due to environmental factors such as temperature, humidity, or operational factors. Li *et al.* (2011) considered component degradation as a linear combination of internal and common external degradations for a single phase.

Suppose that the components of a phase of a PMS are subject to degradation due to different internal cause and common external cause. Let $\mathcal{Y}_1(t)$, $\mathcal{Y}_2(t)$ $\ldots$ $\mathcal{Y}_m(t)$ be independent internal degradation paths and let degradation path due to external factor be denoted by $\mathcal{Z}(t)$. Also, suppose that the j^{th} component is exposed to the external degradation $\mathcal{Z}(t)$ with impact element α_j for $j = 1, 2, \ldots m$. The degradation path, $\mathcal{X}_j(t)$, of j^{th} component $(j = 1, \ldots, m)$ is taken as a linear combination of, $\mathcal{Y}_j(t)$, and, $\mathcal{Z}(t)$, which influences each component.

Wiener process and Gamma process are widely used degradation models in the literature. Wiener process (WP) is used for modelling degradation path in this paper.

It is assumed that internal degradation path as well as external degradation path of a component follows Wiener Process. Thus,

$$\mathcal{Y}_j(t) = \mu_j + \sigma_j W^1(t), \tag{1}$$

where $W^1(t)$ is a standard Wiener Process

$$\Longrightarrow \mathcal{Y}_j(t) \sim N\left(\mu_j, \sigma_j^2\right), \tag{2}$$

and

$$\alpha_j \mathcal{Z}(t) = \alpha_j \mu + \alpha_j \sigma W^1(t), \tag{3}$$

$$\Longrightarrow \alpha_j \mathcal{Z}(t) \sim N\left(\alpha_j \mu, \alpha_j^2 \sigma^2\right), \tag{4}$$

$\mathcal{Y}_j(t)$and $\mathcal{Z}(t)$ are both linear function of t.

Using (1) and (3)

$$\mathcal{X}_j(t) = \mathcal{Y}_j(t) + \alpha_j \mathcal{Z}(t); \qquad j = 1, 2, \ldots, m, \tag{5}$$

where

$$\mathcal{X}_j(t) \sim N\left(\mu_j + \alpha_j \mu, \sigma_j^2 + \alpha_j^2 \sigma^2\right). \tag{6}$$

Let

$$\eta_j = \mu_j + \alpha_j \mu, \text{and } \delta_j^2 = \sigma_j^2 + \alpha_j^2 \sigma^2, \tag{7}$$

then

$$\mathcal{X}_j(t) \sim N\left(\eta_j, \delta_j^2\right). \tag{8}$$

2.2. First passage time distribution of WP

Let $y(t)$ be performance degradation measure at time t defined as:

$$y\left(t\right) = \ \gamma t \ + \ \lambda W^{1}\left(t\right), \ (\gamma \ , \ \lambda) > \ 0. \tag{9}$$

Then, $y(t)$ is a WP with drift γ and diffusion constant λ^2.

When the degradation path follows WP, then the time when the degradation level first reaches a fixed failure critical level, w, has an inverse Gaussian (IG) distribution with reliability function (Chhikara and Folks (1989)):

$$R\left(t\right) = \phi\left[\frac{w-\gamma.t-w_0}{\lambda\sqrt{t}}\right] - exp\left(\frac{2.\gamma.\left(w-w_0\right)}{\lambda^2}\right)\phi\left[-\frac{w+\lambda.t-w_0}{\lambda\sqrt{t}}\right]; (\gamma \ , \lambda) \ > \ 0, \tag{10}$$

where w_0 is the initial degradation level at time zero.

2.3. Copulas

In this paper, copula-based approach is used to model dependency amongst the components in a phase, and also across the phases. The dependence structure relates the first passage time distribution of the Wiener degradation path depicted by the components within the phase to their multivariate distribution. The choice of copula determines the underlying dependence structure (Nelsen (2006)).

Some types of copula functions existing in the literature are **Clayton copula, Frank copula, Gaussian copula, Gumbel-Hougaard (GH) copula, and Student's t-copula**.

In this work ***Gumbel-Hougaard copula, Clayton copula, and Frank copula*** have been used.

Let $X_1, \ X_2, \ldots, \ X_n$ be random variables and let $\overline{\mathcal{G}}_1\left(x_1\right), \ \overline{\mathcal{G}}_2\left(x_2\right), \ldots, \ \overline{\mathcal{G}}_n\left(x_n\right)$ be their respective marginal reliability functions. Further, let $\overline{\mathcal{G}}\left(x_1, x_2, \ldots, x_n\right)$ denote their corresponding joint reliability function. Then, Sklar's Theorem (Nelsen (2006)) states that, $\exists$ a copula reliability function $C(\cdot,\cdot,\cdot)$ s.t $\forall$ $(X_1, \ X_2, \ldots, \ X_n)$ in the defined range:

$$\overline{\mathcal{G}}\left(x_1, x_2, \ldots, x_n\right) = C\left(\overline{\mathcal{G}}_1\left(x_1\right), \overline{\mathcal{G}}_2\left(x_2\right), \ldots, \overline{\mathcal{G}}_n\left(x_n\right)\right), \tag{11}$$

The copulas used in this work are:

(i) Gumbel-Hougaard (GH) copula

$$C\left(\overline{\mathcal{G}}_1\left(x_1\right), \overline{\mathcal{G}}_2\left(x_2\right), \ldots, \overline{\mathcal{G}}_n\left(x_n\right)\right) =$$
$$exp\left[-\left(\left(-log\left(\overline{\mathcal{G}}_1\left(x_1\right)\right)\right)^{\theta} + \left(-log\left(\overline{\mathcal{G}}_2\left(x_2\right)\right)\right)^{\theta} + \cdots + \left(-log\left(\overline{\mathcal{G}}_n\left(x_n\right)\right)\right)^{\theta}\right)^{1/\theta}\right], \tag{12}$$

where $\theta \epsilon \left[1, \infty\right)$

(ii) Clayton copula,

$$C\left(\overline{\mathcal{G}}_1(x_1), \overline{\mathcal{G}}_2(x_2), \ldots, \overline{\mathcal{G}}_n(x_n)\right) = \left(\left(\overline{\mathcal{G}}_1(x_1)\right)^{-\theta} + \left(\overline{\mathcal{G}}_2(x_2)\right)^{-\theta} + \cdots + \left(\overline{\mathcal{G}}_n(x_n)\right)^{-\theta} - n + 1\right)^{-\frac{1}{\theta}}, \quad (13)$$

where $\theta \epsilon (0, \infty)$

(iii) Frank copula

$$C\left(\overline{\mathcal{G}}_1(x_1), \overline{\mathcal{G}}_2(x_2), \ldots, \overline{\mathcal{G}}_n(x_n)\right) = -\frac{1}{\theta} ln\left[1 + \frac{\left(e^{-\theta.\overline{\mathcal{G}}_1(x_1)} - 1\right)\left(e^{-\theta.\overline{\mathcal{G}}_2(x_2)} - 1\right)\ldots\left(e^{-\theta.\overline{\mathcal{G}}_n(x_n)} - 1\right)}{\left(e^{-\theta} - 1\right)^{n-1}}\right], \quad (14)$$

where $\theta \epsilon \mathrm{R} \backslash \{0\}$.

(12)-(14) belong to the Archimedean class of copulas. The Archimedean copula class is easy to work with from a mathematical point of view (Wilson and Y. (2005)), Chapter 8, p.- 114). These copulas find a wide range of applications for several reasons, such as the ease with which they can be constructed and the variety of families of copulas which belong to the class (see Nelsen (2006), Chapter 4). n-dimensional GH copula with each marginal as Weibull is n-variate Weibull distribution (see Lee and Wen (2006)) which is widely used in manufacturing industries.

2.4. Computation of reliability of a phased-mission system

Let T_j be the lifetime of component j, $j = 1, 2, \ldots, m$ of phase i, $i = 1, 2, \ldots, n$ following inverse Gaussian (IG) distribution with parameters γ (> 0) and λ (> 0), and

$$\overline{\mathcal{G}}_{ji}(t) = R_{ji}(t), \quad (15)$$

be its reliability function, where $R_{ji}(t)$ can be obtain from (10) on replacing the failure threshold w by w_j, γ by η_j, and λ by δ_j^2.

Let $\overline{F}_{p1}(t), \overline{F}_{p2}(t), \ldots$ and $\overline{F}_{pn}(t)$ be the respective reliability of phases 1, 2, ..., n, respectively. Then, reliability of PMS is:

$$\overline{F}_{PMS}(t) = \begin{cases} \overline{F}_{p1}(t), & 0 \le t \le t_1 \\ \overline{F}_{p2}(t), & t_1 \le t \le t_2 \\ \vdots & \\ \overline{F}_{pn}(t), & t_{n-1} \le t \le t_n \end{cases}, \quad (16)$$

where $[t_{i-1}, t_i]$ represents time-duration of functioning of phase i of the phased mission system $i = 1, 2, 3, 4, \ldots, n, \quad t_0 = 0$.

Let i^{th} phase be composed of m dependent components and reliability of i^{th} phase be denoted by $\overline{F}_{pi}(t)$ then dependency is modelled using Gumbel-Haugaard, Clayton, and Frank copula giving:

$$\overline{F}_{pi}(t) = C\left(\overline{\mathcal{G}}_{1i}(t), \overline{\mathcal{G}}_{2i}(t), \ldots, \overline{\mathcal{G}}_{mi}(t)\right), \tag{17}$$

and, reliability of PMS:

$$\overline{F}_{PMS}(t) = C\left(\overline{F}_{p1}(t_1), \overline{F}_{p2}(t_2), \overline{F}_{p3}(t_3), \ldots, \overline{F}_{pn}(t_n)\right). \tag{18}$$

The cumulative exposure model Nelson (2009) is used in equation (17), to obtain the reliability of i^{th} phase at t_i. Thus,

$$\overline{F}_{pi}(t_i) = C\left(\overline{\mathcal{G}}_{1i}(t_i - t_{i-1} + l_{1i}), \overline{\mathcal{G}}_{2i}(t_i - t_{i-1} + l_{2i}), \ldots, \overline{\mathcal{G}}_{mi}(t_i - t_{i-1} + l_{mi})\right). \tag{19}$$

where for the component j in the phase i of the PMS, $i = 1, .., n, \ \& \ j = 1, .., m$, l_{ji} is determined in such a way that (Srivastava *et al.* (2022))

$\overline{\mathcal{G}}_{ji}(l_{ji}) = \overline{\mathcal{G}}_{ji-1}(t_{i-1} - t_{i-2} + l_{ji-1})$, and $l_{j1-1} = 0$.

3. Numerical illustration

An Aircraft Flight PMS has been used to explain the model proposed.

Consider an aircraft flight from departure A to destination B and vice versa. The four-phase aircraft flight PMS are depicted in Figure 1 (see also (Srivastava and Rani (2024)).

The taxiing phase takes around 10-20 minutes, depending on airport traffic and distance from the gate to the runaway. The take-off rolls and initial climb to cruising altitude can take about 10-20 minutes. The cruising phase, where the plane is at its highest altitude and flying in a straight line, might last for about 2 hours and 20 minutes. The conditions during landing would be similar to those at the arrival airport B. The decent and landing process can take around 20-30 minutes.

Thus, the following data set has been used:

The duration of

- Taxiing and Take-Off phases has been taken to be 20 minutes,
- cruising phase is 140 minutes, and
- The landing phase is 20 minutes.

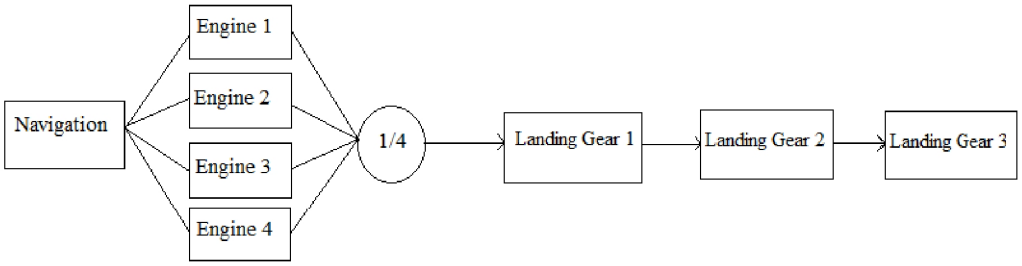


Figure 1(a): *Taxiing Phase*

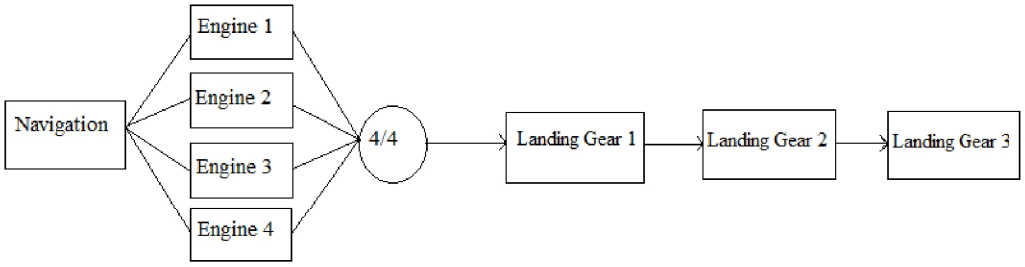


Figure 1(b): *Take-Off Phase*

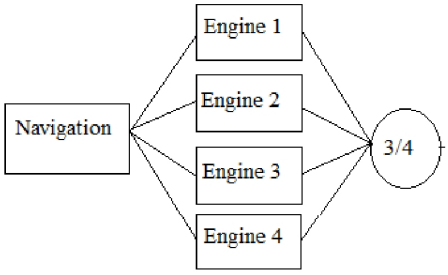


Figure 1(c): *Cruising Phase*

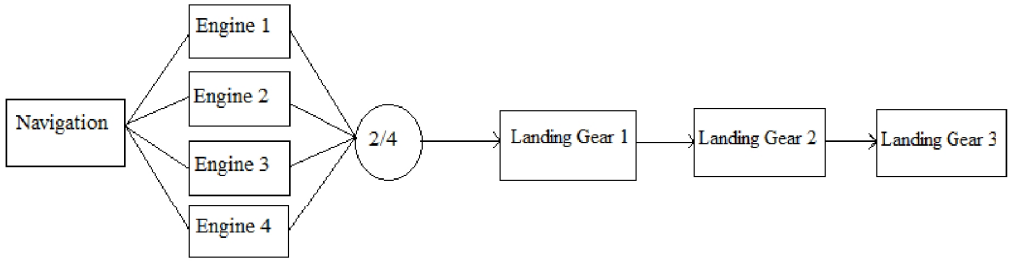


Figure 1(d): *Landing Phase*

Figure 1: 1(a)-(d) Reliability block diagrams for the four- phase aircraft flight

The hypothetical values of the parameters for the eight components, namely, Navigation, Engine 1, 2, 3, and 4; landing gear 1, 2, and 3 are taken as:

$$\mu_1 = 0.6, \quad \mu_2 = \mu_3 = \quad \mu_4 = \quad \mu_5 = 0.5, \ \mu_6 = \mu_7 = \mu_8 = 0.7, \ \mu = 1.4,$$

$\alpha_1 = 0.4, \ \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.5$, $\alpha_6 = \alpha_7 = \alpha_8 = 0.7,$ and

$\sigma_1 = 1.4, \quad \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5 = 1.3$, $\quad \sigma_6 = \sigma_7 = \sigma_8 = 1.2, \quad \sigma = 1.1,$ respectively in (7).

Degradation levels for the three components: Navigation, Engine and landing gear have been taken as

$w_1 = 19,\ w_2 = 18,$ and $w_3 = 17,\ \ w_0 = 1$, copula parameter $\theta = 2.17$.

Table 1 gives reliability of Aircraft Flight PMS for Gumble-Hougaard, Clayton, and Frank copulas. Figure 2 obtained using Table 1 depicts reliability of Aircraft Flight PMS for these three copulas. It is quite evident from the Table and its corresponding Figure that although there is no copula that performs the best in all the phases, the reliability of the PMS is highest for Gumble-Hougaard copula for the hypothetical data set used.

Table 1: Reliability of aircraft flight PMS using Gumbel-Hougaard, Clayton, and Frank copulas

Copula	Phase 1	Phase 2	Phase 3	Phase 4	PMS
Gumble-Hougaard	1	1	0.9997517	0.9859937	0.9859927
Clayton	1	1	0.999883	0.9765497	0.9765496
Frank	1	1	0.9998831	0.9764228	0.9763125

3.1. Sensitivity analysis

Sensitivity analysis is carried out on reliability of PMS using percentage deviation (PD) of the reliability of PMS given as:

$$\left(\frac{|R - R^*|}{R}\right) \times 100,$$

where R is Reliability of PMS obtained with the given design parameters, and R^* is the value of reliability obtained using mis-specified value. $|.|$ denote the absolute value.

Sensitivity analysis is carried out on reliability of PMS with

$(\pm 1\%,\ \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%,\ \pm 30\%,\ \pm 40\%,\ \pm 50\%)$ deviations in:

(i) copula parameter θ keeping other parameters fixed,

(ii) degradation level w_1, w_2, and w_3 of Nevigation, Engines and Landing gears, respectively, with other parameters fixed,

(iii) Parameters, $\mu,\ \sigma$, associated with external degradation,

(iv) impact element $\alpha_j,\ j = 1, 2, \ldots, 8$ of external degradation,

using Gumbel-Hougaard, Clayton, and Frank copulas.

For Gumble-Hougaard copula, Table 2 and Figure 3 show the results for (i), Table 3 and Figure 4 show the results for (ii), Table 4 and Figure 5 show the results for (iii), Table 5 and Figure 6 show the results for (iv). For Clayton copula, Table 6 and Figure 7 show the results for (i), Table 7 and Figure 8 show the results for (ii), Table 8 and Figure 9 show the results for (iii), Table 9 and Figure 10 show the results for (iv). For Frank copula, Table 10

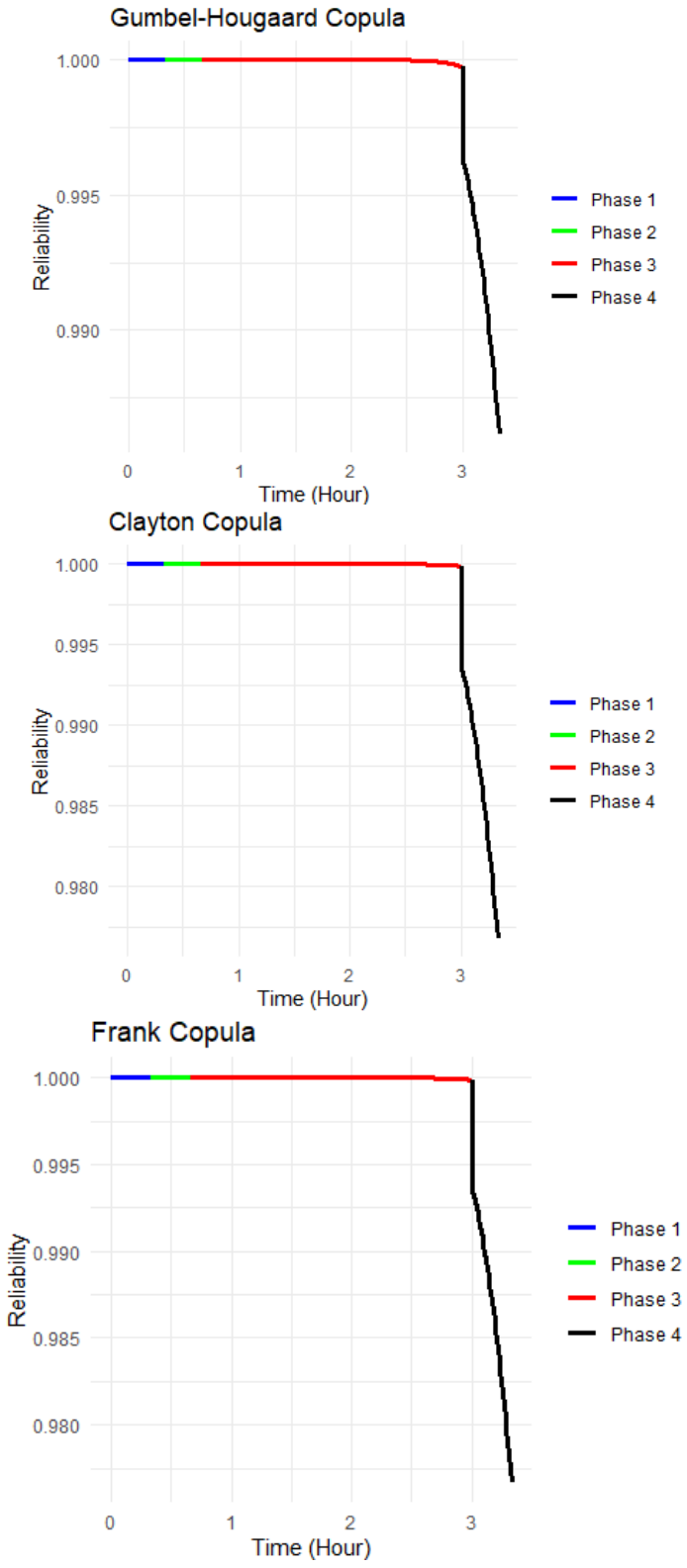


Figure 2: Reliability of aircraft flight PMS using Gumbel-Hougaard, Clayton, and Frank copulas

and Figure 11 show the results for (i), Table 11 and Figure 12 show the results for (ii), Table 12 and Figure 13 show the results for (iii), Table 13 and Figure 14 show the results for (iv).

Table 2: Sensitivity analysis on reliability of aircraft flight PMS w.r.t $(\pm 1\%,\ \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%,\ \pm 30\%,\ \pm 40\%,\ \pm 50\%)$ **deviations in** $\theta = 2.17$ **using Gumbel-Hougaard copula** $w_1 = 19, w_2 = 18, w_3 = 17, \mu = 1.4, \sigma = 1.1, \alpha_1 = 0.4, \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.5, \alpha_6 = \alpha_7 = \alpha_8 = 0.7$

Initial values of copula parameter θ =2.17	True reliability (R)	Obtained reliability (R*)	Percentage deviation (PD) %
+1%	0.9859927	0.9860537	0.006179741
-1%	0.9859927	0.9859302	0.006336702
+ 5 %	0.9859927	0.986283	0.02943944
-5 %	0.9859927	0.9856636	0.03337802
+10 %	0.9859927	0.9865409	0.05559442
- 10 %	0.9859927	0.9852874	0.07153325
+15 %	0.9859927	0.9867715	0.07898237
-15%	0.9859927	0.9848533	0.1155581
+20%	0.9859927	0.9869789	0.1000179
-20%	0.9859927	0.9843471	0.166901
+30%	0.9859927	0.9873368	0.1363148
-30%	0.9859927	0.9830335	0.3001304
+40%	0.9859927	0.9876346	0.1665184
-40%	0.9859927	0.9810778	0.4984735
+50%	0.9859927	0.9878862	0.1920378
-50%	0.9859927	0.9778959	0.8211879

For the Gumbel-Hougaard copula, Tables 2, 3, 4, and 5 indicate that the reliability function is not sensitive to small deviations from the true parameter values.

For the Clayton copula, Tables 6, 7, 8, and 9 indicate that the reliability function is not sensitive to small deviations from the true parameter values.

For Frank copula Tables 10, 11, 12, and 13 indicate that the reliability function is not sensitive to small deviations from the true parameter values.

Further, it can be observed from Table 2 and Figure 3, Table 6 and Figure 7, and Table 10 and Figure 11 that for the Gumbel-Hougaard, Clayton, and Frank copulas, respectively, as the value of copula parameter, θ, increases the reliability of aircraft flight PMS increase, and decreasing the value of parameter, θ, results in decrease in the reliability of aircraft flight PMS, when other parameters are held constant.

Table 3 and Figure 4, Table 7 and Figure 8, and Table 11 and Figure 12, corresponding to the Gumbel-Hougaard, Clayton, and Frank copulas, respectively, show that an increase in the degradation level of components, w_i, results in an increase in the reliability of the aircraft flight PMS. Conversely, a decrease in, w_i, leads to a decrease in reliability.

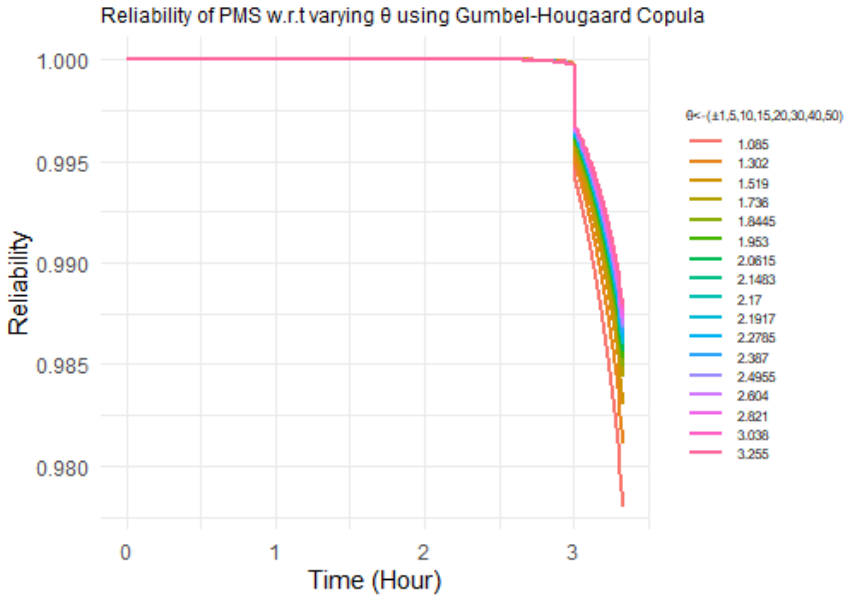


Figure 3: Reliability of aircraft flight PMS w.r.t $(\pm 1\%,\ \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in** $\theta = 2.17$ **using Gumbel-Hougaard copula**

Table 3: Sensitivity analysis on reliability of aircraft flight PMS w.r.t $(\pm 1\%,\ \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$, **deviations in** $w_1 = 19,\ w_2 = 18,\ w_3 = 17$ **using Gumbel-Hougaard copula** $\theta = 2.17, \mu = 1.4, \sigma = 1.1, \alpha_1 = 0.4, \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.5, \alpha_6 = \alpha_7 = \alpha_8 = 0.7$

Degradation level initial values $w_1 = 19, w_2 = 18, w_3 = 17$	True reliability (R)	Obtained reliability (R*)	Percentage deviation (PD) %
+1%	0.9859927	0.9878902	0.1924424
-1%	0.9859927	0.9838344	0.218895
+ 5 %	0.9859927	0.9933866	0.7498937
-5 %	0.9859927	0.9719673	1.422466
+10 %	0.9859927	0.99705	1.121439
- 10 %	0.9859927	0.9470123	3.953421
+15 %	0.9859927	0.9987573	1.294592
- 15 %	0.9859927	0.9055028	8.163342
+20%	0.9859927	0.9995059	1.370514
-20%	0.9859927	0.8413032	14.6745
+30%	0.9859927	0.9999345	1.413982
-30%	0.9859927	0.4639339	52.94753
+40%	0.9859927	0.9999932	1.419932
-40%	0.9859927	0.1378241	86.02179
+50%	0.9859927	0.9999994	1.42057
-50%	0.9859927	0.02954872	97.00315

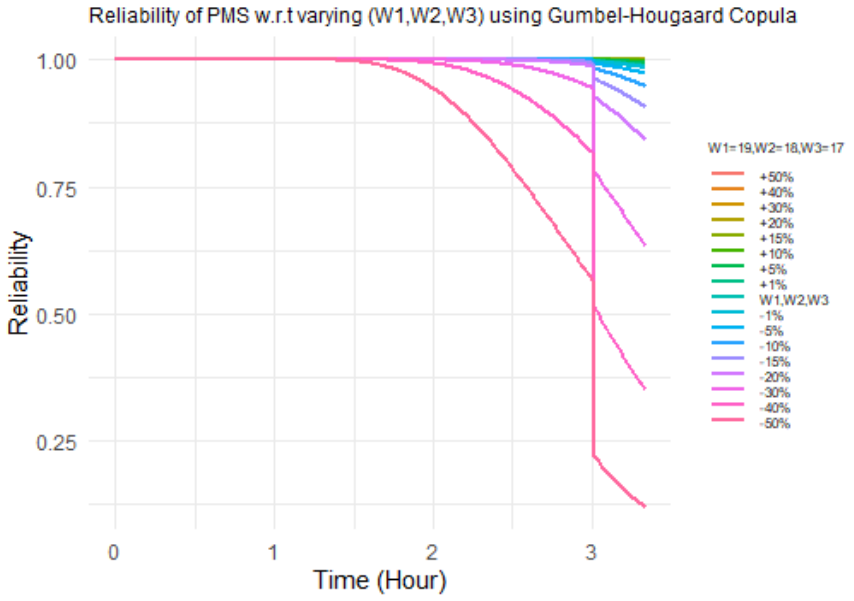


Figure 4: Reliability of aircraft flight PMS w.r.t $(\pm 1\%,\ \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in** $w_1 = 19,\ w_2 = 18,\ w_3 = 17$ **using Gumbel-Hougaard copula**

Table 4: Sensitivity analysis on reliability of aircraft flight PMS w.r.t $(\pm 1\%,\ \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in external degradation parameters** $\mu = 1.4,\ \sigma = 1.1$ **using Gumbel-Hougaard copula** $\theta = 2.17, w_1 = 19, w_2 = 18, w_3 = 17, \alpha_1 = 0.4, \alpha_2 = \alpha_3 = \ \alpha_4 = \alpha_5 = 0.5, \alpha_6 = \alpha_7 = \alpha_8 = 0.7$

Initial values of external degradation parameters $\mu = 1.4,\ \sigma = 1.1$	True reliability (R)	Obtained reliability (R*)	Percentage deviation (PD) %
+1%	0.9859927	0.9849241	0.1083774
-1%	0.9859927	0.9869985	0.1020064
+ 5 %	0.9859927	0.9799717	0.6106535
-5 %	0.9859927	0.990443	0.4513527
+10 %	0.9859927	0.972077	1.411342
- 10 %	0.9859927	0.9936299	0.7745633
+15 %	0.9859927	0.9620341	2.429903
- 15 %	0.9859927	0.9958411	0.9988327
+20	0.9859927	0.9496229	3.688652
-20	0.9859927	0.9973298	1.149814
+30	0.9859927	0.9172041	6.976582
-30	0.9859927	0.9989263	1.311736
+40	0.9859927	0.7950699	19.36351
-40	0.9859927	0.9995567	1.375664
+50	0.9859927	0.7333875	25.61938
-50	0.9859927	0.9997959	1.399928

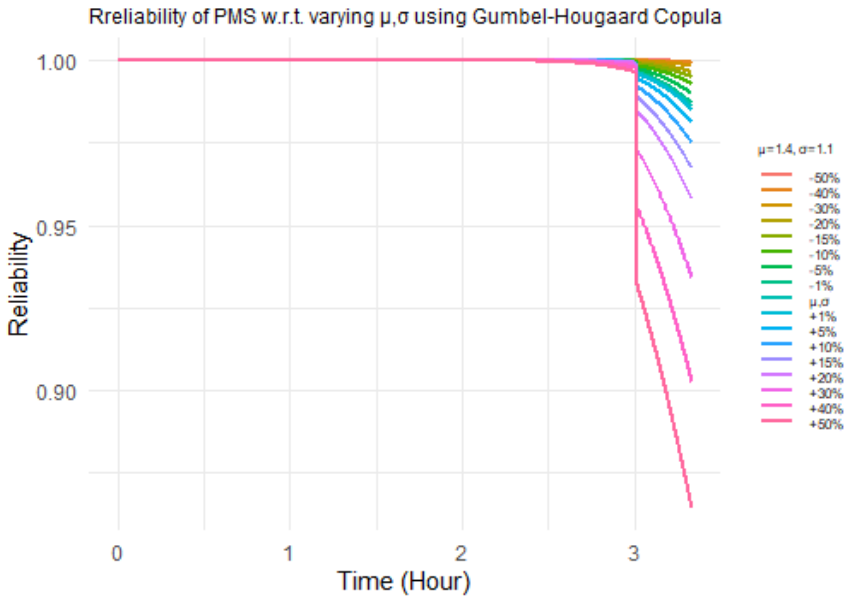


Figure 5: Reliability of aircraft flight PMS w.r.t $(\pm 1\%,\ \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in** $\mu = 1.4,\ \sigma = 1.1$**, using Gumbel-Hougaard copula**

Table 5: Sensitivity analysis on reliability of aircraft flight PMS w.r.t $(\pm 1\%,\ \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in impact element of external degradation** $\alpha_1 = 0.4, \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.5, \alpha_6 = \alpha_7 = \alpha_8 = 0.7$ **using Gumbel-Hougaard copula** $\theta = 2.17, w_1 = 19, w_2 = 18, w_3 = 17, \mu = 1.4, \sigma = 1.1$

Initial values of common factors $\alpha_1 = 0.4, \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.5, \alpha_6 = \alpha_7 = \alpha_8 = 0.7$	True reliability (R)	Obtained reliability (R*)	Percentage deviation (PD) %
+1%	0.9859927	0.9850911	0.0914448
-1%	0.9859927	0.9868494	0.08688919
+ 5 %	0.9859927	0.9810052	0.5058418
-5 %	0.9859927	0.9898571	0.3919228
+10 %	0.9859927	0.97471	1.144305
- 10 %	0.9859927	0.9927842	0.6887922
+15 %	0.9859927	0.9669341	1.932937
- 15 %	0.9859927	0.9949515	0.9086028
+20	0.9859927	0.9575276	2.886954
-20	0.9859927	0.9965205	1.06773
+30	0.9859927	0.9333932	5.334674
-30	0.9859927	0.9984033	1.258685
+40	0.9859927	0.9018792	8.530849
-40	0.9859927	0.9992809	1.347693
+50	0.9859927	0.7776503	21.13022
-50	0.9859927	0.9996672	1.386876

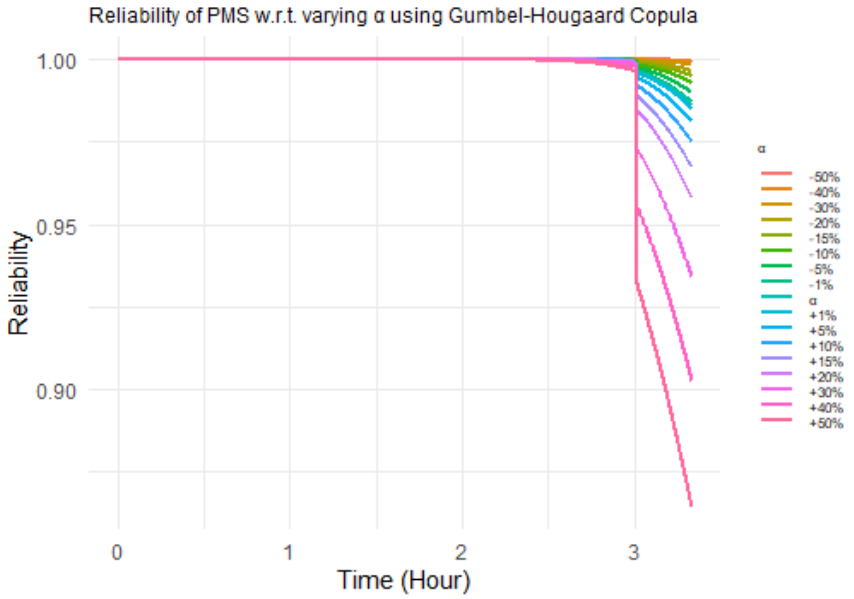


Figure 6: Reliability of aircraft flight PMS w.r.t $(\pm 1\%, \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in** $\alpha_1 = 0.4, \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.5, \alpha_6 = \alpha_7 = \alpha_8 = 0.7$ **using Gumbel-Hougaard copula**

Table 6: Sensitivity analysis on reliability of aircraft flight PMS w.r.t $(\pm 1\%, \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in** $\theta = 2.17$ **using Clayton copula.** $w_1 = 19, w_2 = 18, w_3 = 17, \mu = 1.4, \sigma = 1.1, \alpha_1 = 0.4, \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.5, \alpha_6 = \alpha_7 = \alpha_8 = 0.7$

Initial values of copula parameter θ =2.17	True reliability (R)	Obtained reliability (R*)	Percentage deviation (PD) %
+1%	0.9765496	0.9765533	0.0003794482
-1%	0.9765496	0.9765459	0.0003798237
+ 5 %	0.9765496	0.9765681	0.00189386
-5 %	0.9765496	0.976531	0.001903387
+10 %	0.9765496	0.9765865	0.003780273
- 10 %	0.9765496	0.9765123	0.00382023
+15 %	0.9765496	0.9766049	0.005660352
-15%	0.9765496	0.9764934	0.005757995
+20%	0.9765496	0.9766232	0.007534697
-20%	0.9765496	0.9764741	0.00773097
+30%	0.9765496	0.9766597	0.01126736
-30%	0.9765496	0.9764332	0.01191837
+40%	0.9765496	0.9766959	0.01497956
-40%	0.9765496	0.9763828	0.01708787
+50%	0.9765496	0.976732	0.01867179
-50%	0.9765496	0.9762956	0.02601556

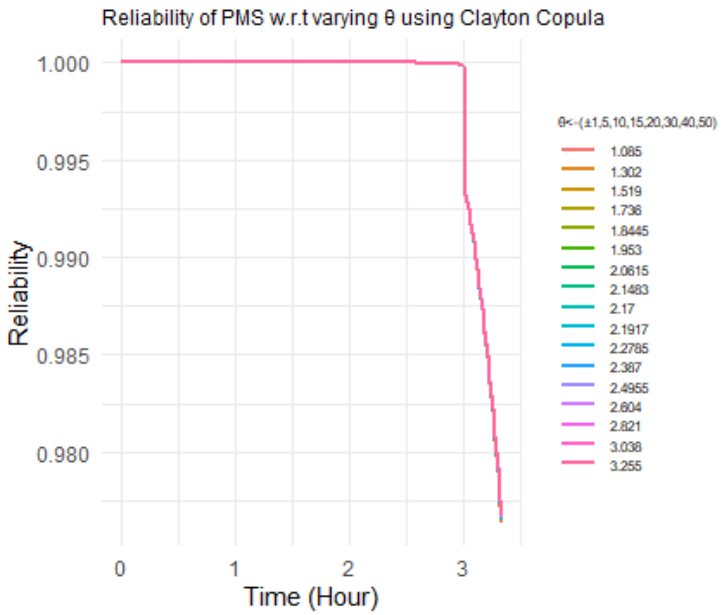


Figure 7: Reliability of aircraft flight PMS w.r.t $(\pm 1\%, \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in** $\theta = 2.17$ **using Gumbel-Hougaard copula**

Table 7: Sensitivity analysis on reliability of aircraft flight PMS w.r.t $(\pm 1\%, \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in** $w_1 = 19, w_2 = 18, w_3 = 17$ **using Clayton copula** $\theta = 2.17, \mu = 1.4, \sigma = 1.1, \alpha_1 = 0.4, \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.5, \alpha_6 = \alpha_7 = \alpha_8 = 0.7$

Degradation level initial values $w_1 = 19$, $w_2 = 18$, $w_3 = 17$	True reliability (R)	Obtained reliability (R*)	Percentage deviation (PD) %
+1%	0.9765496	0.9796191	0.3143206
-1%	0.9765496	0.9730903	0.3542398
+ 5 %	0.9765496	0.9886741	1.241559
-5 %	0.9765496	0.9546094	2.246707
+10 %	0.9765496	0.9948695	1.875979
- 10 %	0.9765496	0.9181977	5.975312
+15 %	0.9765496	0.9978156	2.177667
- 15 %	0.9765496	0.8632117	11.60595
+20%	0.9765496	0.9991249	2.311739
-20%	0.9765496	0.7878829	19.31972
+30%	0.9765496	0.9998828	2.389354
-30%	0.9765496	0.4396652	54.97769
+40%	0.9765496	0.9999877	2.40009
-40%	0.9765496	0.1508246	84.55536
+50%	0.9765496	0.999999	2.401247
-50%	0.9765496	0.04002497	95.90139

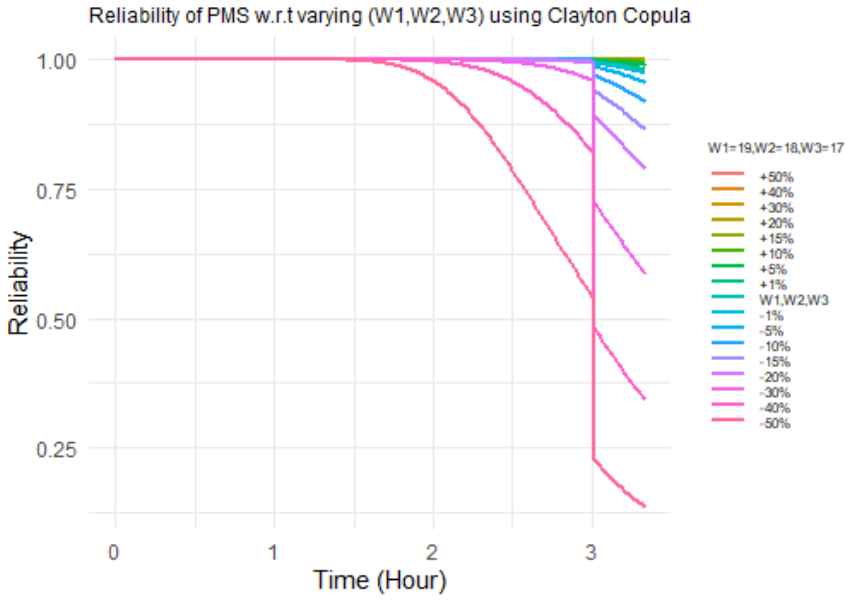


Figure 8: Reliability of aircraft flight PMS w.r.t $(\pm 1\%,\ \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in** $w_1 = 19, w_2 = 18, w_3 = 17$ **using Clayton copula**

Table 8: Sensitivity analysis on reliability of aircraft flight PMS w.r.t $(\pm 1\%,\ \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in external degradation parameters** $\mu = 1.4, \sigma = 1.1$**using Clayton copula** $\theta = 2.17, w_1 = 19, w_2 = 18, w_3 = 17, \alpha_1 = 0.4, \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.5, \alpha_6 = \alpha_7 = \alpha_8 = 0.7$

Initial values of external degradation parameters $\mu = 1.4, \sigma = 1.1$	True reliability (R)	Obtained reliability (R*)	Percentage deviation (PD) %
+1%	0.9765496	0.9746644	0.1819646
-1%	0.9765496	0.9781185	0.1717792
+ 5 %	0.9765496	0.9665029	1.017805
-5 %	0.9765496	0.9838988	0.7637562
+10 %	0.9765496	0.9537403	2.32486
- 10 %	0.9765496	0.9892922	1.31611
+15 %	0.9765496	0.9379365	3.943372
- 15 %	0.9765496	0.9930545	1.701409
+20	0.9765496	0.9190432	5.878283
-20	0.9765496	0.9955899	1.961072
+30	0.9765496	0.8727034	10.62407
-30	0.9765496	0.9982945	2.238055
+40	0.9765496	0.7229708	25.95859
-40	0.9765496	0.9993404	2.345168
+50	0.9765496	0.6612145	32.28323
-50	0.9765496	0.9997218	2.384225

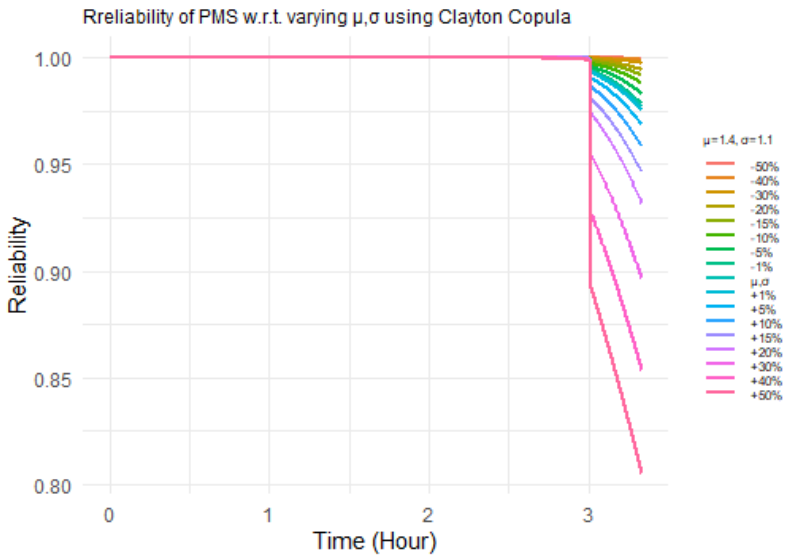


Figure 9: Reliability of aircraft flight PMS w.r.t $(\pm 1\%,\ \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in** $\mu = 1.4, \sigma = 1.1$ **using Clayton copula**

Table 9: Sensitivity analysis on reliability of aircraft flight PMS w.r.t $(\pm 1\%,\ \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in impact element of external degradation** $\alpha_1 = 0.4, \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.5, \alpha_6 = \alpha_7 = \alpha_8 = 0.7$ **using Clayton copula** $\theta = 2.17, w_1 = 19, w_2 = 18, w_3 = 17, \mu = 1.4, \sigma = 1.1$

Initial values of common factors $\alpha_1 = 0.4, \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.5, \alpha_6 = \alpha_7 = \alpha_8 = 0.7$	True reliability (R)	Obtained reliability (R*)	Percentage deviation (PD) %
+1%	0.9765496	0.9749421	0.1535289
-1%	0.9765496	0.9778692	0.1462483
+ 5 %	0.9765496	0.9681982	0.8441897
-5 %	0.9765496	0.9829097	0.6624599
+10 %	0.9765496	0.9579671	1.891978
- 10 %	0.9765496	0.9878536	1.168775
+15 %	0.9765496	0.9455936	3.159183
- 15 %	0.9765496	0.9915347	1.545763
+20 %	0.9765496	0.9310032	4.653427
-20 %	0.9765496	0.9942061	1.819356
+30 %	0.9765496	0.8306779	8.306779
-30 %	0.9765496	0.9974064	2.147105
+40 %	0.9765496	0.8521386	12.73017
-40 %	0.9765496	0.998881	2.298122
+50 %	0.9765496	0.7047044	27.8293
-50 %	0.9765496	0.9995144	2.362989

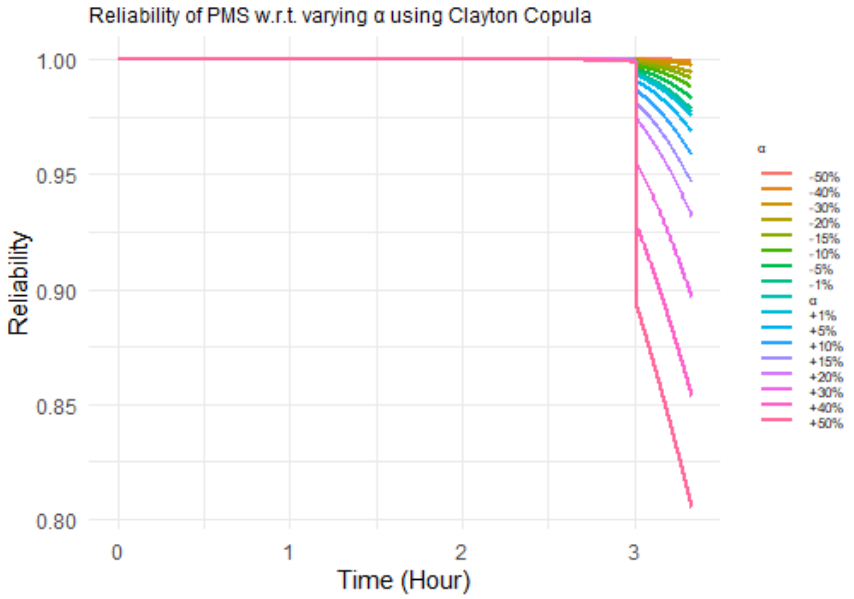


Figure 10: Reliability of aircraft flight PMS w.r.t $(\pm 1\%, \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in** $\alpha_1 = 0.4, \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.5, \alpha_6 = \alpha_7 = \alpha_8 = 0.7$ **using Clayton copula**

Table 10: Sensitivity analysis on reliability of aircraft flight PMS w.r.t $(\pm 1\%, \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in** $\theta = 2.17$ **using Frank copula** $w_1 = 19, w_2 = 18, w_3 = 17, \mu = 1.4, \sigma = 1.1, \alpha_1 = 0.4, \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.5, \alpha_6 = \alpha_7 = \alpha_8 = 0.7$

Initial values of copula parameter $\theta = 2.17$	True reliability (R)	Obtained reliability (R*)	Percentage deviation (PD) %
+1%	0.9763125	0.9763156	0.0003141117
-1%	0.9763125	0.9763094	0.0003133606
+ 5 %	0.9763125	0.9763279	0.001577897
-5 %	0.9763125	0.9762973	0.001559119
+10 %	0.9763125	0.9763435	0.003173496
- 10 %	0.9763125	0.9762822	0.003098377
+15 %	0.9763125	0.9763592	0.004785734
-15%	0.9763125	0.9762674	0.004616689
+20 %	0.9763125	0.9763751	0.006413563
-20 %	0.9763125	0.9762528	0.006112972
+30 %	0.9763125	0.9764073	0.009711903
-30 %	0.9763125	0.9762243	0.009035165
+40 %	0.9763125	0.97644	0.01306057
-40 %	0.9763125	0.9761967	0.01185655
+50 %	0.9763125	0.9764731	0.01645208
-50 %	0.9763125	0.9761703	0.01456915

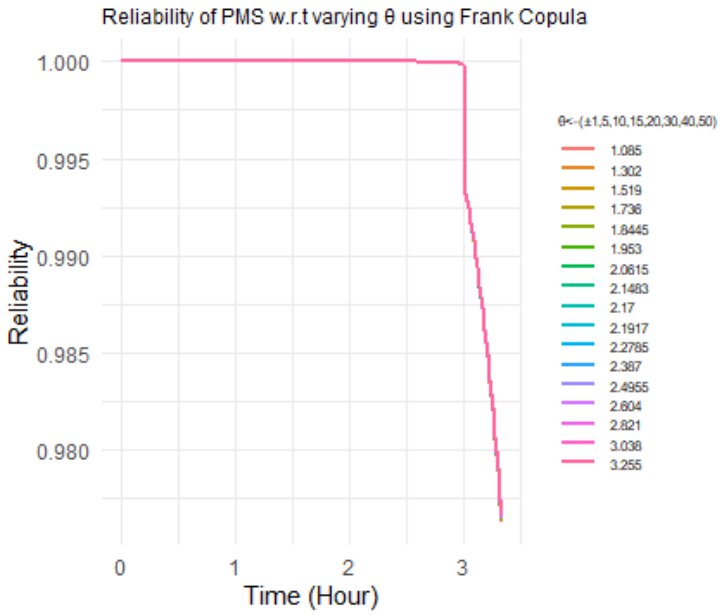


Figure 11: Reliability of aircraft flight PMS w.r.t $(\pm 1\%, \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in** $\theta = 2.17$ **using Clayton copula**

Table 11: Sensitivity analysis on reliability of aircraft flight PMS w.r.t $(\pm 1\%, \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in** $w_1 = 19, w_2 = 18, w_3 = 17$ **using Frank copula** $\theta = 2.17, \mu = 1.4, \sigma = 1.1, \alpha_1 = 0.4, \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.5, \alpha_6 = \alpha_7 = \alpha_8 = 0.7$

Degradation level initial values $w_1 = 19, w_2 = 18, w_3 = 17$	True reliability (R)	Obtained reliability (R*)	Percentage deviation (PD) %
+1%	0.9763125	0.9794347	0.3197983
-1%	0.9763125	0.9727865	0.3611527
+ 5 %	0.9763125	0.988609	1.259486
-5 %	0.9763125	0.9538239	2.303418
+10 %	0.9763125	0.994853	1.899037
- 10 %	0.9763125	0.9158703	6.190866
+15 %	0.9763125	0.9978117	2.202086
- 15 %	0.9763125	0.8571313	12.20728
+20 %	0.9763125	0.999124	2.3365
-20 %	0.9763125	0.7740416	20.71784
+30 %	0.9763125	0.9998828	2.414218
-30 %	0.9763125	0.3703219	62.06933
+40 %	0.9763125	0.9999877	2.424961
-40 %	0.9763125	0.08096258	91.70731
+50 %	0.9763125	0.999999	2.426118
-50 %	0.9763125	0.009363906	99.04089

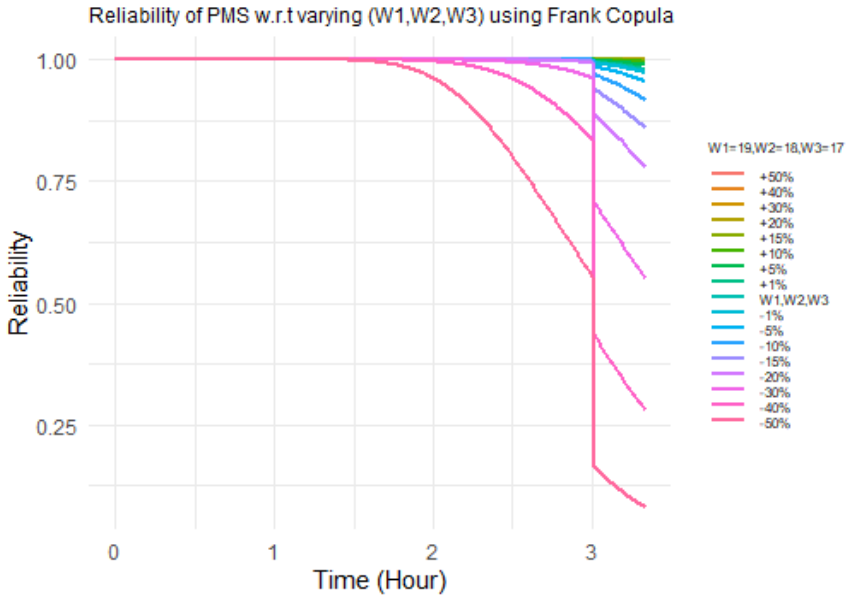


Figure 12: Reliability of aircraft flight PMS w.r.t $(\pm 1\%,\ \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in** $w_1 = 19, w_2 = 18, w_3 = 17$ **using Frank copula**

Table 12: Sensitivity analysis on reliability of aircraft flight PMS w.r.t $(\pm 1\%,\ \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in external degradation parameters** $\mu = 1.4, \sigma = 1.1$**using Frank copula** $\theta = 2.17, w_1 = 19, w_2 = 18, w_3 = 17, \alpha_1 = 0.4, \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.5, \alpha_6 = \alpha_7 = \alpha_8 = 0.7$

Initial values of external degradation parameters $\mu = 1.4, \sigma = 1.1$	True reliability (R)	Obtained reliability (R*)	Percentage deviation (PD) %
+1%	0.9763125	0.9745152	0.1840875
-1%	0.9763125	0.9780078	0.1736391
+ 5 %	0.9763125	0.9662398	1.031709
-5 %	0.9763125	0.9838395	0.7709588
+10 %	0.9763125	0.953233	2.363941
- 10 %	0.9763125	0.9892664	1.326817
+15 %	0.9763125	0.9370136	4.025242
- 15 %	0.9763125	0.9930438	1.713723
+20 %	0.9763125	0.9174555	6.028502
-20 %	0.9763125	0.9955857	1.974083
+30 %	0.9763125	0.8686875	11.02362
-30 %	0.9763125	0.9982939	2.251474
+40 %	0.9763125	0.7026077	28.03455
-40 %	0.9763125	0.9993403	2.358652
+50 %	0.9763125	0.6303879	35.43175
-50 %	0.9763125	0.9997217	2.397721

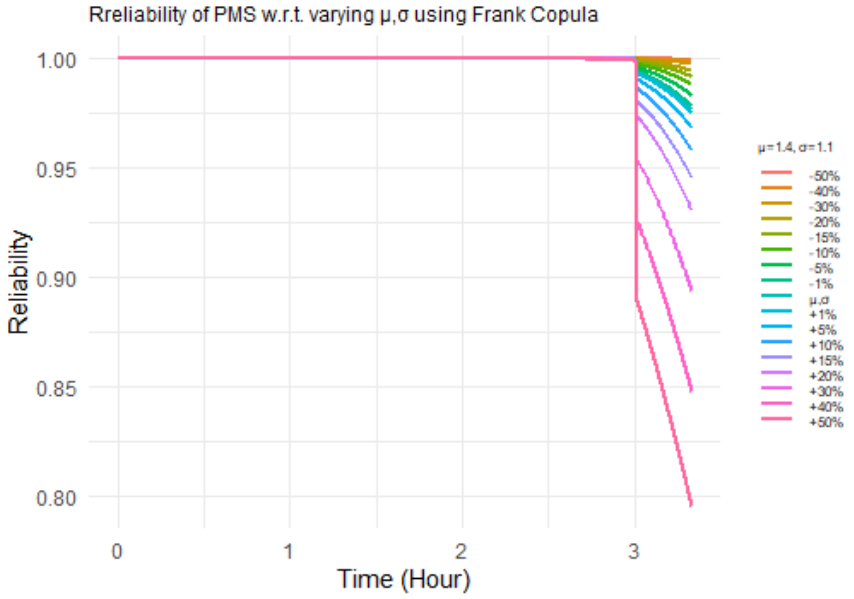


Figure 13: Reliability of aircraft flight PMS w.r.t $(\pm 1\%, \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in** $\mu = 1.4, \sigma = 1.1$ **using Frank copula**

Table 13: Sensitivity analysis on reliability of aircraft flight PMS w.r.t $(\pm 1\%, \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in impact element of external degradation** $\alpha_1 = 0.4, \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.5, \alpha_6 = \alpha_7 = \alpha_8 = 0.7$ **using Frank copula** $\theta = 2.17, w_1 = 19, w_2 = 18, w_3 = 17, \mu = 1.4, \sigma = 1.1$

Initial values of common factors $\alpha_1 = 0.4, \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.5, \alpha_6 = \alpha_7 = \alpha_8 = 0.7$	True reliability (R)	Obtained reliability (R*)	Percentage deviation (PD) %
+1%	0.9763125	0.9747962	0.1553092
-1%	0.9763125	0.9777559	0.1478403
+ 5 %	0.9763125	0.9679614	0.855368
-5 %	0.9763125	0.9828427	0.668864
+10 %	0.9763125	0.9575498	1.921793
- 10 %	0.9763125	0.9878201	1.178685
+15 %	0.9763125	0.9448881	3.218686
- 15 %	0.9763125	0.9915186	1.557509
+20 %	0.9763125	0.929858	4.758156
-20 %	0.9763125	0.9941988	1.832024
+30 %	0.9763125	0.8926443	8.569819
-30 %	0.9763125	0.997405	2.160427
+40 %	0.9763125	0.8466733	13.27845
-40 %	0.9763125	0.9988808	2.311584
+50 %	0.9763125	0.6814569	30.20094
-50 %	0.9763125	0.9995144	2.376479

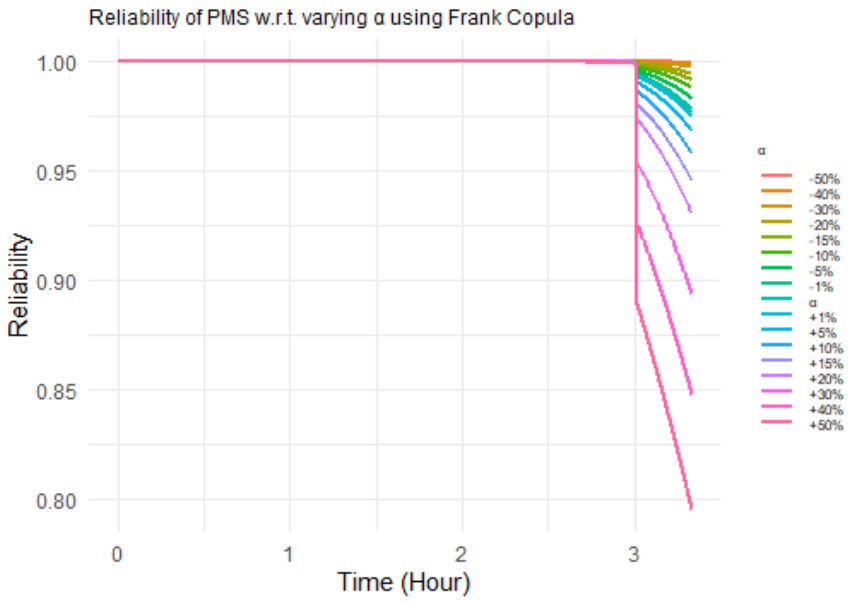


Figure 14: Reliability of aircraft flight PMS w.r.t $(\pm 1\%, \pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 40\%, \pm 50\%)$ **deviations in** $\alpha_1 = 0.4, \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.5, \alpha_6 = \alpha_7 = \alpha_8 = 0.7$ **using Frank copula**

Table 4 and Figure 5, Table 8 and Figure 9, and Table 12 and Figure 13 corresponding to the Gumbel-Hougaard, Clayton, and Frank copulas, respectively, show that increasing the parameters associated with external degradation of components, μ, σ, result in decrease in the reliability of aircraft flight PMS, and decrease in their values result in increase in the reliability of aircraft flight PMS.

Finally, Table 5 and Figure 6, Table 9 and Figure 10, and Table 13 and Figure 14 corresponding to the Gumbel-Hougaard, Clayton, and Frank copulas, respectively, show that increasing impact element, α_j, of external degradation of components results in decrease in the reliability of aircraft flight PMS, and decrease in their values result in increase in the reliability of aircraft flight PMS.

4. Concluding remarks

In the present paper reliability analysis of a PMS with each component in a phase subject to internal and external degradation is studied. A linear combination of the internal degradation and a proportional common external degradation constitute a degradation path of a component in a phase. Wiener process is used to model both internal and common external degradation. The cumulative exposure model has been used to model a PMS, with the dependency amongst the components in a phase and that across the phases modelled using different copulas: Gumbel-Hougaard, Clayton, and Frank. An aircraft flight PMS is used to demonstrate the methodology proposed and comparative study amongst PMS models with different copulas carried out. The results of the sensitivity analyses show that the model proposed is not sensitive to small deviations in the values of the selected parameters. The method developed can be generalized to missions with any number of phases.

Acknowledgements

This research work is financially supported by University of Delhi, Delhi-7, INDIA. The authors are grateful to the reviewers for their valuable comments.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Alam, M. and Al-Saggaf, U. M. (1986). Quantitative reliability evaluation of repairable phased-mission systems using Markov approach. *IEEE Transactions on Reliability*, **35**, 498–503.

Alam, M., Song, M., Hester, S., and Seliga, T. (2006). Reliability analysis of phased-mission systems: a practical approach. In *RAMS'06. Annual Reliability and Maintainability Symposium, 2006.* 551–558, IEEE.

Bing, L. and Xiao-yue, W. (2012). Mission reliability modeling of missile defense system based on phase mission Bayesian networks. *Journal of the Academy of Equipment Command & Technology*, **23**, 75–78.

Bondavalli, A., Chiaradonna, S., Di Giandomenico, F., and Mura, I. (2004). Dependability modeling and evaluation of multiple-phased systems using deem. *IEEE Transactions on Reliability*, **53**, 509–522.

Chhikara, R. and Folks, J. L. (1989). *The Inverse Gaussian Distribution.* Marcel Dekker, NewYork.

Eruguz, A. S., Tan, T., and van Houtum, G.-J. (2017). Optimizing usage and maintenance decisions for k-out-of-n systems of moving assets. *Naval Research Logistics (NRL)*, **64**, 418–434.

Hu, Y., Wang, C., and Xing, L. (2021). Reliability analysis of k-out-of-n phased-mission systems with phase-and requirement. In *IOP Conference Series: Materials Science and Engineering*, volume 1043. 032038, IOP Publishing.

Huang, X., Aslett, L. J., and Coolen, F. P. (2019). Reliability analysis of general phased mission systems with a new survival signature. *Reliability Engineering & System Safety*, **189**, 416–422.

Kim, K. and Park, K. S. (1994). Phased-mission system reliability under Markov environment. *IEEE Transactions on Reliability*, **43**, 301–309.

Lee, C. K. and Wen, M.-J. (2006). A multivariate Weibull disitribution. *arXiv preprint math/0609585*, .

Levitin, G., Xing, L., and Dai, Y. (2018). Optimal work distribution and backup frequency for two non-identical work sharing elements. *Reliability Engineering & System Safety*, **170**, 127–136.

Li, J., Coit, D. W., and Elsayed, E. A. (2011). Reliability modeling of a series system with correlated or dependent component degradation processes. In *2011 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering.* 388–393, IEEE.

Li, X.-Y., Huang, H.-Z., Li, Y.-F., and Xiong, X. (2021). A Markov regenerative process model for phased mission systems under internal degradation and external shocks. *Reliability Engineering & System Safety*, **215**, 107796.

Li, Y.-F., Liu, Y., Huang, T., Huang, H.-Z., and Mi, J. (2020). Reliability assessment for systems suffering common cause failure based on Bayesian networks and proportional hazards model. *Quality and Reliability Engineering International*, **36**, 2509–2520.

Liu, C., KrAmer, A., and NeumANN, S. (2020). Reliability assessment of repairable phased-mission system by monte carlo simulation based on modular sequence-enforcing fault tree model. *Eksploatacja i Niezawodność*, **22**, 272–281.

Meshkat, L., Xing, L., Donohue, S., and Ou, Y. (2003). An overview of the phase-modular fault tree approach to phased mission system analysis. In *Proceedings of the 1st International Conference on Space Mission Challenges for Information Technology (SMC-IT)*. Pasadena, CA, 393–398.

Mura, I. (2021). Stochastic modeling and analysis of phased-mission systems dependability. In *2021 Annual Reliability and Maintainability Symposium (RAMS)*. 1–6, IEEE.

Mural, I., Bondavalli, A., Zang, X., and Trivedi, K. (1999). Dependability modeling and evaluation of phased mission systems: a dspn approach. In *Dependable Computing for Critical Applications 7*. 319–337, IEEE.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer.

Nelson, W. B. (2009). *Accelerated Testing: Statistical Models, Test Plans, and Data Analysis*. John Wiley & Sons.

Ou, Y. and Dugan, J. B. (2004). Modular solution of dynamic multi-phase systems. *IEEE Transactions on Reliability*, **53**, 499–508.

Shrestha, A., Xing, L., and Dai, Y. (2009). Reliability analysis of multi-state phased-mission systems. In *2009 Annual Reliability and Maintainability Symposium*. 151–156, IEEE.

Shrestha, A., Xing, L., and Dai, Y. (2010). Reliability analysis of multistate phased-mission systems with unordered and ordered states. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, **41**, 625–636.

Si, X.-S., Hu, C.-H., Zhang, Q., and Li, T. (2015). An integrated reliability estimation approach with stochastic filtering and degradation modeling for phased-mission systems. *IEEE Transactions on Cybernetics*, **47**, 67–80.

Somani, A. K., Ritcey, J. A., and Au, S. H. (1992). Computationally-efficient phased-mission reliability analysis for systems with variable configurations. *IEEE Transactions on Reliability*, **41**, 504–511.

Srivastava, P. W. and Rani, S. (2024). Predictive maintenance scheme for phased mission systems. *Reliability: Theory & Applications*, **19**, 675–687.

Srivastava, P. W., Rani, S., et al. (2022). Copula-based approach to reliability analysis of phased-mission systems. *International Journal of Reliability, Risk and Safety: Theory and Application*, **5**, 49–62.

Wang, D. and Trivedi, K. S. (2007). Reliability analysis of phased-mission system with independent component repairs. *IEEE Transactions on reliability*, **56**, 540–551.

Wilson, A., L.-N. K.-M. S. and Y., A. (2005). *Modern Statistical and Mathematical Methods in Reliability*, volume 10. World Scientific.

Wu, X. and Hillston, J. (2015). Mission reliability of semi-Markov systems under generalized operational time requirements. *Reliability Engineering & System Safety*, **140**, 122–129.

Xing, L. (2007). Reliability evaluation of phased-mission systems with imperfect fault coverage and common-cause failures. *IEEE Transactions on Reliability*, **56**, 58–68.

Xing, L. and Amari, S. V. (2008). Reliability of phased-mission systems. In *Handbook of Performability Engineering*. 349–368, Springer.

Yang, X. and Wu, X. (2014). Mission reliability assessment of space TT&C system by discrete event system simulation. *Quality and Reliability Engineering International*, **30**, 1263–1273.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 24, No. 1, 2026 (New Series), pp 275–296
http://www.ssca.org.in/journal

A Hyperspectral and Deep Learning Approach for Wheat Yield Prediction

Mohit Kumar[1,2], Alka Arora[2], Sudeep Marwaha[2], Viswanathan Chinnusamy[3], Sudhir Kumar[3], Soumen Pal[2], Mrinmoy Ray[2] and Rajkumar Dhakar[3]

[1] *The Graduate School,*
ICAR-Indian Agricultural Research Institute, New Delhi, 110012, India
[2] *ICAR-Indian Agricultural Statistics Research Institute, New Delhi, 110012, India*
[3] *ICAR-Indian Agricultural Research Institute, New Delhi, 110012, India*

Received: 04 June 2025; Revised: 10 October 2025; Accepted: 03 November 2025

Abstract

Accurate detection of wheat spikes and reliable yield prediction are critical for optimizing crop production and resource management. This study presents an integrated framework for spike detection and yield estimation using pseudo-RGB images derived from hyperspectral data. A YOLOv8 model was trained on 1,050 images, achieving high precision, recall, and mean average precision values. The bounding boxes and masks generated by YOLOv8 were used to quantify spike count and spike area, while six vegetation indices were extracted from hyperspectral images acquired at the booting stage. Three multiple linear regression models were developed for yield prediction: one based on spike features, another on vegetation indices, and a third combining both. The combined model achieved the highest accuracy, with a five-fold cross-validation R^2 of 0.902 $\pm$ 0.007, RMSE of 1.739 $\pm$ 0.133 g, and MAE of 1.289 $\pm$ 0.066 g. Compared with previous approaches, the proposed framework demonstrated improved performance, highlighting the value of integrating spike morphology and spectral data for yield prediction. Overall, the study shows that hyperspectral imaging can simultaneously provide morphological and physiological traits, reducing reliance on high-resolution RGB data in wheat phenotyping.

Key words: Deep learning; YOLOv8; Hyperspectral imaging; Yield prediction; Vegetation indices.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Wheat (*Triticum aestivum L.*) is one of the most widely cultivated cereal crops globally, serving as a staple food source and playing a central role in global food security (Curtis and Halford, 2014). Accurate and timely prediction of wheat yield is essential for agricultural planning, market forecasting, and food policy development. Traditional approaches for

Corresponding Author: Alka Arora
Email: : alka.office27@gmail.com

predicting yield are based on manual sampling and statistical modelling, which are labour-intensive, time-consuming, and often limited in spatial and temporal coverage (Kumar *et al.*, 2016; Sishodia *et al.*, 2020). Recent advances in remote sensing and image-based phenotyping have enabled non-destructive, scalable, and data-driven approaches to monitor crop traits and forecast yield (Khaki and Wang, 2019; Kumar *et al.*, 2024; Muruganantham *et al.*, 2022).

One of the most direct morphological indicators of wheat yield is the spike, which contains the grain-producing florets. Traits such as spike count and spike area are closely related to grain number and biomass, making them valuable phenotypic markers for yield prediction (Misra *et al.*, 2022; Qiongyan *et al.*, 2017). High-resolution RGB imagery captured by ground-based platforms or unmanned aerial vehicles (UAVs) has been widely used to detect and quantify wheat spikes using deep learning and computer vision techniques (Hasan *et al.*, 2018; Misra *et al.*, 2020; Zang *et al.*, 2022). A study by Misra *et al.* (2022) highlighted the potential of the spike count and spike area extracted from RGB images as predictors of wheat yield. However, these approaches are highly dependent on high-resolution RGB images, which can be computationally expensive and time-consuming due to large data volumes (Arora *et al.*, 2023; Dagar *et al.*, 2024). Furthermore, RGB-based methods often overlook important physiological parameters such as plant stress, canopy structure, and biochemical composition that are critical to yield evaluation (Zhou *et al.*, 2023).

The spectral indices derived from remote sensing data offer a valuable alternative for yield prediction, as they serve as representations of the physiological and structural traits of the plant. Indices such as the Normalized Difference Vegetation Index (NDVI), DVI, IPVI, and SAVI capture information on chlorophyll concentration, canopy density, and photosynthetic efficiency, all of which are closely related to biomass accumulation and grain yield (Aboelghar *et al.*, 2014; Xue and Su, 2017). Some studies demonstrated that NDVI is a reliable predictor of wheat yield (Liu *et al.*, 2023) and emphasize that combining multiple vegetation indices can improve prediction accuracy by capturing complementary aspects of crop physiology and canopy structure (Su *et al.*, 2023; Zhou *et al.*, 2023). However, a key limitation of spectral indices is that they do not directly capture spike-related morphological traits, such as spike count and spike area, which have a direct and strong correlation with grain yield.

To bridge this gap, integrating spectral and morphological information could lead to more comprehensive and accurate yield prediction models (Patrignani and Ochsner, 2015). While one option is to combine high-resolution RGB imagery with hyperspectral or multi-spectral data, this approach significantly increases data volume and processing complexity. Moreover, acquiring high-resolution RGB images requires additional equipment and setup. To address these challenges, this study proposes using hyperspectral imagery (HSI) alone to extract both morphological and physiological traits. Specifically, HSI data spanning the 400–1000 nm spectral range is used to generate low-resolution pseudo-RGB images for spike detection and counting. This approach eliminates the need for high-resolution RGB cameras and simplifies the workflow by leveraging a single data source for both morphological and spectral information. In addition to spike detection, the study proposes to compute multiple vegetation indices from the same HSI data and combine these with spike count and spike area to enhance yield prediction accuracy. This integrated approach contributes to the development of scalable, cost-effective phenotyping tools and explores the potential of HSI-derived imagery to complement or replace high-resolution RGB-based methods in precision

agriculture.

Recent advancements in deep learning have further enhanced plant trait extraction from imagery. YOLOv8, the state-of-the-art version of the You Only Look Once (YOLO) architecture, offers improved speed and accuracy in real-time object detection and instance segmentation (Jocher *et al.*, 2023). Its enhanced performance makes it suitable for agricultural applications where fine-scale object localization is needed. In wheat phenotyping, YOLO-based models have shown strong results for spike detection using aerial and ground-level RGB images (Fang and Yang, 2024; Wen *et al.*, 2024). Therefore, this study proposes the use of YOLOv8 to segment wheat spikes from pseudo-RGB images generated from hyperspectral data, enabling accurate spike count and area estimation that are the key features for yield modelling.

To provide an appropriate environment for evaluating the proposed approach, this study was conducted in a phenomics facility. Such facilities are designed for experiments where large numbers of genotypes are grown with replications and screened within a single season or multiple growing seasons, creating a demand for rapid and accurate assessment of yield-related traits. The integration of hyperspectral imaging with deep learning offers an automated, scalable, and precise framework that reduces the need for labour-intensive manual phenotyping and facilitates efficient yield prediction. By addressing the bottleneck of large-scale genotype screening, the methodology directly supports crop improvement and breeding programs. The specific objectives of this study are to: (1) Evaluate the feasibility of spike detection and counting using low-resolution pseudo-RGB images generated from HSI data. (2) Extract key morphological spike traits, particularly spike count and spike area. (3) Compute spectral reflectance indices from the hyperspectral data. (4) Assess the performance of combined morphological and spectral indices for predicting the wheat yield.

2. Materials and methods

The materials and methods section details the acquisition of a hyperspectral imaging (HSI) dataset of wheat and the extraction of key features for yield prediction. These features include spectral vegetation indices and spike-based morphological traits, all derived from the HSI data cube. Pseudo-RGB images were generated from the HSI cube and used for spike detection and segmentation by using the YOLOv8 model to extract spike morphological traits, spike count, and area. In parallel, spectral indices were computed directly from the hyperspectral bands to capture the physiological and spectral characteristics of the plants. A detailed workflow outlining the overall methodology is presented in Figure 1.

2.1. Dataset

The experiment was conducted at the Nanaji Deshmukh Plant Phenomics Centre (NDPPC), ICAR-IARI, Pusa, New Delhi, India, during three winter wheat-growing seasons (2021–22, 2022–23, and 2023–24). In phenomics facilities, wheat varieties and genotypes are grown under controlled environments that minimize variability and ensure uniform plant development.

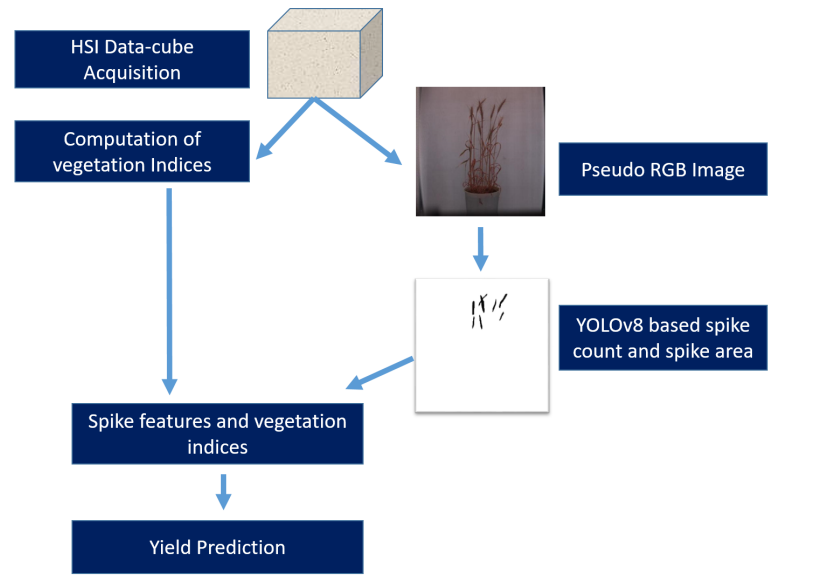


Figure 1: Workflow of the proposed methodology for wheat yield prediction using hyperspectral imagery

A total of 66 wheat varieties and genotypes were included, representing both released cultivars (*e.g.*, HD-2687, PBW343, LOK1, DBW51, Raj4037) and breeding lines (*e.g.*, Core-set accessions, LSP3043, KACHU#1, C518RF), and the detailed names are provided in Annexure Table 5. Each entry was grown in six replications, while the widely cultivated checks HD-2967, HD-3086, HD-3237, and HD-3271 were represented with more than six replications to strengthen comparisons. The six replications were grown under contrasting conditions, comprising two non-stressed controls, two subjected to nitrogen deficiency, and two exposed to drought stress, enabling robust assessment of genotypic performance across different environmental conditions.

Across the three years, the dataset comprised 440 plants in 2023–24, 90 plants in 2022–23, and 90 plants in 2021–22, giving a total of 620 plants. Each plant was imaged at two developmental stages, booting and physiological maturity, producing 1240 hyperspectral images (620 plants $\times$ 2 stages).

Hyperspectral imaging was performed using a Headwall E-Series VNIR camera, operating in band-sequential format (BSL). The system covered the 400–1000 nm spectral range and recorded reflectance across 940 spectral bands, providing detailed information on canopy structure, pigment composition, and physiological traits. The dataset served two main purposes: (i) computation of spectral vegetation indices from booting-stage images, and (ii) conversion of maturity-stage hyperspectral data into pseudo-RGB images for spike detection.

For spike detection model development, 350 maturity-stage images were used for training. The trained model was applied to the remaining 270 maturity-stage test images to detect and segment spikes, ensuring that the images used for yield prediction were completely independent of the training set. Spike features extracted from these test images were then combined with vegetation indices derived from the booting stage and used in the regression analysis for yield prediction.

2.2. Data pre-processing

One of the essential steps after acquiring raw hyperspectral data is to perform geometric and radiometric correction to ensure data consistency and accuracy. Radiometric calibration was performed using Equation 1 (Xue and Su, 2017). Where r_0 represents the corrected reflectance data, r_{raw} denotes the raw hyperspectral data, r_{dark} is the dark reference, and r_{white} refers to the reference white panel data. The Lemnatech Imaging System captures dark reference frames (shutter closed, no light) and white reference frames (Spectralon panel, 99% reflectance) for each imaging cycle. These references are directly linked to the corresponding plant images, ensuring that environmental noise, illumination variability, and sensor drift are minimized.

$$r_0 = \frac{r_{raw} - r_{dark}}{r_{white} - r_{dark}} \tag{1}$$

This calibration step normalized reflectance values across all images, correcting for variations caused by environmental factors and instrumental inconsistencies. Dark frame subtraction reduced electronic noise and sensor offsets, while normalization with the white reference ensured spectral consistency across sessions. To assess image quality, hyperspectral cubes were visually inspected for striping artifacts, saturation, or band misalignment. As the imaging was performed in a standardized LemnaTec facility, no defective images were identified. Additionally, signal-to-noise ratios (SNR) were monitored across representative bands (*e.g.*, 550, 680, 800 nm), confirming spectral stability and reliability throughout the dataset. In the next step, pseudo-RGB images generated from spectral bands (Figure 2) were used for the spike detection and morphological trait extraction (spike count and projected spike area). The image had a width of 400 pixels and a height of 348 pixels.

Figure 2: Pseudo RGB of the wheat plant generated from HSI data

2.3. Spike detection and segmentation

YOLOv8 (Varghese and Sambath, 2024), a state-of-the-art deep learning framework, was used for wheat spike detection and segmentation. This latest version of YOLO intro-

duces architectural improvements such as anchor-free detection, dynamic label assignment, and transformer-based feature aggregation (Jocher *et al.*, 2023; Fang and Yang, 2024), making it well-suited for complex plant phenotyping tasks. In this study, pseudo-RGB images generated from hyperspectral data were provided as input to YOLOv8, and the model was trained on a manually annotated dataset to ensure accurate localization and mask generation.

The YOLOv8-seg architecture optimizes a composite loss function consisting of (i) Complete IoU (CIoU) loss for bounding box regression, (ii) Binary Cross-Entropy (BCE) loss for objectness confidence, and (iii) BCE loss for class prediction. For segmentation, an additional mask loss is computed as the pixel-wise BCE between predicted and ground-truth masks. These loss components collectively balance localization accuracy, object classification, and mask generation. Weighting of the losses followed the YOLOv8 default configuration (box: 7.5, cls: 0.5, dfl: 1.5), empirically optimized by Ultralytics for robust detection and segmentation. The YOLOv8 model was trained for 150 epochs with a batch size of 16, an initial learning rate of 0.001, and the stochastic gradient descent (SGD) optimizer with a momentum of 0.937. Anchor-free detection and dynamic label assignment were enabled by default, and early stopping with a patience of 30 epochs was applied to prevent over-fitting.

2.4. Dataset preparation

The VGG Image Annotator (VIA) tool (Dutta and Zisserman, 2019), developed by the Visual Geometry Group at the University of Oxford, was used to manually label pseudo-RGB images to make the training dataset for wheat spike detection and segmentation. VIA is a web-based, open-source annotation tool that supports polygon, rectangle, and point-based annotations, making it well-suited for detailed labelling of images. Each wheat spike was manually outlined using a polygonal to capture its shape precisely (Figure 3a). All annotated spikes were labelled under the class "spike" to support both detection and segmentation tasks in YOLOv8. Before annotation, the dataset was split into a training set of 350 images and a test set of 270 images. The training images (Figure 3b) were augmented through horizontal and vertical rotations (Figure 3c), horizontal flipping (Figure 3d), and brightness adjustments (Figure 3e), thereby enhancing data diversity and expanding the training dataset to 1,050 images. The final annotations were exported in COCO JSON format, which is compatible with the YOLOv8 training pipeline and includes image IDs, polygon coordinates, and class labels.

2.5. Evaluation metrics for spike detection and segmentation

The performance of the YOLOv8 model in spike detection and segmentation was assessed using standard object detection and segmentation evaluation metrics (Everingham *et al.*, 2010; Lin *et al.*, 2014).

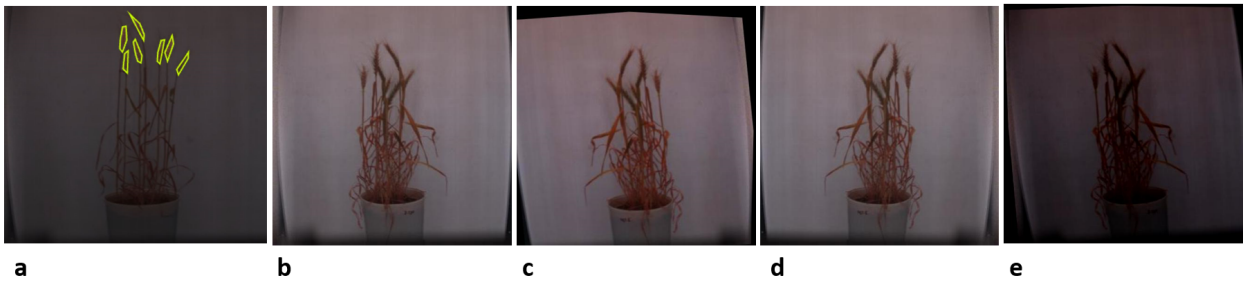


Figure 3: Illustration of the annotation and augmentation process for wheat spike detection. (a) Manual polygonal annotation of wheat spikes using the VIA tool; (b) sample pseudo-RGB image; (c) rotation augmentation; (d) horizontal flip; (e) brightness adjustment. These steps expanded the training dataset to 1,050 images and ensured greater variability for robust YOLOv8 model training

2.5.1. Spike detection metrics

Precision (P) measures the proportion of correctly identified spikes (true positives) among all spikes predicted by the model.

$$P = \frac{TP}{TP + FP} \tag{2}$$

Recall (R) measures the proportion of actual spikes correctly identified by the model.

$$R = \frac{TP}{TP + FN} \tag{3}$$

F1 Score is the harmonic mean of precision and recall, providing a balanced measure of accuracy.

$$F1 = \frac{2 \times (P \times R)}{P + R} \tag{4}$$

Mean Average Precision (mAP) evaluates detection performance across multiple Intersection over Union (IoU) thresholds. It is computed by averaging the precision over recall levels and then averaging across classes (in this case, spikes).

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{5}$$

2.5.2. Spike segmentation metrics

The *Jaccard Index*, commonly referred to as Intersection over Union (IoU), is a widely used metric for evaluating the accuracy of segmentation models. It quantifies the overlap between the predicted segmentation mask and the ground truth mask. The *Dice coefficient* is a metric for overlap between predicted and ground truth segmentation masks.

$$JaccardIndex = \frac{A \cap B}{A \cup B} \tag{6}$$

$$Dice = \frac{2 \cdot |A \cap B|}{|A| + |B|} \tag{7}$$

Where A is the set of pixels (or area) in the predicted spike mask. B is the set of pixels (or area) in the ground truth spike mask.

2.6. Features calculation for yield prediction

2.6.1. Spike count and spike area

The trained YOLOv8 model was applied to the test set of pseudo-RGB images to detect and segment wheat spikes. The model generated binary masks and bounding boxes for each spike. The number of spikes in each image was counted, and the area of each spike was calculated by counting the pixels in its mask. For each image, the spike count and total spike area were recorded, creating a dataset for yield prediction.

2.6.2. Spectral vegetation indices

For the HSI images collected at the booting stage, six spectral reflectance indices were calculated (Table 1). The selection of these indices was based on their proven association with wheat growth dynamics and yield prediction in earlier studies (Aboelghar *et al.*, 2014). Vegetation indices such as NDVI, RVI, SAVI, DVI, and IPVI are widely recognized for their ability to capture canopy vigour, chlorophyll content, and biomass accumulation, which are physiological traits that strongly correlate with final grain yield. The booting stage was specifically chosen because it represents a critical developmental phase where the crop's photosynthetic capacity and canopy structure are highly indicative of yield potential. Previous studies have demonstrated that indices derived at this stage can reliably predict wheat yield at both field and regional scales (Ren *et al.*, 2008; Sultana *et al.*, 2014; Xie *et al.*, 2020). These works collectively establish that early-stage spectral information complements later-stage morphological traits and thereby enhances predictive accuracy.

2.7. Yield prediction based on computed features

Grain yield was determined individually for 270 plants included in the test set. The dataset comprised 90 plants per year across three consecutive growing seasons (2021–22, 2022–23, and 2023–24). Each plant was harvested separately, and the yield was recorded using a precision balance to ensure measurement accuracy.

An ablation study was designed to evaluate the relative contribution of morphological and spectral features to yield prediction. Three feature sets were derived from the hyperspectral data: (i) morphological traits obtained from maturity-stage images, specifically spike count and spike area; (ii) spectral features represented by six vegetation indices computed from booting-stage hyperspectral imagery (Table 1); and (iii) a combined feature set in-

Table 1: Spectral reflectance indices and their formulas

Spectral reflectance index	Formula	References
NDVI	$\frac{NIR - Red}{NIR + Red}$	Baret and Guyot (1991)
SAVI	$\frac{(800\,\text{nm} - 670\,\text{nm})}{(800\,\text{nm} + 670\,\text{nm} + 0.5)}(1 + 0.5)$	Huete (1988)
GVI	$\frac{NIR - Green}{NIR + Green}$	Aboelghar *et al.* (2014)
DVI	$NIR - Red$	Richardson and Everitt (1992)
RVI	$\frac{NIR}{Red}$	Jordan (1969)
IPVI	$\frac{NIR}{NIR + Red}$	Crippen (1990)

tegrating both morphological and spectral traits to capture complementary structural and physiological information.

For each feature set, a separate regression model was constructed: Model 1 (morphological features only), Model 2 (spectral features only), and Model 3 (combined features). All models were developed using multiple linear regression, and model training was performed under a five-fold cross-validation scheme to obtain robust estimates of predictive accuracy and reduce the risk of over-fitting.

Uncertainty in predictions was quantified using 95% prediction intervals derived from the fitted regression models. These intervals provide estimates of the range within which future yield observations are expected to fall, incorporating both model error and residual variability. To address potential multicollinearity among vegetation indices, LASSO regression was used as a regularized comparator. This approach penalizes redundant predictors, thereby facilitating feature selection and improving model stability.

Model performance was evaluated using three standard metrics: the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). Comparative analysis of the ablation models, the LASSO regression, and the prediction interval outputs enabled a rigorous assessment of the relative importance of individual feature sets and their combined predictive capacity.

3. Results

3.1. Spike detection and segmentation

The YOLOv8 model was trained using pseudo-RGB images derived from hyperspectral data for the task of wheat spike detection and segmentation. Model training was conducted on an NVIDIA V100 GPU with 32 GB RAM, ensuring adequate computational resources for efficient learning and convergence. The training dataset consisted of 1,050 augmented images, while 270 images were reserved for testing. During model training, 20% of the training images (210 images) were used as the validation set.

Model performance was monitored through distinct sets of loss and evaluation metrics,

which are presented in five categories: (i) training loss, (ii) validation loss, (iii) box and mask precision/recall, and (iv) mAP@50 and mAP@50:95 (v) learning rate (Figures 4–8).

(i) Training loss

The training loss curves, including box loss, segmentation loss, classification loss, and distribution focal loss (DFL), showed a consistent downward trend across epochs, demonstrating that the model converged smoothly during training (Figure 4).

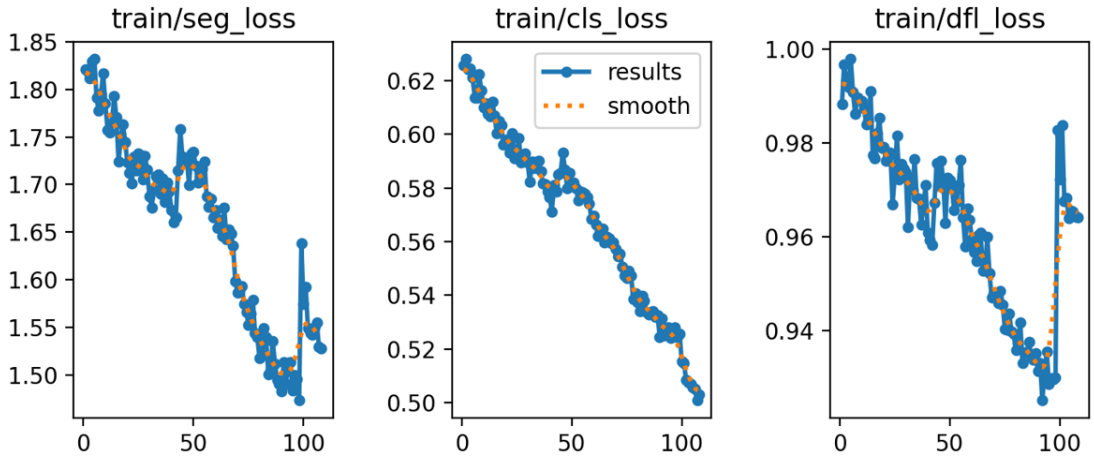


Figure 4: Training loss curves of the YOLOv8 model, including segmentation loss, classification loss, and distribution focal loss (DFL)

(ii) Validation loss

Validation losses for box regression, segmentation, and classification exhibited some fluctuations but remained overall stable throughout the training process, reflecting good generalization of the model to unseen data (Figure 5).

(iii) Box and mask precision and recall

Evaluation metrics for bounding box (B) and mask (M) predictions showed high reliability in spike detection. Precision and recall values ranged between 0.85 to 0.88 for bounding boxes and 0.82 to 0.87 for masks, indicating consistent and accurate predictions across both detection and segmentation tasks (Figure 6).

(iv) mAP@50 and mAP@50:95

The mean average precision at IoU=0.50 (mAP50) reached approximately 0.86 for bounding boxes and 0.83 for masks, while the stricter mAP@50:95 values were slightly lower but remained consistent, confirming robust detection performance across multiple IoU thresholds (Figure 7).

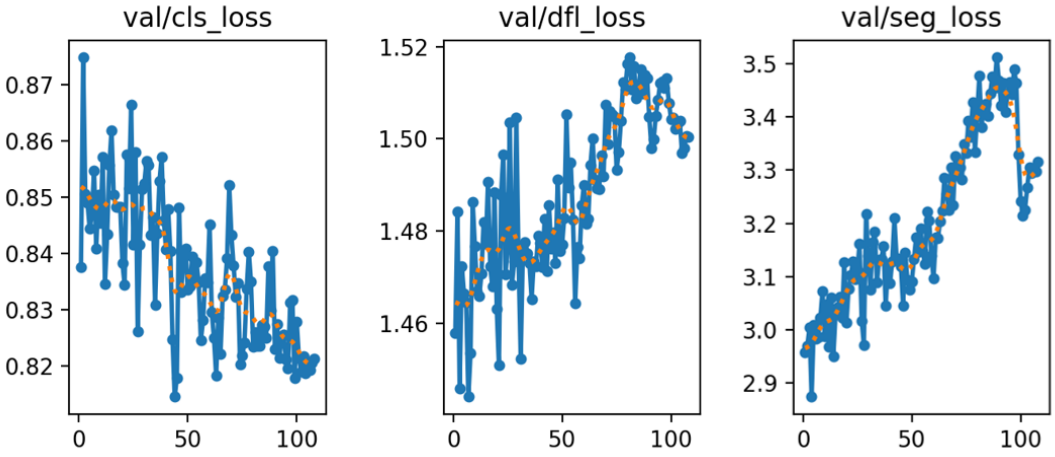


Figure 5: Validation loss curves of the YOLOv8 model, including classification loss, distribution focal loss (DFL), and segmentation loss

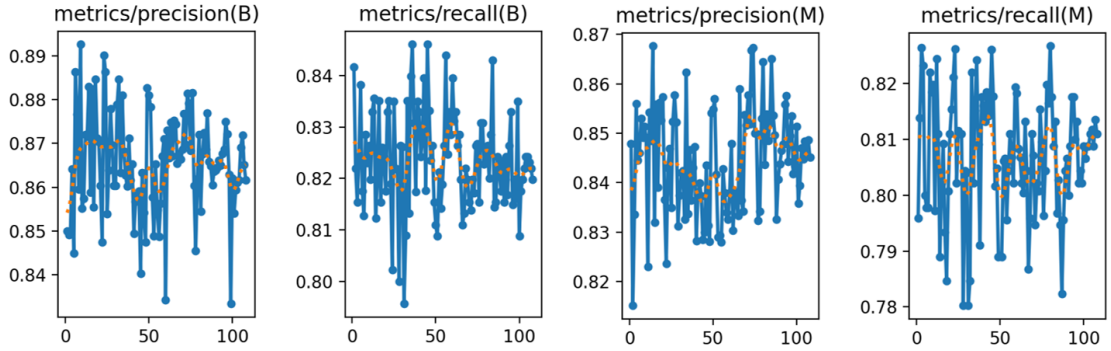


Figure 6: Precision and recall metrics for bounding box (B) and mask (M) predictions during YOLOv8 training, demonstrating high reliability in wheat spike detection and segmentation

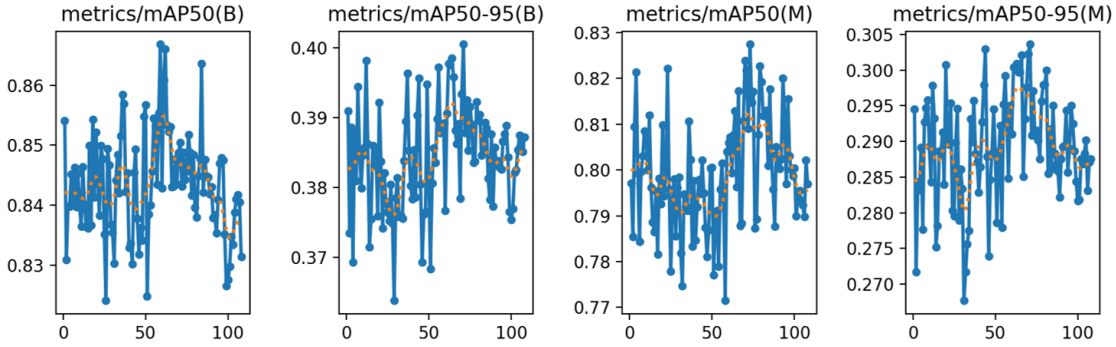


Figure 7: Mean average precision at IoU thresholds (mAP@50 and mAP@50:95) for bounding box (B) and mask (M) predictions

(v) Learning rate

The learning rate schedules for the parameter groups pg0, pg1, and pg2, which correspond to different layers of the YOLOv8 architecture (backbone, neck, and head), exhibited a smooth decay pattern throughout training (Figure 8). A stable and gradually decreasing learning rate is critical to avoid oscillations in the loss landscape and ensures convergence toward an optimal solution. The smooth decay observed across all parameter groups indicated that the optimization process was well-regularized, reducing the likelihood of over-fitting while improving generalization on unseen data..

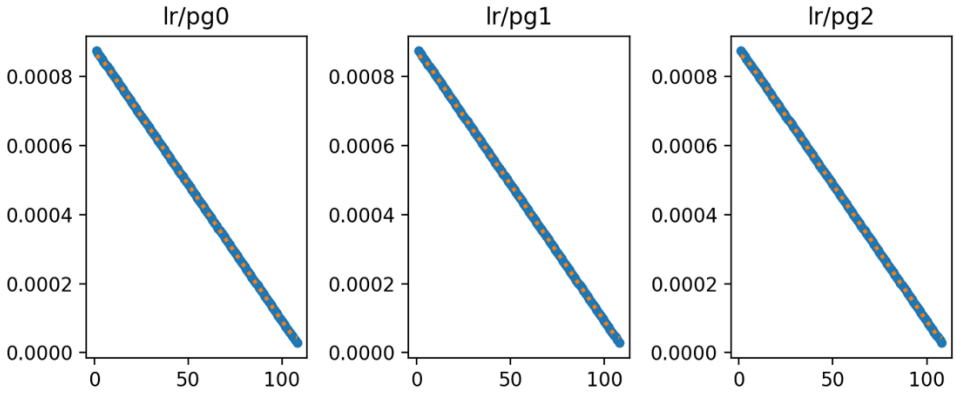


Figure 8: Learning rate schedules for the parameter groups pg0, pg1, and pg2 in the YOLOv8 model

Following training, the YOLOv8 model was evaluated on the test set, with results summarized in Table 2. The model achieved a precision of 0.83, a recall of 0.89, and an F1 score of 0.86, reflecting reliable detection performance. The mAP@50 was 0.89, while the Dice coefficient (0.94) and Jaccard Index (0.88) further confirmed the high accuracy of segmentation. These findings demonstrate that the trained YOLOv8 model effectively learned discriminative features for wheat spike detection and segmentation, achieving robust performance on unseen test images.

Table 2: YOLOv8 evaluation metrics on the test set

Metric	Precision	Recall	F1 Score	mAP@50	Dice coefficient	Jaccard Index
Value	0.83	0.89	0.86	0.89	0.94	0.88

Figure 9 presents qualitative examples of the model's predictions on four test images, where bounding boxes (B) and masks (M) are superimposed on the spikes. The results show that the model consistently detected and segmented wheat spikes, with bounding boxes and masks closely aligned with actual spike regions. The qualitative outputs complement the quantitative performance metrics in Table 2, together confirming that the YOLOv8 model can reliably detect and segment wheat spikes, which is essential for accurate spike counting and area estimation.

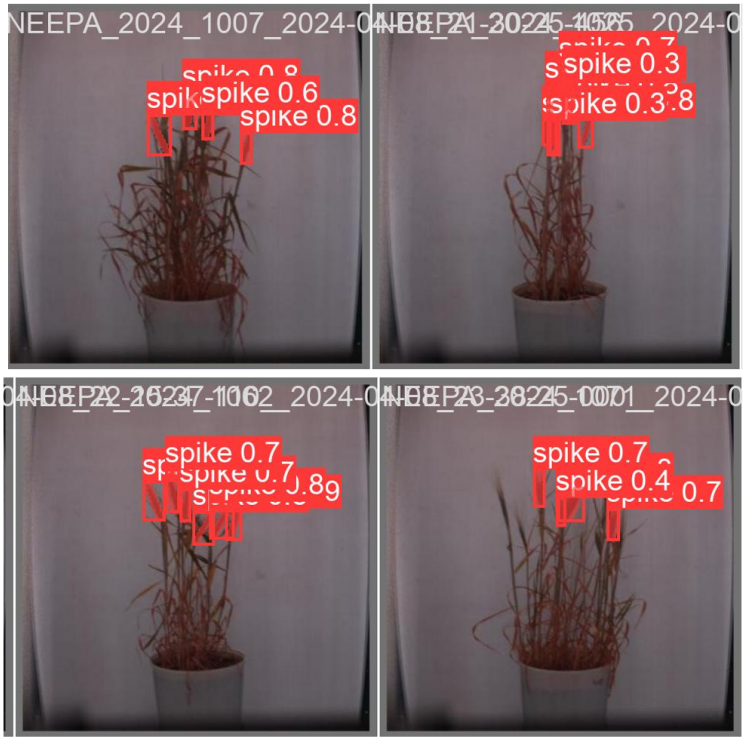


Figure 9: YOLOv8 predictions on four test images, showing bounding boxes and segmentation masks superimposed on wheat spikes. The model accurately localized and segmented spikes, aligning closely with ground-truth regions

The results from training, validation, and test metrics indicate that the model successfully captured the structural characteristics of wheat spikes, enabling its reliable application for downstream spike trait analysis. The YOLO model demonstrated computational efficiency, requiring only approximately 6 milliseconds per image on an NVIDIA V100 GPU with 32 GB RAM, while the same task took 0.5 seconds per image on an Intel Core i7 processor with 16 GB RAM, evaluated over a set of 200 images for inferencing. Although GPUs are typically used for model training, inference is often performed on CPUs in practical scenarios; thus, the CPU performance underscores the model's feasibility for downstream applications under standard computational environments, ensuring faster predictions suitable for high-throughput analysis.

3.2. Yield prediction

The trained YOLOv8 model was applied to a test set of 270 pseudo-RGB images to extract spike-based features for yield prediction. These features included the number of spikes per plant and the total projected spike area. In parallel, vegetation indices were computed from the booting stage hyperspectral images, providing spectral indicators of plant health and canopy structure. After harvesting, the grain yield (in grams) was recorded for the 270 test set plants, allowing for a comparison of model performance. The descriptive statistics of the features and plant yield are provided in Table 3.

Table 3: Descriptive statistics of the features for the yield prediction

Statistic	NDVI	SAVI	GVI	DVI	RVI	IPVI	Spike count	Spike area	Yield (g)
Min	0.61	0.46	0.44	0.46	3.10	0.78	0	0	0
Max	0.73	0.59	0.59	0.61	4.36	0.85	12	12916	36
Mean	0.64	0.49	0.49	0.50	3.33	0.80	7.35	7549.62	17.92
SD	0.02	0.02	0.02	0.02	0.19	0.01	2.32	2343.84	5.61

Figure 10 presents the Pearson correlation heat map illustrating the relationships among vegetation indices, spike features, and grain yield. Spike area and spike count exhibit strong positive correlations with grain yield, with spike area showing the highest correlation coefficient (r = 0.66). This highlights their critical role in yield prediction. Among the vegetation indices, NDVI, SAVI, DVI, RVI, and IPVI demonstrate moderate positive correlations with yield (r = 0.52–0.62), confirming their utility as spectral predictors. Notably, the vegetation indices themselves are highly intercorrelated, particularly SAVI, DVI, RVI, and IPVI ($r > 0.70$), suggesting redundancy among some spectral features. In contrast, the weak or negative correlations observed between vegetation indices and spike features indicate that these two groups of predictors provide largely complementary information, which may improve model robustness when integrated for yield estimation.

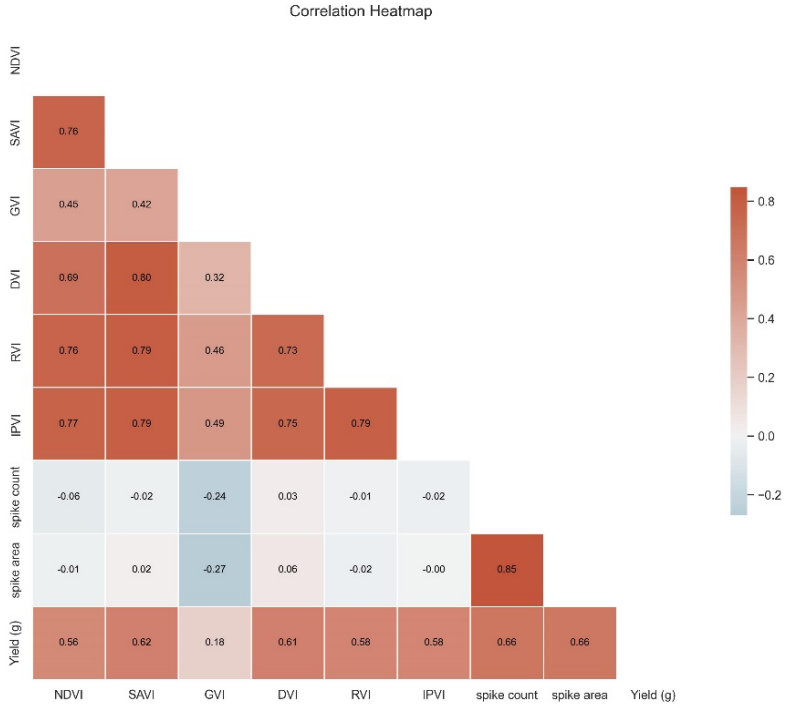


Figure 10: Pearson correlation heat-map showing relationships among vegetation indices, spike features, and grain yield (g)

Three regression models were constructed to evaluate the predictive ability of spike-based morphological features and hyperspectral vegetation indices for wheat yield estimation. The first model used spike traits (spike count and spike area), the second relied solely on vegetation indices, and the third integrated both feature types.

The spike feature-based model (Figure 11a) showed a moderate relationship between predicted and actual yield, with an R^2 of 0.453 ± 0.068, an RMSE of 4.1 ± 0.30 g, and an MAE of 3.291 ± 0.194 g (Table 4). The 95% prediction intervals were relatively wide, indicating variability in prediction accuracy when using morphological traits alone.

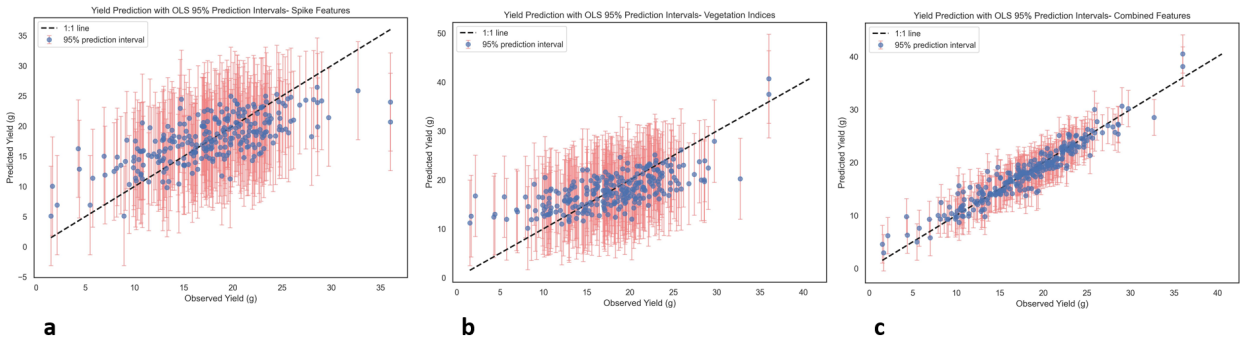


Figure 11: Actual versus predicted yield plot for (a) based on spike features, (b) based on vegetation indices, (c) based on combined spike and vegetation indices features

The vegetation index-based model (Figure 11b) demonstrated slightly weaker predictive performance, with an R^2 of 0.416 ± 0.036, an RMSE of 4.247 ± 0.284 g, and an MAE of 3.377 ± 0.321 g. The prediction intervals in this model were also broad, reflecting greater uncertainty and scatter around the 1:1 line.

Table 4: Cross-validation results (mean ± standard deviation) for yield prediction using vegetation indices, spike features, and their combination across three years of data

Model	R^2 (mean ± std)	RMSE (mean ± std)	MAE (mean ± std)
Indices-only	0.416 ± 0.036	4.247 ± 0.284	3.377 ± 0.321
Spikes-only	0.453 ± 0.068	4.100 ± 0.300	3.291 ± 0.194
Combined	0.902 ± 0.007	1.739 ± 0.133	1.289 ± 0.066

By contrast, the combined model integrating both vegetation indices and spike features (Figure 11c) achieved the best performance, with an R^2 of 0.902 ± 0.007, an RMSE of 1.739 ± 0.133 g, and an MAE of 1.289 ± 0.066 g. Importantly, the prediction intervals in this model were much narrower and closely aligned with the 1:1 line, demonstrating both improved accuracy and reduced uncertainty in yield estimation. To further examine prediction reliability, an error analysis was conducted across the test dataset. Overall, most predictions closely matched the observed yields, with residuals distributed symmetrically around zero. However, larger deviations were evident at the extremes of the yield range, particularly for plants with very low or very high grain yield.

This pattern can be clearly observed in the scatter plots of predicted versus observed yield (Figures 11a–c), where the data points at the lower and upper ends deviate more widely from the 1:1 line. These findings indicate that while the combined model achieved

high overall accuracy, prediction uncertainty increased for outlier cases, likely due to the limited representation of extreme-yielding plants in the training dataset.

As observed in the correlation heat-map (Figure 10), vegetation indices were highly correlated with each other, indicating potential multicollinearity that could affect model stability. To address this, we further applied Lasso regression for feature selection and coefficient shrinkage (Figure 12). The analysis confirmed that spike traits were the dominant predictors, with spike count (2.41) and spike area (1.66) showing the largest coefficients. Among vegetation indices, SAVI (1.04), DVI (0.85), and NDVI (0.80) retained moderate importance, while RVI (0.66), IPVI (0.48), and GVI (0.40) contributed smaller but still non-zero effects. Notably, none of the coefficients were reduced to zero, suggesting that despite their high interdependence, all vegetation indices added complementary spectral information for yield prediction.

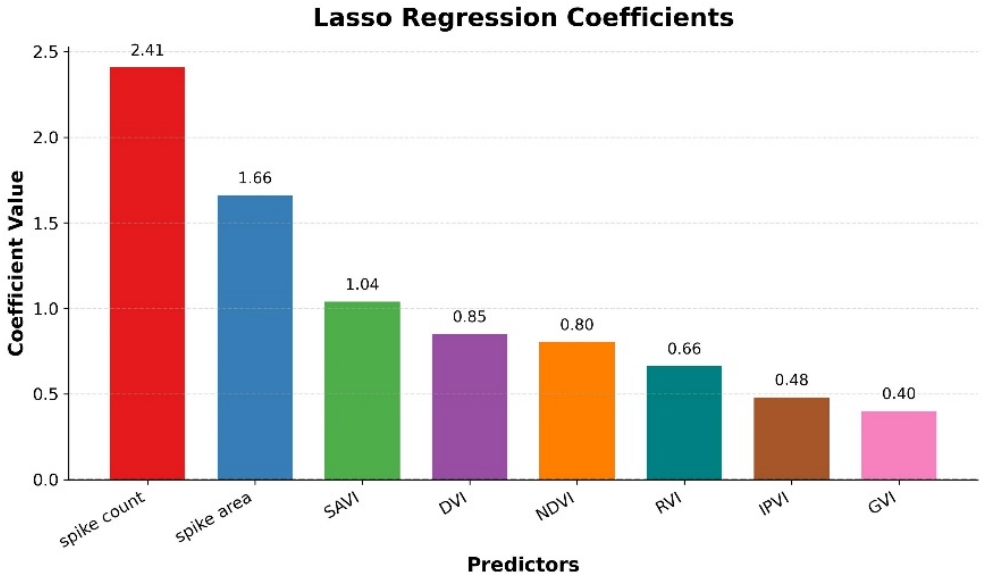


Figure 12: LASSO regression coefficients showing the relative importance of spike traits and vegetation indices

In the post-lasso OLS analysis, each predictor was evaluated under the null hypothesis that its coefficient equals zero ($H_0 : \beta = 0$), against the alternative hypothesis that the coefficient is non-zero ($H_1 : \beta \neq 0$). The null hypothesis was rejected for all predictors, confirming that NDVI, SAVI, GVI, DVI, RVI, IPVI, spike count, and spike area each exerted a statistically significant influence on yield. These findings indicate that both spike-related traits and spectral vegetation indices contribute meaningfully to yield prediction, with the integration of both trait- and index-based information enhancing model robustness.

Taken together, the linear regression and Lasso analyses demonstrate that spike-derived morphological traits explain the majority of yield variability, while vegetation indices enhance predictive robustness by providing additional spectral cues. Overall, the results (Table 4, Figures 11–12) highlight the strength of multi-modal data fusion for precise, reliable, and high-throughput wheat yield estimation.

4. Discussion

Recent studies have predominantly utilized high-resolution RGB images for plant phenotyping and yield estimation. In contrast, this study demonstrates that the YOLOv8 model can achieve strong performance using low-resolution pseudo-RGB images generated from hyperspectral data. This is particularly noteworthy, as it underscores the model's robustness and the potential for reducing reliance on high-resolution imaging systems, thereby making automated crop analysis more scalable and accessible by reducing computation time. The study presented an approach for wheat spike detection and yield prediction using YOLOv8, trained on pseudo-RGB images. The model effectively detected and segmented wheat spikes across a diverse set of images, supporting its application for spike counting and area estimation. These features were subsequently used for yield modelling, in combination with vegetation indices derived from the hyperspectral data cube. To assess the relative contribution of different features, three regression models were developed: one based solely on spike features, another on vegetation indices, and a third combining both. While the individual models showed reasonable predictive capacity, the combined model outperformed them, highlighting the synergistic value of integrating both morphological and spectral features. This fusion significantly improved yield prediction accuracy and demonstrated that spike morphology and vegetation indices contribute complementary information. Similar findings have been reported in previous studies (Al-Gaadi *et al.*, 2016; Xiong *et al.*, 2020). A comparison with existing literature further confirms the strength of our approach. Global Wheat Head Dataset (David *et al.*, 2020) used by Fang and Yang (2024) reported effective spike detection using high-resolution RGB data. Our model not only achieved comparable or better performance but did so using lower-resolution pseudo-RGB inputs. Similarly, our results surpassed the accuracy reported by Fang and Yang (2024); Wen *et al.* (2024), further validating the model's reliability. In the context of yield prediction, our approach showed notable improvements over Misra *et al.* (2022), where only RGB image-based spike traits were used.

This study demonstrated that combining hyperspectral imaging with deep learning provides a powerful framework for non-destructive wheat phenotyping and yield prediction. By integrating both spectral vegetation indices and spike-based morphological traits, the approach captures a holistic representation of plant health and productivity. Importantly, the findings are directly relevant to high-throughput phenomics facilities, where large numbers of genotypes are evaluated, as the proposed framework can substantially reduce the time and effort required for yield estimation. The present work was carried out under controlled conditions using pseudo-RGB images derived from hyperspectral data, with a limited dataset and without the complexity introduced by variable illumination, heterogeneous backgrounds, or environmental noise. While this represents a challenge for scaling the approach to large-scale field applications, it is not a limitation within phenomics facilities where standardized conditions prevail. In this context, the approach is immediately scalable for large-scale phenotyping and will be released as a software framework that can be seamlessly integrated into existing phenomics platforms.

Nonetheless, translating this framework to field-scale deployment will require addressing additional challenges, including illumination variability, canopy occlusion, soil–plant spectral interactions, and larger spatial heterogeneity. Illumination variability can be mitigated through radiometric calibration and deep learning approaches that learn illumination-

invariant features (Misra *et al.*, 2020). Soil and heterogeneous background effects can be reduced using soil-adjusted vegetation indices (such as SAVI, MSAVI, or TSAVI) or soil-line–based corrections, while topographic influences can be minimized with terrain-based corrections, including C-correction or Minnaert correction. Future work should therefore extend the methodology to field-based hyperspectral imaging under diverse agro-climatic conditions, incorporate auxiliary variables such as soil properties, weather data, and management practices, and integrate data from multiple sensing platforms. Coupled with advances in deep learning architectures and hardware efficiency, these efforts could enable robust, real-time, in-field monitoring of crop performance, thereby enhancing scalability and contributing significantly to precision agriculture.

5. Conclusion

This study developed and validated a YOLOv8-based framework for wheat spike detection and yield prediction using pseudo-RGB images derived from hyperspectral data. The model achieved high accuracy in spike detection and segmentation, demonstrating that reliable spike traits such as count and area can be extracted even from low-resolution pseudo-RGB inputs. When integrated with vegetation indices, these morphological features significantly enhanced yield prediction, with the combined regression model achieving superior performance ($R^2 = 0.902$), surpassing models based on individual feature sets. The framework was further strengthened through rigorous validation, including error analysis and LASSO regression, which confirmed the dominance of spike traits while highlighting the complementary role of vegetation indices. By deriving both morphological and physiological traits from hyperspectral data, the approach reduces dependency on high-resolution RGB systems and provides a cost-effective, scalable solution for high-throughput phenotyping. Importantly, this framework addresses key challenges in phenomics facilities, where large numbers of genotypes are routinely evaluated, by enabling rapid and automated yield estimation. Beyond its immediate utility, the approach offers a pathway toward accelerating breeding programs, improving genetic gain, and contributing to the broader goal of sustainable wheat production.

Acknowledgements

The facilities provided by the Indian Council of Agricultural Research - Indian Agricultural Statistics Research Institute (ICAR- IASRI), New Delhi, and the funding granted to the first author by ICAR in the form of an ICAR-SRF fellowship are duly acknowledged for carrying out this study, which is a part of his doctoral research being pursued at ICAR-IASRI. In addition, thanks to the NDPPC, ICAR-IARI, New Delhi, for conducting the experiment and providing the data for the study.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

Code and data availability

The processed datasets (pseudo-RGB images, vegetation index values, and yield records) and the YOLOv8 training scripts with analysis codes are available from the corresponding author upon reasonable request.

References

Aboelghar, M., Ali, A.-R., and Arafat, S. (2014). Spectral wheat yield prediction modeling using spot satellite imagery and leaf area index. *Arabian Journal of Geosciences*, **7**, 465–474.

Al-Gaadi, K. A., Hassaballa, A. A., Tola, E., Kayad, A. G., Madugundu, R., Alblewi, B., and Assiri, F. (2016). Prediction of potato crop yield using precision agriculture techniques. *PLOS ONE*, **11**, e0162219.

Arora, A., Misra, T., Kumar, M., Marwaha, S., Kumar, S., and Chinnusamy, V. (2023). Computer vision approaches for plant phenotypic parameter determination. In *Digital Ecosystem for Innovation in Agriculture*, 263–270. Springer Nature Singapore, Singapore.

Baret, F. and Guyot, G. (1991). Potentials and limits of vegetation indices for lai and apar assessment. *Remote Sensing of Environment*, **35**, 161–173.

Crippen, R. E. (1990). Calculating the vegetation index faster. *Remote Sensing of Environment*, **34**, 71–73.

Curtis, T. and Halford, N. G. (2014). Food security: the challenge of increasing wheat yield and the importance of not compromising food safety. *Annals of Applied Biology*, **164**, 354–372.

Dagar, P., Arora, A., Ray, M., Kumar, S., Chourasia, H., Kumar, M., and Chinnusamy, V. (2024). High-resolution reconstruction of images for estimation of plant height in wheat using rgb-d camera and machine learning approaches. *Current Science*, **127**, 1440–1446.

David, E., Madec, S., Sadeghi-Tehran, P., Aasen, H., Zheng, B., Liu, S., Kirchgessner, N., Ishikawa, G., Nagasawa, K., Badhon, M. A., Pozniak, C., de Solan, B., Hund, A., Chapman, S. C., Baret, F., Stavness, I., and Guo, W. (2020). Global wheat head detection (gwhd) dataset: A large and diverse dataset of high-resolution rgb-labelled images to develop and benchmark wheat head detection methods. *Plant Phenomics*, 3521852.

Dutta, A. and Zisserman, A. (2019). The via annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2276–2279.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, **88**, 303–338.

Fang, C. and Yang, X. (2024). Lightweight YOLOv8 for wheat head detection. *IEEE Access*, **12**, 66214–66222.

Hasan, M. M., Chopin, J. P., Laga, H., and Miklavcic, S. J. (2018). Detection and analysis of wheat spikes using convolutional neural networks. *Plant Methods*, **14**, 100.

Huete, A. R. (1988). A soil-adjusted vegetation index (savi). *Remote Sensing of Environment*, **25**, 295–309.

Jocher, G., Chaurasia, A., Qiu, J., and Stoken, A. (2023). YOLO by ultralytics. GitHub repository.

Jordan, C. F. (1969). Derivation of leaf-area index from quality of light on the forest floor. *Ecology*, **50**, 663–666.

Khaki, S. and Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, **10**, 621.

Kumar, M., Arora, A., Marwaha, S., Chinnusamy, V., Kumar, S., Jain, R., and Pal, S. (2024). Machine learning based approach for wheat plant senescence quantification. *Plant Physiology Reports*, **29**, 823–835.

Kumar, S., Raju, D., Sahoo, R. N., and Chinnusamy, V. (2016). Phenomics: unlocking the hidden genetic variation for breaking the barriers in yield and stress tolerance. *Indian Journal of Plant Physiology*, **21**, 409–419.

Lin, T. Y., Maire, M., Belongie, J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Lecture Notes in Computer Science*, volume 8693, 740–755.

Liu, Y., Sun, L., Liu, B., Wu, Y., Ma, J., Zhang, W., Wang, B., and Chen, Z. (2023). Estimation of winter wheat yield using multiple temporal vegetation indices derived from uav-based multispectral and hyperspectral imagery. *Remote Sensing*, **15**, 4800.

Misra, T., Arora, A., Marwaha, S., Chinnusamy, V., Rao, A. R., Jain, R., Sahoo, R. N., Ray, M., Kumar, S., Raju, D., Jha, R. R., Nigam, A., and Goel, S. (2020). Spikesegnet-a deep learning approach utilizing encoder-decoder network with hourglass for spike segmentation and counting in wheat plant from visual imaging. *Plant Methods*, **16**, 40.

Misra, T., Arora, A., Marwaha, S., Ranjan Jha, R., Ray, M., Kumar, S., Kumar, S., and Chinnusamy, V. (2022). Yield-spikesegnet: An extension of spikesegnet deep-learning approach for the yield estimation in the wheat using visual images. *Applied Artificial Intelligence*, **36**, 2137642.

Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N. H., and Islam, N. (2022). A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sensing*, **14**, 1990.

Patrignani, A. and Ochsner, T. E. (2015). Canopeo: A powerful new tool for measuring fractional green canopy cover. *Agronomy Journal*, **107**, 2312–2320.

Qiongyan, L., Cai, J., Berger, B., Okamoto, M., and Miklavcic, S. J. (2017). Detecting spikes of wheat plants using neural networks with laws texture energy. *Plant Methods*, **13**, 83.

Ren, J., Chen, Z., Zhou, Q., and Tang, H. (2008). Regional yield estimation for winter wheat with modis-ndvi data in shandong, china. *International Journal of Applied Earth Observation and Geoinformation*, **10**, 403–413.

Richardson, A. J. and Everitt, J. H. (1992). Using spectral vegetation indices to estimate rangeland productivity. *Geocarto International*, **7**, 63–69.

Sishodia, R. P., Ray, R. L., and Singh, S. K. (2020). Applications of remote sensing in precision agriculture: A review. *Remote Sensing*, **12**, 3136.

Su, X., Wang, J., Ding, L., Lu, J., Zhang, J., Yao, X., Cheng, T., Zhu, Y., Cao, W., and Tian, Y. (2023). Grain yield prediction using multi-temporal UAV-based multispectral vegetation indices and endmember abundance in rice. *Field Crops Research*, **299**, 108992.

Sultana, S. R., Ali, A., Ahmad, A., Mubeen, M., Zia-Ul-Haq, M., Ahmad, S., and Jaafar, H. Z. (2014). Normalized difference vegetation index as a tool for wheat yield estimation: A case study from faisalabad, pakistan. *The Scientific World Journal*, 725326.

Varghese, R. and Sambath, M. (2024). YOLOv8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, 1–6.

Wen, C., Ma, Z., Ren, J., Zhang, T., Zhang, L., Chen, H., Su, H., Yang, C., Chen, H., and Guo, W. (2024). A generalized model for accurate wheat spike detection and counting in complex scenarios. *Scientific Reports*, **14**, 24189.

Xie, Y., Wang, C., Yang, W., Feng, M., Qiao, X., and Song, J. (2020). Canopy hyperspectral characteristics and yield estimation of winter wheat (triticum aestivum) under low temperature injury. *Scientific Reports*, **10**, 244.

Xiong, C., Fan, C., and Huang, X. (2020). Multipath exploitation with time reversal waveform covariance matrix for SNR maximization. *Remote Sensing*, **12**, 3565.

Xue, J. and Su, B. (2017). Significant remote sensing vegetation indices: A review of developments and applications. *Journal of Sensors*, 1–17.

Zang, H., Wang, Y., Ru, L., Zhou, M., Chen, D., Zhao, Q., Zhang, J., Li, G., and Zheng, G. (2022). Detection method of wheat spike improved YOLOv5s based on the attention mechanism. *Frontiers in Plant Science*, **13**, 993244.

Zhou, H., Yang, J., Lou, W., Sheng, L., Li, D., and Hu, H. (2023). Improving grain yield prediction through fusion of multi-temporal spectral features and agronomic trait parameters derived from UAV imagery. *Frontiers in Plant Science*, **14**, 1217448.

ANNEXURE

Table 5: Plant varieties and genotypes used for imaging

S.No.	Variety/Genotype	S.No.	Variety/Genotype	S.No.	Variety/Genotype
1	BarhamRF	23	DBW51	45	LOKBOLD
2	C518RF	24	DL1266-2	46	LSP3043
3	CG1029	25	Drysdale	47	MACS6222
4	CHIRIYA3	26	HD3271	48	MP1358
5	Coreset110	27	HD2967	49	NIAW34
6	Coreset112	28	HD3086	50	NP4
7	Coreset128	29	HD308673RF	51	NP852
8	Coreset129	30	HD3237	52	NW1014
9	Coreset141	31	HD3271	53	PBW343
10	Coreset17	32	HD2329	54	PBW502
11	Coreset32	33	HD2687	55	PBW644
12	Coreset38	34	HD2833	56	PBW681
13	Coreset41	35	HD2864	57	PBW825
14	Coreset42	36	HD3171	58	PISSI-LOCAL
15	Coreset53	37	HI1500RF	59	RAJ4229
16	Coreset58	38	HI1544	60	Raj3077
17	Coreset66	39	HS611	61	Raj4037
18	Coreset7	40	IHD-29672	62	UAS446
19	Coreset74	41	K9465	63	UASBW10453
20	Coreset93	42	KACHU#1	64	UP2425
21	DBW110	43	LOK1	65	UP262
22	DBW222	44	LOK67	66	VL892

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 24, No. 1, 2026 (New Series), pp 297–321
http://www.ssca.org.in/journal

Bayesian Nonparametrics for Gene-Gene and Gene-Environment Interactions in Case-Control Studies: A Synthesis and Extension

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: 09 July 2025; Revised: 27 April 2026; Accepted: 04 May 2026

Abstract

Gene-gene and gene-environment interactions are widely believed to play significant roles in explaining the variability of complex traits. While substantial research exists in this area, a comprehensive statistical framework that addresses multiple sources of uncertainty simultaneously remains lacking. In this article, we synthesize and propose extension of a novel class of Bayesian nonparametric approaches that account for interactions among genes, loci, and environmental factors while accommodating uncertainty about population substructure. Our contribution is threefold: (1) We provide a unified exposition of hierarchical Bayesian models driven by Dirichlet processes for genetic interactions, clarifying their conceptual advantages over traditional regression approaches; (2) We shed light on new computational strategies that combine transformation-based MCMC with parallel processing for scalable inference; and (3) We present enhanced hypothesis testing procedures for identifying disease-predisposing loci. Through applications to myocardial infarction data, we demonstrate how these methods offer biological insights not readily obtainable from standard approaches. Our synthesis highlights the advantages of Bayesian nonparametric thinking in genetic epidemiology while providing practical guidance for implementation. Our parallel C code for implementing our key Bayesian nonparametric model based on hierarchies of Dirichlet processes, is available at `https://github.com/Sourabh-Bhattacharya/HDP-REALDATA-CODE/blob/main/HDP_REALDATA.zip`.

Key words: Dirichlet process; Disease predisposing loci; Epistasis; Mixture model; Parallel computing; Transformation based Markov Chain Monte Carlo.

AMS Subject Classifications: 62K05, 62F15, 92D10

The video recording of the paper made under the SSCA's Online Lecture series is available at the Youtube channel URL `https://youtu.be/3xK2CWg9EA0`.

Corresponding Author: Durba Bhattacharya
Email: durba@sxccal.edu

1. Introduction

1.1. The challenge of genetic interactions in complex diseases

Complex diseases such as cardiovascular disorders, diabetes, and psychiatric conditions result from intricate networks of genetic and environmental factors. While genome-wide association studies (GWAS) have identified numerous single nucleotide polymorphisms (SNPs) associated with disease risk, these explain only a small fraction of heritability (Larson and Schaid, 2013). The "missing heritability" problem has spurred interest in gene-gene (epistasis) and gene-environment interactions as potential explanations. Traditional approaches to studying these interactions face several significant limitations that our work aims to address.

Most existing methods rely on linear or additive models that may not adequately capture the complex biological pathways through which genetic factors interact (Wang *et al.*, 2010). These simplified modeling assumptions often fail to represent the intricate biochemical networks that characterize many complex diseases. Furthermore, the failure to properly account for population stratification–the presence of genetic substructure within study populations–can lead to inflated false positive rates in association studies (Bhattacharjee *et al.*, 2010). This issue is particularly problematic in genetically diverse populations where different subgroups may have distinct allele frequencies unrelated to disease risk.

The computational burden represents another substantial challenge in studying genetic interactions. Testing all possible pairwise SNP-SNP interactions becomes infeasible for genome-wide data, leading researchers to adopt heuristic screening methods that may miss important interactions or identify spurious ones. Additionally, many current approaches provide point estimates without adequately characterizing the uncertainty in model structure, particularly regarding the number of underlying sub-populations or the complexity of interaction networks. This lack of comprehensive uncertainty quantification limits the reliability and interpretability of findings from genetic interaction studies.

1.2. Our contribution: a Bayesian nonparametric synthesis

This article synthesizes and extends a series of Bayesian nonparametric models developed to address these challenges. Unlike previous works that presented these models in isolation (Bhattacharya and Bhattacharya 2018, 2020, 2024), we provide a unified framework connecting gene-gene interaction models, gene-environment extensions, and hierarchical Dirichlet process formulations. Our synthesis represents a comprehensive overview of this methodological approach, highlighting both theoretical foundations and practical implementation considerations.

We develop enhanced computational strategies that leverage parallel processing and transformation-based MCMC (Dutta and Bhattacharya, 2014) for practical implementation of these complex models. These computational innovations make it feasible to apply Bayesian nonparametric methods to realistic genetic datasets of meaningful size. Additionally, we present new hypothesis testing procedures for identifying disease-predisposing loci in the presence of population stratification, offering more robust alternatives to traditional association tests. Finally, we provide comprehensive applications demonstrating biological insights from myocardial infarction data, showing how these methods can uncover relation-

ships not readily apparent using standard approaches.

Our approach fundamentally departs from standard logistic regression by modeling genotypes conditional on disease status using Dirichlet process mixtures. This inversion of the typical modeling relationship allows us to capture several important features simultaneously. We can model uncertainty in population substructure nonparametrically, allowing the data to inform about the number and characteristics of genetic subgroups. We capture gene-gene interactions through covariance structures rather than regression coefficients, providing a more flexible representation of genetic dependencies. Furthermore, we accommodate subject-specific environmental effects through hierarchical modeling, enabling personalized assessment of genetic risk factors.

1.3. Article structure

The remainder of this article is organized as follows. Section 2 introduces our modeling philosophy and contrasts it with traditional approaches, explaining the conceptual shift from regression-based to conditional genotype modeling. Section 3 presents the gene-gene interaction model with computational details, including our parallel implementation strategy. Section 4 extends this framework to incorporate gene-environment interactions, discussing both modeling extensions and enhanced hypothesis testing procedures. Section 5 introduces the hierarchical Dirichlet process formulation, which addresses limitations of previous models by allowing more flexible sharing patterns. Section 6 presents comprehensive applications to myocardial infarction data, demonstrating the practical utility of our methods. Section 7 provides extensive sensitivity analyses to assess the robustness of our results to key modeling choices across all three frameworks. Finally, Section 8 discusses comparisons with existing methods, addresses limitations and robustness considerations, and outlines future research directions.

2. A unified Bayesian nonparametric framework

2.1. Modeling philosophy: from regression to conditional genotype models

Traditional GWAS analyze disease status Y given genotypes X using logistic regression models of the form $P(Y = 1 \mid X) = \text{logit}^{-1}(\beta_0 + \sum \beta_j X_j + \sum \beta_{jk} X_j X_k)$. This approach, while interpretable and widely used, faces the "curse of dimensionality" when considering interactions and makes strong parametric assumptions about the relationship between genotypes and disease risk. The logistic regression framework requires specifying which interactions to include, typically limiting consideration to pairwise effects due to computational constraints. Furthermore, these models assume that genetic effects are additive on the log-odds scale, which may not reflect the true biological mechanisms underlying complex diseases.

Our approach inverts this relationship by modeling genotypes conditional on disease status using finite mixture models: $[\boldsymbol{X} \mid Y = k] = \sum_{m=1}^{M} \pi_m f(\boldsymbol{X} \mid \boldsymbol{\theta}_{mk})$, where the mixture components represent latent sub-populations. By placing Dirichlet process priors on the parameters $\{\boldsymbol{\theta}_{mk}\}$, we allow the number of sub-populations to be inferred from the data rather than specified in advance. This modeling strategy represents a fundamental shift in perspective: instead of asking how genotype affects disease risk, we ask how the distribution

of genotypes differs between cases and controls. This approach naturally accommodates population structure, as mixture components can correspond to genetic subgroups with distinct allele frequencies.

The conditional modeling approach offers several conceptual advantages. First, it directly addresses the issue of population stratification by explicitly modeling genetic heterogeneity. Second, it provides a natural framework for identifying disease-associated genetic variants through comparison of genotype distributions between cases and controls. Third, it allows for flexible modeling of dependencies among genetic loci through the specification of the mixture components. Rather than assuming independence or simple correlation structures, we can incorporate complex dependence patterns that may reflect biological relationships among genes.

2.2. Key advantages of our framework

Our Bayesian nonparametric framework offers several key advantages over traditional approaches. First, it provides flexible modeling of population structure through Dirichlet processes, which automatically discover population substructure without requiring pre-specified ancestry information. This is particularly valuable in studies of admixed populations or when detailed ancestry information is unavailable. The nonparametric nature of Dirichlet processes allows the number of sub-populations to be determined by the data, avoiding both underfitting and overfitting that can occur with fixed finite mixture models.

Second, our approach conceptualizes gene-gene interactions as statistical dependence rather than additive effects on disease risk. We model these interactions through covariance structures in hierarchical priors, capturing complex dependencies that may not be well represented by linear or additive models. This representation aligns with biological understanding that genetic factors often act in networks rather than independently. By focusing on covariance structures, we can identify sets of genes that co-vary in their effects on disease risk, potentially revealing functional pathways or biological modules.

Third, the conditional independence structure of our models enables scalable computation through parallel implementation. Different components of the model can be updated independently, allowing us to leverage modern parallel computing architectures. This computational efficiency makes it feasible to apply our methods to datasets with large numbers of genetic variants and samples, addressing a key limitation of many Bayesian nonparametric approaches.

Fourth, our fully Bayesian approach provides comprehensive uncertainty quantification through posterior distributions for all quantities of interest. This includes not only point estimates of genetic effects but also measures of uncertainty about population structure, interaction strengths, and model complexity. Proper characterization of uncertainty is particularly important in genetic studies, where sample sizes are often limited relative to the number of genetic variants considered.

2.3. Notation and data structure

We consider case-control data with N_k subjects in group $k \in \{0, 1\}$, where $k = 0$ denotes controls and $k = 1$ denotes cases. For each subject i, we observe genotypes

represented as $x_{ijr}^s \in \{0,1\}$ for chromosome $s \in \{1,2\}$, gene $j \in \{1,\ldots,J\}$, and locus $r \in \{1,\ldots,L_j\}$. Here, $x_{ijr}^s = 1$ indicates the presence of the minor allele on chromosome s at locus r of gene j for individual i, while $x_{ijr}^s = 0$ indicates its absence. Additionally, we may observe environmental covariates $\boldsymbol{E}_i$ for each individual, which could include factors such as sex, age, smoking status, or other exposures. Disease status is denoted by $Y_i \in \{0,1\}$. The total number of genes is J, and L_j represents the number of loci within gene j. This notation provides a comprehensive framework for representing the complex hierarchical structure of genetic data, with variations at the chromosome, locus, gene, individual, and group levels.

3. Gene-gene interaction model

3.1. Model specification: a roadmap

Our model for gene-gene interactions comprises three key components that work together to capture the complex structure of genetic data. First, we employ subject-level mixture models that represent genotypes as arising from finite mixtures, with each mixture component corresponding to a latent sub-population. This approach allows us to account for population stratification without requiring prior specification of ancestry groups. Second, we place Dirichlet process priors on the parameters of these mixture models, enabling flexible sharing of mixture components across subjects and automatic determination of the number of sub-populations. The Dirichlet process framework provides a principled Bayesian nonparametric approach to modeling population structure, with the precision parameter controlling the degree of sharing among subjects. Third, we introduce a hierarchical interaction structure through matrix-normal priors that capture dependencies among genes. This hierarchical structure allows us to model gene-gene interactions at multiple levels, from pairwise correlations to higher-order dependencies.

The combination of these three components creates a comprehensive modeling framework that addresses several limitations of traditional approaches. The mixture model component handles population stratification, the Dirichlet process prior provides flexibility in modeling the number and characteristics of sub-populations, and the hierarchical interaction structure captures complex dependencies among genetic factors. Importantly, these components are integrated in a coherent Bayesian framework that allows for uncertainty quantification at all levels, from individual genotype probabilities to overall population structure.

3.2. Detailed formulation

For each gene j and group k, we model the genotype vector $\boldsymbol{X}_{ijk} = (x_{ijk1},\ldots,x_{ijkL_j})$ using a finite mixture distribution: $[\boldsymbol{X}_{ijk}] = \sum_{m=1}^{M} \pi_{mjk} \prod_{r=1}^{L_j} \text{Bernoulli}(x_{ijkr} \mid p_{mjkr})$. Here, M represents the maximum number of mixture components, which we set to a sufficiently large value to ensure adequate model flexibility. The mixing weights π_{mjk} are fixed at $1/M$ for all m, j, and k. This choice of fixed equal weights, while somewhat unconventional, has been shown in previous work (Majumdar *et al.*, 2013) to yield better performance in estimating the true number of mixture components compared to using Dirichlet priors. The robustness of this choice is discussed further in Section 8.2.

The allele frequency parameters p_{mjkr} follow a hierarchical structure based on Dirichlet process designed to capture both locus-specific effects and dependencies among genes.

Specifically, it holds that the marginal prior distribution of the allele frequency parameters is $p_{mjkr} \sim \text{Beta}(\nu_{1jkr}, \nu_{2jkr})$, where the Beta distribution parameters are themselves modeled as $\nu_{1jkr} = \exp(u_r + \lambda_{jk})$ and $\nu_{2jkr} = \exp(v_r + \lambda_{jk})$. The parameters u_r and v_r are locus-specific effects assumed to follow independent standard normal distributions: $u_r, v_r \sim N(0, 1)$. Allowing u_r and v_r to differ ensures that the mean of the Beta distribution for p_{mjkr} depends on the specific locus r, capturing variation in allele frequencies across the genome.

Gene-gene interactions are captured through a matrix-normal prior on the parameter matrix $\boldsymbol{\lambda} = \{\lambda_{jk}\}$, which represents gene- and group-specific effects. We assume $\boldsymbol{\lambda} \sim N(\boldsymbol{\mu}, \boldsymbol{A} \otimes \boldsymbol{\Sigma})$, where $\boldsymbol{A}$ represents the covariance matrix capturing dependencies among genes, and $\boldsymbol{\Sigma}$ represents the covariance matrix capturing dependency between the genotype distribution of case and control, given any gene. The Kronecker product structure $\boldsymbol{A} \otimes \boldsymbol{\Sigma}$ allows for separable covariance structures across genes and disease status, providing a computationally tractable yet flexible representation of dependencies. This formulation enables us to model complex interaction patterns while maintaining computational feasibility through the separable covariance structure.

3.3. Computational strategy

The conditional independence structure of our model enables efficient computational implementation through a combination of parallel processing and specialized MCMC techniques. The key insight is that, given the interaction parameters $\boldsymbol{\lambda}$, the mixture models for different gene-group pairs (j, k) are conditionally independent. This conditional independence allows us to update the mixture parameters for each (j, k) pair in parallel across multiple processors. Each processor handles a subset of the gene-group pairs, updating the allocation variables z_{ijk}, the allele frequency parameters p_{mjkr}, and the other mixture-related parameters independently of the others.

For updating the interaction parameters $\boldsymbol{\lambda}$, we employ transformation-based MCMC (TMCMC), which updates all elements of $\boldsymbol{\lambda}$ simultaneously through deterministic transformations of a low-dimensional random variable. TMCMC is particularly well-suited for high-dimensional parameter spaces, as it can propose moves in directions that respect the correlation structure of the posterior distribution. In our implementation, we use a single random variable to generate proposals for all elements of $\boldsymbol{\lambda}$, with the transformation designed to maintain reasonable acceptance rate and mixing.

The Dirichlet process mixture components are updated using a fast Gibbs sampling algorithm specifically designed for finite mixtures with Dirichlet process priors. This algorithm leverages the Polya urn representation of the Dirichlet process to efficiently update the allocation of subjects to mixture components and the parameters associated with each component. The algorithm alternates between updating the allocation variables given the parameters and updating the parameters given the allocations, with both steps being computationally efficient due to conjugacy relationships.

Figure 3.1 displays the schematic diagram for our gene-gene interaction model. For each gene and case-control status, genotype data are modeled using Dirichlet process-based mixtures that capture sub-population structure. SNP-level dependencies and gene-gene interactions are introduced through a matrix-normal prior on latent interaction parameters. The modular design of the model allows efficient parallel computation: gene-specific mixture

components are updated independently across processors, while the interaction parameters are updated centrally using TMCMC. Our parallel codes are written in C, in accordance with the Message Passing Interface (MPI) protocol.

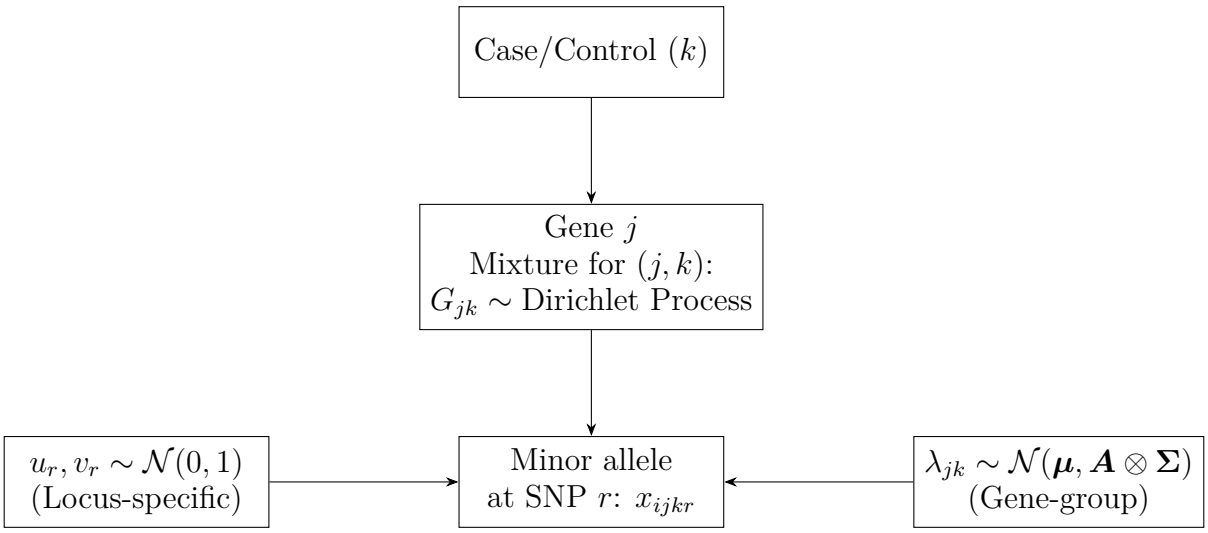


Figure 3.1: Schematic representation of the Bayesian model for gene-gene interactions.

3.4. Hypothesis testing and DPL identification

We develop comprehensive Bayesian hypothesis testing procedures to evaluate various aspects of genetic associations within our framework. For assessing gene effects, we compare clustering patterns between cases and controls using metrics based on posterior distributions of the mixture components. Significant differences in clustering patterns indicate that the genetic structure differs between cases and controls, suggesting an association between the gene and disease status. This approach provides a more nuanced assessment than traditional association tests, as it considers the entire distribution of genotypes rather than just summary statistics.

To test for gene-gene interactions, we examine elements of the covariance matrix $\boldsymbol{A}$ in the matrix-normal prior on $\boldsymbol{\lambda}$. Non-zero off-diagonal elements in $\boldsymbol{A}$ indicate dependencies between the effects of different genes, which we interpret as statistical interactions. We compute posterior probabilities that specific elements of $\boldsymbol{A}$ are non-zero, providing a quantitative measure of interaction strength with associated uncertainty. This approach allows us to identify pairs or groups of genes that show coordinated effects on disease risk, potentially revealing biological pathways or functional modules.

Our key idea for identifying disease-predisposing loci (DPL) may be likened to computing posterior probabilities that specific SNPs show differential distributions between cases and controls. Specifically, for each SNP r in gene j, our idea is analogous, in principle, to computing $P(\text{SNP } r \text{ is DPL} \mid \text{Data}) = P\left(\left|p_{.jr}^{\text{case}} - p_{.jr}^{\text{control}}\right| > \delta \mid \text{Data}\right)$, where δ is a clinically meaningful threshold for allele frequency differences. The probabilities $p_{.jr}^{\text{case}}$ and $p_{.jr}^{\text{control}}$ represent the allele frequencies in cases and controls, respectively. This approach provides a principled Bayesian method for DPL identification that accounts for multiple sources of

uncertainty, including population stratification and estimation error.

Our hypothesis testing framework extends beyond simple significance testing to include measures of effect size and uncertainty. For each test, we report posterior probabilities along with credible intervals for relevant parameters, providing a comprehensive picture of the evidence for various genetic associations. This Bayesian approach naturally incorporates multiple testing considerations through the prior distributions and posterior probabilities, avoiding the need for ad hoc corrections that can be problematic in high-dimensional settings.

4. Gene-environment interaction model

4.1. Extending the framework

To incorporate environmental covariates $\boldsymbol{E}_i$, we extend our modeling framework to allow for subject-specific mixture distributions that depend on environmental factors. The extended model represents genotypes as $[\boldsymbol{X}_{ijk} \mid \boldsymbol{E}_i] = \sum_{m=1}^{M} \pi_{mijk} \prod_{r=1}^{L_j} \text{Bernoulli}(x_{ijkr} \mid p_{mijkr})$, where now the mixing weights and component parameters can vary across individuals based on their environmental exposures. As in the gene-gene interaction model, we fix $\pi_{mijk} = 1/M$ for all (i, j, k, m), maintaining computational simplicity while allowing flexibility through the component parameters.

The key extension in the gene-environment interaction model lies in the parameterization of the Beta distribution parameters for the allele frequencies. We now model these as $\nu_{1ijkr} = \exp(u_{jr} + \lambda_{ijk} + \mu_{jk} + \boldsymbol{\beta}_{jk}^{\top}\boldsymbol{E}_i)$ and $\nu_{2ijkr} = \exp(v_{jr} + \lambda_{ijk} + \mu_{jk} + \boldsymbol{\beta}_{jk}^{\top}\boldsymbol{E}_i)$. This expanded parameterization includes several new terms: u_{jr} and v_{jr} are gene- and locus-specific effects; λ_{ijk} represents individual-specific genetic effects that capture residual variation not explained by other factors; μ_{jk} are gene- and group-specific intercept terms; and $\boldsymbol{\beta}_{jk}$ are vectors of coefficients capturing the effects of environmental covariates on the genetic parameters for gene j in group k.

The inclusion of environmental effects through the term $\boldsymbol{\beta}_{jk}^{\top}\boldsymbol{E}_i$ allows the distribution of genotypes to vary systematically with environmental exposures. When $\boldsymbol{\beta}_{jk} \neq \mathbf{0}$, environmental factors modify the genetic parameters for gene j in group k, representing gene-environment interaction. Importantly, this formulation allows environmental factors to affect not only the mean genetic effects but also the covariance structure among genes, as the environmental terms enter into the hierarchical model that generates the λ_{ijk} parameters. This enables the model to capture how environmental exposures might modify not just individual genetic effects but also the patterns of interaction among genes.

The covariance structure in this extended model continues to capture how environment modifies gene-gene interactions. The individual-specific effects λ_{ijk} inherit dependence structure from the hierarchical model, with environmental factors potentially influencing both the mean and covariance of these effects. This allows for rich patterns of gene-environment interaction, including scenarios where environmental exposures alter the strength or pattern of genetic correlations. For example, an environmental factor might strengthen the correlation between two genes in cases but weaken it in controls, or it might induce correlations among genes that are independent in the absence of the exposure.

4.2. Enhanced hypothesis testing

We extend our hypothesis testing framework to address questions specific to gene-environment interactions. After testing for overall genetic effect by extending the previous method, we test for the presence of gene-environment interaction by evaluating the null hypothesis $H_0 : \boldsymbol{\beta}_{jk} = \mathbf{0}$ for specific genes and environmental factors. We compute posterior probabilities that $\boldsymbol{\beta}_{jk}$ differs from zero, providing a direct measure of evidence for gene-environment interaction. This approach allows us to identify genes whose effects on disease risk are modified by environmental exposures, which is of particular interest for understanding disease etiology and targeted interventions.

We also test for joint effects involving both gene-gene and gene-environment interactions. This involves evaluating whether environmental factors modify the patterns of interaction among genes, which we assess by examining how environmental covariates affect the covariance structure in the model. Specifically, we test whether parameters governing the dependence of the covariance structure on environmental factors differ from zero. Significant findings indicate that environmental exposures alter the network of genetic interactions, potentially revealing mechanisms through which environment influences disease risk.

Third, we develop methods for stratified DPL identification that account for environmental heterogeneity. Rather than identifying DPL that show average differences between cases and controls, we identify SNPs whose effects depend on environmental exposure. This involves computing posterior probabilities that the difference in allele frequencies between cases and controls varies across levels of environmental factors. For example, we might identify SNPs that show strong associations in exposed individuals but weak or no associations in unexposed individuals, or vice versa. This stratified approach can reveal genetic factors that are important only under specific environmental conditions, providing insights into the context-dependence of genetic risk.

Apart from these, we additionally develop visualization tools to help interpret complex interaction patterns, including heatmaps of posterior interaction probabilities and network diagrams showing how genes and environmental factors are connected through interaction effects.

Figure 4.1 provides the schematic diagram for our gene-environmental interaction model. Environmental covariates influence individual-level Dirichlet process mixtures, allowing the model to account for personalized effects of environmental exposures on genotype distributions. The prior structure integrates locus-specific, gene-specific, and environment-dependent parameters. Parallel computation is employed by updating gene-environment specific components in parallel and interaction parameters centrally via TMCMC. As before, our parallel codes are written in C, leveraging the MPI protocol.

5. Hierarchical Dirichlet process model

5.1. Motivation and limitations of previous models

The gene-environment interaction model presented in Section 4 makes the simplifying assumption that environmental effects act uniformly on gene-gene interactions across all individuals. This assumption may not hold in many practical settings for several rea-

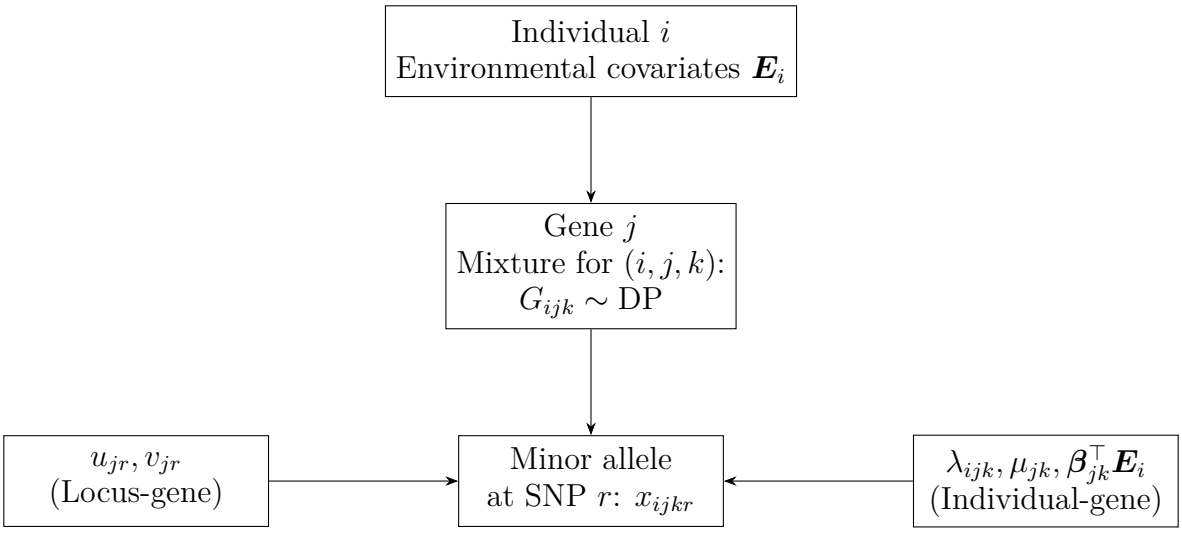


Figure 4.1: Diagram of the extended Bayesian framework incorporating gene-environment interactions.

sons. First, different individuals may experience different levels or types of environmental exposure, leading to heterogeneous effects on genetic interactions. For example, the effect of smoking on genetic risk factors for cardiovascular disease may depend on duration and intensity of smoking, which varies across individuals. Second, environmental effects may be heterogeneous across the population due to unmeasured factors or effect modifiers. Genetic background, other environmental exposures, or lifestyle factors might modify how a particular environmental factor influences genetic interactions. Third, gene-environment interactions may be context-dependent, with effects manifesting only under specific conditions or in specific subgroups. The uniform effect assumption fails to capture this context-dependence, potentially missing important biological relationships.

These limitations motivate the development of a more flexible modeling approach that can accommodate heterogeneous and context-dependent gene-environment interactions. The hierarchical Dirichlet process (HDP) model provides a principled nonparametric framework for capturing complex sharing patterns among individuals, genes, and groups in a data-driven manner. Rather than imposing a parametric structure on how environmental factors influence genetic dependencies, the HDP model learns these relationships nonparametrically from the data. This approach can discover complex patterns of interaction that might be missed by more restrictive parametric models, while still providing a coherent probabilistic framework for inference.

5.2. HDP formulation

We address the limitations of previous models through a three-level hierarchical Dirichlet process formulation that introduces flexible, nonparametric dependence structures among genes, environmental variables, and case-control status. Our model represents a significant extension of the traditional HDP framework (Teh *et al.*, 2006) by incorporating an additional level of hierarchy that specifically captures case-control dependence while allowing for subject-specific gene-gene interactions influenced by environmental factors.

For each individual i in group k, gene j, and mixture component m, we assume that the minor allele frequency vector $\boldsymbol{p}_{mijk} = (p_{mijk1}, p_{mijk2}, \ldots, p_{mijkL})$ is generated from a hierarchy of Dirichlet processes:

$$\boldsymbol{p}_{1ijk}, \boldsymbol{p}_{2ijk}, \ldots, \boldsymbol{p}_{Mijk} \overset{iid}{\sim} \boldsymbol{G}_{ijk} \tag{5.1}$$

$$\boldsymbol{G}_{ijk} \sim \mathrm{DP}(\alpha_{G,ik}\boldsymbol{G}_{0,jk}) \tag{5.2}$$

where $\mathrm{DP}(\alpha_{G,ik}\boldsymbol{G}_{0,jk})$ denotes a Dirichlet process with base measure $\boldsymbol{G}_{0,jk}$ and precision parameter $\alpha_{G,ik}$. The environmental dependence enters through the precision parameter:

$$\log(\alpha_{G,ik}) = \mu_G + \boldsymbol{\beta}_G^\top \boldsymbol{E}_{ik} \tag{5.3}$$

where $\boldsymbol{E}_{ik}$ is a d-dimensional vector of environmental variables for individual i in group k, $\boldsymbol{\beta}_G$ is a d-dimensional vector of regression coefficients, and μ_G is an intercept term.

At the second level of the hierarchy, we assume:

$$\boldsymbol{G}_{0,jk} \overset{iid}{\sim} \mathrm{DP}(\alpha_{G_0,k}\boldsymbol{H}_k); \quad j = 1, \ldots, J \tag{5.4}$$

with

$$\log(\alpha_{G_0,k}) = \mu_{G_0} + \boldsymbol{\beta}_{G_0}^\top \bar{\boldsymbol{E}}_k \tag{5.5}$$

where $\bar{\boldsymbol{E}}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} \boldsymbol{E}_{ik}$ is the average environmental variable in group k.

The third level of hierarchy creates dependence between case and control groups:

$$\boldsymbol{H}_k \overset{iid}{\sim} \mathrm{DP}(\alpha_H \tilde{\boldsymbol{H}}); \quad k = 0, 1 \tag{5.6}$$

with

$$\log(\alpha_H) = \mu_H + \boldsymbol{\beta}_H^\top \bar{\bar{\boldsymbol{E}}} \tag{5.7}$$

where $\bar{\bar{\boldsymbol{E}}} = (\bar{\boldsymbol{E}}_0 + \bar{\boldsymbol{E}}_1)/2$ is the overall average environmental variable.

The global base measure $\tilde{\boldsymbol{H}}$ is specified as:

$$p_{mijkr} \overset{iid}{\sim} \mathrm{Beta}(\nu_1, \nu_2) \quad \text{under } \tilde{\boldsymbol{H}} \tag{5.8}$$

where $\nu_1, \nu_2 > 0$ are fixed hyperparameters.

This hierarchical structure has a natural interpretation in terms of genetic architecture. The subject-level distributions $\boldsymbol{G}_{ijk}$ capture individual-specific genetic patterns, which may differ due to unique genetic backgrounds, environmental exposures, or other individual factors. These subject-level distributions share components through the gene-group level distributions $\boldsymbol{G}_{0,jk}$, which represent common patterns for each gene in each disease group. The group-level distributions $\boldsymbol{H}_k$ capture overall genetic patterns for cases and controls separately, allowing for differences in genetic architecture between affected and unaffected individuals. Finally, the global distribution $\tilde{\boldsymbol{H}}$ represents the overall genetic background of the population.

5.3. Chinese restaurant process analogy

The hierarchical structure can be understood through an extended Chinese restaurant process analogy (Teh *et al.*, 2006). For each group $k = 0, 1$, imagine J restaurants (genes). Each individual i visits these restaurants, where at the j-th restaurant, they are seated at tables (mixture components) that serve dishes (allele frequency parameters). The dishes at different tables within the same restaurant are drawn from $\boldsymbol{G}_{0,jk}$, which itself is drawn from $\boldsymbol{H}_k$. This creates sharing of dishes across tables within a restaurant (within a gene across individuals).

Now consider that all restaurants share a global menu of dishes from $\tilde{\boldsymbol{H}}$. Different restaurants (genes) may serve different selections from this global menu, with the selections drawn from $\boldsymbol{H}_k$. This creates sharing of dishes across restaurants (across genes), with the degree of sharing controlled by α_H. The environmental variables influence how likely customers (individuals) are to sit at new tables through $\alpha_{G,ik}$, and how likely restaurants are to offer new dishes through $\alpha_{G_0,k}$ and α_H.

This analogy clarifies how our model creates dependence: individuals sharing the same table at a restaurant share the same dish (allele frequency parameters) for that gene, creating dependence among individuals. Different restaurants (genes) serving the same dish creates dependence among genes. The sharing of dishes between case and control restaurants ($k = 0$ and $k = 1$) creates case-control dependence. Environmental factors influence these sharing probabilities, thereby modulating the dependence structure.

5.4. Dependence structure induced by the HDP model

Our HDP model induces three key types of dependence that are crucial for understanding gene-gene and gene-environment interactions:

5.4.1. Dependence among individuals

From the Polya urn representation, the joint distribution of $\boldsymbol{p}_{Mijk} = \{\boldsymbol{p}_{1ijk}, \ldots, \boldsymbol{p}_{Mijk}\}$ shows that individuals share dish parameters ϕ_{tijk} drawn from $\boldsymbol{G}_{0,jk}$. The probability of sharing depends on $\alpha_{G,ik}$, which in turn depends on individual environmental variables $\boldsymbol{E}_{ik}$ through equation (5.3). This creates environmental-modulated dependence among individuals: individuals with similar environmental exposures are more likely to share genetic patterns.

Importantly, the marginal distribution of $\boldsymbol{p}_{mijk}$ is $\boldsymbol{G}_{0,jk}$, which does not depend on $\boldsymbol{E}_{ik}$. This is biologically desirable: population minor allele frequencies should not be affected by environmental variables, although environmental exposure may influence how individuals cluster together in their genetic patterns.

5.4.2. Dependence among genes

Gene-gene dependence arises through the sharing of dishes (parameters) across restaurants (genes). The parameters ϕ_{tijk} for different genes j share common values from $\boldsymbol{H}_k$, creating dependence among genes. The degree of this dependence is influenced by $\alpha_{G_0,k}$, which depends on the group average environment $\bar{\boldsymbol{E}}_k$ through equation (5.5), and also indirectly

through $\alpha_{G,ik}$ which affects the number of tables τ_{ijk}.

This structure ensures that gene-gene interactions are specific to individuals and are influenced by both individual environmental variables ($\boldsymbol{E}_{ik}$) and group averages ($\bar{\boldsymbol{E}}_k$), while the marginal distributions of individual genes remain unaffected by environment.

5.4.3. Case-control dependence

The sharing of dishes between case and control restaurants (through $\boldsymbol{H}_0$ and $\boldsymbol{H}_1$ sharing from $\tilde{\boldsymbol{H}}$) creates dependence between case and control groups. This dependence captures factors that affect both cases and controls but may not be explicitly measured, such as population stratification or unmeasured environmental factors.

The degree of case-control sharing is controlled by α_H, which depends on the overall average environment $\bar{\bar{\boldsymbol{E}}}$ through equation (5.7). This allows environmental factors to modulate the similarity between cases and controls in their genetic patterns.

5.5. Advantages of HDP approach

The hierarchical Dirichlet process model offers several advantages over the parametric models discussed previously. First, it provides flexible, nonparametric dependence structure: unlike previous matrix-normal based models that assume uniform environmental effects on covariance structures, our HDP model allows environment to influence dependence structures in a flexible, nonparametric manner that varies across individuals. Second, it enables subject-specific gene-gene interactions: each individual can have their own pattern of gene-gene interactions, with environmental factors influencing these individual-specific patterns. This is more realistic than assuming a single correlation structure for all individuals. Third, it maintains biological interpretability: the model preserves the biologically important property that environmental variables influence dependence structures but not marginal allele frequency distributions. Only dependence structures (how genes interact) are affected by environment, not the genes themselves. Fourth, it offers automatic complexity control: the Dirichlet process priors automatically determine the appropriate number of mixture components at each level of the hierarchy, avoiding the need to pre-specify the number of sub-populations or interaction patterns. Fifth, it achieves computational efficiency through parallelization: the conditional independence structure of the HDP model enables efficient parallel implementation, with different genes and individuals updated independently given the higher-level parameters.

5.6. Computational implementation

We implement the HDP model using a novel parallel MCMC algorithm that combines Gibbs sampling steps with transformation-based MCMC (TMCMC). The algorithm leverages the conditional independence structure of the model: given the higher-level distributions ($\boldsymbol{H}_k$ and $\boldsymbol{G}_{0,jk}$), the subject-level parameters can be updated independently across individuals and genes. This allows for parallel computation across multiple processors. Key components of our computational strategy include retrospective sampling (Papaspiliopoulos and Roberts, 2008) for updating parameters from Dirichlet process mixtures, which avoids infinite-dimensional representations by generating parameters as needed. The al-

location variables z_{ijk} and allele frequency parameters p_{mijkr} are updated in parallel across different (i, j, k) combinations using Gibbs steps that exploit conjugacy. The parameters $\mu_G, \boldsymbol{\beta}_G, \mu_{G_0}, \boldsymbol{\beta}_{G_0}, \mu_H, \boldsymbol{\beta}_H$ are updated in a single block using a mixture of additive and multiplicative TMCMC (Dey and Bhattacharya, 2016), which efficiently explores correlated parameter spaces. Our parallel implementation in C, as before, uses MPI for communication between processors, with careful load balancing to ensure efficient parallel scaling.

The schematic diagram of our HDP model is presented in Figure 5.1. This fully nonparametric framework models dependencies across individuals, genes, and groups through a three-level hierarchy of Dirichlet processes. Environmental covariates influence the precision parameters at each level, allowing flexible, individualized representation of interaction structures. The base distribution is a Beta prior on allele frequencies. This hierarchy enables rich modeling of stratification and interaction while maintaining computational scalability.

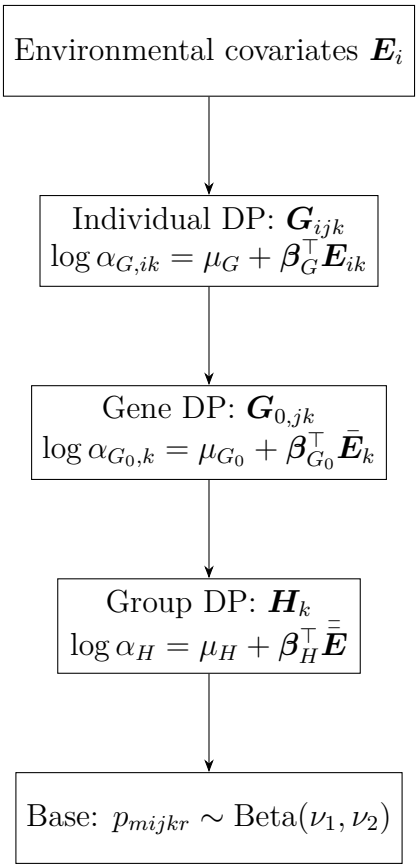


Figure 5.1: Schematic of the hierarchical Dirichlet process (HDP) model for gene-gene and gene-environment interactions.

5.7. Hypothesis testing in the HDP framework

We extend our hypothesis testing procedures to the HDP framework with specific tests for: gene effects, testing $H_0 : h_{0j} = h_{1j}$ for $j = 1, \ldots, J$, where h_{0j} and h_{1j} are the marginal distributions of genotypes for controls and cases respectively for gene j; environmental effects, testing $H_0 : \boldsymbol{\beta}_G = \mathbf{0}$, $H_0 : \boldsymbol{\beta}_{G_0} = \mathbf{0}$, and $H_0 : \boldsymbol{\beta}_H = \mathbf{0}$ for the environmental coefficients at different levels of the hierarchy; gene-gene interactions, defining subject-specific gene-gene

interaction measures $C(i, j_1, j_2, k)$ as covariances between logit-transformed average allele frequencies for genes j_1 and j_2 for individual i in group k, and testing $H_0 : C(i, j_1, j_2, k) = 0$; and case-control dependence, assessing whether the sharing between cases and controls (through α_H) is significant, indicating common factors affecting both groups. The interpretation of these tests follows a logical framework: if genes have no effect but environment affects dependence structures, then environment influences gene-gene interactions without affecting disease status. If genes have effects but environment doesn't affect interactions, then purely genetic factors determine disease. If both genes and environment are significant and affect interactions, then gene-environment interactions influence disease risk.

6. Applications to myocardial infarction data

6.1. Data description

We apply our Bayesian nonparametric framework to a myocardial infarction (MI) case-control dataset to demonstrate its practical utility and biological insights. The dataset comprises genotype information from the MI Gen study, obtained from the dbGaP database `http://www.ncbi.nlm.nih.gov/gap`. This dataset includes multiple sub-populations: Caucasian, Han Chinese, Japanese, and Yoruban. For our analysis, we considered a set of SNPs that are found to be individually associated with different cardiovascular endpoints in various GWAS, along with SNPs marginally associated with MI in the MIGen study.

The data involve 32 genes covering 1251 loci, including 33 previously identified loci associated with myocardial infarction. The dataset includes both cases (individuals who experienced myocardial infarction) and controls. An important environmental covariate available is sex (male/female), which is incorporated to investigate gene-environment interactions. This provides an opportunity to demonstrate how our framework can uncover both genetic main effects and interactions with environmental factors.

The genetic data consist of diploid genotypes, which we convert to binary indicators for each chromosome for compatibility with our modeling framework. This representation allows us to model the two chromosomes separately while maintaining the diploid nature of the data in our likelihood function.

6.2. Implementation details

We implement our Bayesian nonparametric models using our MPI-based C codes and parallel computing resources. For the HDP model, we set the maximum number of mixture components $M = 30$ based on preliminary analyses to provide sufficient flexibility without excessive computational cost. We set $\nu_1 = \nu_2 = 1$, making $\tilde{\boldsymbol{H}}$ a uniform distribution on $[0, 1]$. The precision parameters are specified as:

$$\begin{aligned}\alpha_{G,ik} &= 0.1 \times \exp(100 + \mu_G + \boldsymbol{\beta}_G^\top \boldsymbol{E}_{ik}) \\ \alpha_{G_0,k} &= 0.1 \times \exp(100 + \mu_{G_0} + \boldsymbol{\beta}_{G_0}^\top \bar{\boldsymbol{E}}_k) \\ \alpha_H &= 0.1 \times \exp(100 + \mu_H + \boldsymbol{\beta}_H^\top \bar{\bar{\boldsymbol{E}}})\end{aligned}$$

with $\mu_G, \mu_{G_0}, \mu_H \overset{iid}{\sim} U(0, 1)$ and $\boldsymbol{\beta}_G, \boldsymbol{\beta}_{G_0}, \boldsymbol{\beta}_H \overset{iid}{\sim} U(-1, 1)$ as priors. This structure ensures adequate numbers of sub-populations and satisfactory MCMC mixing.

We perform MCMC sampling with 30,000 iterations, discarding the first 10,000 as burn-in. Convergence is assessed primarily using trace plots. The parallel implementation distributes computation across 50 cores, with total runtime of approximately 7 days for the full analysis.

6.3. Results and biological insights

Application of our HDP model to the myocardial infarction data yields several important insights:

6.3.1. Effect of sex variable

We find strong evidence that sex influences genetic patterns: $P(|\boldsymbol{\beta}_G| < \epsilon_{\beta_G}|\text{Data}) \approx 0$ and $P(|\boldsymbol{\beta}_{G_0}| < \epsilon_{\beta_{G_0}}|\text{Data}) \approx 0$, indicating that individual-level ($\boldsymbol{E}_{ik}$) and group-average ($\bar{\boldsymbol{E}}_k$) sex effects are highly significant. However, $P(|\boldsymbol{\beta}_H| < \epsilon_{\beta_H}|\text{Data}) \approx 1$, suggesting the overall average sex effect ($\bar{\bar{\boldsymbol{E}}}$) is not significant. This pattern indicates that sex plays an important role in influencing subject-specific and group-level genetic patterns, but not the overall case-control similarity.

6.3.2. Roles of individual genes

Our clustering-based hypothesis tests indicate significant overall genetic effects. However, individual gene tests show that none of the 32 genes are individually significant. This apparent paradox is explained by gene-gene interactions: when genes are correlated, the maximum of their individual distances can be significant even when individual distances are not, similar to how $\max(X_1, X_2)$ from a bivariate normal distribution can have a non-zero median even when X_1 and X_2 have zero medians individually.

6.3.3. Gene-gene interactions

Our HDP model reveals subject-specific gene-gene interaction patterns not detected by previous models. Two genes, *AP006216.10* and *C6orf106*, show significant interactions with other genes in most subjects. Interestingly, the only subjects with no significant gene-gene interactions involving these genes were male cases, suggesting that lack of protective gene-gene interactions may contribute to MI risk in males.

The gene-gene interactions appear to have a protective effect: in subjects where *AP006216.10* and/or *C6orf106* interact with other genes, the risk of MI seems reduced. This beneficial effect of gene-gene interactions contrasts with the traditional view that interactions primarily increase disease risk.

It is important to mention that the subject-wise gene–gene interactions are all non-negative, but range between the orders of 10^{-5} and 10^{-4} for all subjects, irrespective of case and control status; see also Figures 6.2 and 6.3 of (Bhattacharya and Bhattacharya, 2024) and the detailed description. That in spite of such small values *AP006216.10* and *C6orf106* exhibit significant interactions with other genes in most individuals may seem counterintuitive. However, the appropriate Bayesian test of hypotheses is implemented by comparison with a "null model" (see Section 4.2.4 of (Bhattacharya and Bhattacharya, 2024)

for the detailed description), which encapsulates even smaller positive posterior values for these correlations.

6.3.4. Population structure

The posterior distribution of the number of sub-populations shows modes at 3 and 4 components, supporting the known four sub-populations in the data (Caucasian, Han Chinese, Japanese, Yoruban). The model correctly identifies that these populations cannot be further subdivided genetically, consistent with the biological understanding of these population groups.

7. Sensitivity analysis

7.1. Sensitivity to mixture component specifications

For the gene-gene interaction model of Section 3 and the gene-environment interaction model of Section 4, we examine sensitivity to the maximum number of mixture components M and the fixed mixing weights $\pi_{mjk} = 1/M$. Following the rule-of-thumb established in prior work (Majumdar *et al.*, 2013), our primary analyses set $M = 30$, which provides sufficient capacity while maintaining computational efficiency. Alternative specifications with $M = 20$ and $M = 50$ yield qualitatively similar results; the posterior distribution of the effective number of mixture components remains stable once M exceeds a threshold (approximately 20 in our applications), as the Dirichlet process prior automatically determines the number of distinct components through the Pólya urn scheme. This robustness renders the exact value of M largely inconsequential beyond ensuring adequate model flexibility.

7.2. Sensitivity to priors on base measure parameters

For the Beta base measure parameters u_r and v_r, which are specified as independent standard normal distributions in our primary analyses, we investigate robustness using alternatives: $N(0, 10^2)$, $N(0, 0.1)$, and Cauchy$(0, 1)$. Posterior distributions of the interaction parameters λ_{jk} and the resulting gene-gene correlation estimates exhibit remarkable stability across these specifications. As noted in Section 3, we find that Gaussian priors on u_r and v_r with other means and variances do not yield significantly different results, indicating inherent prior robustness in our modeling strategy. The Cauchy prior produces slightly heavier tails but does not alter conclusions of any hypothesis tests regarding gene significance or gene-gene interactions.

7.3. Sensitivity to priors on covariance matrices

For the gene-gene interaction matrix $\boldsymbol{A}$ and case-control dependence matrix $\boldsymbol{\Sigma}$ in the matrix-normal prior $\boldsymbol{\lambda} \sim N(\boldsymbol{\mu}, \boldsymbol{A} \otimes \boldsymbol{\Sigma})$, we specify inverse-Wishart priors: $\boldsymbol{A} \sim \mathcal{IW}(\xi, \boldsymbol{A}_0)$ with $\xi = J + 2$, and $\boldsymbol{\Sigma} \sim \mathcal{IW}(\zeta, \boldsymbol{\Sigma}_0)$ with $\zeta = 4$. Sensitivity is assessed across degrees of freedom ($\xi = J + 1, J + 2, J + 5, J + 10$; $\zeta = 3, 4, 5, 10$) and scale matrix specifications (estimated from data versus identity matrices). Posterior inferences regarding which genes are marginally significant and which gene-gene interactions are present prove highly robust to variations in degrees of freedom. However, the magnitudes of posterior correlations show some sensitivity: larger degrees of freedom, which correspond to stronger prior information,

produce shrinkage toward the prior mean and yield slightly attenuated correlation estimates. The choice of scale matrices has minimal impact when degrees of freedom are set to the minimum values ensuring proper priors ($\xi = J + 2$, $\zeta = 4$).

7.4. Sensitivity to hypothesis testing thresholds

A critical component of our Bayesian testing procedure, described in Section 3, is the specification of thresholds ϵ for distance measures. Following our established approach, we set $\epsilon = F^{-1}(0.55)$ where F is the posterior distribution function under the null model. The choice of the 55th percentile rather than the median is deliberate: for the median, the posterior probability of the true null hypothesis is 0.5, whereas under zero-one loss the true null will be accepted only if its posterior probability exceeds 1/2. Using the median results in borderline decisions in null simulations, while the 55th percentile provides appropriate operating characteristics. Sensitivity analyses using the 50th, 60th, and 75th percentiles demonstrate that the 60th and 75th percentiles produce more conservative tests with reduced false positives but increased false negatives, while the 50th percentile yields unacceptably high false positive rates. Our choice of the 55th percentile represents a balanced compromise that maintains power while controlling Type I error.

7.5. Sensitivity to environmental covariance structure

In our gene-environment interaction model of Section 4, we examine sensitivity to the specification of the environmental covariance structure. The posterior probability $P(\phi < \epsilon_\phi \mid$ Data) is robust across alternative kernel specifications, consistently indicating no differential effect of sex on genetic interactions in the myocardial infarction application. Posterior estimates of ϕ show some sensitivity to prior variance, with more diffuse priors yielding wider credible intervals, but the 95% credible interval consistently contains zero. The smoothness parameter in the kernel is difficult to identify from the data, with posterior distributions largely reflecting the prior when sample sizes are moderate. However, this lack of identifiability does not affect primary conclusions regarding gene significance or gene-gene interactions, as ϕ is consistently estimated to be negligible.

7.6. Sensitivity to hierarchical Dirichlet process hyperparameters

For our hierarchical Dirichlet process model of Section 5, we assess sensitivity to the specification of precision parameters and their associated regression coefficients. Our primary specification, with $\alpha_{G,ik} = 0.1 \times \exp(100 + \mu_G + \boldsymbol{\beta}_G^\top \boldsymbol{E}_{ik})$ and analogous structures for $\alpha_{G_0,k}$ and α_H, ensures adequate numbers of sub-populations and satisfactory MCMC mixing. Removing the constant offset of 100 results in significantly smaller α values, leading to fewer distinct mixture components and poorer mixing. Alternative prior distributions for μ_G, μ_{G_0}, μ_H and $\boldsymbol{\beta}_G, \boldsymbol{\beta}_{G_0}, \boldsymbol{\beta}_H$ (normal versus uniform) perform similarly, with uniform priors showing slightly better convergence properties. The key conclusions regarding gene-gene interactions and the influence of environmental factors on dependence structures are robust across all specifications examined.

7.7. Sensitivity to linkage disequilibrium and locus label permutation

A critical concern associated with Section 3 is whether sharing the parameters u_r and v_r across all genes implies non-exchangeability of locus labels. To address this, we conduct permutation experiments wherein the labels of loci within each gene are randomly permuted prior to re-analysis. The results are entirely consistent with the original analyses based on non-permuted data. For the gene-gene interaction simulation, the posterior probability measures of genetic effect remain strongly suggestive of significance; for the null simulation, these measures correctly indicate no genetic influence. These results confirm that our model does not impose non-exchangeability of locus labels.

7.8. Summary of sensitivity analysis

Across all sensitivity investigations, consistent patterns emerge. Hypothesis test conclusions regarding which genes are marginally significant, which gene-gene interactions are present, and whether environmental variables affect genetic interactions are highly robust to reasonable variations in prior specifications and modeling choices across all three modeling frameworks. Posterior magnitudes of correlation parameters and distance measures show some sensitivity to prior informativeness, but these variations do not affect the binary decisions in our Bayesian testing framework. Convergence and mixing of MCMC algorithms are sensitive to certain choices, particularly the Dirichlet process precision parameters and the offset parameter in HDP specifications, but our primary specifications consistently provide adequate performance. The hierarchical nature of our Bayesian models provides natural protection against prior misspecification, with data information dominating prior information for key parameters of interest. These findings collectively demonstrate that our Bayesian nonparametric approaches are not unduly sensitive to specific prior choices and modeling assumptions, lending credibility to the conclusions drawn from both simulation studies and the real myocardial infarction data analysis.

8. Discussion

8.1. Comparison with existing methods

Our Bayesian nonparametric approach differs fundamentally from standard GWAS methods in multiple aspects. Traditional GWAS typically employ logistic regression models that examine disease status conditional on genotypes ($P(Y|X)$). In contrast, our approach models genotypes conditional on disease status ($P(\boldsymbol{X}|Y)$), which provides a different perspective on genetic associations. This inversion of the modeling relationship allows us to directly address population stratification through mixture models while naturally accommodating complex dependence structures among genetic variants.

For handling population structure, traditional methods often rely on principal component analysis (PCA) for adjustment or conduct stratified analyses by presumed ancestry groups. Our approach uses Dirichlet process mixtures to nonparametrically model population substructure, allowing the number and characteristics of sub-populations to be inferred from the data rather than pre-specified. This can provide more accurate adjustment for population stratification, particularly in admixed populations where discrete ancestry categories may not adequately represent the continuum of genetic backgrounds.

In modeling interactions, traditional approaches typically use regression coefficients for specific interaction terms (*e.g.*, product terms in logistic regression). Our HDP approach represents interactions through hierarchical sharing of parameters, capturing dependencies without requiring specification of particular interaction forms. This allows us to discover complex interaction patterns that might not correspond to simple product terms, potentially revealing richer biological relationships.

Compared to our previous matrix-normal based models (Bhattacharya and Bhattacharya, 2020), the HDP model offers several advantages: subject-specific rather than population-wide gene-gene interactions, nonparametric rather than parametric dependence on environment, and more interpretable sharing structures through the Chinese restaurant process analogy.

Uncertainty quantification differs substantially between the approaches. Traditional methods provide confidence intervals based on frequentist inference, while our fully Bayesian approach yields posterior distributions for all quantities of interest. This includes not only point estimates but also complete characterizations of uncertainty about population structure, interaction patterns, and model complexity. The Bayesian approach naturally incorporates multiple testing considerations through prior distributions and posterior probabilities, avoiding the need for arbitrary significance thresholds or correction procedures.

Computational scaling presents different challenges. Traditional pairwise testing scales as $O(p^2)$ where p is the number of SNPs, becoming prohibitive for genome-wide data. Our gene-level analysis scales as $O(J^2)$ where J is the number of genes, which is typically much smaller than the number of SNPs. While our HDP model is computationally intensive per test, the reduction in dimensionality through gene-level modeling and efficient parallel implementation makes it feasible for realistically sized datasets.

Table 8.1: Comparison of methodological approaches for genetic association studies. Our Bayesian nonparametric framework offers complementary strengths to traditional methods, particularly for complex interaction analysis and uncertainty quantification.

Aspect	**Traditional GWAS**	**Our Bayesian Nonparametric Approach**
Modeling framework	Logistic regression: $P(Y\|\boldsymbol{X})$	Conditional genotype: $P(\boldsymbol{X}\|Y)$
Population structure	PCA adjustment or stratified analysis	Dirichlet process mixtures
Interaction modeling	Regression coefficients	Covariance structures
Uncertainty quantification	Confidence intervals	Posterior distributions
Computational scaling	$O(p^2)$ for pairwise testing	$O(J^2)$ for gene-level analysis
Key advantage	Interpretable effect sizes	Flexible dependence structures

Our approach complements rather than replaces traditional methods in several ways. Researchers might use logistic regression for initial screening of marginal associations, providing interpretable effect sizes for individual variants. Our Bayesian nonparametric methods could then be applied to selected genes or pathways to uncover complex interaction pat-

terns and population structure that might modify or explain the marginal associations. The findings from both approaches could be integrated through meta-analysis or hierarchical modeling frameworks that combine estimates from different methodological perspectives.

For genome-wide discovery, traditional methods remain essential due to their computational efficiency and interpretability. For focused investigation of specific biological pathways or complex traits where interactions are suspected, our approach offers additional insights that might be missed by marginal association testing alone. The two approaches can also inform each other: findings from our interaction analyses might suggest specific interaction terms to test in traditional models, while significant marginal associations from traditional analyses might guide the selection of genes for more detailed investigation with our methods.

8.2. Limitations and robustness considerations

8.2.1. Fixed mixture weights

The choice of fixed mixture weights $\pi_m = 1/M$ warrants careful consideration, as it represents a departure from the more common approach of placing a Dirichlet prior on the mixture weights. Our choice is empirically justified by findings from previous work, which demonstrated that fixed equal weights outperform Dirichlet priors in estimating the true number of mixture components in finite mixture models with Dirichlet process priors. The Dirichlet prior tends to produce many very small weights, effectively underestimating the number of components that contribute meaningfully to the mixture. Fixed equal weights avoid this shrinkage toward fewer components, allowing better recovery of the true mixture structure.

From a computational perspective, fixed weights simplify the Gibbs sampling updates by eliminating the need to sample the weight parameters. This reduces computational complexity and improves mixing of the Markov chains, particularly in high-dimensional settings with many mixture components. The simplification comes with the cost of assuming equal prior weight for all components, but in practice, the posterior distribution adapts to the data through the allocation of subjects to components, effectively learning the relative importance of different components despite the equal prior weights.

Theoretical support for fixed weights comes from results on posterior consistency of Dirichlet process mixture models (Mukhopadhyay and Bhattacharya, 2021). Under mild conditions, both random weights (from a Dirichlet prior) and fixed equal weights lead to the same asymptotic posterior inference as the sample size increases.

8.2.2. Model complexity and interpretability

While our HDP model offers substantial flexibility in capturing complex genetic architectures, this flexibility comes with challenges related to model complexity and interpretability. The three-level hierarchical structure with Dirichlet processes at each level involves many parameters with complex dependencies. The Chinese restaurant process analogy helps with conceptual understanding, but detailed interpretation of posterior results requires careful examination of sharing patterns across multiple levels.

Substantial computational resources are required to fit the HDP model, particularly for datasets with large numbers of genes or samples. Our parallel implementation helps address this challenge, but users still need access to computing clusters or high-performance workstations. The MCMC algorithms require careful tuning and convergence assessment, with runtimes measured in days for full analyses. While these computational demands are substantial, they are becoming increasingly feasible with modern computing infrastructure and algorithmic improvements.

Careful prior specification is essential for obtaining stable and meaningful results from the HDP model. The precision parameters $\alpha_{G,ik}, \alpha_{G_0,k}, \alpha_H$ and their regression coefficients require thoughtful specification based on domain knowledge or sensitivity analysis. Inappropriate prior choices can lead to poor mixing, convergence issues, or biased inference. We provide default settings based on our experience with genetic data, but users should assess sensitivity to these choices in their specific applications.

Expertise in Bayesian nonparametrics is valuable for properly implementing and interpreting the HDP model. Concepts such as Dirichlet processes, Chinese restaurant processes, hierarchical modeling, and MCMC diagnostics may be unfamiliar to researchers trained in traditional genetic epidemiology. To improve accessibility, we shall provide software with user-friendly interfaces, detailed documentation, and tutorial examples. We shall also offer workshops and training materials to help researchers develop the necessary skills to apply these methods effectively.

Despite these challenges, the interpretability of the HDP model can be enhanced through appropriate visualization and summary measures. Posterior distributions of key quantities–such as the number of mixture components at each level, gene-gene correlation patterns for individual subjects, or sharing probabilities between cases and controls–can be presented in intuitive graphical formats. Comparative analyses showing how results differ from simpler models can highlight the value added by the additional complexity. Biological interpretation can be facilitated by connecting statistical findings to known pathways and functional annotations.

8.3. Future directions

Several promising directions exist for extending and improving our Bayesian nonparametric framework for genetic interaction analysis. First, high-dimensional extensions are needed to make the methods scalable to whole-genome data. Current implementations focus on gene-level analysis with selected genes, but applications to genome-wide data would require further computational enhancements. Developing sparse versions of our models that encourage parsimony in interaction structures could also improve scalability while maintaining interpretability.

Second, integration with functional genomic data could enhance biological interpretation and statistical power. Incorporating information from gene expression, methylation, chromatin accessibility, or other omics layers could help prioritize genes and interactions based on functional relevance. Hierarchical models that jointly analyze multiple data types could reveal connections between genetic variation, regulatory mechanisms, and phenotypic outcomes. Such integrative approaches align with systems biology perspectives that view biological processes as interconnected networks rather than isolated components.

Third, longitudinal modeling approaches could capture dynamics in genetic and environmental factors over time. Many complex diseases develop through processes that unfold over years or decades, with genetic risk factors potentially interacting differently with environmental exposures at different life stages. Extending our framework to handle repeated measures or time-to-event data would enable investigation of how genetic interactions influence disease progression and timing. This could provide insights into critical periods for intervention and personalized risk prediction over the life course.

Fourth, causal inference extensions could help distinguish correlation from causation in gene-environment interactions. While observational data alone cannot establish causation, incorporating instrumental variables, Mendelian randomization principles, or structural equation modeling approaches could strengthen causal interpretations. Bayesian methods are particularly well-suited for causal inference because they naturally incorporate uncertainty about causal structures and mechanisms. Developing causal versions of our models would enhance their utility for informing interventions and public health decisions.

Additional future directions include extending the models to handle more complex pedigree or family data, incorporating spatial information for geographically structured populations, developing online learning algorithms for streaming genetic data, and creating user-friendly software packages with graphical interfaces for non-statisticians. Cross-disciplinary collaboration between statisticians, geneticists, and biologists will be essential for advancing these methodological developments while ensuring their relevance to substantive scientific questions.

8.4. Conclusion

We have synthesized a comprehensive Bayesian nonparametric framework for studying genetic interactions in case-control studies. Our HDP model represents a significant advance over previous approaches by providing a flexible, nonparametric representation of gene-gene and gene-environment interactions that accommodates subject-specific effects, population structure, and complex dependence patterns. The three-level hierarchy with Dirichlet processes at each level creates a rich dependence structure that can capture how environmental factors modulate genetic interactions at multiple levels.

Key innovations of our approach include: subject-specific rather than population-wide gene-gene interactions, allowing for heterogeneity in how genes interact across individuals; environmental modulation of dependence structures at multiple levels (individual, gene-group, and case-control), providing a nuanced representation of gene-environment interactions; automatic determination of population structure through Dirichlet process mixtures, avoiding the need to pre-specify the number of sub-populations; and efficient computational implementation through parallel MCMC algorithms that leverage the conditional independence structure of the model.

Applications to myocardial infarction data demonstrate how these methods can uncover biological insights not readily obtainable from standard approaches. We identified significant gene-gene interactions with a protective effect, discovered subject-specific patterns of genetic risk, and revealed how sex modifies genetic architecture. These findings illustrate the value of moving beyond marginal association testing to consider interactions and population structure in genetic epidemiology.

The hierarchical Dirichlet process model represents a particularly flexible extension that accommodates heterogeneous and context-dependent gene-environment interactions. By allowing sharing patterns to vary across individuals, genes, and groups in a data-driven manner, this model can discover complex relationships that might be missed by parametric approaches. The application to myocardial infarction data revealed differences in genetic clustering patterns and latent subgroups with distinct risk profiles, suggesting potential etiological heterogeneity.

While our methods require substantial computational resources and statistical expertise, we believe these barriers are diminishing as computing power increases and statistical training improves. By making our implementation publicly available and providing educational resources, we hope to facilitate wider adoption of Bayesian nonparametric methods in genetic epidemiology. The increasing recognition of complexity in genetic architectures–with interactions, heterogeneity, and context-dependence playing important roles–creates a growing need for methodological approaches that can address this complexity in a principled manner.

Future methodological developments should focus on improving scalability, integrating multiple data types, extending to longitudinal settings, and strengthening causal interpretations. Substantive applications should explore diverse complex diseases and populations, potentially revealing new biological insights and therapeutic targets. As genetic data continue to grow in size and complexity, Bayesian nonparametric methods offer a flexible framework for discovery that respects the inherent uncertainty and interdependence in biological systems.

Acknowledgment

We thank the reviewer for constructive comments that improved this manuscript.

Conflict of interest

The authors declare no conflicts of interest.

Data availability statements

The MI Gen data has been obtained from the dbGaP database `http://www.ncbi.nlm.nih.gov/gap`, and our parallel C code for implementing our HDP model is available at `https://github.com/Sourabh-Bhattacharya/HDP-REALDATA-CODE/blob/main/HDP_REALDATA.zip`.

References

Bhattacharjee, S., Wang, Z., Ciampa, J., Kraft, P., Chanock, S., Yu, K., and Chatterjee, N. (2010). Using principal components of genetic variation for robust and powerful detection of gene-gene interactions in case-control and case-only studies. *The American Journal of Human Genetics*, **86**, 331–342.

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.

Bhattacharya, D. and Bhattacharya, S. (2024). Gene-gene and gene-environment interactions in case-control studies based on hierarchies of Dirichlet processes. *Statistics and Applications*, **22**, 327–360. Special Issue in Memory of Prof. C. R. Rao.

Dey, K. K. and Bhattacharya, S. (2016). On geometric ergodicity of additive and multiplicative transformation based Markov chain Monte Carlo in high dimensions. *Brazilian Journal of Probability and Statistics*, **30**, 570–613. Also available at "http://arxiv.org/pdf/1312.0915.pdf".

Dutta, S. and Bhattacharya, S. (2014). Markov chain Monte Carlo based on deterministic transformations. *Statistical Methodology*, **16**, 100–116. Also available at http://arxiv.org/abs/1106.5850. Supplement available at http://arxiv.org/abs/1306.6684.

Larson, N. B. and Schaid, D. J. (2013). A kernel regression approach to gene-gene interaction detection for case-control studies. *Genetic Epidemiology*, **37**, 695–703.

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.

Mukhopadhyay, S. and Bhattacharya, S. (2021). Bayesian MISE convergence rates of mixture models based on the Pólya urn model: asymptotic comparisons and choice of prior parameters. *Statistics: A Journal of Theoretical and Applied Statistics*, **55**, 120–151. Also available at http://arxiv.org/abs/1205.5508.

Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, **95**, 169–186.

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.

Wang, X., Elston, R. C., and Zhu, X. (2010). The meaning of interaction. *Human Heredity*, **70**, 269–277.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 24, No. 1, 2026 (New Series), pp 323–349
http://www.ssca.org.in/journal

The Man Who Tamed Uncertainty: Andrei Kolmogorov and the Mathematics of Chance

Jyotishka Datta

Department of Statistics, Virginia Tech

This is a special invited paper on request from the Editor-in-Chief.

Received: 22 February 2026; Revised: 20 March 2026; Accepted: 30 May 2026

Abstract

This essay explores the life and contributions of Andrei Nikolaevich Kolmogorov (1903–1987), one of the twentieth century's most influential mathematicians. Beginning with the Borel–Kolmogorov paradox, we examine how Kolmogorov transformed probability theory from a collection of informal methods into a rigorous mathematical framework. We trace his remarkable journey through the tumultuous Soviet era, his historic visit to the Indian Statistical Institute, Kolkata, and his profound contributions spanning probability theory, turbulence, complexity theory, topology, and mathematical education. Kolmogorov's ideas continue to shape modern science, from stochastic modeling and statistical inference to turbulence theory and algorithmic information theory. This essay explores both the mathematical contributions that reshaped probability theory and the historical context in which those ideas emerged.

Key words: Borel–Kolmogorov paradox; Grundbegriffe; History of probability; Kolmogorov complexity; Measure theory; Soviet mathematics.

AMS Subject Classifications: 01A60, 60-03, 60A05

'I, for one, have followed all my life the precept that truth is sacred, that it is our duty to seek it out and to defend it, regardless of whether it is pleasant or not.'

Andrei Nikolaevich Kolmogorov
(25 April 1903 - 20 October 1987)

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

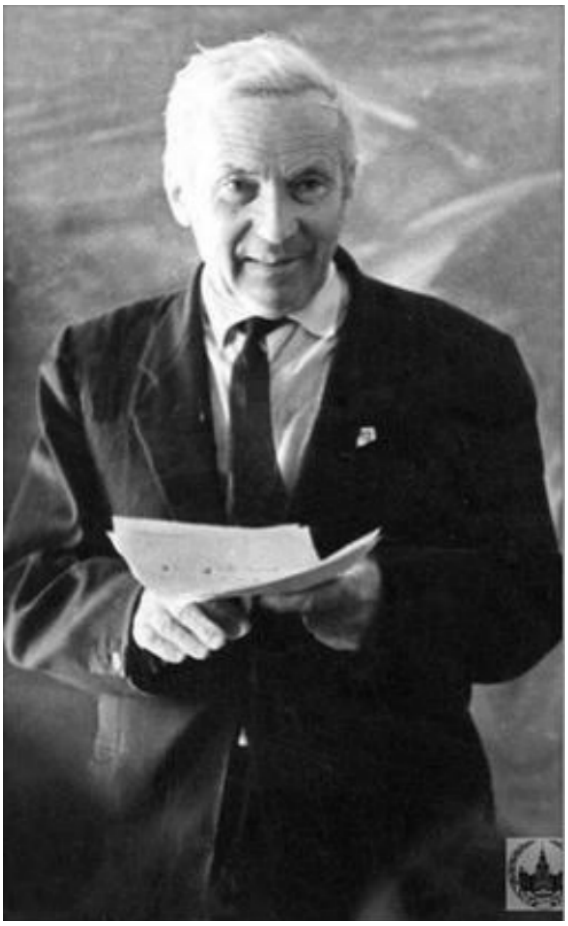


Figure 1: A. N. Kolmogorov. Source: Wikimedia Commons

1. Do aliens prefer hotter climates?

Suppose aliens land uniformly at random on a perfectly spherical Earth[1]. By "uniformly," we mean that every patch of equal surface area is equally likely. Now we are told that the landing point lies on a 'great circle.' Is the distribution along that circle uniform?

To answer this, imagine that an alien is standing somewhere on Earth, somewhere on the planet's curved surface. Now, suppose you randomly select a point on Earth that lies on the equator. The question takes a simpler form: if you walk along the equator, is every spot equally likely to be the chosen point? The answer seems obvious: yes, of course. Every point on the equator should be equally probable.

Now let us look at the same problem from a different 'angle'. Suppose the chosen point lies on the Prime Meridian, the longitude line running from the North Pole through Greenwich to the South Pole. If instead of the equator you walk along this line, is every spot equally likely? Again, it is an obvious 'yes'.

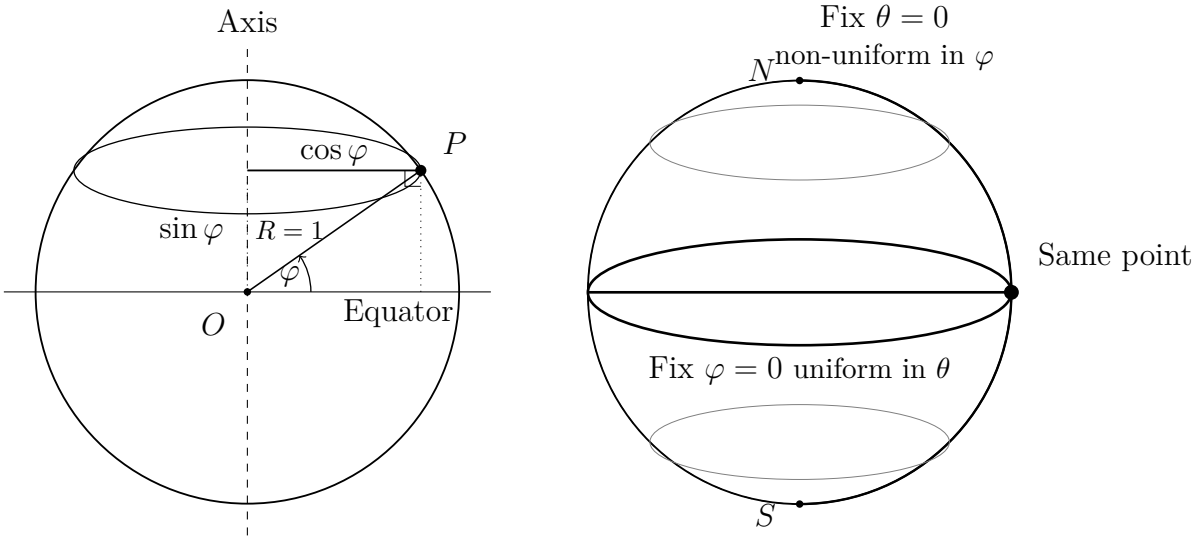
But thereby hangs a tale: this is the set-up of a famous puzzle that troubled mathematicians in the early 20th century. The equator and the Prime Meridian *intersect*: they

[1]This analogy used in a recent popular account of Kolmogorov's work (Gerovitch, 2023)

share a common point. In fact, by rotating the globe, any great circle can be made to coincide with the equator, and any point on that circle can be described in different coordinate systems. And yet, when we calculate the probabilities mathematically, we get *different answers* depending on which description we use. This is the Borel-Kolmogorov paradox, named after Émile Borel and Andrei Nikolaevich Kolmogorov. To understand how Kolmogorov resolved it, and in doing so, transformed probability from a philosophical puzzle into rigorous mathematics, we need to dive into the geometry of spheres and conditional probability.

The mathematics of the sphere. Any point on a unit sphere (a sphere with radius 1) can be described using two angles. The first is φ, the latitude angle, measuring how far north or south of the equator you are (ranging from $-90°$ to $+90°$). The second is θ, the longitude angle, measuring how far east or west you have traveled (ranging from $0°$ to $360°$). If points are scattered uniformly on the sphere, meaning every patch of equal area is equally likely, and if latitude $\varphi \in [-\pi/2, \pi/2]$ and longitude $\theta \in [0, 2\pi]$, the joint density of a uniformly chosen point on the sphere is:

$$f(\varphi, \theta) = \frac{\cos\varphi}{4\pi}. \tag{1}$$



(a) Latitude circle radius $\cos\varphi$ **(b) Different conditional limits**

Figure 2: Geometry of the sphere and the Borel–Kolmogorov paradox

The $\cos\varphi$ term in (1) reflects the shrinking circumference of circles of latitude as one moves away from the equator. The equator ($\varphi = 0°$) is the widest circle, with radius 1. As you move toward the poles, the circles shrink. At $60°$ north latitude, the circle has radius $\cos(60°) = 0.5$, exactly half the size of the equator. At the North Pole ($\varphi = 90°$), the circle shrinks to a single point, because $\cos(90°) = 0$. This means a small angular patch near the equator covers *more actual surface area* than the same angular patch near the poles. To keep the distribution uniform over *surface area*, the probability density per degree must be higher near the equator.

Now we can calculate the probability when we 'condition'[2] on being at a particular location. If we condition on being at a fixed longitude (say, $\theta = 0°$)[3], the limiting conditional density is:

$$f(\varphi \mid \theta = 0°) = \frac{\cos\varphi}{2}, \tag{2}$$

which is *not* uniform! Points are more likely near the equator than near the poles. However, if we condition on being at a fixed latitude, (say, $\varphi = 0°$):

$$f(\theta \mid \varphi = 0°) = \frac{1}{2\pi}, \tag{3}$$

which *is* uniform: every point around the equator is equally likely. But, you can rotate the sphere and turn the equator into a meridian. That is, a great circle can be described as "the circle where $\varphi = 0°$" or equally well as the meridian that is "the intersection of all longitude lines." It's the same circle! How can the probability distribution on it be both uniform and non-uniform? This is the Borel–Kolmogorov paradox (Meehan and Zhang, 2021; Gal, 2014; Sankaran, 2013).

Simply speaking, when we condition on a zero-probability event (like "being exactly on the equator," which is a one-dimensional curve on a two-dimensional surface and thus has zero area), we are really conditioning on a 'measure-zero' set, and that operation is not well-defined without further specification. From a modern measure-theoretic viewpoint, the paradox arises because conditional probabilities on measure-zero sets are not uniquely defined unless the conditioning σ-algebra and the associated regular conditional distribution are specified.

But we can still approach the same problem via taking a *limit*. Think of horizontal bands getting thinner and thinner around the equator. As the bands shrink toward zero width, we approach '$\varphi = 0°$' and get a uniform distribution. On the other hand, if you take a 'vertical approach', taking vertical slices getting thinner and thinner around a particular longitude, as the slices shrink, we approach "$\theta = 0°$" and get the $\cos\varphi$ distribution.

These are different approaches involving different limits, and they naturally lead to different answers. The conditioning event 'on the great circle' is *ambiguous* until we specify *how* we are approaching this zero-probability event. One of Kolmogorov's great contributions to probability was to place probability on rigorous mathematical foundations using measure theory, where these subtleties could be precisely defined and understood.

2. The man behind the Mathematics

2.1. A tragic beginning

Andrei Nikolaevich Kolmogorov was born on April 25, 1903, in Tambov, Russia, about 500 kilometers southeast of Moscow. His birth was marked by tragedy: his unmarried mother, Maria Yakovlevna Kolmogorova, died in childbirth. Little is known about his father,

[2]In plain English, conditioning means: "How should I update my probabilities given I have some information?"

[3]Conditioning on $\theta = 0°$ involves a measure-zero event; the limiting conditional density is obtained by conditioning on a narrowing band around $\theta = 0°$.

but he was probably named Nikolai Matveyevich Katayev, who was an agronomist involved in revolutionary politics who disappeared and was likely killed during the Russian Civil War in 1919 (Kendall *et al.*, 1990; Shiryaev, 1989; Rioul, 2022). The orphaned child was raised by his two aunts at his grandfather's estate in Tunoshna, near Yaroslavl. His grandfather was a well-to-do nobleman, and young Andrei received an excellent education, first at a village school run by his Aunt Vera Yakovlevna. Despite the early tragedies, Kolmogorov said that he had a happy childhood, surrounded by love, kindness, and attention (Tikhomirov, 1988).

Even as a child, Kolmogorov showed remarkable mathematical ability. At age five, he was appointed "editor" of the mathematical section of the school journal, charmingly titled *The Swallow of Spring*. At age six, he made his first mathematical discovery: the pattern in sums of odd numbers. The sum of the first n odd numbers, $1 + 3 + \ldots + (2n - 1)$, always equals n^2. A six-year-old had independently discovered a theorem known to the ancient Greeks. Shiryaev recounts another remarkable story from Kolmogorov's childhood.[4] As a five-year-old boy, Kolmogorov solved this and arrived at an absolutely correct figure. As a school-student, he invented 'perpetual motion machines' multiple times, and his designs were so elaborate that he was able to fool his school teachers every time.

2.2. Moscow and mathematical awakening

In 1910, at the age of seven, Kolmogorov moved to Moscow with his aunt. He graduated from high school in 1920, during the tumultuous early years of Soviet Russia, and enrolled simultaneously at Moscow State University (studying mathematics and history) and the Mendeleev Institute of Technology (studying metallurgy). When he was 10-14 years old, Kolmogorov was deeply fascinated by Biology and Physics, and later by History and Sociology. He apparently thought of and even drafted a utopian constitution for a community that could ensure higher justice in practice. Interestingly, he was captivated by chess, and participated in competitions, but soon abandoned it forever.

His interest in History was profound, and lifelong, and for a time the young Kolmogorov was torn between mathematics and history. How he came to choose mathematics remains an interesting, possibly apocryphal story, though repeated across multiple sources.

When he was a 17-year-old student, he made his first scientific report to Prof. S. V. Bakhrushin's seminar, about the landholding practices in 15th–16th century Novgorod. Prof. Bakhrushin, a leading historian, recognized his discovery, and Kolmogorov asked whether it could be published. When the professor discovered he'd based his analysis on a sample of just five landowners, he said, "You have found only one proof, that is very little for a historian. You need at least five!" Kolmogorov reportedly replied that he'd analyzed all the data available.[5] Besides drafting a thesis on Russian history, he wrote a treatise on Newtonian mechanics, and somewhere in between also interrupted his studies to become a train conductor, all as a teenager.

Kolmogorov was 17 years old when he entered Moscow State University in 1920, and

[4]In his own words, 'There is a button with four holes in it. Thread should go through at least two of them to fix the button. In how many ways can this be done?"

[5]The report he wrote was discovered in his papers and published posthumously; it seems Kolmogorov did rely on statistical methods for his conclusion.

immediately began making original contributions as an undergraduate. At the time, one of Luzin's central open problems was whether the Fourier series of every square-integrable function converges almost everywhere. In 1922, aged just 19, Kolmogorov constructed a function in $L_1([-\pi, \pi])$, the strictly larger class of merely integrable functions, whose Fourier series diverges almost everywhere (Kolmogorov, 1923), and sharpened this in 1926 to divergence everywhere. Rather than resolving Luzin's conjecture, this result clarified how delicate it was: L_1 integrability is simply too weak to guarantee convergence, and the L_2 question remained open for another four decades. It was finally settled by Carleson (Carleson, 1966) for L_2, and extended by Hunt (Hunt, 1968) to all L_p with $p > 1$. Kolmogorov's counterexample and the Carleson–Hunt theorem together constitute the definitive boundaries of this circle of problems in harmonic analysis, and the original result brought Kolmogorov immediate international recognition at an age when most mathematicians are still learning the subject.

He graduated from Moscow State University in 1925, and began his thesis with Nikolai Luzin, and fell under his influence. Luzin was a charismatic mathematician who led a circle of brilliant students they jokingly called "Luzitania", a pun on Luzin's name and the British ocean liner. The students would gather for what they called "joint beating of hearts," passionate discussions of mathematics late into the night. They developed playful terminology: "partial irreverential equations" instead of partial differential equations, "fine night differences" instead of finite differences.

Beginning in 1925, while working on formalizing 'intuitionist logic' with Luzin, Kolmogorov started publishing pathbreaking and foundational papers on probability. This included the famous Kolmogorov inequality that strengthens Chebyshev's inequality, and the Kolmogorov three series theorem that provides the necessary and sufficient conditions for almost sure convergence of a random series $\sum_n X_n$ by requiring convergence of three deterministic series. These results would later inspire the theory of martingales. When Kolmogorov graduated in 1929, he had already published 18 papers on logic, analysis, and probability.

The year 1929 would also be a remarkable epoch in his life: it marked the beginning of his lifelong friendship with the mathematician Pavel Alexandrov. That summer they took a long boat trip on the Volga, to lake Sevan in Armenia, and once again in 1931. In 1935, they bought a house together in Komarovka, outside Moscow, where they hosted many renowned mathematicians for discussions.

3. A polymath's legacy

Kolmogorov's influence extended across an astonishing range of fields. He was, undoubtedly, one of the last universalists. When asked if there can be someone like him in the future, versatile in so many fields with so much impact, Shiryaev (1989) lists only four encyclopedic mathematicians: Poincaré, Hilbert, von Neumann, and Kolmogorov. In pure mathematics, he worked on topology with Pavel Alexandrov, wrote his influential 1925 paper "On the principle of the excluded middle" in intuitionistic logic, and contributed to functional analysis and approximation theory. He made important advances in trigonometric series and developed the Kolmogorov-Arnold-Moser (KAM) theorem in dynamical systems. Working with his student Vladimir Arnold, he produced a partial solution to Hilbert's 13th problem in 1957 (Arnold, 1993; Tikhomirov, 1988). In applied mathematics, his work touched celestial mechanics, differential equations, and mathematical linguistics. He developed ap-

plications to biology, geology, and metallurgy. In statistics and probability, he published over 300 research papers and supervised more than 60 PhD students. It is hard to think of another mathematician of the 20th century with such breadth and depth. When the famous American statistician J. Wolfowitz stood up to address the audience at the 1963 conference at Tbilisi, he said: 'I came to the USSR with the specific purpose of finding out whether Nikolaevich Kolmogorov is an individual or an institution' (Parthasarathy, 2019).

Shiryaev (1989) captures this breadth memorably:

> "If we take a Russian mathematical encyclopedia, we find Kolmogorov axioms, K duality, K integral, K criterion, K inequality, K space, K equation, K–Smirnov criterion, K–Chapman equations. If you take any encyclopedia on probability and mathematical statistics, you will find Kolmogorov axiomatization, K self-similarity, K law of two-thirds, K criterion, K matrix, K model, K distribution, K statistic, K law of five-thirds, K spectral theory."

Kolmogorov's contributions to mathematics are, in a word, unfathomable: spanning more than sixty years and encompassing fields as disparate as topology, logic, celestial mechanics, turbulence, and algorithmic information theory, they defy any compact summary. What follows is necessarily a selective account, focusing on those results in mathematics, statistics, and probability that have proved most enduring and that best illuminate the singular character of his mathematical vision.

3.1. The foundations of probability (1933)

> "It is difficult to overstate the impact of the Grundbegriffe on the development of the subject; essentially the history of probability theory splits in 1933 between pre-Kolmogorov and post-Kolmogorov"
>
> *N. H. Bingham*
> *(Kendall et al., 1990)*

Kolmogorov's most famous work was his slim 1933 monograph *Grundbegriffe der Wahrscheinlichkeitsrechnung* (Foundations of the Theory of Probability) (Kolmogorov, 1933). In about 62 pages[6], he transformed probability from a collection of informal methods into a rigorous mathematical theory.

His approach was simple yet revolutionary: probability is a measure on a set. He defined three axioms. First, non-negativity: for any event A, the probability $P(A)$ must be greater than or equal to zero. Second, normalization: the probability of the entire sample

[6]The original 1933 Springer monograph had 62 main pages. The 1950 English translation (Foundations of the Theory of Probability, Chelsea) runs longer because of added material and formatting differences.

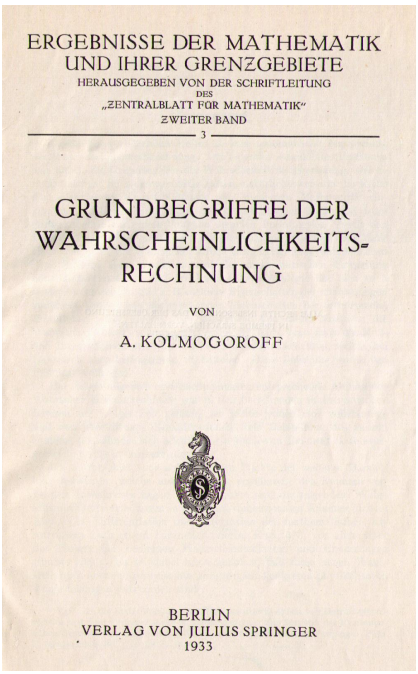


Figure 3: Title page of 'Grundbegriffe der Wahrscheinlichkeitsrechnung' by A. N. Kolmogoroff. Berlin, Springer, 1933. Source: Wikimedia Commons

space is 1, written as $P(\Omega) = 1$. Third, additivity[7]: for mutually exclusive events A and B, the probability that A or B occurs equals $P(A) + P(B)$.

From these simple axioms, all of probability theory could be derived. It was as if Euclid had returned to organize geometry, but this time, for the mathematics of uncertainty. As Parthasarathy (1988) notes, 'To this day, it has stood the test of time and constitutes the cornerstone around which the entire edifice of statistical theory and computation is erected.'

To appreciate what this achieved, it helps to recall the state of probability before 1933. The classical foundation, dominant for two centuries, defined probability as a ratio of equally likely cases. This worked well for dice and cards. In continuous settings, it became circular or ambiguous. Bertrand's paradoxes (1889) exposed this sharply: ask for the probability that a random chord of a circle exceeds the side length of the inscribed equilateral triangle, and you get different answers depending on what "random" means, with no principled way to choose. The Borel–Kolmogorov paradox in §1 is a close cousin. Frequentist approaches, championed by von Mises, grounded probability in limiting relative frequencies, but required an idealized infinite sequence as a primitive object. Conditioning on zero-probability events had no rigorous basis in either framework. Borel had introduced countable additivity into probability in 1909, and Fréchet, Steinhaus, Bernstein, and Cantelli had pushed toward measure-theoretic treatments in the decades that followed. But no one had assembled these pieces into a complete, self-contained axiomatic system. As Shafer and Vovk (2006) put it,

[7]One technically important detail: the additivity axiom in the *Grundbegriffe* is σ-additivity (countable additivity), not merely finite additivity. That is, for any countable collection of mutually exclusive events $A_1, A_2, \ldots$, we have $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$. This is the ingredient that makes limit theorems—the strong law of large numbers, the central limit theorem—derivable within the framework. Finite additivity alone is insufficient.

the *Grundbegriffe* was a work of synthesis as much as of new mathematics: its achievement was to recognize that the right pieces were already on the table, and to take responsibility for declaring the theory complete.

3.2. Random walks and Markov processes

In 1931, Kolmogorov published "Analytic Methods in Probability Theory," laying the foundations for the theory of Markov processes (Kolmogorov, 1931). A Markov process is one where the future depends only on the present state, not on the history of how you got there.

Think of a person taking random steps: at each moment, they step north, south, east, or west with equal probability. Where they step next depends only on where they are now, not on the path they took to get there. This is a random walk, and random walks are everywhere. Stock prices (approximately) follow random walks. Molecules diffusing through a liquid follow random walks. The electrical signals in your neurons, the fluctuations in population sizes, the spread of diseases, all can be modeled as stochastic (random) processes.

The mathematical framework Kolmogorov developed includes the Chapman-Kolmogorov equations, which describe how probabilities evolve over time:

$$P_{ik}(t_1 + t_2) = \sum_j P_{ij}(t_1) P_{jk}(t_2) \tag{4}$$

In words: the probability of getting from state i to state k in time $t_1 + t_2$ equals the sum over all intermediate states j of the probability of going $i \to j$ in time t_1, then $j \to k$ in time t_2. Nearly a hundred years later, virtually every domain of applied probability was reshaped by the developments set in motion by the simple, elegant, and fundamental 1931 paper.

3.2.1. Approximation of distributions by infinitely divisible laws

A remarkable consequence of Kolmogorov's 1956 work (Kolmogorov, 1956) is that, as Arak (1980) succinctly puts it, "*the sum of a large number of independent identically distributed random variables has a distribution which is close to being infinitely divisible.*"

Recall that a distribution is *infinitely divisible* if, for every $n \in \mathbb{N}$, it can be written as the n-fold convolution of some distribution with itself; denote the class of all such distribution functions by $\mathcal{D}$. One of the central problems in this area was to characterise all infinitely divisible laws. Kolmogorov obtained a canonical representation of the log-characteristic function for the finite-variance case (Tikhomirov, 1988). Using a different method, Lévy later removed the finite-variance restriction, and Khinchin subsequently showed that Lévy's general result could be recovered by Kolmogorov's original method, leading to what is now called the Lévy–Khinchin canonical representation.

Kolmogorov's 1956 theorem (Kolmogorov, 1956) asks how uniformly well $\mathcal{D}$ approximates n-fold convolutions of an arbitrary distribution. Let $\Phi_n(\,\cdot\,, F)$ denote the distribution function of the sum of n i.i.d. copies drawn from $F(\cdot)$.

Theorem 1: There exists a universal constant C such that for every distribution function

$F(\cdot)$ and every $n \in \mathbb{N}$, there exists $\Psi(\cdot) \in \mathcal{D}$ with

$$\sup_t \left|\Phi_n(t, F) - \Psi(t)\right| \leq C\, n^{-1/5}.$$

The exponent 1/5 was subsequently improved in a sequence of contributions: Prokhorov (1955) and Meshalkin (1961) obtained intermediate bounds, Kolmogorov himself returned to the problem in Kolmogorov (1962), and the sharp exponent 2/3 was eventually established by Arak (1982a,b), and shown to be optimal.

3.3. The Kolmogorov–Smirnov test and Kolmogorov's distribution

Any standard course in nonparametric statistics introduces the Kolmogorov–Smirnov test: the one-sample goodness-of-fit test for a completely specified null hypothesis, and the two-sample test for comparing two continuous distributions that may differ in any way. Both rest on a single fundamental result in mathematical statistics due to Kolmogorov.

Let ξ be a random variable with continuous distribution function $F(\cdot)$, and let $x_1 \leq x_2 \leq \cdots \leq x_n$ be an ordered sample of n independent realisations. The empirical distribution function is:

$$F_n(x) = \begin{cases} 0 & x < x_1, \\ k/n & x_k \leq x < x_{k+1}, \\ 1 & x \geq x_n. \end{cases}$$

Glivenko established that $\sup_x |F_n(x) - F(x)| \to 0$ almost surely. In the same journal issue, Kolmogorov went considerably further, deriving the exact limiting distribution of the rescaled statistic $D_n = \sqrt{n}\, \sup_x |F_n(x) - F(x)|$:

Theorem 2: Let $F(\cdot)$ be continuous. Then as $n \to \infty$, uniformly in $\lambda > 0$,

$$P\{D_n < \lambda\} \;\longrightarrow\; \Phi(\lambda) = \sum_{k \in \mathbb{Z}} (-1)^k \exp\Big(-2k^2\lambda^2\Big).$$

The function Φ is now called *Kolmogorov's distribution.* It is worth noting that the question of how to quantify the discrepancy between F_n and F had attracted von Mises and Cramér, yet neither obtained the limiting law. It is a measure of the result's practical importance that the Kolmogorov–Smirnov test has become one of the most widely used procedures in nonparametric statistics.

3.4. Turbulence: The chaos in fluids

In 1941, while much of the world was consumed by war, Kolmogorov published revolutionary work on turbulence, the chaotic, swirling motion of fluids. When you watch smoke rise from a candle, it initially rises smoothly, then breaks into swirling, unpredictable patterns. When you observe a fast-flowing river, you see eddies within eddies, chaos at every scale. This is turbulence, and it's one of the hardest problems in physics.

Kolmogorov developed a statistical theory of turbulence (Kolmogorov, 1941b,c,a), including what is now called the "Kolmogorov 5/3 law" (K41), which predicts a $-5/3$ energy

spectrum. For turbulent flow at high Reynolds numbers, the energy spectrum follows a power law:

$$E(k) \propto k^{-5/3} \tag{5}$$

where k is the wave number (spatial frequency) and $E(k)$ is the energy at that scale.

This means energy cascades from large eddies to smaller and smaller eddies in a specific mathematical way. The 5/3 law has been verified experimentally in everything from atmospheric flows to ocean currents to laboratory experiments. Nearly 85 years later, Kolmogorov's turbulence theory remains fundamental to fluid dynamics.

3.5. Kolmogorov's superposition theorem and Hilbert's 13th problem

Hilbert's 13th problem asked whether every continuous function of three variables could be represented as a superposition of continuous functions of *two* variables. The conjecture seemed self-evident to many, and Hilbert even proposed a specific analytic function he believed would serve as a counterexample to reducibility. In 1956 Kolmogorov proved that any continuous function of *four* variables is representable as a superposition of continuous functions of three, a result he considered his most technically demanding, requiring the longest sustained concentration of his career. The final step was taken shortly after by his third-year undergraduate student V. I. Arnol'd, who resolved the case of functions on universal trees in $\mathbb{R}^3$, thereby disproving Hilbert's conjecture entirely. Kolmogorov then found a cleaner construction yielding the following sharp result:

Theorem 3: For any integer $n \geq 2$ there exist continuous functions $\psi_{ij} : [0,1] \to \mathbb{R}$, $1 \leq i \leq 2n+1$, $1 \leq j \leq n$, such that every continuous $f : [0,1]^n \to \mathbb{R}$ admits the representation

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{2n+1} \chi_i \left(\sum_{j=1}^{n} \psi_{ij}(x_j) \right),$$

where the outer functions $\chi_i : \mathbb{R} \to \mathbb{R}$ are continuous and depend on f.

In other words, any multivariate continuous function can be built from univariate functions and a single bilinear operation (addition), refuting the expectation that genuine multivariate complexity is unavoidable. This structural insight has been used directly in deep learning: Liu *et al.* (2025) proposed the Kolmogorov–Arnold Networks (KAN), an architecture that utilizes the theorem by placing learnable univariate spline functions on the edges of the network graph rather than fixing scalar activation functions at nodes, offering improved interpretability and promising results on problems with known compositional structure.

3.6. Information and complexity

During his 1962 visit to the ISI in Kolkata, Kolmogorov developed what we now call *Kolmogorov complexity*, independently discovered around the same time by Ray Solomonoff and Gregory Chaitin. The complexity of an object is the length of the shortest computer program that can generate it.

Consider these two sequences of 64 digits:

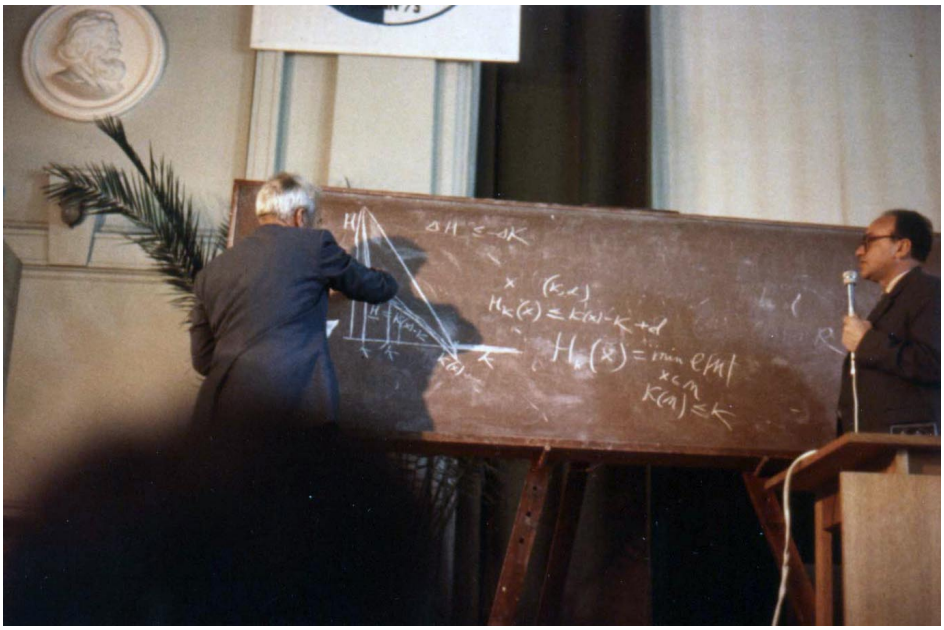


Figure 4: Kolmogorov giving a lecture on Kolmogorov complexity. Source: Wikimedia Commons

Sequence A: 01010101010101010101010101010101010101010101010101010101010101010101
Sequence B: 11001001011100111001111010111001111011110000100101011011010110 11

Which is more complex? Sequence A can be generated by a simple program: "Print '01' thirty-two times." Very short! Sequence B has no obvious pattern. The shortest program that generates it might be: "Print '1100100101110011...[all 64 digits]..."' The program is as long as the sequence itself. In Kolmogorov's framework, K(Sequence A) is small: a simple, compressible pattern. K(Sequence B) is large: complex, incompressible, random.

This idea has profound implications. It gives a rigorous definition of randomness, explains why some things compress well and others don't, and provides a measure for how much "information" an object contains. Kolmogorov complexity even offers an answer to a philosophical question: what is the difference between a law of nature and a coincidence? A law is a simple pattern (low Kolmogorov complexity) that describes many observations. A coincidence is a complex pattern (high Kolmogorov complexity) with no shorter description.

4. Kolmogorov at ISI, Kolkata

In 1962, Kolmogorov visited the Indian Statistical Institute in Kolkata (then Calcutta) from April 14 to May 12, conducting seminars and a series of lectures in Calcutta and at the summer course in Bangalore (Ganguly, 2018). The ISI was a remarkable institution, founded in 1931 by P. C. Mahalanobis, and had become a world center for statistics and probability. In February 1962, the ISI was planning its first convocation. C. R. Rao asked the young probabilists at the institute, S. R. S. Varadhan, K. R. Parthasarathy, and R. Ranga Rao, who they would like to invite from the USSR. The answer was immediate: Kolmogorov.

Rao approached Mahalanobis, who "put the right word in at the right time" to the USSR Academy of Sciences. Kolmogorov agreed to come, but with one condition. He would

Figure 5: P. C. Mahalanobis welcomes A. N. Kolmogorov at Amrapali on April 24, 1962. Source: P. C. Mahalanobis Memorial Museum & Archives, ISI Kolkata. Reproduced with permission for academic use only.

not fly. He had problems with his ears and refused to travel by air. Kolmogorov found an elegant solution. He would join a Soviet oceanographic expedition, and when the ship reached Bombay (now Mumbai), he would disembark, take a train to Calcutta (now Kolkata), and give his lectures at the ISI.

And so he did. During the long sea voyage, Kolmogorov worked on a problem that had fascinated him: ***what is a random number***? How do you define randomness? How do you generate random numbers, and what is the complexity involved in describing them? On that boat, somewhere on the Indian Ocean, Kolmogorov developed the ideas that would later become *Kolmogorov complexity*, a fundamental concept in computer science and information theory (*vide* §3.6). When he arrived in Calcutta, he prepared a note on his work and submitted it to *Sankhyā̃*, the ISI's journal, where it was published in 1963 (Kolmogorov, 1963; Parthasarathy, 2019).

Kolmogorov was also an avid swimmer, and the long sea voyage allowed him to take a plunge every now and then, for relaxation and stimulation (Parthasarathy, 1988). Parthasarathy also recounts that Kolmogorov went on to suggest that institutes and universities in India should be along the coastline, so that the students and faculty can go for a swim before plunging into deeper mathematics.

The institute held its first convocation on February 12, 1962, in the mango grove on campus, chaired by Prof. K. B. Madhava (Vice-President) in the absence of President Sir C. D. Deshmukh. The convocation address was delivered by R. A. Fisher himself. Honorary degrees of Doctor of Science were conferred on five eminent persons: S. N. Bose, R. A. Fisher, Jawaharlal Nehru, W. A. Shewhart, and Kolmogorov. As Kolmogorov had not yet arrived in India, his degree, along with that of Prime Minister Nehru, was conferred in absentia (Ganguly, 2018). Notably, K. R. Parthasarathy and J. Sethuraman, two of the young probabilists who had proposed inviting Kolmogorov, received their PhDs at this same ceremony.

When Kolmogorov arrived in Kolkata on April 14, a Special Convocation was organ-

ised on April 28, 1962, presided over by S. N. Bose (Vice-President of the ISI), to formally present the degree to him in person. On the occasion, P. C. Mahalanobis addressed the gathering:

> "I welcome Academician Kolmogorov on behalf of the Indian Statistical Institute and I should like to greet him in the Indian way. At our First Convocation in last February, we announced the award of the honorary degree to two persons who could not be present, our Prime Minister Jawaharlal Nehru, and Academician Kolmogorov. We are very happy that Academician Kolmogorov could come here, although somewhat behind schedule, and we are glad to have this opportunity to welcome him."

The citation was then read by C. R. Rao, Head of the Research and Training School (Anonymous, 1962, p. 85).

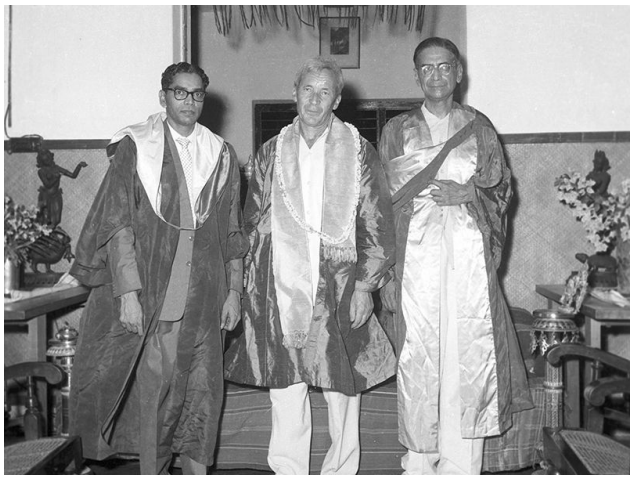


Figure 6: A. N. Kolmogorov with C. R. Rao and P.C. Mahalanobis at the Special Convocation of the Institute on April 28, 1962. Source: P. C. Mahalanobis Memorial Museum & Archives, ISI Kolkata. Reproduced with permission for academic use only.

The ISI in the 1960s was an extraordinary place. Mahalanobis's residence, Amrapali, had become a meeting place of great minds from across the world. Among those who visited were R. A. Fisher, W. A. Shewhart, Norbert Wiener, J. B. S. Haldane, Niels Bohr, P. M. S. Blackett, J. D. Bernal, Frédéric Joliot-Curie and Irène Joliot-Curie, Jan Tinbergen, Ragnar Frisch, Joan Robinson, and Julian Huxley (Ganguly, 2018). It had installed the first indigenous computer in India (1953) and operated two of the first digital computers in South Asia, an HEC-2M from England (1956) and a URAL from the Soviet Union (1959). The institute also employed translators who converted Russian mathematical papers into English, reflecting India's position in the Soviet orbit during the Cold War.

There is a funny, possibly anecdotal, story about Kolmogorov's visit, that I first heard from my advisor (Prof. J. K. Ghosh) and later read on Prof. Debraj Ray's blog (Ray, 2013). The story goes that ISI Kolkata had translators among their employees whose job was to translate Russian works into English for the faculty, and one of them was present during

the speech. After Kolmogorov spoke, there was awed silence during the question-and-answer session. Finally, the translator stood up and delivered an elaborate homily, in Russian, that went on for a good minute or two:

> "Professor Andrei Nikolaevich Kolmogorov, it is because of the presence of individuals such as yourself that I owe my livelihood. Not just mine, but that of my wife and children, who, but for the grace of your genius, would never have had the opportunities they enjoy. . . "

He went on in this vein, a heartfelt tribute to the great mathematician who had made his career as a translator of Russian mathematics possible. When the translator finished, Kolmogorov said: "Excuse me, but my English isn't very good. Could you repeat please?" The translator had been speaking in Russian. Kolmogorov, perhaps not catching that this was meant as a public tribute rather than a question, asked for an English translation of the Russian homage (Ray, 2013).

Figure 7: Andrei N. Kolmogorov during his 1962 visit to the Indian Statistical Institute, Kolkata. Seated in front; standing behind him (left to right) are K. R. Parthasarathy, B. P. Adhikari, S. R. S. Varadhan, J. Sethuraman, C. R. Rao, and P. K. Pathak. Source: Autobiography of S. R. Srinivasa Varadhan

As K. R. Parthasarathy, one of the probabilists who suggested inviting Kolmogorov, later reflected: "It is interesting in the history of Indian mathematics that at the birth of a very important concept, *Sankhyā* played an important role." Indian mathematical journals of the period were venues of genuine international significance; Arthur Wightman's foundational paper on axiomatic quantum field theory, for instance, appeared in the Journal of the Indian Mathematical Society.

The Abel laureate probabilist S. R. S. Varadhan recalls in a 2018 interview (Zeitouni, 2018) that when Kolmogorov visited the Indian Statistical Institute in 1962, he agreed to serve as an external examiner for his doctoral thesis. At the arranged lecture, which ran

far longer than planned, members of the audience began quietly leaving; Kolmogorov, angered by what he perceived as disrespect, threw down the chalk and walked out. He later remarked that in Moscow seminars ran for hours and that "when Kolmogorov speaks, people should listen" (Zeitouni, 2018). At the time of Kolmogorov's visit, S. R. S. Varadhan, K. R. Parthasarathy, R. Ranga Rao, and J. Sethuraman were young research scholars at ISI, in the early stages of their doctoral work, forming a remarkable cohort that would go on to shape modern probability theory.

Figure 8: A. N. Kolmogorov Bhavan at the Indian Statistical Institute, Kolkata. Construction of the building began in 2001 (Ganguly, 2018). Picture courtesy: Prof. Mrinal Kanti Das

To this day, one of the main buildings at ISI Kolkata is named "A. N. Kolmogorov Bhavan" in his honor, with construction having begun in 2001 (Ganguly, 2018).

5. The Soviet context: Mathematics under Stalin

5.1. The Luzin affair

> 'The past is a foreign country; they do things differently there.'
>
> L. P. Hartley, The Go-Between, 1953.

When talking about the genius of Kolmogorov, it is perhaps not irrelevant to talk about the extraordinarily hostile environment for science that prevailed in the Soviet Union under Stalin's rule, and especially after the advent of Lysenkoism (Lorentz, 2002). The trials

that scientists had to endure often forced upon them difficult choices — choices that should not be judged from our place and time. The past is distant, intangible, and fundamentally different from the present. Whatever window we have onto it shapes what we see through it.

In 1936, the beloved teacher Nikolai Luzin, one of Moscow University's outstanding mathematicians, was accused of plagiarism and of being a "servant to fascistoid science" and an enemy of the Soviet people. This was during Stalin's Great Purge, when such accusations typically led to imprisonment or execution (Lorentz, 2002; Rioul, 2022; Khimchenko, 2001).

Figure 9: Nikolai Luzin (9 December 1883 – 28 February 1950). Source: Wikimedia Commons

Kolmogorov and other former students were called to testify. They did testify against Luzin. Whether they were coerced, whether they believed the charges, or whether personal acrimony played a role remains a matter of historical debate. Archival evidence suggests that political pressures were intense, and many mathematicians faced difficult choices under the Soviet regime. None of the students ever spoke publicly about the affair afterward (Arnold, 1993; Shiryaev, 1989). Curiously, Luzin was not arrested or expelled from the Academy of Sciences. He lost some positions but continued working. Recent archival evidence from the 1990s suggests that Stalin personally concluded Luzin posed no real threat.

The bitterness, however, lingered for years; in 1946, after Luzin voted against Alexandrov's election to the Academy of Sciences, Kolmogorov struck his former teacher in the face on the floor of the Academy. The immediate provocation was a remark by Luzin that was understood by all present as a personal attack on Kolmogorov and Alexandrov's relationship (Graham and Kantor, 2009). This episode was later reported to the Kremlin, where Stalin recommended no action, though the Academy itself stripped Kolmogorov of his administrative positions, including the directorship of the Mathematical Research Institute (Lorentz, 2002; Graham and Kantor, 2009).

The affair left its mark on Kolmogorov. He became known as one of the very few "non-political mathematicians with real power" in the Soviet system, someone who could navigate between scientific discovery and political constraint, advancing mathematics while avoiding the fate of so many intellectuals in Stalin's Russia.

It would be too simple, however, to portray Kolmogorov only as a victim of ideological coercion. His relationship with Luzin was probably already strained well before the events of 1936. In his last interview (Khimchenko, 2001), Kolmogorov recalled how Luzin had, with great insistence, convinced the young Pavel Alexandrov that he was destined to solve the Continuum Problem, planting in him an expectation so heavy that it led to a personal crisis and a temporary desire to abandon mathematics. Personal tensions and intellectual disagreements thus predated the political drama, and the political events of 1936 fell on ground that was already disturbed.

5.2. Science in service of the State

The Soviet state demanded that science serve practical purposes. The official stance was that the government had all correct answers; competing ideas were rejected. Yet somehow, Kolmogorov and his students developed probability and statistics into powerful tools.

At the height of Lysenkoism, Soviet biology was reshaped by the doctrine that heredity could be molded by environment and that classical Mendelian genetics was a "bourgeois" fraud. Trofim Lysenko was an agronomist who rose to extraordinary political power under Stalin, effectively controlling Soviet biological research from the 1930s through the 1960s, and whose rejection of genetics set Soviet biology back by a generation (Lorentz, 2002). Lysenko's slogan that "**science is the enemy of chance**" expressed not merely a biological claim but an ideological suspicion toward probabilistic reasoning itself (Rioul, 2022). In February 1940 Kolmogorov published a short note defending Mendelian laws, arguing on mathematical grounds that the experimental claims advanced by Lysenko's supporters were fully consistent with classical genetics; when Lysenkoism became official doctrine in 1948, he was compelled to retract publicly (Rioul, 2022). After a 1936 *Pravda* denunciation of Soviet mathematicians for publishing their best work abroad, Kolmogorov arranged for a Russian edition of the *Grundbegriffe*, and in the ensuing years major scientific work increasingly appeared in Soviet journals (Rioul, 2022).

Yet, as G. G. Lorentz later remarked, the position of mathematicians, precarious though it could be, was often less catastrophic than that of writers and poets, whose work was more immediately exposed to political interpretation (Lorentz, 2002). The broader Stalinist period had already witnessed the execution of Nikolai Gumilev in 1921, the death of Osip Mandelstam in a transit camp in 1938, and the suicide of Marina Tsvetaeva in 1941; Vladimir Mayakovsky had taken his own life in 1930. Mathematics, by contrast, sometimes afforded a narrow shelter in abstraction, even if it could not entirely shield its practitioners from ideological pressure. As we discussed earlier, during the 1936 Luzin affair, Kolmogorov testified to Luzin's scientific achievements while simultaneously referring to his alleged "moral and political decadence," a formulation that illustrates the uneasy balance between loyalty and conformity (Rioul, 2022).

During World War II, Kolmogorov applied his statistical theories to artillery fire and developed a stochastic distribution scheme for barrage balloons to protect Moscow during the

fierce Battle of Moscow. His 1938 paper on "smoothing and predicting stationary stochastic processes" would later find major applications during the Cold War, though these were, of course, classified.

His contributions were recognized even within the Soviet system. In 1939, he was elected Academician of the USSR Academy of Sciences. In 1941, he received the Stalin Prize, followed by the Lenin Prize in 1965. Later, international recognition followed: the Wolf Prize in 1980, one of the highest honors in mathematics, and the Lobachevsky Prize in 1987.

5.3. Pavel Alexandrov: Partnership and protection

> "In the country house where P. S. Alexandrov and A. N. Kolmogorov held their famous gatherings for many years, there was a seminar room with a blackboard. Several years after Kolmogorov's death it had not been erased, and still had in his hand, in English, the motto:
>
> **MEN ARE CRUEL, BUT MAN IS KIND.**"
>
> Khimchenko (2001)

Throughout his life, Kolmogorov's closest relationship was with mathematician Pavel Alexandrov. The friendship that started during a boat trip on the Volga, would last 53 years, until Kolmogorov's death. In 1935, Alexandrov and Kolmogorov bought a house together in Komarovka, outside Moscow. They invited many renowned mathematicians there for discussions. A year before his death, Kolmogorov confided: "For me, these 53 years of intimate and indissoluble friendship were the reason why my whole life was filled with happiness, and the basis of this happiness was the permanent consideration that Alexandrov made for me" (Rioul, 2022).

Their relationship was most probably romantic (Graham and Kantor, 2009, p. 170), though they lived in a time and place where homosexuality was not accepted, let alone openly discussed. This fact would become a vulnerability that the Soviet authorities exploited throughout their lives. Rioul considers it plausible that the authorities threatened to reveal their relationship in order to compel their compliance (Rioul, 2022). Graham and Kantor note more broadly that the Soviet secret police gathered information on all prominent people, including scholars, noting their sexual and personal habits, and that such knowledge was routinely used to obtain the behaviour they wished (Graham and Kantor, 2009).

The political pressures of the Soviet era left marks on Kolmogorov's public record that are difficult to ignore. Kolmogorov and Alexandrov participated in the 1936 campaign against Luzin, the very man who had shaped both of their careers. Kolmogorov, who had previously questioned the scientific foundations of Trofim Lysenko's theories of heredity, did

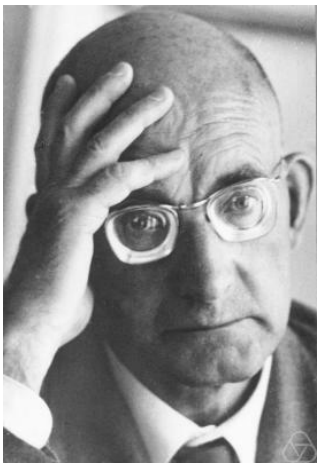


Figure 10: Pavel Sergeyevich Alexandrov (7 May 1896 – 16 November 1982). Source: Wikimedia Commons

not sustain that position publicly when the state required otherwise; Lysenko's work is now regarded as pseudoscience (Rioul, 2022). In the postwar years, he and Alexandrov put their names to a letter in *Pravda* denouncing Alexandr Solzhenitsyn.

NO PARDON FOR TREASON

"We learned with deep satisfaction that Solzhenitsyn has been deprived of Soviet citizenship and bounced out of our country. Soviet intelligentsia are characterized above all by their high civic consciousness, feeling of duty to people and state, respect for traditions and emblems of the people and pride in the high achievements of the Soviet people building communism. In his creations, published in the West, A. Solzhenitsyn blackens our social structure, desecrates the memory of those fallen in the battles of the Great Patriotic War and purposefully gives a distorted picture of the life of Soviet people. In his way he not only violates Soviet laws but also trespasses in the inner sanctum of our people. He has put himself outside of our society. Such persons have no place in our country."

Excerpt from a joint letter by Pavel Alexandrov and Andrei Kolmogorov condemning Alexander Solzhenitsyn, published in Pravda*, February 16, 1974. The letter was written under pressure from Soviet authorities, who used pressure to compel the mathematicians' participation.*

Rioul (2022) as well as Graham and Kantor (2009) reported that Kolmogorov on several occasions tried to explain his inconsistencies and what must have seemed like disloyalties to colleagues, saying, "Sometime I will explain everything to you." Shortly before his death he stated that he would "fear 'them' [the secret police] to his last day". This might help the reader today to make sense of what otherwise seems inexplicable: how could one of the greatest minds of the century participate in supporting junk science, and condemning a courageous dissident?

5.4. The weight of difficult choices

Theoretical computer scientist Scott Aaronson has written thoughtfully about what he calls "the Kolmogorov option" (Aaronson, 2017): the strategy of building fortresses of truth in places the ideological authorities do not particularly understand or care about, like pure mathematics, while avoiding direct confrontation with beliefs a culture considers necessary for its operation. The idea is not a moral verdict but a description of a strategy, one that many scientists under totalitarian regimes have navigated in different ways.

The historical record shows that Kolmogorov testified against Luzin in 1936, complied with demands to endorse Lysenko's doctrine, and co-signed the 1974 letter in *Pravda* condemning Solzhenitsyn. It also shows that in 1940, at personal risk, he published a note defending Mendelian genetics on mathematical grounds, that he financed the research stays of young mathematicians from his own funds, that he founded a school for mathematically gifted children, and that, according to Rioul (Rioul, 2022), he quietly protected Jewish researchers during periods of intense antisemitism in Soviet mathematics. These things coexisted.

What the record does not resolve, and what this essay cannot resolve, is the question of how these coexistences should be weighed. The circumstances that Soviet scientists faced under Stalin and after, where a person's most intimate relationships could be used as instruments of state compulsion, are not ones that admit easy comparisons across cultures or historical periods. Kolmogorov built mathematics of enduring value under conditions of extreme constraint, and he protected, as best he could, the people and the work he cared most about. The rest belongs to history.

6. The human side

Parthasarathy described Kolmogorov as "a great humanist," recalling how he was visibly moved by the poverty he witnessed in India, photographed ordinary people at work, reflected publicly on the coexistence of poverty and plenty, and showed deep concern for children and their education (Parthasarathy, 1988). Parthasarathy also remembered him as "an outdoor mathematician," to whom key ideas came while walking in the woods, swimming in the sea or a lake, rowing, or skiing down mountain slopes (Parthasarathy, 1988). Mathematics for Kolmogorov was not confined to the desk; it unfolded in motion, in landscape, and in conversation.

The Kolmogorov School: Beyond his research, Kolmogorov was passionate about education. In 1963, he founded a specialized mathematical boarding school, Kolmogorov School No. 18 (now the Kolmogorov School at Moscow State University). He personally taught up to 26 hours per week, leading not just classroom lessons but famous Sunday hikes, 40-kilometer walks with students, filled with mathematical discussions, philosophical debates, and literary conversations. These would end with dinner at his dacha in Komarovka, which he shared with his lifelong friend, the mathematician Pavel Alexandrov. Kolmogorov reformed the mathematics curriculum across the Soviet Union and wrote textbooks for grades 6–10 that influenced mathematical education for generations. He contributed over 80 articles to the Great Soviet Encyclopedia, making advanced mathematics accessible to general readers.

Kolmogorov never spoke about his personal life in public, but we know he married

Anna Dmitrievna Egorova in 1942. Trained as a teacher, she shared his deep interest in education and remained a steady presence in his life for more than four decades. Those who knew him remarked that she managed the practical affairs of daily life with quiet devotion, allowing him the freedom to pursue his wide-ranging intellectual projects. The marriage endured until his death in 1987, and she later played an important role in preserving his papers and personal archive. His relationship with Pavel Alexandrov, with whom he bought a house in 1935 and lived for decades, was likely more than friendship, though this remains a matter of historical interpretation.

What is clear is that Kolmogorov was a man of broad cultural interests: history, literature, philosophy, and poetry. He was not a narrow specialist but a Renaissance figure who saw mathematics as part of the larger tapestry of human knowledge. In 1971, at age 68, he joined an oceanographic expedition aboard the research vessel *Dmitri Mendeleev*. He worked with students well into old age, even as Parkinson's disease took its toll. In his final years, nearly blind, he continued thinking about mathematics.

He died on October 20, 1987, in Moscow, and was buried in Novodevichy Cemetery, the resting place of many of Russia's most celebrated artists, writers, and scientists.

7. Envoi: The harmony of uncertainty

A portrait that hangs in Komarovka. (Artist unknown.)

Figure 11: Source: Khimchenko (2001) as well as the video library, 'On the centenary of the great Russian scientist Andrei Nikolaevich Kolmogorov (25.IV.1903–20.X.1987)'.

Vladimir Arnold, one of Kolmogorov's most famous students, once said: "Kolmogorov, Poincaré, Gauss, Euler, Newton, are only five lives separating us from the source of our science" (Arnold, 1993). Five lives. From Newton, who created calculus and discovered the laws of motion and gravitation, to Kolmogorov, who created the mathematics of randomness and uncertainty, just five lifetimes span the distance.

Kolmogorov spent his life showing that randomness isn't the absence of pattern but a deeper kind of order. The Borel–Kolmogorov paradox, far from being a flaw in mathematics, revealed the subtle structure underlying probability. Turbulent fluids, seemingly chaotic, follow precise statistical laws. Random sequences can be defined by their complexity. Uncertainty itself can be axiomatized. Kolmogorov believed, as he once said, that "every mathematician believes that he is ahead of the others. The reason none state this belief in public is because they are intelligent people."

Perhaps he was ahead. Perhaps his genius lay in seeing that the universe, even in its randomness, has an inner harmony, and that mathematics could reveal it.

From the geometry of the sphere to the turbulence of fluids, from the random motion of particles to the complexity of information, from the tragedy of his birth to the triumph of his ideas, Kolmogorov's life was itself a kind of proof: that even amid chaos, uncertainty, and historical upheaval, the human mind can discern patterns, create beauty, and seek truth.

I close with a brief exchange from his final interview with the filmmaker Aleksandr Nikolaevich Marutyan, recorded for the 1983 film *Stories on Kolmogorov*:

> **K.:** Are you familiar with Shklovskii's *The Universe, Life, and the Mind*?
> **M.:** Yes.
> **K.:** He maintains that the development of every culture, if it is not aborted by some catastrophic events—and we all know what might befall humankind now—culminates in a stage of *loss of interest in technology*. Perhaps he really is right.
> **M.:** What does "loss of interest in technology" mean? You mean that people occupy themselves more with humanistic problems?
> **K.:** Not really humanistic problems. But it must be possible to return to a more basic and child-like joy in living. Do you know the German writer Hesse?
> **M.:** Yes.
> **K.:** In *Das Glasperlenspiel*, Hesse depicts such a society, and quite brilliantly, I would say. A society which has lost interest in technological progress.

Appendix

1903–1920: Childhood & revolution

- Born April 25, 1903 in Tambov. Mother dies in childbirth.
- Father disappears in Civil War (1919).
- Age 6: Discovers pattern $1 + 3 + 5 + \cdots + (2n - 1) = n^2$.
- Raised by aunts at grandfather's estate.

1920–1935: Student & rising star

- **1920**: Enters Moscow State University.
- **1922** (age 19): Fourier series diverging almost everywhere $\rightarrow$ international fame.
- **1925**: 8 papers including intuitionist logic, probability theory.
- **1929**: Meets Pavel Alexandrov – 53-year partnership begins.
- **1931**: Professor. Publishes Markov processes, Chapman-Kolmogorov equation.
- **1933**: *Grundbegriffe der Wahrscheinlichkeitsrechnung* – revolutionizes probability with three axioms.

1936–1945: Terror & war

- **1936**: Luzin Affair – political pressures force difficult choices; testifies against mentor. Stalin's Great Purge begins.
- **1940**: Defends Mendelian genetics against Lysenkoism.
- **1941**: Germany invades USSR. Publishes turbulence theory (K41, 5/3 law). Applies mathematics to Moscow defense.
- **1941–44**: Siege of Leningrad (872 days, $\sim$1 million dead).
- **1942**: Marries Anna Dmitrievna Egorova.

1945–1962: Post-war achievements

- **1948**: Lysenko triumphs – forced to retract genetics support.
- **1950s**: Introduces ε-entropy. KAM theorem (dynamical systems).
- **1953**: Stalin dies – gradual thaw.
- **1957**: With Arnold, partial solution to Hilbert's 13th problem.
- **1962**: Visits ISI Kolkata (April–May) by ship; develops Kolmogorov complexity during voyage. Special Convocation April 28. Lectures in Calcutta and Bangalore. Publishes in *Sankhyā* (1963).

1963–1987: Elder statesman

- **1963**: Founds School No. 18 for mathematically gifted children.
- **1970s**: Sophistication theory (refinement of Kolmogorov complexity).
- **1974**: With Alexandrov, signs *Pravda* letter condemning Solzhenitsyn (under pressure).
- **1980**: Wolf Prize.
- **1987**: Dies October 20 (age 84) from Parkinson's disease.
- **1991**: Soviet Union dissolves (4 years after his death).

Major historical events he lived through

WWI (1914–18)	Age 11–15
Russian Revolution (1917)	Age 14
Civil War (1918–21)	Age 15–18; father killed
Stalin's Purges (1936–38)	Age 33–35; directly affected
WWII (1941–45)	Age 38–42; applied work for defense
Siege of Leningrad (1941–44)	Age 38–41
Stalin's Death (1953)	Age 50
Cold War (1947–91)	Age 44–84; entire mature career

Acknowledgment

The author thanks the Editor-in-Chief for the kind invitation to contribute this essay, and Professor Abhyuday Mandal for his encouragement in developing this essay as a journal article. The author is grateful to Avik and Dr. Satpath of the P. C. Mahalanobis Memorial Museum & Archives, Indian Statistical Institute, Kolkata, for providing high-resolution photographs and for clarifying the historical record of Kolmogorov's 1962 visit. Two photographs (the welcome at Amrapali and the Special Convocation) are reproduced from the collection of the P. C. Mahalanobis Memorial Museum & Archives, ISI Kolkata, with permission for academic and non-commercial use only; copyright remains with the Indian Statistical Institute, Kolkata.

References

Aaronson, S. (2017). The Kolmogorov option. Blog post on *Shtetl-Optimized*.

Anonymous (1962). 30th Annual Report of the Indian Statistical Institute (April 1961–March 1962). Technical report, Indian Statistical Institute, Kolkata.

Arak, T. V. (1980). On the approximation of n-fold convolutions of distributions having non-negative characteristic functions with accompanying laws. *Theory of Probability and its Applications*, **25**, 221–243.

Arak, T. V. (1982a). On the convergence rate in Kolmogorov's uniform limit theorem. i. *Theory of Probability & Its Applications*, **26**, 219–239.

Arak, T. V. (1982b). On the convergence rate in Kolmogorov's uniform limit theorem. ii. *Theory of Probability & Its Applications*, **26**, 437–451.

Arnold, V. I. (1993). On A. N. Kolmogorov. In Zdravkovska, S. and Duren, P. A., editors, *Golden Years of Moscow Mathematics*, pages 129–153. American Mathematical Society, Providence, R.I.

Carleson, L. (1966). On convergence and growth of partial sums of Fourier series. *Acta Mathematica*, **116**, 135–157.

Gal, Y. (2014). The Borel–Kolmogorov paradox. Short talk presentation.

Ganguly, N., editor (2018). *An Annotated Chronological History of Indian Statistical Institute*. Library, Documentation and Information Science Division, Indian Statistical Institute, Kolkata.

Gerovitch, S. (2023). How random chance changed the man who invented modern probability. Originally published in Nautilus.

Graham, L. and Kantor, J.-M. (2009). *Naming Infinity: A True Story of Religious Mysticism and Mathematical Creativity*. Harvard University Press, Cambridge, MA.

Hunt, R. A. (1968). On the convergence of Fourier series. In *Orthogonal Expansions and their Continuous Analogues*, pages 235–255. Southern Illinois University Press, Carbondale, IL. Proc. Conf., Edwardsville, Ill., 1967.

Kendall, D. G. et al. (1990). Andrei Nikolaevich Kolmogorov (1903–1987). *Bulletin of the London Mathematical Society*, **22**, 31–100.

Kendall, D. G., Batchelor, G. K., Bingham, N. H., Hayman, W. K., Hyland, J. M. E., Lorentz, G. G., Moffatt, H. K., Parry, W., Razborov, A. A., Robinson, C. A., et al. (1990). Andrei Nikolaevich Kolmogorov (1903–1987).

Khimchenko, N. G. (2001). From the "Last Interview" with A. N. Kolmogorov. *The Mathematical Intelligencer*, **23**, 30–38.

Kolmogorov, A. N. (1923). Une série de Fourier–Lebesgue divergente presque partout. *Fundamenta Mathematicae*, **4**, 324–328.

Kolmogorov, A. N. (1931). On analytical methods in probability theory. *Mathematische Annalen*, **104**, 415–458. German title: Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung.

Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer-Verlag, Berlin. English translation: *Foundations of the Theory of Probability*, 1950.

Kolmogorov, A. N. (1941a). Dissipation of energy in locally isotropic turbulence. *Doklady Akademii Nauk SSSR*, **32**, 16–18. Reprinted in *Proceedings of the Royal Society of London, Series A*, **434** (1991), 15–17.

Kolmogorov, A. N. (1941b). The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers. *Doklady Akademii Nauk SSSR*, **30**, 301–304. Reprinted in *Proceedings of the Royal Society of London, Series A*, **434**, (1991), 9–13.

Kolmogorov, A. N. (1941c). On degeneration (decay) of isotropic turbulence in an incompressible viscous liquid. *Doklady Akademii Nauk SSSR*, **31**, 538–540.

Kolmogorov, A. N. (1956). Two uniform limit theorems for sums of independent random variables. *Theory of Probability and its Applications*, **1**, 384–394.

Kolmogorov, A. N. (1962). On the approximation of distributions of sums of independent random variables by infinitely divisible distributions. *Trudy Moskovskogo Matematicheskogo Obshchestva*, **12**, 437–451. In Russian.

Kolmogorov, A. N. (1963). On tables of random numbers. *Sankhyā Ser. A*, **25**, 369–375.

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljacic, M., Hou, T. Y., and Tegmark, M. (2025). KAN: Kolmogorov–Arnold networks. In *The Thirteenth International Conference on Learning Representations*.

Lorentz, G. G. (2002). Mathematics and politics in the Soviet Union from 1928 to 1953. *Journal of Approximation Theory*, **116**, 169–223.

Meehan, A. and Zhang, S. (2021). The Borel-Kolmogorov paradox is your paradox too: A puzzle for conditional physical probability. *Philosophy of Science*, **88**, 1011–1032.

Meshalkin, L. D. (1961). On approximation of distribution functions of sums by infinitely divisible laws. *Theory of Probability and its Applications*, **6**, 233–252.

Parthasarathy, K. R. (1988). Obituary: Andrei Nikolaevich Kolmogorov. *Journal of Applied Probability*, **25**, 444–450.

Parthasarathy, K. R. (2019). A different kind of mind, K.R. Parthasarathy in conversation with B.V. Rajarama Bhat. Interview excerpt published in *Bhāvanā* online magazine.

Prokhorov, Y. V. (1955). On sums of identically distributed random variables. *Doklady Akademii Nauk SSSR*, **105**, 645–647. In Russian.

Ray, D. (2013). Cryptic tales from the ISI. Blog post on *Chhota Pegs*.

Rioul, O. (2022). The life and work of Kolmogorov. Technical report, HAL open archive. HAL Id: hal-03718689.

Sankaran, K. (2013). Borel-Kolmogorov paradox.

Shafer, G. and Vovk, V. (2006). The sources of Kolmogorov's Grundbegriffe. *Statistical Science*, **21**, 70–98.

Shiryaev, A. N. (1989). A. N. Kolmogorov: Life and creative activities. *Annals of Probability*, **17**, 866–944.

Tikhomirov, V. M. (1988). The life and work of Andrei Nikolaevich Kolmogorov. *Russian Mathematical Surveys*, **43**, 1–39.

Zeitouni, O. (2018). A conversation with S. R. S. Varadhan. *Statistical Science*, **33**, 126–137.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 24, No. 1, 2026 (New Series), pp 351–352
http://www.ssca.org.in/journal

Corrigendum on "Prabhu-Ajgaonkar's 1967 Result Revisited"

Bikas Kumar Sinha[1] **and Manisha Pal**[2]
[1]*Retired Professor, Stat-Math Unit, Indian Statistical Institute, Kolkata, India*
[2]*Department of Statistics, St. Xavier's University, Kolkata, India*

Received: 23 November 2025; Revised: 02 March 2026; Accepted: 04 March 2026

In the publication entitled "Prabhu-Ajgaonkar's 1967 Result Revisited", that appeared in Statistics and Applications, Vol. 23, No. 2, 323-329, 2025, under Shorter Communications, a few critical errors have crept in. We express our deepest concern for any discomfort that the readers may experience because of these undesirable errors/mistakes.

1. Page 324: Table 1 should read as

Table 1: Population of size $N = 4$

Unit i	1	2	3	4	Total
Y_i	0.5	1.2	2.1	3.2	7.0
p_i	0.1	0.2	0.3	0.4	1.0

2. Page 326, Section 4: The correct expression of $\mathrm{Var}(\hat{T}_L(Y))$ is

$$\begin{aligned}\mathrm{Var}\big(\hat{T}_L(Y)\big) &= \sum_{\substack{i,j=1\\ i<j}}^{N} \frac{(Y_i+Y_j)^2}{(p_i+p_j)(N-1)} - [T(Y)]^2 \\ &= \sum_{\substack{i,j=1\\ i\neq j}}^{N} \frac{(Y_i+Y_j)^2}{2(p_i+p_j)(N-1)} - [T(Y)]^2.\end{aligned}$$

3. Page 326, 3rd line from bottom: The inequality should be numbered as

$$(N-2)\left[\frac{Y_i^2}{p_i}+\frac{Y_j^2}{p_j}\right] + \frac{1}{p_ip_j(p_i+p_j)}(p_jY_i - p_iY_j)^2 \geq 0, \quad \text{for all } i<j. \tag{1}$$

4. Page 327, section 6, lines 7-8: The expression

   "Note that the cross product terms can be written as

   ....."

should be written as

"Note that the sum of squares term can be written as

$$\sum_{i=1}^{N} Y_i^2 = \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \frac{Y_i^2 + Y_j^2}{2(N-1)}."$$

Publisher
Society of Statistics, Computer and Applications
Mailing Address: B-133, Ground Floor, C.R. Park, New Delhi-110019, INDIA
Tele: 011-40517662
https://ssca.org.in/
statapp1999@gmail.com
2026

Printed by : Galaxy Studio & Graphics
Mob: +91 9818 35 2203, +91 9582 94 1203