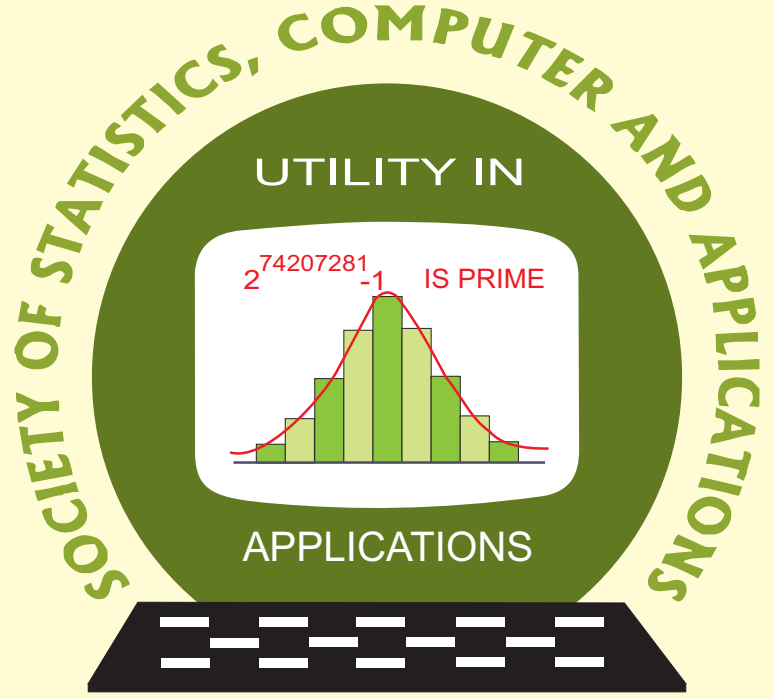
ISBN 978-81-950383-8-1

Special Proceedings (27)
(Based upon the 27th (Annual) International Conference of the Society of Statistics, Computer and Applications (SSCA) - 2025 held at the Department of Statistics, North-Eastern Hill University, Shillong - 793022, Meghalaya, India, during, February 21-23, 2025)



Society of Statistics, Computer and Applications
https://ssca.org.in/
2025

Society of Statistics, Computer and Applications

Council and Office Bearers

Founder President
Late M.N. Das

President
V.K. Gupta

Executive President
Rajender Parsad

Patrons

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

Vice Presidents

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

Secretary
D. Roy Choudhury

Foreign Secretary
Abhyuday Mandal

Treasurer
Ashish Das

Joint Secretaries

Aloke Lahiri, Shibani Roy Choudhury, Vishal Deo

Council Members

B. Re. Victor Babu, Banti Kumar, Bishal Gurung, Imran Khan, Mukesh Kumar
Parmil Kumar, Piyush Kant Rai, Rajni Jain, Rakhi Singh, Raosaheb V. Latpate
Renu Kaul, Shalini Chandra, Sukanta Dash, V.M. Chacko, Vishnu Vardhan R.

Ex-Officio Members (By Designation)

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

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

ISBN 978-81-950383-8-1

Special Proceedings (27)
(Based upon the 27th (Annual) International Conference of the Society of Statistics, Computer and Applications (SSCA) - 2025 held at the Department of Statistics, North-Eastern Hill University, Shillong - 793022, Meghalaya, India, during, February 21-23, 2025)

Editors

V.K. Gupta
Baidya Nath Mandal
Durba Bhattacharya
R. Vishnu Vardhan
Ranjit Kumar Paul
Rajender Parsad
Dipak Roy Choudhury

Date of Publication: October 25, 2025

Published By:
Institutional Publisher: Society of Statistics, Computer and Applications
Registered Office: I-1703, Chittaranjan Park, New Delhi- 110019, INDIA
Mailing Address: B-133, Ground Floor, C.R. Park, New Delhi-110019, INDIA
Tele: 011-40517662
https://ssca.org.in/
statapp1999@gmail.com
2025

Printed by: Galaxy Studio & Graphics
A-181, Weavers Colony, Ashok Vihar, Phase-IV, New Delhi-110052
Mob: +91 98183 52203, +91 95829 41203
Email: galaxystudio08@gmail.com

Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025

CONTENTS

	Preface	i-ii
1	Failure Time Prediction Model for an Injection Molding System *Pritam Ranjan, Arnab Koley, Sandip K. Pal and Debasis Kundu*	1–11
2	On Bivariate Weibull Frailty Model *Debasis Kundu*	13–30
3	Fondly Remembering Two Interesting Collaborations *A. Goswami*	31–41
4	On Exploring Tails via Tail Equivalence *Sreenivasan Ravi*	43–50
5	Advances in Stepped Wedge Design: A Comprehensive Review *Soumadeb Pain and Satya Prakash Singh*	51–66
6	Challenges in Flexible and Scalable Modeling of Point-referenced Spatial Data *Suman Guha*	67–75
7	Use of Designs for Statistical Experiments in Constructing Cryptographic Schemes *Mausumi Bose*	77–86
8	Large Language Models in Practice: Training Paradigms, Knowledge Systems, and Production-Scale Deployments *Utkarsh Tripathi*	87–100
9	A Comparative Study of Estimation Methods for Nakagami Distribution in Reliability Analysis *Rahul Gupta and Bhagwati Devi*	101–120
10	Predicting Stock Market Crash with Bayesian Generalised Pareto Regression *Sourish Das*	121–134
11	Unraveling Biological Complexity: AI and Statistical Approaches to Multi-Omics Data Integration *D. C. Mishra, Shesh Nath Rai, Mamatha Y. S., K. K. Chaturvedi, Sudhir Srivastava, Neeraj Budhlakoti1 and Girish Kumar Jha*	135–152
12	An Overview of Bayesian Semiparametric Approaches for Genetic Association Studies *Durba Bhattacharya and Sourabh Bhattacharya*	153–166

13 Predictive Modeling of Maize Yield in Jammu Subtropical Zone using Weather Data and Penalized Regression 167–176
Manish Sharma, Amandeep Verma, Nishant Jasrotia, Sushil Sharma and Divyam Sharma

Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025

PREFACE

We are pleased to present the Special Conference Proceedings 2025 from the twenty-seventh annual conference of the Society of Statistics, Computer, and Applications (SSCA), organized by the Department of Statistics at North-Eastern Hill University, Shillong, Meghalaya, India, from February 21-23, 2025. This conference was a significant part of the broader international event focused on "Advances of Interdisciplinary Statistics and Applications in AI & ML (AISAAM-2025)."

Founded in 1998, with its inaugural gathering at Haryana Agricultural University, Hisar, SSCA has consistently organized annual conferences at various educational institutions across the country. The society's core mission is to foster research at the intersection of Statistics and Information Technology, supporting both theoretical and applied statisticians dedicated to technological advancements for societal progress. SSCA also promotes open access to knowledge through its journal, Statistics and Applications, which facilitates free downloads, saving, and printing of full-length papers. In addition to regular issues, the journal periodically releases special volumes addressing significant global and national themes.

The recent 27th conference aimed to provide a unified platform for discussions on regional and global statistical issues. Distinguished experts in theoretical and applied statistics from India and abroad, particularly from the USA, actively participated in knowledge dissemination. Speakers represented prestigious Indian institutions, including the Indian Statistical Institute, IITs, ICAR, RBI, universities, and government offices. The conference featured several notable events, including a pre-conference workshop and various technical sessions. These sessions included the M.N. Das Memorial Lecture and a dedicated session on Financial Statistics, where renowned statisticians and leading practitioners shared insights on finance-related topics. Additionally, three endowment lectures were presented: the B.K. Kale Memorial Endowment Lecture, the J.K. Ghosh Memorial Endowment Lecture, and the Bikas Kumar Sinha Endowment Lecture. These lectures were delivered by speakers closely associated with, or students of, the respective honorees. Furthermore, the V.K. Gupta Endowment Award Lecture for Achievements in Statistical Thinking and Practice was presented by Shyamal Peddada from the USA on June 22, 2025. In 2024, the SSCA established the Aditya Shastri Memorial Lecture in memory of the late Aditya Shastri, Vice Chancellor of Banasthali Vidyapith, who passed away in 2021 due to COVID-19. The first lecture was delivered by Navin M. Singhi, and this year, the second lecture was presented by R.K. Sharma.

The Executive Council of SSCA resolved to compile the "Special Conference Proceedings 2025," highlighting selected presentations, including those from the specialized Financial Statistics session. The Guest Editors appointed by the Council—V.K. Gupta, Baidya Nath Mandal, Durba Bhattacharya, R. Vishnu Vardhan, Ranjit Kumar Paul, Rajender Parsad, and Dipak Roy Choudhury—meticulously curated these proceedings. Although constraints limited the inclusion of all invited papers, esteemed speakers were invited to submit their research contributions. Following a rigorous review process, a distinguished selection of 13 papers was accepted for publication in the special proceedings. We extend our sincere gratitude to all authors for the prompt submission of high-quality research papers. We owe a special debt of thanks to all the reviewers whose dedicated efforts ensured a swift and thorough review process, completing it within a short timeframe. Special acknowledgment is also due to all members and office bearers of the SSCA Executive Council for their steadfast support. We are particularly grateful to Ashish Das, Treasurer of the SSCA, for arranging funds for publishing these special proceedings. We also thank Ms. Jyoti Gangwani for meticulously formatting the papers. Furthermore, our deepest appreciation goes to all the family members and students who sponsored the Endowment Lectures. The special proceedings from these sessions have been assigned the ISBN #: 978-81-950383-8-1.

We are confident that the contents of these special proceedings will be immensely beneficial to our readership, fostering further insights and advancements in the field of Statistics and its applications in AI and beyond. We welcome any suggestions for improving future conferences and special proceedings as we continually strive to better serve the statistical community.

V.K. Gupta
Baidya Nath Mandal
Durba Bhattacharya
R. Vishnu Vardhan
Ranjit Kumar Paul
Rajender Parsad

September 30, 2025

Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025; pp 1-11

Failure Time Prediction Model for an Injection Molding System

Pritam Ranjan[1], **Arnab Koley**[1], **Sandip K. Pal**[1] **and Debasis Kundu**[2]

[1]*OMQT Area, Indian Institute of Management Indore, MP, India*
[2]*Dept. of Mathematics and Statistics, Indian Institute of Technology Kanpur, UP, India*

Received: 01 May 2025; Revised: 11 May 2025; Accepted: 15 May 2025

Abstract

In industrial sectors, understanding machine behaviour in real-time is crucial for minimizing unscheduled downtime and maximizing production with expected quality. Advanced machines like injection molding machine used for manufacturing plastic bottles for soft drinks are equipped with sensors that record event log times. We adopt a hierarchical parametric model to predict machine failure time based on its current state, that are, "running with alerts," "running without alerts," and system breakdown. The model utilizes Weibull distribution for the event duration to predict failure times.

Key words: Generalized linear regression; Predictive modeling; Count data distribution; Lifetime distribution.

AMS Subject Classifications: 62N05, 90B25

1. Injection molding machine

Running manufacturing equipment involves maintenance of machines on a regular basis. Preventive maintenance is a popular and well accepted approach, however, such tasks are carried out according to a timetable, and not always done when the equipment specifically calls for them. Thus, it is crucial to predict machine failures with enough lead time.

Several predictive models have been proposed by different authors to predict the failure time using the sequence of events. Li *et al.* (2007) developed a Cox-proportional hazard (CPH) model to predict the time to failure model. Luo *et al.* (2014) proposed a framework which consists of three stages: data pre-processing, event extraction, and correlation analysis. In the data pre-processing stage. A few other works on the correlation based event prediction model are Motahari-Nezhad *et al.* (2011); Wu *et al.* (2010); Zhu and Shasha (2002); Lou *et al.* (2010). Agrawal *et al.* (1993) use association mining to learn a pattern based on a historical sequence of past events to predict the probable occurrence of next event(s). In retail sector, the market basket analysis has been recognized as a proven

Corresponding Author: Pritam Ranjan
Email: pritamr@iimidr.ac.in

and successful application of association rule mining for cross selling, product placement, promotion affinity analysis, and product promotion and targeting (Kohavi *et al.* , 2004), for mining gene sequence expression (Jiang and Gruenwald , 2005) and for web-log mining (Huang and An , 2002; Rudin *et al.* , 2011).

The main objective of this study is to propose a time to failure model of an injection molding (IM) machine for a plastic soft drink bottle (see Figure 1).

Figure 1: An injection molding (IM) machine schematic diagram (source: https://prototechasia.com/en/injection-molding/questions-injection-molding)

Industry 4.0 brings forth intelligent machines equipped with sophisticated sensors, embedded software, and robotics which gather and store data as machine logs in a semi-structured format. These data are usually collected while machine is in running condition, and primarily consists of operation events, performance counters, and alert messages, among others. This research focuses on the system logs data that are captured through various sensors mounted in the IM machine (see Figure 2).

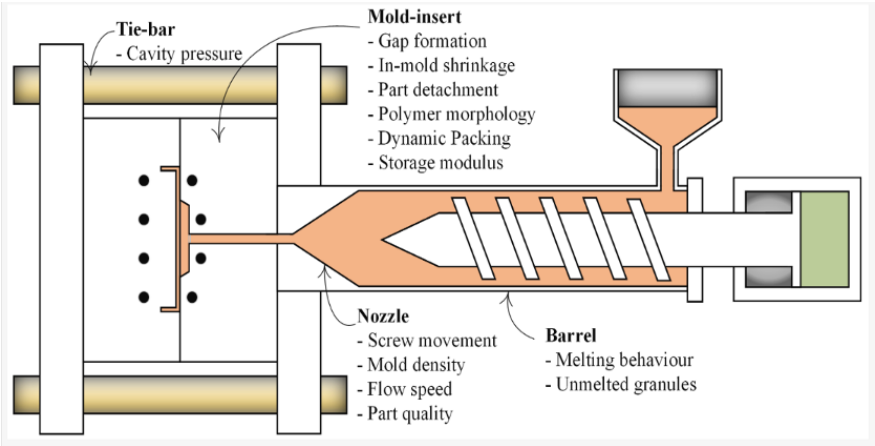


Figure 2: A schematics of IM machine with various sensors.

Figure 3 depicts different operational sequence of a few events of an IM machine which typically provide sufficient information for engineers to diagnose the working condition of equipment.

The alert messages have been clubbed into three groups. When the machine is running smoothly and does not produce any message or alert we label it as “running without alert”. Alternatively, when the machine is running but generate some warning or requires human

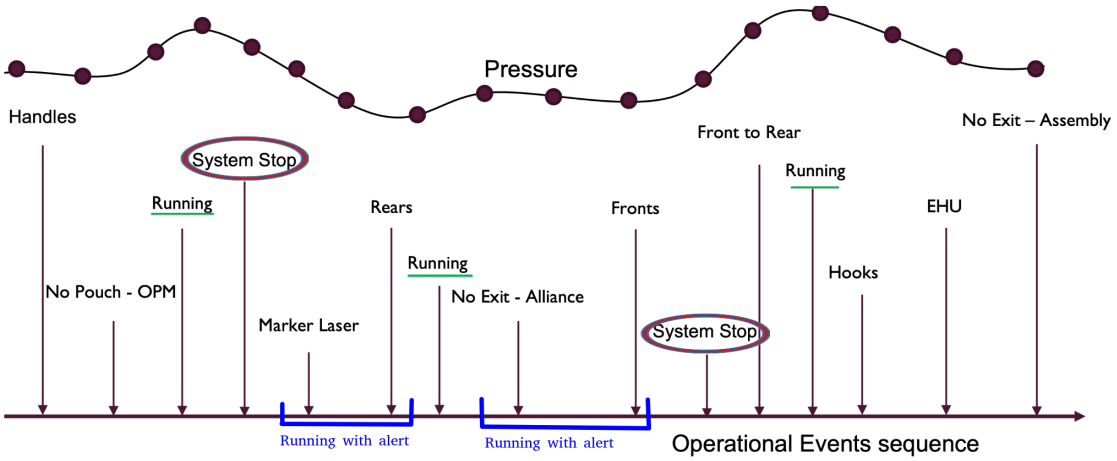


Figure 3: An example of sequence of events while machine is in operational condition.

intervention, we call it as "running with alert". Finally, "failure" refers to the state of the machine when the system is down and requires maintenance.

Furthermore, if there are two or more consecutive occurrences of the same type of events (say, "running with alert") then Pal *et al.* (2024) clubbed them together as one event. This implied that "running without alert" and "running with alert" will occur alternatively followed by "failure". It may also be possible that the machine experiences only one type of events, say "running without alert" or "running with alert" in an epoch before "failure". One sequence of events until the failure is also referred to as an epoch. In this illustrative image only two epochs are shown for the purpose of understanding. Figure 4 illustrate the sequence of events leading to failure.

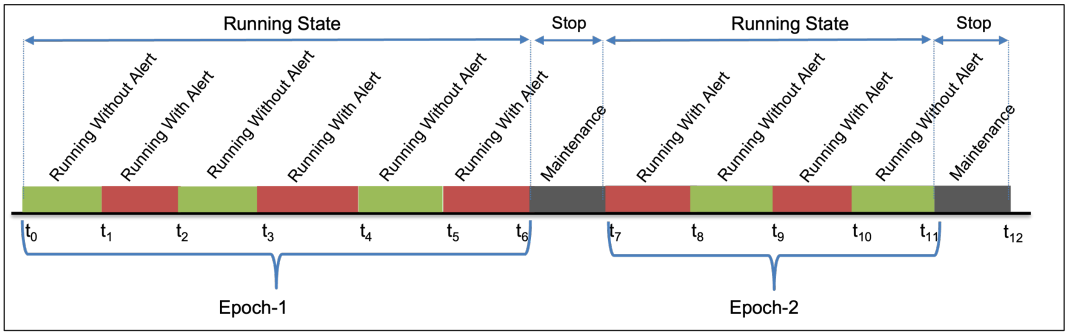


Figure 4: Illustration of a snapshot of different states of the machine (*i.e.*, "running without alert", "running with alert" and "failure").

The data considered by Pal *et al.* (2024) consists of 45 epochs, in which the total number of "running with alert" events is 1584, whereas the total number of "running without alert" events is 1606 (*i.e.*, 3190 running events and 45 failures). Moreover, the IM machine considered here consists of 72 different sensors that may explain the reasons behind the time spent on the three states. These sensors are majorly related to *mold surface temperature*, *cooling rate of cavities*, *post gate cavity pressure*, *filled area of post gate cavity*, *filled area of molding*, *injection fill time*, *screw runtime*, *etc.*

2. Model developed by Pal *et al.* (2024)

This section summarizes the key aspects of the failure time prediction methodologies developed by Pal *et al.* (2024).

1. Pre-processing of the data: the machine states are labelled into three categories: failure, running with alert, and running without alert. Furthermore, the consecutive (different) alerts with the same label (*e.g.*, running with alert) are clubbed together as the same state / event.

2. Distributional assumption of the key variables:

   (a) Number of events per epoch (R_i): Since $R_i \geq 1$ (0 is not possible), a shifted Poisson distribution is assumed. The probability mass function (PMF) is

   $$P(R_i = r_i) = e^{-\mu} \frac{\mu^{(r_i-1)}}{(r_i-1)!} \; ; \; r_i = 1, 2, \cdots . \tag{1}$$

   (b) Duration of a "running without alert" event j in epoch i (denoted by X_{ij}^1): exponential with rate parameter λ_1

   (c) Duration of a "running with alert" event j in epoch i (denoted by X_{ij}^2): exponential with rate parameter λ_2

3. Let N_i^1 be the number of "running without alert" events in the i-th epoch, and N_i^2 refers to the number of "running with alert" event in the i-th epoch. Grouping of epochs into four situations:

   (a) **Situation 1**: The epoch starts with the event "running without alert" and the number of events r_i is odd. Hence $N_i^1 = \frac{r_i+1}{2}$ and $N_i^2 = \frac{r_i-1}{2}$.

   (b) **Situation 2**: The epoch starts with the event "running without alert" and the number of events r_i is even. Hence $N_i^1 = \frac{r_i}{2}$ and $N_i^2 = \frac{r_i}{2}$.

   (c) **Situation 3**: The epoch starts with the event "running with alert" and the number of events r_i is odd. Hence $N_i^1 = \frac{r_i-1}{2}$ and $N_i^2 = \frac{r_i+1}{2}$.

   (d) **Situation 4**: The epoch starts with the event "running with alert" and the number of events r_i is even. Hence $N_i^1 = \frac{r_i}{2}$ and $N_i^2 = \frac{r_i}{2}$.

4. Likelihood calculation: the likelihood for Situation 1 can be written as:

   $$L_1(\theta) = c_1 \prod_{i \in S_1} \left[P(R_i = r_i) \times p \times \prod_{j=1}^{\frac{r_i+1}{2}} f^1(x_{ij}^1) \times \prod_{j=1}^{\frac{r_i-1}{2}} f^2(x_{ij}^2) \right], \tag{2}$$

   where, $f^k(.)$ is the probability density function (PDF) of exponential distribution with mean $1/\lambda_k$, for $k = 1, 2$, p is the probability of the first event being "running without alert", and c_1 is the proportionality constant independent of the parameters θ. After ignoring the constant, using appropriate PDFs and PMF in the above likelihood

function, and taking natural-log we get the log-likelihood,

$$\begin{aligned}\mathcal{L}_1(\theta) = &- n_1\ \mu + ln(\mu) \sum_{i\in S_1}(r_i-1) - \sum_{i\in S_1} ln((r_i-1)!) + ln(\lambda_1) \sum_{i\in S_1}\left(\frac{r_i+1}{2}\right) \\ &- \lambda_1 \sum_{i\in S_1}\sum_{j=1}^{\frac{r_i+1}{2}} x_{ij}^1 + ln(\lambda_2) \sum_{i\in S_1}\left(\frac{r_i-1}{2}\right) - \lambda_2 \sum_{i\in S_1}\sum_{j=1}^{\frac{r_i-1}{2}} x_{ij}^2 + n_1 ln(p).\end{aligned} \tag{3}$$

For other situations, the likelihood expression will be similar and the readers can refer to Appendix A1 in Pal *et al.* (2024). Subsequently, the log-likelihood of the data from all n epochs and four situations can be written as, $\mathcal{L}(\theta) = \mathcal{L}_1(\theta)+\mathcal{L}_2(\theta)+\mathcal{L}_3(\theta)+\mathcal{L}_4(\theta)$. The parameter vector $\theta = (\lambda_1, \lambda_2, p, \mu)$ is estimated by maximizing $\mathcal{L}(\theta)$. By defining $N_{il}^s = \frac{r_i+a_l^s}{2}$ and

$$a_l^s = \begin{cases}(-1)^{s+1}, \text{ if } l = 1\\ 0, \text{ if } l = 2, 4\\ (-1)^s, \text{ if } l = 3\end{cases}$$

for $s = 1, 2$, the closed form analytical expression of the maximum likelihood estimators (MLEs) are given by

$$\widehat{\lambda}_s = \frac{\sum_{l=1}^{4}\sum_{i\in S_l} N_{il}^s}{\sum_{l=1}^{4}\sum_{i\in S_l}\sum_{j=1}^{N_{il}^s} x_{ij}^s} \text{ for } s = 1, 2, \quad \widehat{p} = \frac{n_1+n_2}{n} \text{ and } \widehat{\mu} = \frac{1}{n}\sum_{l=1}^{4}\sum_{i\in S_l}(r_i-1). \tag{4}$$

5. Since there are 72 sensors, important ones that might influence the current state of the machine are identified via variable importance method within the random forest model framework.

6. Subsequently, these m important sensor-based covariates are introduced in the model via generalized linear regressors. That is, the generalized linear model (GLM) considered for λ_1, λ_2 and μ in i-th epoch can be written as,

$$\lambda_{1i} = \exp\left(\beta_0 + \sum_{k=1}^{m} F_{ki}\beta_k\right),\ \lambda_{2i} = \exp\left(\gamma_0 + \sum_{k=1}^{m} F_{ki}\gamma_k\right),\ \mu_i = \exp\left(\eta_0 + \sum_{k=1}^{m} F_{ki}\eta_k\right),$$

where $\beta_k, \gamma_k, \eta_k$ for $k = 0, 1, \cdots, m$ denote the unknown regression coefficients and F_{ki} denotes the k-th sensor value in the i-th epoch.

7. Next, the MLEs of these regression parameters are obtained using numerical optimization. Additionally, uncertainty bounds for these estimates are obtained through non-parametric (asymptotic and Bootstrap) confidence intervals.

8. Finally, Pal *et al.* (2024) addressed the main objective of the paper, *i.e.*, the derivation of the expected time to fail for the IM machine. Given that the epochs in four situations are based on whether the number of events is even or odd, and whether the first event

is of "running with alert" or "running without alert", the expected time to fail can be written as:

$$E[\text{Time to fail}] = \frac{(1-e^{-2\mu})(\mu+1)}{4}\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}\right) + \frac{(1+e^{-2\mu})}{4}\left[\frac{\mu+2p}{\lambda_1}+\frac{\mu+2(1-p)}{\lambda_2}\right]. \quad (5)$$

In practice, the values of $\lambda_1, \lambda_2, \mu$ and p are required, which in-turn requires the values of covariates, to compute the expected time to fail for an out-of-sample epoch. Pal *et al.* (2024) have taken the epoch-wise average value of covariates for comparison the model performance. Alternatively, one can take the average sensor values across all 45 epochs (*i.e.*, over 3190 events) to estimate the expected time to fail. Of course, if we know the true values of the sensors, one can use that instead, however these values are typically not known in advance.

Pal *et al.* (2024) implemented the methodology on the data obtained from the IM machine that manufactures softdrink bottles. The performance comparison of the actual time to fail with the expected time to fail derived in Step 8, and the popular Cox-proportional hazard (CPH) model is presented in Figure 5.

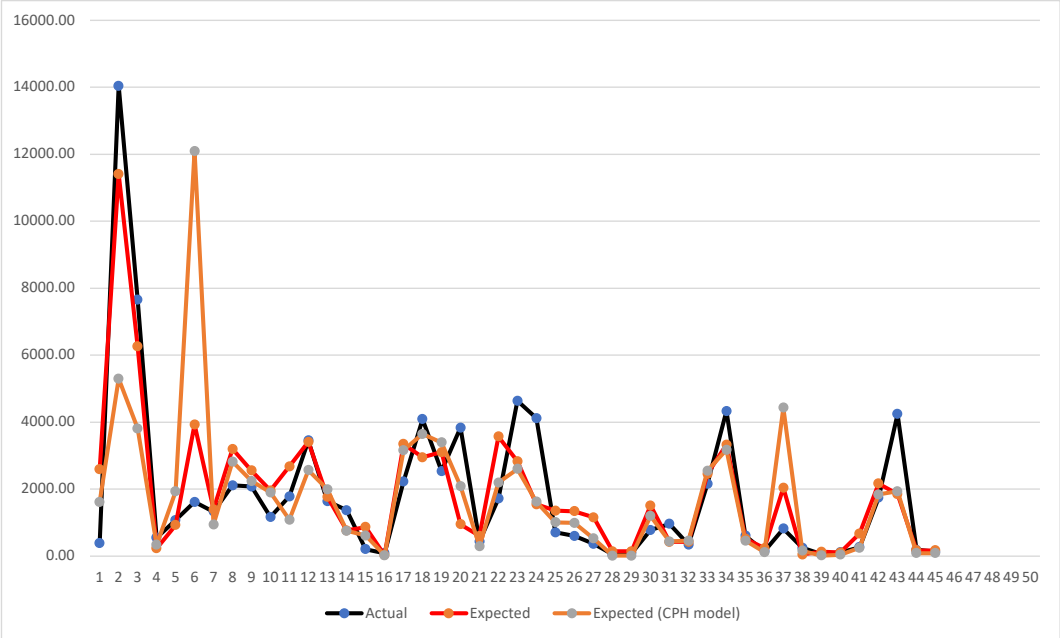


Figure 5: Plot of epoch-wise actual and expected time to fail by the proposed model and the CPH model.

The visual comparison between the three sets of values in Figure 5 clearly show the superior performance of the proposed model as compared to the CPH model. However, one can compute various goodness of fit measures for quantitative comparison as well. Table 1 presents the mean square error (MSE), mean absolute error (MAE), maximum error (MaxE) and correlation between actual data and the proposed model.

Table 1: Performance of the proposed model vs CPH model

Model	MSE	MAE	MaxE	Correlation
Proposed	1299394.36	796.38	2881.12	0.89
CPH	5389542.03	1102.77	10477.12	0.48

3. Proposed extension

Although the model proposed by Pal *et al.* (2024) demonstrates superior performance than the popular CPH model, there is a room for further investigation and possible improvement. For instance, the distribution of X_{ij}^k (j-th event of type k ($k = 1, 2$) in epoch i), the time duration spent by the machine on a given state (*i.e.*, duration of "running without alert" or "running with alert") was assumed to be exponential because of popularity and simplicity. It turns out that Weibull distribution is more general and hence a better choice than exponential for modeling X_{ij}^k. This paper discusses the key expressions of Pal *et al.* (2024) that need to be modified as per the Weibull distribution.

Let $X_{ij}^k \sim Weibull(\alpha_k, \lambda_k)$, for $k = 1, 2$. For simplicity, one can take identical shape parameters, *i.e.*, $\alpha_k = \alpha$. As a result, the PDF of X_{ij}^k is given by

$$f_k(x) = \lambda_k \alpha x^{\alpha-1} e^{-\lambda_k x^\alpha},$$

with mean $\frac{1}{\lambda^{1/k}}\Gamma(1 + \frac{1}{\alpha})$.

First, the likelihood in (2), and for other situations, will be modified as

$$L_1(\theta) = c_1 \prod_{i \in S_1} \left[P(R_i = r_i) \times p \times \prod_{j=1}^{\frac{r_i+1}{2}} \left\{ \lambda_1 \alpha (x_{ij}^1)^{\alpha-1} e^{-\lambda_1 (x_{ij}^1)^\alpha} \right\} \times \prod_{j=1}^{\frac{r_i-1}{2}} \left\{ \lambda_2 \alpha (x_{ij}^2)^{\alpha-1} e^{-\lambda_2 (x_{ij}^2)^\alpha} \right\} \right].$$

This leads to the update of the log-likelihood expression in (3) as

$$\begin{aligned}
\mathcal{L}_1(\theta) &= -n_1\ \mu + ln(\mu) \sum_{i \in S_1} (r_i - 1) - \sum_{i \in S_1} ln((r_i - 1)!) + ln(\alpha) \sum_{i \in S_1} r_i + n_1 ln(p) \\
&\quad + ln(\lambda_1) \sum_{i \in S_1} \left(\frac{r_i + 1}{2} \right) - \lambda_1 \sum_{i \in S_1} \sum_{j=1}^{\frac{r_i+1}{2}} (x_{ij}^1)^\alpha + (\alpha - 1) \sum_{i \in S_1} \sum_{j=1}^{\frac{r_i+1}{2}} ln(x_{ij}^1) \\
&\quad + ln(\lambda_2) \sum_{i \in S_1} \left(\frac{r_i - 1}{2} \right) - \lambda_2 \sum_{i \in S_1} \sum_{j=1}^{\frac{r_i-1}{2}} (x_{ij}^2)^\alpha + (\alpha - 1) \sum_{i \in S_1} \sum_{j=1}^{\frac{r_i-1}{2}} ln(x_{ij}^2).
\end{aligned}$$

AS earlier, the total log-likelihood of the data from all n epochs and four situations, $\mathcal{L}(\theta) = \mathcal{L}_1(\theta) + \mathcal{L}_2(\theta) + \mathcal{L}_3(\theta) + \mathcal{L}_4(\theta)$, can be maximized to obtain

$$\widehat{\lambda}_s = \frac{\sum_{l=1}^{4} \sum_{i \in S_l} N_{il}^s}{\sum_{l=1}^{4} \sum_{i \in S_l} \sum_{j=1}^{N_{il}^s} (x_{ij}^s)^{\widehat{\alpha}}} \text{ for } s = 1, 2, \quad \widehat{p} = \frac{n_1 + n_2}{n} \text{ and } \widehat{\mu} = \frac{1}{n} \sum_{l=1}^{4} \sum_{i \in S_l} (r_i - 1), \tag{6}$$

where, $\widehat{\alpha}$ can be obtained by maximizing the profile log-likelihood function of α, given by,

$$\begin{aligned} g(\alpha) &= ln(\alpha)\sum_{l=1}^{4}\sum_{i\in S_l} r_i - \sum_{s=1}^{2}\sum_{l=1}^{4}\sum_{i\in S_l}\left(\frac{r_i+a_l^s}{2}\right) ln\left(\sum_{l=1}^{4}\sum_{i\in S_l}\sum_{j=1}^{N_{il}^s}(x_{ij}^s)^\alpha\right) \\ &\quad +\alpha\sum_{s=1}^{2}\sum_{l=1}^{4}\sum_{i\in S_l}\sum_{j=1}^{N_{il}^s} ln(x_{ij}^s). \end{aligned} \tag{7}$$

The uniqueness of $\widehat{\alpha}$ can be established with the help of the following two theorems

Theorem 1: The profile log-likelihood of α, given by, $g(\alpha)$ in (7) is a concave function.

Proof: We skip the derivation of the first derivative of $g(\alpha)$ and directly jump to the second derivative of $g(\alpha)$, *i.e.*,

$$\frac{d^2g(\alpha)}{d\alpha^2} = -\frac{1}{\alpha^2}\sum_{l=1}^{4}\sum_{i\in S_l} r_i - \sum_{s=1}^{2}\sum_{l=1}^{4}\sum_{i\in S_l}\left(\frac{r_i+a_l^s}{2}\right)(D_s(\alpha)-E_s(\alpha))\left(\sum_{l=1}^{4}\sum_{i\in S_l}\sum_{j=1}^{N_{il}^s}(x_{ij}^s)^\alpha\right)^{-2},$$

where, $D_s = \sum_{l=1}^{4}\sum_{i\in S_l}\sum_{j=1}^{N_{il}^s}(x_{ij}^s)^\alpha \sum_{l=1}^{4}\sum_{i\in S_l}\sum_{j=1}^{N_{il}^s}(x_{ij}^s)^\alpha (ln(x_{ij}))^2$ and $E_s = \left(\sum_{l=1}^{4}\sum_{i\in S_l}\sum_{j=1}^{N_{il}^s}(x_{ij}^s)^\alpha ln(x_{ij})\right)^2$.

Using Cauchy-Schwarz inequality, we get $D_s(\alpha)-E_s(\alpha)\geq 0$ confirming $d^2g(\alpha)/d\alpha^2 \leq 0$. Hence $g(\alpha)$ is a concave function. □

Theorem 2: The profile log-likelihood of α, given by, $g(\alpha)$ in (7) has a unique maximum.

Proof: Given Theorem 1, we only need to show that the first order derivative of $g(\alpha)$ has a unique root. Note that $dg(\alpha)/d\alpha = 0$ can be written as $G(\alpha) - H(\alpha) = 0$,

$$\text{where, } G(\alpha) = \frac{\sum_{l=1}^{4}\sum_{i\in S_l} r_i}{\alpha} + \sum_{s=1}^{2}\sum_{l=1}^{4}\sum_{i\in S_l}\sum_{j=1}^{N_{il}^s} ln(x_{ij}^s),$$

$$\text{and } H(\alpha) = \sum_{s=1}^{2}\sum_{l=1}^{4}\sum_{i\in S_l}\left(\frac{r_i+a_l^s}{2}\right)\frac{\sum_{l=1}^{4}\sum_{i\in S_l}\sum_{j=1}^{N_{il}^s}(x_{ij}^s)^\alpha ln(x_{ij}^s)}{\left(\sum_{l=1}^{4}\sum_{i\in S_l}\sum_{j=1}^{N_{il}^s}(x_{ij}^s)^\alpha\right)}. \tag{8}$$

Clearly $G(\alpha)$ is a decreasing function of α. Also the first order derivative of $H(\alpha)$ is

$$\frac{d}{d\alpha}H(\alpha) = \sum_{s=1}^{2}\sum_{l=1}^{4}\sum_{i\in S_l}\left(\frac{r_i+a_l^s}{2}\right)\frac{D_s(\alpha)-E_s(\alpha)}{\left(\sum_{l=1}^{4}\sum_{i\in S_l}\sum_{j=1}^{N_{il}^s}(x_{ij}^s)^\alpha\right)^2},$$

where, $D_s(\alpha)$ and $E_s(\alpha)$ are defined in Theorem 1. By using Cauchy-Schwarz inequality, it can be noted that $dH(\alpha)/d\alpha \geq 0$, ensuring $H(\alpha)$ is an increasing function of α. Since $G(\alpha)$ is a decreasing function of α and the function $g(\alpha)$ has at-least one maximum, it is clear that $G(\alpha)$ and $H(\alpha)$ intersect at only one point ensuring unique solution of (8). Hence $g(\alpha)$ has the unique maximum value. □

Using Theorem 1 and Theorem 2, it is proved that $\widehat{\alpha}$ exists and is unique. By using invariance property of MLE, $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ are also unique.

As proposed by Pal *et al.* (2024), the sensors-based covariates are used to form generalized linear regression with the same re-parametrization as follows,

$$\lambda_{1i} = \exp\left(\beta_0 + \sum_{k=1}^{m} F_{ki}\beta_k\right),\ \lambda_{2i} = \exp\left(\gamma_0 + \sum_{k=1}^{m} F_{ki}\gamma_k\right),\ \mu_i = \exp\left(\eta_0 + \sum_{k=1}^{m} F_{ki}\eta_k\right).$$

For the sake of simplicity, the parameter α is not parametrized in terms of the covariates although one can re-parametrize it if needed as $\alpha_i = \exp\left(\zeta_0 + \sum_{k=1}^{m} F_{ki}\zeta_k\right)$, for some unknown regression coefficients ζ_k for $k = 0, 1, \cdots m$.

Following the similar approach by Pal *et al.* (2024), the expression for the expected time to fail can be written as:

$$\begin{aligned} E[\text{Time to fail}] \;=\; & \Gamma\left(1+\frac{1}{\alpha}\right)\left[\frac{(1-e^{-2\mu})(\mu+1)}{4}\left(\frac{1}{\lambda_1^{1/\alpha}}+\frac{1}{\lambda_2^{1/\alpha}}\right)\right. \\ & \left.+\frac{(1+e^{-2\mu})}{4}\left[\frac{\mu+2p}{\lambda_1^{1/\alpha}}+\frac{\mu+2(1-p)}{\lambda_2^{1/\alpha}}\right]\right]. \end{aligned} \tag{9}$$

Therefore after substituting the values of $\widehat{\alpha}$, $\widehat{\lambda}_1$, $\widehat{\lambda}_2$ and $\widehat{p}$ in (9), the estimated time to fail of the machine is obtained.

4. Concluding remarks

This study extends the model proposed by Pal *et al.* (2024) to analyze sequential data from an IM machine, focusing on alternating periods of operation with alerts and without alerts, resulting in machine failure. The durations with alerts is assumed to follow Weibull distribution with scale parameter λ_1 and shape parameter α, while durations without alerts is assumed to follow Weibull distribution independently with scale parameter λ_2 and the same shape parameter α, allowing for flexible modeling. Notably setting $\alpha = 1$ recovers the earlier model by Pal *et al.* (2024) as a special case. The number of events before failure is modeled using a conditional Poisson distribution, given at least one has happened prior to failure. We have derived maximum likelihood estimators for the parameters and used these to formulate the expected time to machine failure. However we have not reported the numerical findings of the proposed model and we at this stage leave it for a future work.

Acknowledgements

Our sincere thanks to the organizers of the conference for providing a wonderful opportunity to discuss the paper with eminent researchers around the globe. We are also grateful to the Editors and reviewers for their valuable 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

Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, **22**, 207-216.

Ding, J., Liu, Y., Zhang, L., and Wang, J. (2015). Modeling the process of event sequence data generated for working condition diagnosis. *Mathematical Problems in Engineering*, https://doi.org/10.1155/2015/693450.

Efron, B., and Tibshirani, R.J. (1994). *An Introduction to the Bootstrap*. Chapman and Hall/CRC, New York.

Glazier, S. W. (2019). *Sequential Survival Analysis with Deep Learning*. Brigham Young University, Master of Science Thesis.

Hastie, T., Tibshirani, R., and Friedman, J. (2013). *The Elements of Statistical Learning*. Springer, New York.

Huang, X., and An, A. (2002). Discovery of interesting association rules from livelink web log data. *IEEE International Conference on Data Mining, 2002 Proceedings, Maebashi City, Japan.*

Jiang, X-R., and Gruenwald, Le. (2005). Microarray gene expression data association rules mining based on BSC-Tree and FIS-Tree. *Data & Knowledge Engineering*, **53**, 3–29.

Kohavi, R., Mason, L., Parekh, R., and Zheng, Z. (2004). Lessons and challenges from mining retail e-commerce data. *Machine Learning*, **57**, 83-113.

Li, Z., Zhou, S., Choubey, S., and Sievenpiper, C. (2007). Failure event prediction using the Cox proportional hazard model driven by frequent failure signatures. *IIE transactions*, **39**, 303-315.

Lou, J. G., Fu, Q., Wang, Y., and Li, J. (2010). Mining dependency in distributed systems through unstructured logs analysis. *ACM SIGOPS Operating Systems Review*, **44**, 91-96.

Luo, C., Lou, J. G., Lin, Q., Fu, Q., Ding, R., Zhang, D., and Wang, Z. (2014). Correlating events with time series for incident diagnosis. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to Linear Regression Analysis*. Wiley.

Motahari-Nezhad, H. R., Saint-Paul, R., Casati, F., and Benatallah, B. (2011). Event correlation for process discovery from web service interaction logs. *The VLDB Journal*, **20**, 417-444.

Pal, S.K., Koley, A., Ranjan, P., and Kundu, D. (2024). Modeling time to failure using a temporal sequence of events, *Quality Engineering*, 1-20, https://doi.org/10.1080/08982112.2024.2441367

Rudin, C., Letham, B., Salleb-Aouissi, A., Kogan, E., and Madigan, D. (2011). Sequential event prediction with association rules. *JMLR: Workshop and Conference Proceedings* **19**, 615–634.

Wu, D., Ke, Y., Yu, J. X., Yu, P. S., and Chen, L. (2010). Detecting leaders from correlated time series. In *Kitagawa, H., Ishikawa, Y., Li, Q., Watanabe, C. (eds) Database Systems for Advanced Applications. DASFAA 2010. Lecture Notes in Computer Science, vol 5981. Springer, Berlin, Heidelberg.*

Zhu, Y., and Shasha, D. (2002). StatStream: Statistical monitoring of thousands of data. In *Proceedings 2002 VLDB Conference: 28th International Conference on Very Large Databases (VLDB). Elsevier.*

Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025; pp 13-30

On Bivariate Weibull Frailty Model

Debasis Kundu

Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, Pin 208016, India

Received: 14 April 2025; Revised: 18 May 2025; Accepted: 20 May 2025

Abstract

In this paper we have considered bivariate Weibull frailty (BWF) model. It has a singularity along the line $x = y$. There is a positive probability that the two marginals can be equal similar to the Marshall-Olkin bivariate exponential or Marshall-Olkin bivariate Weibull models. The Marshall-Olkin bivariate exponential or Weibull distribution can be obtained as a limiting case of the proposed model. It is a very flexible model. The joint probability density function can take variety of shapes. Different properties of the BWF model have been studied. Different dependency measures have been investigated. The model has five parameters. Computing the maximum likelihood estimators of the unknown parameters involves solving a five dimensional optimization problem. An effective EM algorithm has been proposed and it can be implemented quite conveniently in practice. Extensive simulations have been performed to show the effectiveness of the proposed method. One diabetic retinopathy data set has been analyzed. We have further proposed to analyze dependent competing risks data, and one competing risks data set has been analyzed. The results are quite satisfactory.

Key words: Marshall-Olkin bivariate exponential distribution; Marshall-Olkin bivariate Weibull distribution; Bivariate singular distribution; Bivariate copula; Positive dependence, Maximum likelihood estimators; EM algorithm.

AMS Subject Classifications: 62F10, 62F03, 62H12.

1. Introduction

The motivation of this work came when we were trying to analyze a diabetic retinopathy data set. The diabetic retinopathy is a medical condition of the eyes. This particular eye conditions may depend on various factors of the individual namely age, sex, blood sugar level, cholesterol level etc. One major issue of this disease is that unless somebody goes for a regular eye check-up this may not be detected at the early stage. The final outcome of this disease is blindness. Till today we do not have any treatment available so that this disease can be cured. The available treatment can only delay the onset of blindness. Due to this

Corresponding Author: Debasis Kundu
Email: kundu@iitk.ac.in

reason an extensive amount of work is going on to develop new treatment so that the onset of the blindness can be delayed. One such treatment which has been recently introduce is the laser treatment. Different experiments have been conducted to test whether the laser treatment has any significant effect in delaying the onset of blindness or not, compared to the traditional treatment.

We will be discussing two such data sets. One data set is of the form (X, Y). Here X and Y denote the time to blindness of the laser treated eye and the other eye, respectively, of the same individual. Clearly, here both $X > 0$ and $Y > 0$. The other data set is of the form (T, Δ), here $T = \min\{X, Y\}$, where X and Y are same as defined above, and $\Delta = 1$, $\Delta = 2$ or $\Delta = 3$, if $T = X$, $T = Y$ or $T = X = Y$, respectively. One important feature of these data sets is that there is a significant portion of the data points where $X = Y$, hence it cannot be ignored. Due to this reason, several authors have analyzed these data sets based on the assumptions that (X, Y) follows Marshall-Olkin bivariate Weibull (MOBW) distribution, see for example Feizjavadian and Hashemi (2015), Cai et al. (2017), Shen and Xu (2018), Kundu (2022) and Samanta and Kundu (2023).

In the first data set few covariates (age, sex and blood sugar level) are available, where as in the second data set no covariates is available. It is quite possible that many other variables also influence these survival times. Such factors are usually unknown and thus cannot be explicitly included in the analysis. Vaupel et al. (1979) suggested a mathematical model for this. They have introduced a random variable, which is not observed, for each individual to the associated survival function. Since it is not observed it is integrated out. There is some identifiability issue but it can be sorted out.

The main aim of this paper is to introduce the frailty for the MOBW model in a very natural way. We call this new model as the Bivariate Weibull Frailty (BWF) model. It may be mentioned that the MOBW distribution has received a considerable amount of attention in analyzing bivariate data with ties. Extensive work has been done in establishing different properties and developing both the classical and Bayesian inference procedures for MOBW model, see for example the review article by Kundu (2023) and the references cited there in. The MOBW model can be obtained as a limiting case of the BWF mode. The MOBW distribution has four parameters, where as the proposed BWF model has five parameters. The BWF model is a very flexible model and it also has a singularity along the line $x = y$ similar to the MOBW model. Hence, this model also can be used quite effectively if there are ties in the data set. The joint probability density function (JPDF) can be take variety of shapes, and we have derived different properties of the BWF model. Different dependency properties and dependency measures also have been established. The maximum likelihood estimators of the unknown parameters cannot be obtained in explicit forms. It involves solving five dimensional optimization process. Moreover, finding the five dimensional initial guesses also is not a trivial issue. To avoid that we have proposed to use EM algorithm, which can be implemented quite conveniently in practice. Extensive simulations have been performed to show the effectiveness of the proposed method and one bivariate diabetic retinopathy data set has been analyzed based on the proposed model. We have further proposed a competing risks model based on the BWF model, and one diabetic retinopathy competing risks data has been analyzed based on this model.

The rest of the paper is organized as follows. In Section 2 we have defined the BWF

model and provided several of its properties. The maximum likelihood estimators based on the EM algorithm has been proposed in Section 3. The analysis of a bivariate diabetic retinopathy data set has been provided in Section 4. In Section 5 we have indicated how the proposed BWF model can be used effectively to analyze competing risks data and the analysis of a data set has also been presented. Finally we have concluded the paper in Section 6.

2. Bivariate Weibull frailty model

2.1. Notations and preliminaries

We will use the following notations for the rest of the paper. The two-parameter Weibull distribution with the shape parameter $\alpha > 0$ and $\lambda > 0$ has the following probability density function (PDF);

$$f_{WE}(x;\alpha,\lambda) = \alpha\lambda x^{\alpha-1}e^{-\lambda x^{\alpha}}; \quad \text{for} \quad x > 0, \tag{1}$$

and zero, otherwise. The corresponding cumulative distribution function (CDF), survival function (SF) and the hazard function (HF) will be denoted by $F_{WE}(x;\alpha,\lambda)$, $S_{WE}(x;\alpha,\lambda)$ and $h_{WE}(x;\alpha,\lambda)$, respectively, and for $x > 0$, they are as follows:

$$F_{WE}(x;\alpha,\lambda) = 1 - e^{-\lambda x^{\alpha}}, \quad S_{WE}(x;\alpha,\lambda) = e^{-\lambda x^{\alpha}}, \quad h_{WE}(x;\alpha,\lambda) = \alpha\lambda x^{\alpha-1}.$$

From now on a Weibull distribution with the shape parameter α and scale parameter λ will be denoted by $\text{WE}(\alpha,\lambda)$. The two-parameter gamma distribution with the shape parameter $\alpha > 0$ and $\lambda > 0$ has the following probability density function (PDF);

$$f_{GA}(x;\alpha,\lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}; \quad \text{for} \quad x > 0, \tag{2}$$

and zero, otherwise. It will be denoted by $\text{GA}(\alpha,\lambda)$.

Let $U_1 \sim$ (follows) $\text{WE}(\alpha,\lambda_1)$, $U_2 \sim \text{WE}(\alpha,\lambda_2)$, $U_3 \sim \text{WE}(\alpha,\lambda_3)$ and they are independently distributed, then (X,Y), where $X = \min\{U_1,U_3\}$, $Y = \min\{U_2,U_3\}$ follows Marshall-Olkin bivariate Weibull (MOBW) distribution with parameters $\alpha,\lambda_1,\lambda_2,\lambda_3$. From now on it will be denoted by $\text{MOBW}(\alpha,\lambda_1,\lambda_2,\lambda_3)$. The joint PDF of (X,Y) can be written as

$$f_{MOBW}(x,y) = \begin{cases} f_{WE}(x;\alpha,\lambda_1)f_{WE}(y;\alpha,\lambda_2+\lambda_3) & \text{if} \quad x < y \\ f_{WE}(x;\alpha,\lambda_1+\lambda_3)f_{WE}(y;\alpha,\lambda_2) & \text{if} \quad x > y \\ \frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3}f_{WE}(z;\alpha,\lambda_1+\lambda_2+\lambda_3) & \text{if} \quad x = y = z. \end{cases} \tag{3}$$

2.2. Model descriptions

Definition: Suppose $U_1 \sim \text{WE}(\alpha,\lambda_1)$, $U_2 \sim \text{WE}(\alpha,\lambda_2)$, $U_3 \sim \text{WE}(\alpha,\lambda_3)$, $V \sim \text{GA}(\beta,\beta)$ and they are all independently distributed. Let us define

$$X = \min\left\{\frac{U_1}{V^{1/\alpha}}, \frac{U_3}{V^{1/\alpha}}\right\} \quad \text{and} \quad Y = \min\left\{\frac{U_2}{V^{1/\alpha}}, \frac{U_3}{V^{1/\alpha}}\right\}. \tag{4}$$

Then (X,Y) is said to have BWF distribution with parameters $\alpha,\lambda_1,\lambda_2,\lambda_3,\beta$. It will be denoted by $\text{BWF}(\alpha,\lambda_1,\lambda_2,\lambda_3,\beta)$.

The joint SF of (X,Y) for $x>0$ and $y>0$ becomes

$$\begin{aligned}
S_{BWF}(x,y) &= P(X>x,Y>y)\\
&= \int_0^\infty P(U_1>v^{1/\alpha}x, U_2>v^{1/\alpha}y, U_3>v^{1/\alpha}\max\{x,y\})f_{GA}(v;\beta,\beta)dv\\
&= \int_0^\infty P(U_1>v^{1/\alpha}x, U_2>v^{1/\alpha}y, U_3>v^{1/\alpha}\max\{x,y\})f_{GA}(v;\beta,\beta)dv\\
&= \int_0^\infty e^{-v\lambda_1 x^\alpha}e^{-v\lambda_2 y^\alpha}e^{-v\lambda_3\max\{x^\alpha,y^\alpha\}}f_{GA}(v;\beta,\beta)dv\\
&= \frac{\beta^\beta}{\Gamma(\beta)}\int_0^\infty v^{\beta-1}e^{-v(\beta+\lambda_1 x^\alpha+\lambda_2 y^\alpha+\lambda_3\max\{x^\alpha,y^\alpha\})}dv\\
&= \left[1+\frac{\lambda_1}{\beta}x^\alpha+\frac{\lambda_2}{\beta}y^\alpha+\frac{\lambda_3}{\beta}\max\{x^\alpha,y^\alpha\}\right]^{-\beta}\\
&= \begin{cases}\left[1+\theta_1 x^\alpha+(\theta_2+\theta_3)y^\alpha\right]^{-\beta} & \text{if} \quad x<y\\ \left[1+(\theta_1+\theta_3)x^\alpha+\theta_2 y^\alpha\right]^{-\beta} & \text{if} \quad y\le x.\end{cases}
\end{aligned}$$

Here we have denoted $\theta_1=\lambda_1/\beta$, $\theta_2=\lambda_2/\beta$ and $\theta_3=\lambda_3/\beta$. Hence, the marginal SFs of X and Y become

$$P(X>x)=\left[1+(\theta_1+\theta_3)x^\alpha\right]^{-\beta} \quad \text{and} \quad P(Y>y)=\left[1+(\theta_2+\theta_3)y^\alpha\right]^{-\beta}.$$

The PDFs of X and Y for $x>0$ and $y>0$ become

$$f_X(x)=\frac{\alpha\beta(\theta_1+\theta_3)x^{\alpha-1}}{\left[1+(\theta_1+\theta_3)x^\alpha\right]^{\beta+1}} \quad \text{and} \quad f_Y(y)=\frac{\alpha\beta(\theta_2+\theta_3)x^{\alpha-1}}{\left[1+(\theta_2+\theta_3)y^\alpha\right]^{\beta+1}},$$

respectively. We introduce the following notation. A random variable is said to have a univariate Weibull frailty (UWF) distribution with parameters α,β,θ, if it has the following PDF for $x>0$

$$f_{WF}(x;\alpha,\beta,\theta)=\frac{\alpha\beta\theta x^{\alpha-1}}{\left[1+\theta x^\alpha\right]^{\beta+1}}, \tag{5}$$

and zero, otherwise. It will be denoted by UWF(α,β,θ). The generation of random sample from a UWF model is quite simple by using the inverse transformation. Hence, the generation of random sample from a BWF model can be performed very easily.

Following exactly the same procedure as in Kundu and Gupta (2009) the joint PDF of (X,Y) if (X,Y) can be obtained from the joint SF of the BWF. Alternatively, it can be obtained as follows. Observe that $\{(X,Y)|V=v\}\sim$ MOBW$(\alpha,\lambda_1 v,\lambda_2 v,\lambda_3 v)$. Hence, using (3) we can write the joint PDF of (X,Y) given $V=v$ as follows;

$$f_{(X,Y)|V=v}(x,y)=\begin{cases} f_{WE}(x;\alpha,\lambda_1 v)f_{WE}(y;\alpha,(\lambda_2+\lambda_3)v) & \text{if} \quad x<y\\ f_{WE}(x;\alpha,(\lambda_1+\lambda_3)v)f_{WE}(y;\alpha,\lambda_2 v) & \text{if} \quad y<x\\ \frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3}f_{WE}(z;\alpha,(\lambda_1+\lambda_2+\lambda_3)v) & \text{if} \quad x=y=z.\end{cases} \tag{6}$$

Hence, the joint PDF of (X,Y) becomes

$$f_{BWF}(x,y)=\begin{cases}\frac{\alpha^2\beta(\beta+1)\theta_1\theta_2 x^{\alpha-1}y^{\alpha-1}}{(1+\theta_1 x^\alpha+(\theta_2+\theta_3)y^\alpha)^{\beta+2}} & \text{if} \quad x<y\\ \frac{\alpha^2\beta(\beta+1)\theta_1\theta_2 x^{\alpha-1}y^{\alpha-1}}{(1+(\theta_1+\theta_3)x^\alpha+\theta_2 y^\alpha)^{\beta+2}} & \text{if} \quad y<x\\ \frac{\alpha\theta_3\beta z^{\alpha-1}}{(1+(\theta_1+\theta_2+\theta_3)z^\alpha)^{\beta+1}} & \text{if} \quad x=y=z.\end{cases} \tag{7}$$

The BWF distribution has an absolutely continuous part and a singular part. It is absolutely continuous on $0 < x \neq y < \infty$ and it has a singular part on $0 < x = y < \infty$. The joint PDF (7) can be written as

$$f_{BWF}(x,y) = \frac{\theta_1+\theta_2}{\theta_1+\theta_2+\theta_3} f_{ac}(x,y) + \frac{\theta_3}{\theta_1+\theta_2+\theta_3} f_{si}(x,y), \tag{8}$$

where

$$f_{ac}(x,y) = \frac{\theta_1+\theta_2+\theta_3}{\theta_1+\theta_2} \begin{cases} \frac{\alpha^2\beta(\beta+1)\theta_1\theta_2 x^{\alpha-1}y^{\alpha-1}}{(1+\theta_1 x^\alpha+(\theta_2+\theta_3)y^\alpha)^{\beta+2}} & \text{if} \quad x<y \\ \frac{\alpha^2\beta(\beta+1)\theta_1\theta_2 x^{\alpha-1}y^{\alpha-1}}{(1+(\theta_1+\theta_3)x^\alpha+\theta_2 y^\alpha)^{\beta+2}} & \text{if} \quad y<x \end{cases}$$

and

$$f_{si}(x,y) = \begin{cases} \frac{\alpha\beta(\theta_1+\theta_2+\theta_3)z^{\alpha-1}}{(1+(\theta_1+\theta_2+\theta_3)z^\alpha)^{\beta+1}} & \text{if} \quad x=y=z \\ 0 & \text{if} \quad x \neq y. \end{cases}$$

It can be easily seen that

$$P(X<Y) = \frac{\theta_1}{\theta_1+\theta_2+\theta_3}, \quad P(Y<X) = \frac{\theta_2}{\theta_1+\theta_2+\theta_3}, \quad P(X=Y) = \frac{\theta_3}{\theta_1+\theta_2+\theta_3},$$

The correlation coefficient of X and Y varies from zero to one. X and Y are independent if $\theta_3 = 0$ and the correlation tends to one as $\theta_3 \to \infty$.

The following results will be used in the implementation of the EM algorithm and they can be obtained after some calculations.

$$V|\{(X,Y)=(x,y)\} \sim \text{Gamma}(\beta+2, (\beta+\lambda_1 x^\alpha+(\lambda_2+\lambda_3)y^\alpha) \quad \text{if } x<y, \tag{9}$$
$$V|\{(X,Y)=(x,y)\} \sim \text{Gamma}(\beta+2, (\beta+(\lambda_1+\lambda_3)x^\alpha+\lambda_2 y^\alpha) \quad \text{if } x>y, \tag{10}$$
$$V|\{(X,Y)=(x,y)\} \sim \text{Gamma}(\beta+1, (\beta+(\lambda_1+\lambda_2+\lambda_3)z^\alpha) \quad \text{if } x=y=z. \tag{11}$$

Hence,

$$E(V|\{(X,Y)=(x,y)\}) = \frac{\beta+2}{(\beta+\lambda_1 x^\alpha+(\lambda_2+\lambda_3)y^\alpha)} \quad \text{if } x<y, \tag{12}$$
$$E(V|\{(X,Y)=(x,y)\}) = \frac{\beta+2}{(\beta+(\lambda_1+\lambda_3)x^\alpha+\lambda_2 y^\alpha)} \quad \text{if } x>y, \tag{13}$$
$$E(V|\{(X,Y)=(x,y)\}) = \frac{\beta+1}{(\beta+(\lambda_1+\lambda_2+\lambda_3)z^\alpha)} \quad \text{if } x=y=z \tag{14}$$

and

$$E(\ln V|\{(X,Y)=(x,y)\}) = \psi(\beta+2)-\psi(\beta+\lambda_1 x^\alpha+(\lambda_2+\lambda_3)y^\alpha)) \quad \text{if } x<y, \tag{15}$$
$$E(\ln V|\{(X,Y)=(x,y)\}) = \psi(\beta+2)-\psi(\beta+(\lambda_1+\lambda_3)x^\alpha+\lambda_2 y^\alpha)) \quad \text{if } x>y, \tag{16}$$
$$E(\ln V|\{(X,Y)=(x,y)\}) = \psi(\beta+1)-\psi(\beta+(\lambda_1+\lambda_2+\lambda_3)z^\alpha)) \quad \text{if } x=y=z. \tag{17}$$

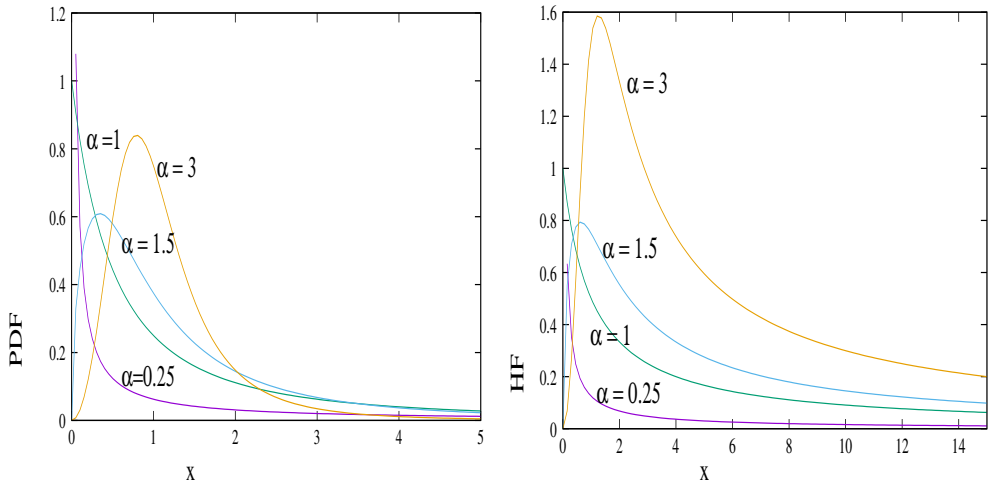
2.3. Properties

In this section we provide some properties of the UWF and BWF models. If $X \sim \text{UWF}(\alpha,\beta,\theta)$, then it can be easily seen that the PDF of X is a decreasing function for

$0 < \alpha \leq 1$ and it is an unimodal function for all values of $\beta > 0$ and $\theta > 0$. The hazard function of X becomes

$$h_{WF}(x) = \frac{\alpha\beta x^{\alpha-1}}{1+\theta x^{\alpha}}; \quad x > 0. \tag{18}$$

It can be easily shown that if $0 < \alpha \leq 1$, the hazard function is a decreasing function and for $\alpha > 1$, the hazard function is an unimodal function. It is clear that the shape (whether it will be decreasing or unimodal) of the PDF or HF does not depend on β and θ, it depends only on α. In Figure 1 we provide the plot of the PDF and HF for different values of α.



(a) (b)

Figure 1: The PDF and HF of WF(α, β, θ) for different values of α, when $\beta = \theta = 1$: (a) PDF and (b) HF

The following results are useful for data analysis purposes or it may have some independent interests also.

Theorem 1: Suppose $(X, Y) \sim \text{BWF}(\alpha, \beta, \theta_1, \theta_2, \theta_3)$. Then we have the following results:

(a) $X \sim \text{UWF}(\alpha, \beta, (\theta_1 + \theta_3))$.

(b) $Y \sim \text{UWF}(\alpha, \beta, (\theta_2 + \theta_3))$.

(c) $\min\{X, Y\} \sim \text{UWF}(\alpha, \beta, (\theta_1 + \theta_2 + \theta_3))$

Proof: The proof can be easily obtained from the joint survival function. □

The following results provide the shapes of the joint PDF of the absolute continuous part of the BWF.

Theorem 2: Suppose $(X, Y) \sim \text{BWF}(\alpha, \beta, \theta_1, \theta_2, \theta_3)$. Then we have the following results:

(a) The joint PDF of the absolute continuous part of (X, Y) is continuous everywhere if $\theta_1 = \theta_2$.

(b) If $\theta_1 \neq \theta_2$, the joint PDF of the absolute continuous part of (X, Y) is continuous everywhere except on the line $x = y$.

(c) If $0 < \alpha \leq 1$, the joint PDF of the absolute continuous part of (X, Y) is a decreasing function for all values of $\beta > 0$, $\theta_1 > 0$, $\theta_2 > 0$ and $\theta_3 > 0$.

(d) If $\theta_1 = \theta_2 = \theta$ and $\alpha > 1$ the joint PDF of the absolute continuous part of (X, Y) has a unique mode at $\left(\left(\frac{\alpha-1}{(2\theta+\theta_3)(\alpha\beta+1)}\right)^{1/\alpha}, \left(\frac{\alpha-1}{(2\theta+\theta_3)(\alpha\beta+1)}\right)^{1/\alpha}\right)$.

(e) If $\theta_1 > \theta_2 + \theta_3$ and $\alpha > 1$ the joint PDF of the absolute continuous part of (X, Y) has a unique mode at $\left(\left(\frac{\alpha-1}{\theta_1(\alpha\beta+2)}\right)^{1/\alpha}, \left(\frac{\alpha-1}{(\theta_2+\theta_3)(\alpha\beta+2)}\right)^{1/\alpha}\right)$.

(f) $\theta_2 > \theta_1 + \theta_3$ and $\alpha > 1$ the joint PDF of the absolute continuous part of (X, Y) has a unique mode at $\left(\left(\frac{\alpha-1}{(\theta_1+\theta_3)(\alpha\beta+2)}\right)^{1/\alpha}, \left(\frac{\alpha-1}{\theta_2(\alpha\beta+2)}\right)^{1/\alpha}\right)$.

(g) If $\theta_2 < \theta_1 < \theta_2 + \theta_3$ or $\theta_1 < \theta_2 < \theta_1 + \theta_3$, the joint PDF of the absolute continuous part of (X, Y) does not have any mode on $0 < x \neq y < \infty$.

Proof: The Proof of Theorem 2 is not very difficult, hence it is avoided. □

In Figure 2 we provide the contour plot of the joint PDF of the absolute continuous part of BWF for different parameter values. It shows that when $\alpha > 1$, the joint PDF is an unimodal function Figures 2 (a),(c),(d), if $\alpha \leq 1$, the joint PDF is an decreasing function Figure 2 (b), if $\theta_1 = \theta_2$, the mode of the joint PDF is on $x = y$, Figure 2 (a), if $\theta_1 > \theta_2 + \theta_3$, the mode of the joint PDF is on $x < y$, Figure 2 (c) and if $\theta_2 > \theta_1 + \theta_3$, the mode of the joint PDF is on $x > y$, Figure 2 (d).

The hazard gradient of BWF model according to Johnson and Kotz (1975) can be defined as follows;

$$h_1(x, y) = -\frac{\partial}{\partial x} \ln S_{X,Y}(x, y) \qquad \text{and} \qquad h_2(x, y) = -\frac{\partial}{\partial y} \ln S_{X,Y}(x, y).$$

Hence, the hazard gradients of BWF are as follows':

$$h_1(x, y) = -\frac{\partial}{\partial x} \ln S_{X,Y}(x, y) = \begin{cases} \frac{\beta\alpha\theta_1 x^{\alpha-1}}{1+\theta_1 x^\alpha+(\theta_2+\theta_3)y^\alpha} & \text{if} \quad x < y \\ \frac{\beta\alpha(\theta_1+\theta_3)x^{\alpha-1}}{1+(\theta_1+\theta_3)x^\alpha+\theta_2 y^\alpha} & \text{if} \quad x > y \end{cases}$$

$$h_2(x, y) = -\frac{\partial}{\partial y} \ln S_{X,Y}(x, y) = \begin{cases} \frac{\beta\alpha(\theta_2+\theta_3)y^{\alpha-1}}{1+\theta_1 x^\alpha+(\theta_2+\theta_3)y^\alpha} & \text{if} \quad x < y \\ \frac{\beta\alpha\theta_2 y^{\alpha-1}}{1+(\theta_1+\theta_3)x^\alpha+\theta_2 y^\alpha} & \text{if} \quad x > y \end{cases}$$

Now we will show that the survival function of BWF satisfies the total positivity of order two (TP_2) property, and also it satisfies some hazard rate ordering properties. Note that a function $g : \mathbb{R} \longrightarrow \mathbb{R}$ is said to have TP_2 property, if for all $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$, $\mathbf{y} =$

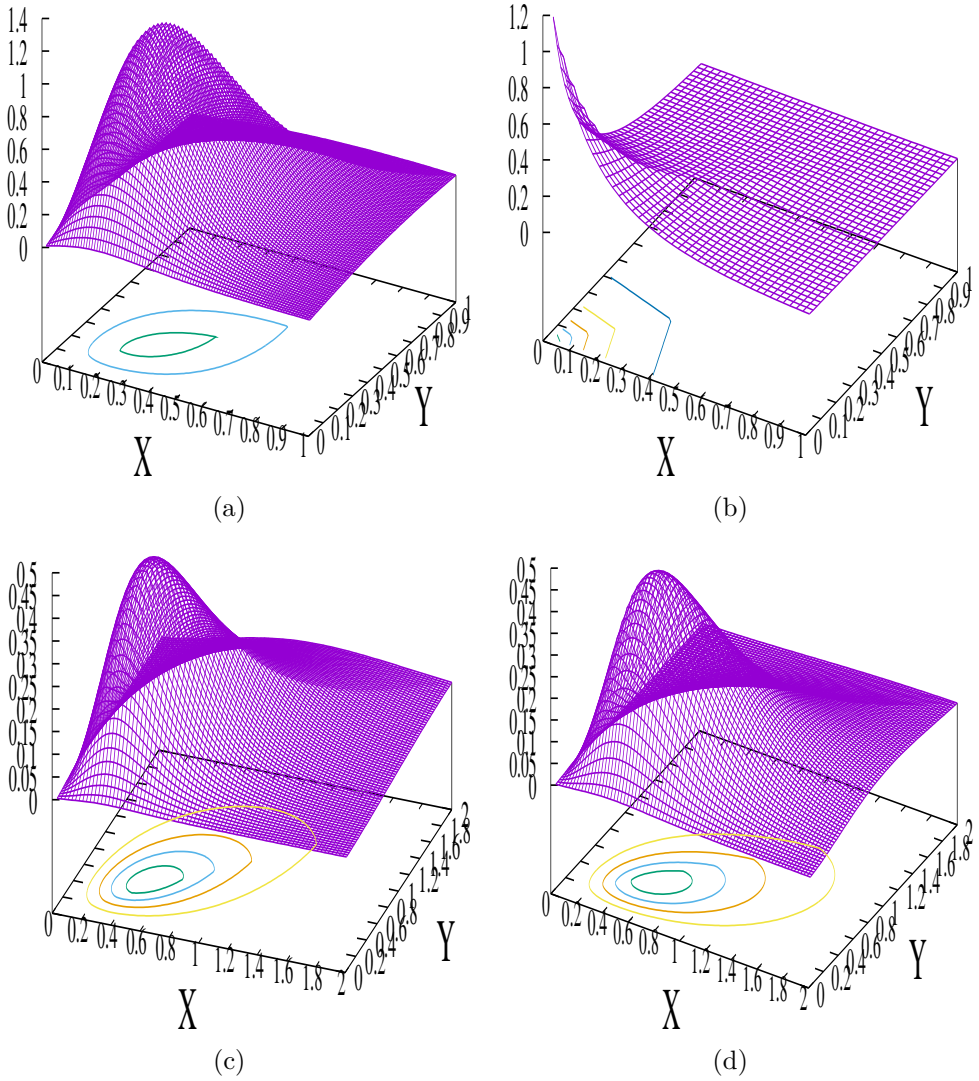


Figure 2: The contour plot of the joint PDF of the absolute continuous part of BWF$(\alpha, \beta, \theta_1, \theta_2, \theta_3)$ for different parameter values: (a) $\alpha = 2$, $\beta = 2$, $\theta_1 = \theta_2 = 1$, $\theta_3 = 2$, (b) $\alpha = 1$, $\beta = 0.5$, $\theta_1 = \theta_2 = 1$, $\theta_3 = 2$, (c) $\alpha = 2.5$, $\beta = 0.5$, $\theta_1 = 3$, $\theta_2 = \theta_3 = 1$, (d) $\alpha = 2.5$, $\beta = 0.5$, $\theta_1 = \theta_3 = 1$, $\theta_2 = 3$

$(y_1, y_2) \in \mathbb{R}^2$, $g(\mathbf{x})g(\mathbf{x}) \leq g(\mathbf{x} \wedge \mathbf{y})g(\mathbf{x} \vee \mathbf{y})$. Here $\mathbf{x} \wedge \mathbf{y} = (\min\{x_1, y_1\}, \min\{x_2, y_2\})$ and $\mathbf{x} \vee \mathbf{y} = (\max\{x_1, y_1\}, \max\{x_2, y_2\})$. Let $\boldsymbol{U}$ and $\boldsymbol{V}$ be two bivariate random vectors with survival function S_U and S_V, respectively. We say that $\boldsymbol{U}$ is smaller that $\boldsymbol{V}$ in the bivariate hazard rate order (denoted by $\boldsymbol{U} \leq_{hr} \boldsymbol{V}$) if

$$S_U(\boldsymbol{x})S_V(\boldsymbol{y}) \leq S_U(\boldsymbol{x} \wedge \boldsymbol{y})S_V(\boldsymbol{x} \vee \boldsymbol{y}); \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2.$$

We say that $\boldsymbol{U}$ is smaller that $\boldsymbol{V}$ in the bivariate weak hazard rate order (denoted by $\boldsymbol{U} \leq_{whr} \boldsymbol{V}$) if

$$S_U(\boldsymbol{x})S_V(\boldsymbol{y}) \leq S_U(\boldsymbol{x})S_V(\boldsymbol{y}); \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2. \tag{19}$$

We have the following results.

Theorem 3:

(a) If $(X, Y) \sim \text{BWF}(\alpha, \beta, \theta_1, \theta_2, \theta_3)$, then the survival function of (X, Y) has the TP_2 property.

(b) If $\boldsymbol{X} = (X_1, X_2) \sim \text{BWF}(\alpha, \beta, \theta_1, \theta_2, \theta_3)$ and $\boldsymbol{Y} = (Y_1, Y_2) \sim \text{BWF}(\alpha, \beta, \theta_1, \theta_2, \widetilde{\theta_3})$. If $\theta_3 > \widetilde{\theta_3}$, then $\boldsymbol{X} \leq_{whr} \boldsymbol{Y}$.

(c) If $\boldsymbol{X} = (X_1, X_2) \sim \text{BWF}(\alpha, \beta, \theta_1, \theta_2, \theta_3)$ and $\boldsymbol{Y} = (Y_1, Y_2) \sim \text{BWF}(\alpha, \beta, \theta_1, \theta_2, \widetilde{\theta_3})$. If $\theta_3 > \widetilde{\theta_3}$, then $\boldsymbol{X} \leq_{hr} \boldsymbol{Y}$.

Proof:

(a) To prove this, we need to show that for all possible values of $\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2$ and $\boldsymbol{y} = (y_1, y_2) \in \mathbb{R}^2$

$$S_{X,Y}(x_1, x_2)S_{X,Y}(y_1, y_2) \leq S_{X,Y}(x_1 \wedge y_1, x_2 \wedge y_2)S_{X,Y}(x_1 \vee y_1, x_2 \vee y_2). \tag{20}$$

Now the above inequality (20) can be shown by considering all possible twenty four cases namely $x_1 < x_2 < y_1 < y_2$, $x_1 < x_2 < y_2 < y_1$, and so on.

(b) To prove this, we need to show (19). It can be shown again by considering all possible twenty four cases as above.

(c) Using (a), (b), and Theorem 2.1 of Hu et al. (2003), the result follows.

The BWF has the following survival copula

$$\overline{C}(u, v) = \begin{cases} \left(\frac{\theta_1}{\theta_1+\theta_3}(u^{-1/\beta} - 1) + v^{-1/\beta}\right)^{-\beta} & \text{if} \quad (\theta_2 + \theta_3)(u^{-1/\beta} - 1) < (\theta_1 + \theta_3)(v^{-1/\beta} - 1) \\ \left(u^{-1/\beta} + \frac{\theta_2}{\theta_2+\theta_3}(v^{-1/\beta} - 1)\right)^{-\beta} & \text{if} \quad (\theta_2 + \theta_3)(u^{-1/\beta} - 1) > (\theta_1 + \theta_3)(v^{-1/\beta} - 1) \end{cases} \tag{21}$$

If $\theta_1 = \theta_2$ and if we denote $\delta = \dfrac{\theta_1}{\theta_1 + \theta_3} = \dfrac{\theta_2}{\theta_2 + \theta_3}$, then (21) becomes

$$\overline{C}(u, v) = \begin{cases} \left(\delta(u^{-1/\beta} - 1) + v^{-1/\beta}\right)^{-\beta} & \text{if} \quad u > v \\ \left(u^{-1/\beta} + \delta(v^{-1/\beta} - 1)\right)^{-\beta} & \text{if} \quad u < v \end{cases} \tag{22}$$

Hence, different dependency properties of BWF may be explored through copula function.

3. Inference

3.1. Modified EM algorithm

In this section we provide the maximum likelihood estimators of the unknown parameters based on the observations $D = \{(x_i, y_i); i = 1, \ldots, n\}$. Let us denote $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3$, where $\mathcal{D}_1 = \{(x_i, y_i) : x_i < y_i\}$,$\mathcal{D}_2 = \{(x_i, y_i) : x_i > y_i\}$, $\mathcal{D}_3 = \{(x_i, y_i) : x_i = y_i = z_i\}$ and the number of elements in the set $\mathcal{D}_j$ is n_j for $j = 1, 2, 3$. The log-likelihood function of the observed data D becomes

$$\begin{aligned} l(\Theta|D) &= (2n_1 + 2n_2 + n_3)\ln\alpha + n\ln\beta + (n_1 + n_2)(\ln(\beta+1) + \ln\theta_1 + \ln\theta_2) + n_3\ln\theta_3 + \\ &\quad (\alpha-1)\left(\sum_{i\in D_1\cup D_2}(\ln x_i + \ln y_i) + \sum_{i\in D_3}\ln z_i\right) - (\beta+1)\ln(1 + (\theta_1+\theta_2+\theta_3)z_i^\alpha) - \\ &\quad (\beta+2)\left[\sum_{i\in D_1}\ln(1 + \theta_1 x_i^\alpha + (\theta_2+\theta_3)y_i^\alpha) + \sum_{i\in D_1}\ln(1 + (\theta_1+\theta_3)x_i^\alpha + \theta_2 y_i^\alpha)\right], \quad (23) \end{aligned}$$

where $\Theta = (\alpha, \beta, \theta_1, \theta_2, \theta_3)$. It may be mentioned that to avoid introducing another notation, we sometime denote $\Theta = (\alpha, \beta, \lambda_1, \lambda_2, \lambda_3)$ and it should be clear from the context. Hence, the MLEs of the unknown parameters can be obtained by maximizing (23) with respect to unknown parameters. It involves a 5-dimensional optimization problem. To avoid that we treat this as a missing value problem. First observe that $\{(X, Y)|V\}$ has MOBW distribution, and there is a very effective EM algorithm has been proposed by Kundu and Dey (2009). It is assumed that the complete observations are coming from the following random vector $(X, Y, \Delta_1, \Delta_2, V)$. Here V is the frailty random variable, and (Δ_1, Δ_2) are defined as follows:

$$\Delta_1 = \begin{cases} 0 & \text{if } X = \frac{U_3}{V^{1/\alpha}} \\ 1 & \text{if } X = \frac{U_1}{V^{1/\alpha}} \end{cases} \qquad \Delta_2 = \begin{cases} 0 & \text{if } Y = \frac{U_3}{V^{1/\alpha}} \\ 2 & \text{if } Y = \frac{U_2}{V^{1/\alpha}} \end{cases}$$

Here U_1, U_2, U_3 are same as defined before. Note that

$$\begin{aligned} P(\Delta = 1, \Delta_2 = 0|X < Y) &= \frac{\theta_3}{\theta_2+\theta_3}, \quad P(\Delta = 1, \Delta_2 = 2|X < Y) = \frac{\theta_2}{\theta_2+\theta_3}, \\ P(\Delta = 0, \Delta_2 = 2|X > Y) &= \frac{\theta_3}{\theta_1+\theta_3}, \quad P(\Delta = 1, \Delta_2 = 2|X > Y) = \frac{\theta_1}{\theta_1+\theta_3}. \end{aligned}$$

It can be easily seen that if we have a complete observations $D_c = \{(x_i, y_, \delta_{1i}, \delta_{2i}, v_i); i = 1, \ldots, n\}$, the MLEs of the unknown parameters can be obtained by solving two one dimensional optimization problems. We use the following notations for further development. At the $k-th$ stage of the EM algorithm, the parameter vector the estimate of the parameter vector Θ will be denoted by $\Theta^{(k)} = (\alpha^{(k)}, \beta^{(k)}, \theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k)})$ and

$$a_1^{(k)} = \frac{\theta_3^{(k)}}{\theta_2^{(k)}+\theta_3^{(k)}}, \quad a_2 = \frac{\theta_2^{(k)}}{\theta_2^{(k)}+\theta_3^{(k)}}, \quad b_1 = \frac{\theta_3^{(k)}}{\theta_1^{(k)}+\theta_3^{(k)}}, \quad b_2 = \frac{\theta_1^{(k)}}{\theta_1^{(k)}+\theta_3^{(k)}},$$

$$c_{1i}^{(k)} = E(V_i|x_i = X < Y = y_i), \quad c_{2i}^{(k)} = E(V_i|x_i = X > Y = y_i), \quad c_{3i}^{(k)} = E(V_i|X = Y = z_i, \Theta^{(k)})$$

$$d_{1i}^{(k)} = E(\ln V_i|x_i = X < Y = y_i), \quad d_{2i}^{(k)} = E(\ln V_i|x_i = X > Y = y_i),$$

$$d_{3i}^{(k)} = E(\ln V_i | X = Y = z_i).$$

At the k-th stage of the EM algorithm, the pseudo log-likelihood contribution of (x_i, y_i) with out the constant for different cases are as follow;
Case 1: $x_i < y_1$

$$2\ln\alpha + \ln\lambda_1 + a_1^{(k)}\ln\lambda_3 + a_2^{(k)}\ln\lambda_2 + \alpha(\ln x_i + \ln y_i) - \lambda_1 c_{1i}^{(k)} x_i^{\alpha} - (\lambda_2+\lambda_3)c_{1i}^{(k)} y_i^{\alpha} + \beta\ln\beta - \ln(\Gamma(\beta)) + \beta(d_{1i}^{(k)} - c_{1i}^{(k)})$$

Case 2: $x_i < y_1$

$$2\ln\alpha + \ln\lambda_2 + b_1^{(k)}\ln\lambda_3 + b_2^{(k)}\ln\lambda_1 + \alpha(\ln x_i + \ln y_i) - (\lambda_1+\lambda_3)c_{2i}^{(k)} x_i^{\alpha} - \lambda_2 c_{2i}^{(k)} y_i^{\alpha} + \beta\ln\beta - \ln(\Gamma(\beta)) + \beta(d_{2i}^{(k)} - c_{2i}^{(k)})$$

Case 3: $x_i = y_1 = z_i$

$$\ln\alpha + \ln\lambda_3 + \alpha\ln z_i - (\lambda_1+\lambda_2+\lambda_3)c_{3i}^{(k)} z_i^{\alpha} + \beta\ln\beta - \ln(\Gamma(\beta)) + \beta(d_{3i}^{(k)} - c_{3i}^{(k)}).$$

Hence, the pseudo log-likelihood function at the $(k+1)-th$ stage can be written as

$$\begin{aligned} l_{pseudo}(\Theta|\Theta^{(k)}) &= (n_1 + n_2 b_2^{(k)})\ln\lambda_1 - \lambda_1\left(\sum_{i\in D_1} c_{1i}^{(k)} x_i^{\alpha} + \sum_{i\in D_2} c_{2i}^{(k)} x_i^{\alpha} + \sum_{i\in D_3} c_{3i}^{(k)} z_i^{\alpha}\right) + \\ &\quad (n_2 + n_1 a_2^{(k)})\ln\lambda_2 - \lambda_2\left(\sum_{i\in D_1} c_{1i}^{(k)} y_i^{\alpha} + \sum_{i\in D_2} c_{2i}^{(k)} y_i^{\alpha} + \sum_{i\in D_3} c_{3i}^{(k)} z_i^{\alpha}\right) + \\ &\quad (n_1 a_1^{(k)} + n_2 b_1^{(k)} + n_3)\ln\lambda_3 - \lambda_3\left(\sum_{i\in D_1} c_{1i}^{(k)} y_i^{\alpha} + \sum_{i\in D_2} c_{2i}^{(k)} x_i^{\alpha} + \sum_{i\in D_3} c_{3i}^{(k)} z_i^{\alpha}\right) + \\ &\quad (2n_1 + 2n_2 + n_3)\ln\alpha + \alpha\left(\sum_{i\in D_1\cup D_2}(\ln x_i + \ln y_i) + \sum_{i\in D_3}\ln z_i\right) + \\ &\quad n\beta\ln\beta - n\ln(\Gamma(\beta)) + \beta\left(\sum_{i\in D_1}(d_{1i}^{(k)} - c_{1i}^{(k)}) + \sum_{i\in D_2}(d_{2i}^{(k)} - c_{2i}^{(k)})\right. \\ &\quad \left. + \sum_{i\in D_3}(d_{3i}^{(k)} - c_{3i}^{(k)})\right). \end{aligned} \tag{24}$$

Now $\Theta^{(k+1)}$ can be obtained by maximizing (24) with respect Θ, and they are as follows. If we denote

$$\begin{aligned} \lambda_1^{(k+1)}(\alpha) &= \frac{(n_1 + n_2 b_2^{(k)})}{\sum_{i\in D_1} c_{1i}^{(k)} x_i^{\alpha} + \sum_{i\in D_2} c_{2i}^{(k)} x_i^{\alpha} + \sum_{i\in D_3} c_{3i}^{(k)} z_i^{\alpha}} \\ \lambda_2^{(k+1)}(\alpha) &= \frac{(n_2 + n_1 a_2^{(k)})}{\sum_{i\in D_1} c_{1i}^{(k)} y_i^{\alpha} + \sum_{i\in D_2} c_{2i}^{(k)} y_i^{\alpha} + \sum_{i\in D_3} c_{3i}^{(k)} z_i^{\alpha}} \\ \lambda_3^{(k+1)}(\alpha) &= \frac{(n_1 a_1^{(k)} + n_2 b_1^{(k)} + n_3)}{\sum_{i\in D_1} c_{1i}^{(k)} y_i^{\alpha} + \sum_{i\in D_2} c_{2i}^{(k)} x_i^{\alpha} + \sum_{i\in D_3} c_{3i}^{(k)} z_i^{\alpha}}, \end{aligned}$$

then first obtain $\alpha^{(k+1)}$ by maximizing $g(\alpha)$ with respect to α, where

$$g(\alpha) = (2n_1 + 2n_2 + n_3)\ln\alpha + \alpha\left(\sum_{i\in D_1\cup D_2}(\ln x_i + \ln y_i) + \sum_{i\in D_3}\ln z_i\right) + (n_1 + n_2 b_2^{(k)}) \quad (25)$$
$$\ln\lambda_1^{(k+1)}(\alpha) + (n_2 + n_1 a_2^{(k)})\ln\lambda_2^{(k+1)}(\alpha) + (n_1 a_1^{(k)} + n_2 b_1^{(k)} + n_3)\ln\lambda_3^{(k+1)}(\alpha).$$

Once $\alpha^{(k+1)}$ is obtained, then obtain $\lambda_1^{(k+1)}$, $\lambda_2^{(k+1)}$, $\lambda_3^{(k+1)}$ as $\lambda_1^{(k+1)}(\alpha^{(k+1)})$, $\lambda_2^{(k+1)}(\alpha^{(k+1)})$ and $\lambda_3^{(k+1)}(\alpha^{(k+1)})$, respectively. Obtain $\beta^{(k+1)}$ by maximizing $h(\beta)$ with respect to β, where

$$h(\beta) = n\beta\ln\beta - n\ln(\Gamma(\beta)) + \beta\left(\sum_{i\in D_1}(d_{1i}^{(k)} - c_{1i}^{(k)}) + \sum_{i\in D_2}(d_{2i}^{(k)} - c_{2i}^{(k)}) + \sum_{i\in D_3}(d_{3i}^{(k)} - c_{3i}^{(k)})\right). \quad (26)$$

The following algorithm can be used for that purpose.

Algorithm:

Step 1: Take some initial values of $\Theta = (\alpha, \beta, \lambda_1, \lambda_2, \lambda_3)$, say $\Theta^{(0)} = (\alpha^{(0)}, \beta^{(0)}, \lambda_1^{(0)}, \lambda_2^{(0)}, \lambda_3^{(0)})$.

Step 2: Based on $\Theta^{(0)}$, compute $a_1^{(0)}, a_2^{(0)}, b_1^{(0)}, b_2^{(0)}$, $\{(c_{1i}^{(0)}, c_{2i}^{(0)}, c_{3i}^{(0)}, d_{1i}^{(0)}, d_{2i}^{(0)}, d_{3i}^{(0)}); i = 1, \ldots, n\}$.

Step 3: Obtain $\alpha^{(1)}$ by maximizing (**??**), obtain $\lambda_1^{(1)}, \lambda_2^{(1)}, \lambda_3^{(1)}$ and obtain $\beta^{(1)}$ by maximizing (26).

Step 4: Go back to Step 1 and replace '0' by '1' and continue the process until convergence takes place.

3.2. Testing of hypothesis

In this section we want to discuss the following testing of hypothesis problem which has some practical importance. We want to test the following hypothesis

$$H_0 : \lambda_1 = \lambda_2 \quad vs. \quad H_1 : \lambda_1 \neq \lambda_2. \quad (27)$$

It manly tests the equality of the two marginals. To test the hypothesis (27), we propose to use the likelihood ratio test. Hence, we need to maximize the log-likelihood function (23) under H_0. In this case also the modified EM algorithm can be used quite effectively with the necessary changes. If we denote the estimate without restriction of the unknown parameter vector Θ as $\widehat{\Theta}$ and by $\widehat{\Theta}_0$ with restriction, then we reject the null hypothesis if $-2(l_0(\widehat{\Theta}_0|D) - l(\widehat{\Theta}|D))$ is greater than the appropriate upper percentage of χ_1^2 value.

4. Data analysis

In this section we analyze one bivariate data set to show how the proposed model and the method can be used in practice. The proposed model has been used on a diabetic retinopathy data set. Diabetic retinopathy is a physical disorder of the eye, and it is observed mainly among the diabetic patients. This particular disease leads to blindness. An

extensive amount of work has been done in developing different treatment for this disease. Among different methods one recent treatment is the laser treatment. The main aim of this experiment is to test whether the laser treatment has significant different effect compared to the traditional treatment in delaying the onset of blindness. The experiment has been conducted as follows. For each patient one eye has been chosen at random and the laser treatment has been given where in the other eye the traditional treatment has been administered. For each patient the time to onset of blindness of both the eyes have been recorded. Here X and Y denote the times for the laser treated eye and the traditionally treated eye, respectively. The data set has been presented in Table 1.

Table 1: Bivariate diabetic retinopathy data set. Here Y_1 denotes the time to blindness of the laser treated eye and Y_2 denotes the same for the other eye

Sl. No.	X	Y	Sl. No.	X	Y	Sl. No.	X	Y	Sl. No.	X	Y
1.	20.17	6.90	2.	10.27	1.63	3.	5.67	13.83	4.	5.77	1.33
5.	5.90	35.53	6.	25.63	21.90	7.	33.90	14.80	8.	1.73	6.20
9.	30.20	22.00	10.	25.80	13.87	11.	5.73	48.30	12.	9.90	9.90
13.	1.73	1.73	14.	1.77	43.03	15.	8.30	8.30	16.	18.70	6.53
17.	42.17	42.17	18.	14.30	48.43	19.	13.33	9.60	20.	14.27	7.60
21.	34.57	1.80	22.	4.10	12.20	23.	21.57	9.90	24.	13.77	13.77
25.	33.63	33.63	26.	63.33	27.60	27.	38.47	1.63	28.	10.33	0.83
29.	13.83	1.57	30.	11.07	1.97	31.	2.10	11.30	32.	12.93	4.97
33.	24.43	9.87	34.	13.97	30.40	35.	6.30	56.97	36.	13.80	19.00
37.	13.57	5.43	38.	42.77	42.77	39.	42.43	46.63	40.	2.70	2.70

For this data set $n = 38$, and 12, 20 and 6 observations for which $X < Y$, $X > Y$ and $X = Y$, respectively. Based on this data set we want to test whether laser treatment has any significant different effect compared to the traditional treatment or not. Before progressing further we provide some of the basic statistics of the data set. We have provided the median, Q1 (first quartile) and Q3 (third quartile) of X, Y and $\min\{X, Y\}$. We have also fitted UWF models to X, Y and $\min\{X, Y\}$, their fitted parameter values, the Kolmogorv-Smirnov distances (KSD) and the associated p-values have been presented in Table 2. Based on the KSD and the associated p-values, it is clear that UWF model fits X, Y and $\min\{X, Y\}$. We have plotted the empirical survival curves of the time to blindness of both the eyes in Figure 3. We have also plotted the best fitted UWF models to the marginals in Figure 4. They fit loots quite quite well in both the cases. Hence, it is reasonable to try to fit BWF model to the bivariate data set (X, Y).

We would like to compute the MLEs of the fitted BWF model using the EM algorithm as proposed in Section 3. Based on the fitted UWF models to the marginals and the minimum, we have obtained the following initial guesses; $\alpha^{(0)} = 1.2597$, $\beta^{(0)} = 4.8359$, $\lambda_1^{(0)} = 0.0016$, $\lambda_2^{(0)} = 0.0089$ and $\lambda_3^{(0)} = 0.0110$. We have stopped the EM algorithm when the relative difference between to consecutive log-likelihood functions is less than $\epsilon = 10^{-6}$. The iteration stops after 12 steps. The MLEs of the unknown parameters and the associated 95% confidence intervals are also obtained from the las t step of the EM algorithm based on the method of Louis (1982). They are as follows: $\widehat{\alpha} = 1.2077(\mp 0.3765)$, $\widehat{\beta} = 11.4772(\mp 2.6754)$,

Table 2: Some basic statistics of the bivariate retinopathy data set. The median, Q1 and Q3 of the marginals and their minimum. The estimated parameters of the UWF models, their respective K-S distances and the associated p-values have been presented

Variable	Median	Q1	Q3	α	β	θ	KSD	p-Value
X	13.82	13.35	25.03	1.4698	5.3205	0.0026	0.1278	0.5640
Y	10.60	5.20	21.95	1.1516	4.5085	0.0099	0.1186	0.6581
$\min\{X, Y\}$	7.25	1.88	13.83	1.1577	4.6788	0.0115	0.1365	0.4787

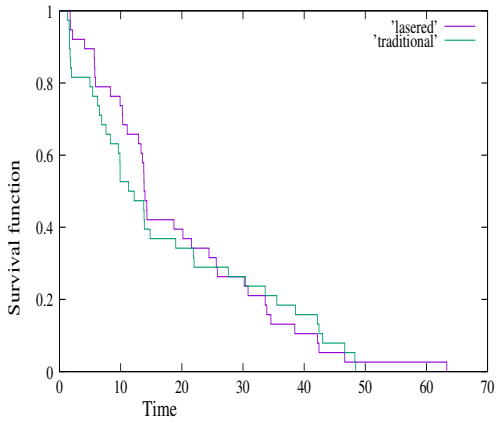


Figure 3: Empirical survival functions of the time to blindness of the two eyes

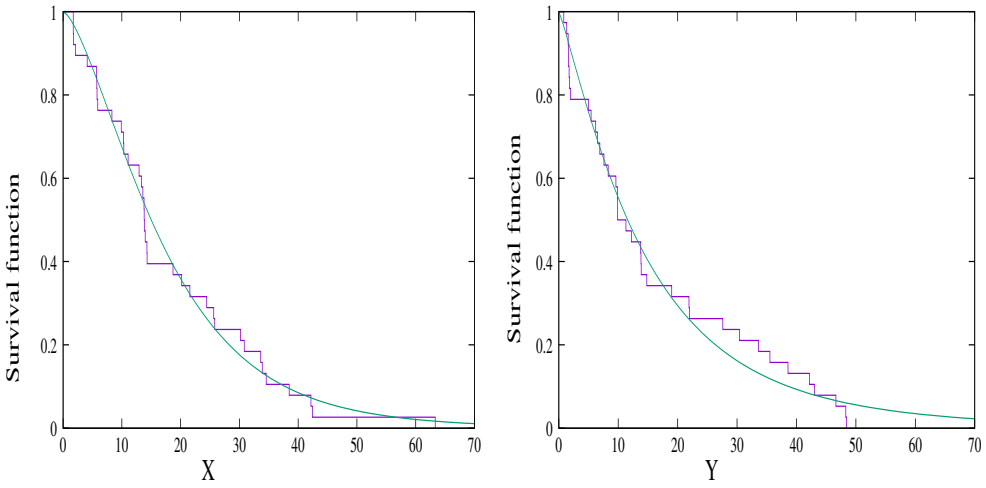
$\widehat{\lambda}_1 = 0.00220(\mp 00034)$, $\widehat{\lambda}_2 = 0.00245(\mp 0.00037)$, $\widehat{\lambda}_3 = 0.00031(\mp 0.00001)$. The corresponding log-likelihood value is -292.9115.

One natural question is whether the proposed BWF distribution provides a good fit to the above bivariate data. For that purpose we consider the following statistic

$$D = \sup_{x,y} |S_n(x, y) - \widehat{S}_{BWF}(x, y)|.$$

Here $S_n(x, y)$ denotes the empirical survival function, *i.e.* $S_n(x, y) = \dfrac{\#\{i; x_i \geq x, y_i \geq y\}}{n}$ and $\widehat{S}_{BWF}(x, y)$ denotes the estimated value of $S_{BWF}(x, y)$ based on MLEs. We obtain the $D = 0.1134$ and based on simulation we obtain the associated p value as 0.681. Hence, it shows that the proposed BWF provides a good fit to the bivariate Retinopathy data set.

Now we want to test whether the laser treatment has any significant effect in delaying the blindness or not. It is equivalent in testing (27). We have obtained the MLEs and the associated 95% confidence intervals of the unknown parameters under the null hypothesis as follows: $\widehat{\alpha}_0 = 1.2041(\mp 0.3668)$, $\widehat{\beta}_0 = 13.2188(\mp 2.5753)$, $\widehat{\lambda}_{10} = 0.00202(\mp 00032)$, $\widehat{\lambda}_{20} = 0.00202(\mp 0.00032)$, $\widehat{\lambda}_{30} = 0.00027(\mp 0.00001)$. The corresponding log-likelihood value is -292.9893. Hence, based on $-2(l_0(\widehat{\Theta}_0|D) - l(\widehat{\Theta}|D))$, we cannot reject the null hypothesis. Therefore, the present data indicates that there is no significant difference between the laser treatment and the traditional treatment.



(a) (b)

Figure 4: Empirical and fitted survival functions of the time to blindness of the two eyes (a) laser treated eye and (b) traditionally treated eye

5. Application: Competing risks

In many life testing experiment the failure might occur due to different causes. In this type of experiment one observes the failure time as well as the cause of failure. It is important to study the effect of one cause in presence of other causes. In the statistical literature it is known as the competing risks problem. There are mainly two approaches to analyze competing risks data; one is latent failure type approach of Cox (1959) and the other one is cause specific hazard function approach by Prentice et al. (1978). In case of exponential or Weibull failure time distributions it has been shown by Kundu (2004) that both the methods provide the same likelihood function, although their interpretations are different. An extensive amount of work has been done in the area of competing risks both in the parametric and non-parametric set up. One is referred to the book by Crowder (2001) for a comprehensive treatment on this topic.

A typical competing risk data is of the form (T, Δ), here T denotes the observed failure time and Δ denotes the cause of failure. The failure time T is usually assumed to be continuous, where as Δ is a discrete random variable. In this paper we have assumed Cox's latent failure approach to analyze the competing risks data. In this case it is assumed that there are total K causes of failures, *i.e.* Δ can take values $1, \ldots, K$. Further, there are K lifetimes say $T_1, \ldots, T_K$ due to K different causes, where $T = \min\{T_1, \ldots, T_K\}$ and $\Delta = j$ if $T_j < \{T_1, \ldots, T_{j-1}, T_{j+1}, \ldots, T_K\}$.

We have another data set from a diabetic retinopathy study has been obtained as before. The experiment was same as before, but here the minimum time to blindness (T) and the indicator specifying whether the laser treated eye ($\Delta = 1$), the traditionally treated eye ($\Delta = 2$) or both the eyes ($\Delta = 3$) have failed simultaneously have been recorded. The data set has been presented in Table 3.

Table 3: Competing risks diabetic retinopathy data set. Here T denotes the minimum time to blindness in days and Δ denotes its causes

T	Δ	T	Δ	T	Δ	T	Δ	T	Δ	T	Δ
266	1	272	3	203	3	91	2	1137	3	84	1
154	2	1484	1	392	1	285	3	315	1	1140	2
583	1	287	2	901	1	547	2	1252	1	1247	3
79	1	717	2	448	2	622	3	642	1	904	2
707	2	141	2	276	1	469	2	407	1	520	1
93	1	356	1	485	2	1313	2	1653	3	248	2
805	1	427	2	503	1	344	1	699	1	423	2
790	2	36	2	285	2	125	2	667	1	315	2
777	2	588	2	727	2	306	1	471	3	210	2
415	1	126	1	409	2	307	2	350	2	584	1
637	2	350	1	355	1	577	2	663	3	1302	1
178	1	567	2	227	2	517	2	966	3		

The problem is same as before, *i.e.* we want to test whether there is any significant between the laser treatment and the traditional treatment in delaying the onset of blindness of the affected eyes. We treat this data as a competing risks data where the two treatments can be considered as the two different causes of failures. Here $T = \min\{T_1, T_2\}$, where T_1 (T_2) denotes the lifetime of the laser (traditionally) treated eye, and $\Delta = 1$, if $T_1 < T_2$, $\Delta = 2$, if $T_1 > T_2$ and $\Delta = 3$ if $T_1 = T_2$. Here both T_1 and T_2 are continuous random variables, but there is a positive probability that $T_1 = T_2$. We have assumed that $(T_1, T_2) \sim$ BWF$(\alpha, \beta, \lambda_1, \lambda_2, \lambda_3)$. Now based on the observations $D = \{(t_i, \delta_i); i = 1, \ldots, n\}$, the log-likelihood function can be written as

$$\begin{aligned} l(\Theta|D) &= n\log\beta + n\ln\alpha + n_1\ln\theta_1 + n_2\ln\theta_2 + n_3\ln\theta_3 + (\alpha-1)\sum_{i=1}^{n}\ln t_i - \\ &\quad (\beta+1)\sum_{i=1}^{n}\ln(1+(\theta_1+\theta_2+\theta_3)t_i^{\alpha}). \end{aligned} \tag{28}$$

Here n_1, n_2 and n_3 denote the number of $\delta_i = 1$, $\delta_i = 2$ and $\delta_i = 3$, respectively. The MLEs of the unknown parameters can be obtained by maximizing (28) with respect to the unknown parameters. The MLEs of the unknown parameters and the associated 95% confidence intervals are: $\widehat{\alpha} = 6.6043(\mp 1.1967)$, $\widehat{\beta} = 0.0242(\mp 0.0061)$, $\widehat{\theta}_1 = 0.5984(\mp 0.1534)$, $\widehat{\theta}_2 = 0.8245(\mp 0.2278)$ and $\widehat{\theta}_3 = 3.0644(\mp 1.0014)$. The corresponding log-likelihood value is -5191.931. We want to test whether there is any significant difference between the laser treatment and the traditional treatment and it is equivalent to test (27). Under the null hypothesis the estimates and the associated 95% confidence intervals are: $\widehat{\alpha}_0 = 6.6113(\mp 1.1913)$, $\widehat{\beta}_0 = 0.0239(\mp 0.0058)$, $\widehat{\theta}_{10} = 0.8154(\mp 0.2256)$, $\widehat{\theta}_{20} = 0.8154(\mp 0.2256)$ and $\widehat{\theta}_{30} = 3.0657(\mp 0.9976)$. The corresponding log-likelihood value is -5203.363. Based on the test statistic $-2(l_0(\widehat{\Theta}_0|D) - l(\widehat{\Theta}|D))$ we reject the null hypothesis. Hence, in this case the laser treatment has a significant different effect than the traditional treatment.

6. Conclusions

In this paper we have proposed a bivariate Weibull frailty model which is a singular distribution. The proposed model has five parameters and the joint PDF can take variety

of shapes. We have derived different properties and developed the classical inference of the unknown parameters. We have used this model to analyze a dependent competing risks data. Although we have developed the classical inference, we have not developed any Bayesian inference of the unknown parameters. It will be interesting to develop the Bayesian inference of the unknown parameters for this model. More work is needed in that direction.

Conflict of interest

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

References

Cai, J., Shi, Y., and Liu, B. (2017). Analysis of incomplete data in the presence of dependent competing risks from Marshall-Olkin bivariate Weibull distribution under progressive hybrid censoring. *Communications in Statistics - Theory and Methods*, **46**, 6497–6511.

Cox, D. R. (1959). The analysis of exponentially distributed lifetimes with two types of failures. *Journal of the Royal Statistical Society Series B*, **21**, 411 – 421.

Crowder, M. J. (2001). *Classical Competing Risks*, Chapman & Hall, Boca Raton, Florida.

Feizjavadian, S. H. and Hashemi, R. (2015). Analysis of dependent competing risks in the presence of progressive hybrid censoring using Marshall-Olkin bivariate Weibull distribution. *Computational Statistics and Data Analysis*, **82**, 19–34.

Hu, T., Khaledi, B-E, and Shaked, M. (2003). Multivariate hazard rate. *Journal of Multivariate Analysis*, **84**, 173 –189.

Johnson, N. L. and Kotz, S. (1975). A vector multivariate hazard rate. *Journal of Multivariate Analysis*, **5**, 53–66.

Kundu, D. (2004). Parameter Estimation of the Partially Complete Time and Type of Failure Data. *Biometrical Journal*, **46**, 165–179.

Kundu, D. (2022). Bivariate semi-parametric singular family of distributions and its applications. *Sankhya*, Ser B, **84**, Part 2, 846 – 872.

Kundu, D. (2023). Bivariate distributions with singular components. *Springer Handbook of Engineering Statistics*, Editor: Hoang Pham, Springer, pp 733 – 761.

Kundu, D. and Dey, A. K. (2009). Estimating the Parameters of the Marshall Olkin Bivariate Weibull Distribution by EM Algorithm. *Computational Statistics and Data Analysis*, **53**, 956 – 965.

Kundu, D. and Gupta, A. K. (2013). Bayes estimation for the Marshall-Olkin bivariate Weibull distribution. *Computational Statistics and Data Analysis*, **57**, 271 – 281.

Kundu, D. and Gupta, R. D. (2009). Bivariate generalized exponential distribution. *Journal of Multivariate Analysis*, **100**, 581 –593.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society*, Series B, **44**, 226–233.

Prentice, R. L., Kalbfleish, J. D., Peterson, Jr. A. V., Flurnoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in presence of competing risks. *Biometrics*, **34**, 541 – 554.

Samanta, D. and Kundu, D. (2023). Bivariate semi-parametric model: Bayesian inference. *Methodology and Computing in Applied Probability*, **25**, Article no. 87.

Shen, Y. and Xu, A. (2018). On the dependent competing risks using Marshall-Olkin bivariate Weibull model: Parameter estimation with different methods. *Communications in Statistics - Theory and Methods*, **47**, 5558–5572.

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.

Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025; pp 31-41

Fondly Remembering Two Interesting Collaborations

A. Goswami
School of Mathematical and Computational Science
Indian Association for the Cultivation of Science, Jadavpur, Kolkata, India

Received: 07 June 2025; Revised: 17 June 2025; Accepted: 19 June 2025

Abstract

This article is based on the talk delivered by the author at the 27th Annual Conference of the Society of Statistics, Computer and Applicationns, held at NEHU, Shillong in February 2025. The talk was the fourth in the prestigeous Bikas K. Sinha Endowment Lecture Series. It is a fond recollection of the exciting journey of active collaboration with Professor Sinha on two separate occasions. These were on problems being investigated by Professor Sinha and the author was called on by him and was presented with the wonderful opportunity of collaborating with him.

Key words: Biodiversity; Species richness and species abundance; Time till full discovery; Social network; Measure of reciprocity; Sequential sampling stategy.

AMS Subject Classifications: 60G40, 62L12.

1. Introduction

1.1. A prologue from the author

It was around 1979-80 that I first met Professor Bikas K. Sinha or "*Bikas-da*" (as Professor Bikas K. Sinha is most popularly known to almost all of his younger colleagues, friends and also a large majority of his ex-students). I was a first year research scholar (Ph.D. student) then in the Stat-Math division at I.S.I., Kolkata, and attended, along with a few others, a research course on Optimal Designs taught by Bikas-da. Had I not left for USA next year to pursue my PhD, there was a distinct possibility that I could have earned the honour of being the first PhD student of Bikasda in ISI. Anyway, it was a little over eight years later when I retuned to ISI to join as a faculty in Stat-Math, that my contact with Bikasda had a restart. And, from then on till today, Bikasda remained my friend, my elder brother, my long time colleague. What I am going to discuss in this article are two interesting problems that we collaborated on. These were problems that originally Bikasda was investigating and at some point, he decided to discuss those with me and get me on board. This is another fascinating thing about Bikasda that when he discusses a problem with someone else, he can instill genuine interest in that person about the problem and, more

Corresponding Author: A.Goswami
Email: alok.goswami@iacs.res.in

often than not, it culminates into a productive collaboration. In fact, I do not personally know of anyone who has had as many collaborators as Bikasda has worked with, and that too from widely different fields of expertise, from pure mathematics to applied statistics, from sociology to biology. After having seen and known Bikasda for a very long time, first as his student and then as his younger colleague, I can only repeat what anyone and everyone, who has come in touch with him in any capacity, will say. He is one of the finest human beings that I have known, always jovial, always in high spirits, always with that reassuring smile on his face that makes a very bad day seem not so bad. During our joint tenure in ISI for a long period of time, there have been many occasions when Bikasda and I travelled together to multiple places to participate in a variety of academic programmes. Each of these occasions presented a wonderful opportunity for me to know Bikasda more closely, outside the framework of our relationship in ISI. They just reinforced my impression of him as an amazingly fascinating person with an unstoppable liveliness that always spread like an infection. Of all these journeys, there was one that deserves a special mention. Sometime during 2017, one day Bikasda asked me casually (at least that's how it seemed to me) '*ki bhaiya, bangladesh jabe naki?*', meaning "brother, interested in a trip to Bangladesh?" I was, of course, very excited and replied with an emphatic 'Yes'. Sure enough, a trip was organized for September that year. The two of us were invited to deliver lectures, first at Rajshahi University for 3-4 days and then at East West University in Dhaka for another 3-4 days. The entire trip was wonderful. The exciting academic interactions with young masters' and PhD students, the fascinating traditional Bangladeshi hospitality, the lazy evenings spent strolling along the serene banks of the river *Padma*, stopping occasionally to enjoy a sip of '*Morich Cha*' (finely brewed tea garnished with a pod of split green pepper to make it just a bit fiery) — all of these made it one of the most enjoyable trips of my life. But the part of the trip that will always remain etched in my memory, happened during our road trip from Rajshahi to Dhaka. We took a slight detour to visit Bikasda's ancestral village. Riding through that village with Bikasda recalling his childhood memories, lots and lots of magical stories about the young twin brothers$\cdots\cdots$I really had a glimpse into a different side of Bikasda that not too many persons I know had perhaps seen$\cdots\cdots$visualizing the twin brothers strolling and running through the muddy roads of the village, occasionally indulging in small mischiefs for fun, that is typical of that age$\cdots\cdots$it was a total surprise that I didn't expect. And Bikasda being Bikasda hadn't informed me in advance of this plan. We finally stopped at his ancestral home where members of their family who stayed back still live. Needless to say we were offered an elaborate lunch, with multiple fish dishes in abundance, before we hit the road again on our way to Dhaka. The detour turned out to be an absolutely memorable surprise indeed! Thank you Bikasda for gifting me this magical experience!!!

1.2. Two interesting problems

The first problem we discuss here falls in the domain of biodiversity analysis. Two quantities that are regarded as central to the measurement of biodiversity are *species richness* and *species abundance*. In the context of developing appropriate sampling strategies to gain understanding of species richness and species evenness, a conjecture, supported also by empirical observations, that has inherently played a crucial role, is that the species evenness distribution which allows for a minimal sample size is the one, in which, for a fixed size of species richness, the abundance rates are all equal [see Gore and Paranjape (1997), (2001)].

This was our focus of study in one of the two collaborations mentioned above and is discussed in greater detail in Section 2. The main objective of our work was to try and provide an analytical proof of thisd conjecture and, luckily, we succeeded. For a fixed species richness size, we considered the distribution of "effort size" for full discovery, as a function of the underlying species abundance rate vector, and showed analytically that, it is stochastiaclly smallest when all the species are "as equally abundant as possible", a phrase that is going to be clarified in Section 2. The analysis is done separately for the two cases, namely, for infinite population and for finite population. The mathematical formulation of the problem and sketches of the proof of the main results are given in Section 2. The details are skipped here, since complete details are available in Goswami and Sinha (2006).

The other problem discussed in this article is in the context of social networks. Concept of *Reciprocity* in a social network is recognized as an important characteristic for study in order to gain understanding of the network. Rao and Bandyopadhayay (1987) suggested a simple natural measure of Reciprocity. Our focus in this work of collaboration was the choice of optimal sampling strategies for unbiased estimation of the above measure of reciprocity. Being a diehard champion of optimality, Bikasda had a firm conviction that, in this particular context, with some of its inherent special characteristics, the standard (SRSWOR, sample mean) strategy cannot be the last word. The strength of our collaboration, with Professor S. Sengupta from Calcutta University also joining the team, was rooted in that conviction. Finally (and despite challenges thrown by some rather complicated algebra), our work was able to illustrate that certain naturally suggested sequential strategies using specific selection rules coupled with appropriately defined unbiased estimators, fare better than the usual (SRSWOR, sample mean) strategy. Outlines and some crucial steps behind the core ideas are discussed in some detail in Section 3. For complete details and more (including numerical illustrations), the reader may see Goswami, Sinha and Sengupta (1990).

2. Analyzing time till discovery of all species

We consider a population consisting of m different species. The number m, called the *species richness*, is assumed to be fixed in this analysis. Clearly, only $m \geq 2$ is of any interest. The variable parameter in our analysis is going to be the *abundance rate vector*

$$\mathbf{p} = (p_1, \cdots, p_m), \tag{1}$$

where p_i, the abundance rate of the ith species, is its proportion in the population (equivalently, the probability that a randomly drawn unit from the population is from ith species).

In case of a finite population, say, of known size N, the admissible abundance rate vectors will have to be necessarily of the form

$$\mathbf{p} = \left(\tfrac{N_1}{N}, \cdots, \tfrac{N_m}{N}\right), \tag{2}$$

where $N_1, \cdots, N_m$ are positive integers adding up to N, while for an infinite population, any set of numbers $p_i \in (0, 1),\ i = 1, \cdots, m$, that add up to 1, will constitute a possible abundance rate vector $\mathbf{p} = (p_1, \cdots, p_m)$.

The focus and aim of our investigation here are as follows. Suppose we keep on drawing units at random from the population until all the m species are "discovered". Denoting T

to be the number of draws needed, it is clear that T is a random variable whose distribution depends on the abundance rate vector $\mathbf{p}$, and also on the sampling scheme − specifically, in case of a finite population, whether the sampling is done with or without replacement. We state the main result that we are able to prove.

Main result: *In both cases of random sampling from an infinite population as well as random sampling, with or without replacement, from a finite population, the random variable T is "stochastically smallest" when all the m species are "almost" equally abundant.*

In case of any concern about the phrase "almost" equally abundant used above, we want to point out that this is relevant only in finite population case and will be elaborated in Section 2.2. We proceed now to present an outline of the steps through which the above result is proved. The cases of infinite population and finite population need to be handled separately and that is what we do. But at the start, let us state a simple result which will be used frequently in the sequel. Noting that relabelling the m species among themselves does not have any impact on the time T till full discovery, the following result is obvious.

Theorem 1: The distribution of T, under an abundance rate vector $\mathbf{p}$, remains invariant over permutations of the coordinates of the vector $\mathbf{p}$.

2.1. Infinite population

In this case, we do not need to distinguish between sampling with or without replacement. As mentioned before, any choice of numbers $p_i \in (0,1)$, $i = 1, \cdots, m$, with $\sum p_i = 1$, will constitute a possible abundance rate vector $\mathbf{p} = (p_1, \cdots, p_m)$. Let us denote

$$\Phi(t, m, \mathbf{p}) = P(T > t \,\big|\, m, \mathbf{p}), \text{ for each } t. \tag{3}$$

The abundance rate vector capturing the "equally abundant" case is denoted $\mathbf{p}_0$, that is. $\mathbf{p}_0 = (\frac{1}{m}, \cdots, \frac{1}{m})$. With these notations, here is our main result.

Theorem 2: For an infinite population with m species, where $m \geq 2$,

$$\Phi(t, m, \mathbf{p}) \geq \Phi(t, m, \mathbf{p}_0), \quad \text{for all } t \geq m, \tag{4}$$

with the inequality in (4) being strict, for every $t \geq m$, unless $\mathbf{p} = \mathbf{p}_0$.

Proof: (outline) The case $m = 2$ is fairly trivial. For any $\mathbf{p} = (p_1, p_2)$, one has $\Phi(t, 2, \mathbf{p}) = p_1^t + p_2^t$, for all $t \geq 2$, and the right-hand-side, for each $t \geq 2$, can easily be shown to have a unique minimum at $p_1 = p_2 = \frac{1}{2}$, subject to the conditions that $p_1 > 0, p_2 > 0$, $p_1 + p_2 = 1$.

A natural idea now is to complete the proof by induction, but a first step towards that would be to get some relation between $\phi(\cdot, m, \cdot)$ and $\phi(\cdot, m-1, \cdot)$. To do this, we need a notation. For $m > 2$ and for any abundance rate vector $\mathbf{p} = (p_1, \cdots, p_m)$, let us denote $\mathbf{p}^{(i)}$, for $1 \leq i \leq m$, to be the abundance rate vector of size $m - 1$, obtained by removing the ith coordinate from $\mathbf{p}$ and normalizing the remaining coordinates, that is, $\mathbf{p}^{(i)} = \left(\frac{p_1}{1-p_i}, \cdots, \frac{p_{i-1}}{1-p_i}, \frac{p_{i+1}}{1-p_i}, \cdots, \frac{p_m}{1-p_i}\right)$. By a fairly straightforward conditioning argument, one can now show that, for any $m > 2$, any $\mathbf{p} = (p_1, \cdots, p_m)$ and any $i \in \{1, \ldots, m\}$,

$$\Phi(t, m, \mathbf{p}) = b(t, p_i; 0) + \sum_{s \geq t-m+2} b(t, p_i; s) + \sum_{s=1}^{t-m+1} b(t, p_i; s)\, \Phi(t-s, m-1, \mathbf{p}^{(i)}), \quad \text{for all } t \geq m, \tag{5}$$

where $b(t, p_i; s)$, $0 \le s \le t$, are the p.m.f. of the Binomial distribution with parameters (t, p_i).

This paves the way for invoking induction. Assuming the assertion (4) to hold for $m-1$ will mean that $\Phi(t', m-1, \mathbf{p}^{(i)}) \ge \Phi\Big(t', m-1, (\frac{1}{m-1}, \cdots, \frac{1}{m-1})\Big)$, for all $t' \ge m-1$, with strict inequality unless $\mathbf{p}^{(i)} = \Big(\frac{1}{m-1}, \cdots, \frac{1}{m-1}\Big)$. Using this in (5) and recombining the terms, one easily gets that, for all $m \ge 3$, all $\mathbf{p} = (p_1, \cdots, p_m)$ and all $i \in \{1, \cdots, m\}$,

$$\Phi(t, m, \mathbf{p}) \ge \Phi\Big(t, m, (\, \tfrac{1-p_i}{m-1}, \cdots, \tfrac{1-p_i}{m-1},\, p_i,\, \tfrac{1-p_i}{m-1}, \cdots, \tfrac{1-p_i}{m-1}\,)\Big), \quad \text{for all } t \ge m, \tag{6}$$

with strict inequality unless the $p_j, j \ne i$ are all equal.

Now, if $\mathbf{p} \ne \mathbf{p}_0$, then there must exist i such that $p_j, j \ne i$ are not all equal. Recalling that $\Phi(t, m, \mathbf{p})$ is invariant under permutations of coordinates of $\mathbf{p}$, we can, without loss of generality, assume $i = 1$. Writing $p_1 = p$, we then have

$$\Phi(t, m, \mathbf{p}) > \Phi\Big(t, m, (p,\, \tfrac{1-p}{m-1}, \cdots, \tfrac{1-p}{m-1}\,)\Big), \quad \text{for all } t \ge m, \tag{7}$$

Denoting the right-hand-side as $g(p)$, our final job is to show that the function g on $(0, 1)$ has a unique global minimum at $p = 1/m$ (for all $t \ge m$!). Having thus reduced the problem to one of ascertaining the global minimum of a function of a single variable, one would like to believe that the job should now be easily completed through simple calculus. Unfortunately, that is not to be! The second derivative g'' seems intractable. Overcoming this hurdle required a completely different idea. Denoting $f(p) = \frac{1-p}{m-1}$, for $p \in (0, 1)$, the inequality (6) can be paraphrased as saying that, for all $m \ge 3$, all $\mathbf{p} = (p_1, \cdots, p_m)$ and all $i \in \{1, \cdots, m\}$,

$$\Phi(t, m, \mathbf{p}) \ge \Phi\Big(t, m, (f(p_i), \cdots, f(p_i),\, p_i,\, f(p_i), \cdots, f(p_i)\,)\Big), \quad \text{for all } t \ge m, \tag{8}$$

with strict inequality unless the $p_j, j \ne i$ are all equal. Now, if $\mathbf{p} \ne \mathbf{p}_0$, then, as argued above, we may assume that $p_j, j \ne 1$ are not all equal, so that, by denoting $p_1 = p$, we get

$$\Phi(t, m, \mathbf{p}) > \Phi\Big(t, m, (p, f(p), \cdots, f(p))\Big), \quad \text{for all } t \ge m. \tag{9}$$

Now applying inequality (8) with $i = 2$ on the right-hand-side of (9) and then using invariance of Φ under permutation of coordinates of abundance rate vector will give

$$\Phi\Big(t, m, (p, f(p), \cdots, f(p))\Big) \ge \Phi\Big(t, m, (f(p), f(f(p)), \cdots, f(f(p)))\Big), \quad \text{for all } t \ge m. \tag{10}$$

Repeating this and denoting $f^{(n)}$ to be the nth iterate of f, one gets the string of inequaities

$$\begin{aligned}\Phi\Big(t, m, (p, f(p), \cdots, f(p))\Big) &\ge \Phi\Big(t, m, (f(p), f(f(p)), \cdots, f(f(p)))\Big) \\ &\ge \cdots\ \cdots\ \cdots\ \cdots \\ &\ge \Phi\Big(t, m, (f^{(n-1)}(p), f^{(n)}(p), \cdots, f^{(n)}(p))\Big) \\ &\ge \cdots\ \cdots\ \cdots\ \cdots\end{aligned}$$

The job is now completed by invoking the Lemma stated below, the proof of which is fairly easy and hence omitted. Using the Lemma and the fact that $\Phi\Big(t, m, (p_1, \cdots, p_m)\Big)$, for every $t \ge m$, is continuous in $p_1, \cdots, p_m$ (an easy exercise!), one can easily put together (9) and the above string of inequalities to complete the proof of assertion (4). □

Lemma 1: The function $f(p) = \frac{1-p}{m-1}$, $p \in [0, 1]$, has a unique fixed point at $p_0 = \frac{1}{m}$. Further, p_0 is a globally attracting fixed point in the sense that $\lim_n f^{(n)}(p) = p_0$, for all $p \in [0, 1]$.

2.2. Finite population

Handling the finite population case follows the same central ideas as in the infinite population case, but, not surprisingly, the actual execution gets somewhat complicated. Recall that in case the population size is finite, say, N, the abundance rate vectors are necessarily of the form $\mathbf{p} = \left(\frac{N_1}{N}, \cdots, \frac{N_m}{N}\right)$, where the N_i are positive integers adding up to N. If we assume N to be fixed, then the actual "abundance vector" $\mathbf{N} = (N_1, \cdots, N_m)$ can be regarded as an equivalent representation of the underlying parameter, in place of the rate vector $\mathbf{p}$. We will follow this viewpoint and formulate everything in terms of the abundance vector $\mathbf{N}$ as the parameter. Using notations similar to the infinite population case, we denote $\Phi(t, N, m, \mathbf{N})$ to be the probability that for a population of size N with abundance vector $\mathbf{N}$ for m species, more than t draws are needed to discover all the species. This probability, of course, depends on whether the draws are with or without replacement, but we will use the same notation. We ask the same question again, namely, is there a special abundance vector that minimizes $\Phi(t, N, m, \mathbf{N})$, uniformly over all $t \geq m$? The answer we get is 'Yes'; there is a special abundance vector $\mathbf{N}_0$, representing "*almost equal*" abundance for the m species, that does the job, irrespective of whether draws are made with or without replacement.

Let us first explain what we mean by "*almost equal*" abundance and why we need this. Note that all the m species being *exactly* equally abundant would mean that the N_i's must all be equal, which, of course, can happen *only if* the population size N is a multple of m. That will certainly be an undesirable restriction on N. So, the right thing to do would be to get, for any given N, as close as possible to equal abundance and that is what is represented by the special vector $\mathbf{N}_0$, which is formally defined below.

For any abundance vector $\mathbf{N} = (N_1, \cdots, N_m)$, let $d(\mathbf{N}) = \max\{|N_i - N_j| : i \neq j\}$. It is obvious that $d(\mathbf{N})$ is invariant under permutations of coordinates of $\mathbf{N}$ (as is $\Phi(t, N, m, \mathbf{N})$). Also, the larger $d(\mathbf{N})$ is, the farther is $\mathbf{N}$ from "equal abundance". A little reflection will convince the reader that the minimum value of $d(\mathbf{N})$ equals 0 only if N is a multiple of m and is attained by $\mathbf{N}_0 = (\frac{N}{m}, \cdots, \frac{N}{m})$; otherwise, the minimum value is 1, attained by the unique (upto permutations) abundance vector of $\mathbf{N}_0 = \Big(\underbrace{N_0 + 1, \cdots, N_0 + 1}_{k}, \underbrace{N_0, \cdots, N_0}_{m-k}\Big)$, where $N = mN_0 + k,\ 0 < k < m$. In other words, $\mathbf{N}_0$ is always the unique (upto permutations) abundance vector satisfying $d(\mathbf{N}_0) \leq 1$. It is clear that $\mathbf{N}_0$ represents, for any given N, an admissible abundance vector where the m species are as equally abundant as possible. Here is our main result for the finite population case.

Theorem 3: For a finite population of size N, consisting of m $(2 \leq m \leq N)$ different species,

$$\Phi(t, N, m, \mathbf{N}) \geq \Phi(t, N, m, \mathbf{N}_0), \quad \text{for all } t \geq m, \tag{11}$$

irrespective of whether units are drawn with or without replacement. Further, the inequality in (11) is strict (for all t, for which the left-hand-side is positive), unless $\mathbf{N} = \mathbf{N}_0$ (upto permutations).

As mentioned earlier, the main ideas of the proof are the same as those in case of infinite population, but the execution of those ideas are far more complicated. That should not be surprising because, in the infinite population case, we were dealing with continuous

variables p_i, whereas now we are handling the problem of minimizing a function of positive integer variables N_i. The difficulty gets a bit more multiplied by the fact that we also have to treat the cases of sampling WR and WOR differently. Since the details are available in Goswami and Sinha (2006), we just briefly outline the steps here.

As before, the idea is to first prove it for $m = 2$ and then use induction on m. For $m = 2$, we can take any abundance vector (N_1, N_2), where (without loss of generality) $N_1 \leq N_2$, and write down explicit formulas for $\Phi\Big(t, N, 2, (N_1, N_2)\Big)$ in both WR and WOR cases. Now, if $N_2 - N_1 \geq 2$, one can, with some work, show that

$$\Phi\Big(t, N, 2, (N_1, N_2)\Big) > \Phi\Big(t, N, 2, (N_1 + 1, N_2 - 1)\Big), \quad \text{for all } t \geq 2.$$

It is now just a matter of repeating this inequality over and over again to finally get (11) for $m = 2$. To proceed with induction now, we need, as before, a formula relating $\Phi(\cdot, \cdot, m, \cdot)$ to $\Phi(\cdot, \cdot, m-1, \cdot)$. This can again be obtained by using a similar kind of conditioning as before. Indeed, for any $m \geq 3$, $N \geq m$ and any abundance vector $\mathbf{N} = (N_1, \cdots, N_m)$, one can get $\Phi(t, N, m, \mathbf{N})$ to be a weighted average of 1 and $\Phi(t-s, N-N_i, m-1, \mathbf{n}_{(i)})$, $1 \leq s \leq t-m+1$, where the weights are p.m.f.s of an appropriate Binomial distribution or Hypergeometric distribution, according as the draws are WR or WOR. Here, $\mathbf{N}_{(i)}$, for any i, is the abundance vector of size $(m-1)$, obtained by just removing N_i from $\mathbf{N}$. Using the induction hypothesis now, one can get an inequality analogous to (6), namely,

$$\Phi(t, N, m, \mathbf{N}) \geq \Phi(t, N, m, \mathbf{N}^{(i)}), \quad \text{for all } t \geq m \text{ and each } 1 \leq i \leq m, \tag{12}$$

where $\mathbf{N}^{(i)}$ is the abundance vector whose ith coordinate is the same as that of $\mathbf{N}$, that is, N_i, while the rest of the coordinates are those of an abundance vector of size $(m-1)$, adding up to $N - N_i$, that represents "as equal abundance as possible" for $m-1$ species in a population of size $N - N_i$. An explicit description is given in Goswami and Sinha (2006). Further, induction hypothesis will also imply strict inequality in (12), unless $d(\mathbf{N}_{(i)}) \leq 1$.

Now comes the final step. Recall that, the final step in the infinite population case presented a little hurdle and we needed to use a result like Lemma 1 to complete the proof. Here also, the job is far from over. The passage from (12) to completion of the proof of Therem 3, poses a significant challenge. Luckily, with some effort, we were able to formulate and prove a result that helped us cross this last hard mile. This result played a role very analogous to that played by Lemma 1 in Section 2.1. Unfortunately (but perhaps not surprisingly), neither the statement of the result nor its proof are as straightforward as Lemma 1. For the sake of brevity of this article, we refrain from stating the result here and also skip the details of how the lemma is used to complete the proof of Theorem 3. An interested reader will find all the details in Goswami and Sinha (2006).

3. Sampling from a social network: optimal strategies

A social network is a population equipped with an irreflexive binary relation. Denoting the binary relation by $\rightsquigarrow$, we say that two distinct units i and j have some "*tie*" if either $i \rightsquigarrow j$ or $j \rightsquigarrow i$. We say that there is a "*symmetric tie*" or a "*reciprocal tie*" between units i and j if both $i \rightsquigarrow j$ and $j \rightsquigarrow i$ hold. Given a social network, one among several quantities of interest for sociologists, is the extent of reciprocity in the network. Many different measures

have been proposed for this, among which we consider here the simplest one, proposed by Rao and Bandyopadhayay (1987). It is defined as

$$\overline{\theta} = \frac{1}{N(N-1)} \sum\sum_{\substack{i,j \in \Omega \\ i \neq j}} \theta(i,j), \tag{13}$$

where Ω denotes the population, N the size of the population and $\theta(i,j)$ is defined to be 1 if both $i \rightsquigarrow j$ and $j \rightsquigarrow i$ hold and defined to be 0 otherwise.

The problem that we investigated in Goswami, Sinha and Sengupta (1990) is that of estimating the population parameter $\overline{\theta}$ unbiasedly on the basis of "data available" from a sample of size n drawn from the population. Note that if SRSWOR is used to draw a sample s of size n, then one can easily see that the usual sample mean

$$\hat{\overline{\theta}}(s) = \frac{1}{n(n-1)} \sum\sum_{\substack{i,j \in s \\ i \neq j}} \theta(i,j) \tag{14}$$

is an unbiased estimator of $\overline{\theta}$. The question we asked is whether one can do better (in the sense of reducing the variance) by adapting an appropriate sequential sampling scheme, coupled with an appropriate unbiased estimator. To keep the question alive, it is very important to turn our attention to a clear understanding of "data available from the sample".

To clarify our point, let us denote $\overline{\theta}_i$, for each $i \in \Omega$, to be the average number of symmetric ties in which i is involved, that is, $\overline{\theta}_i = \frac{1}{N-1} \sum\limits_j \theta(i,j)$. Clearly then, $\overline{\theta}$ is the population mean of the $\overline{\theta}_i$ and therefore, if one stipulates that from a sample drawn from the population, the value of $\overline{\theta}_i$ will be "available" for each sample unit i, then our question stops there. This is because, in that case, the admissibility of the (SRSWOR, sample mean) strategy is a classical result. However, a large number of practitioners in this field are strongly opposed to the stipulation that $\overline{\theta}_i$ for sample units i are "observable". So, let us place here the wide consensus on what is "observable" and what is not.

For each population unit i, the "out-set" and "in-set" of i are defined respectively as $\mathcal{O}(i) = \{j : i \rightsquigarrow j\}$ and $\mathcal{I}(i) = \{j : j \rightsquigarrow i\}$, with their cardinalities, denoted by d_i and e_i respectively, being called the "out-degree" and "in-degree" of i. It then follows that, $\overline{\theta}_i = \frac{1}{N-1} \Big| \mathcal{O}(i) \cap \mathcal{I}(i) \Big|$, for any $i \in \Omega$, and so information on both $\mathcal{O}(i)$ and $\mathcal{I}(i)$ is required to know $\overline{\theta}_i$. The widely held opinion of practitioners is that, while somewhat reliable information on $\mathcal{O}(i)$ may be available from a sample unit i, but information on $\mathcal{I}(i)$ is highly unreliable.

We accept this premise, namely, that the only information "available" from a sample unit i, that is reliable and useful, is its out-set $\mathcal{O}(i)$. This would mean that the admissibility of (SRSWOR, sample mean) strategy is no longer guaranteed and therefore, our search for a better strategy becomes valid and meaningful.

Indeed, what we were able to achieve in Goswami, Sinha and Sengupta (1990) is to stitch up an alternative strategy that performs **uniformly better** than the (SRSWOR, sample mean) strategy. The main idea behind our proposed strategy essentially orginated from a detailed examination of the case with sample size $n = 2$. So, we are going to describe

that special case in detail here and follow it up by just giving an outline of the scheme for general sample size n.

Just to keep things simple, we make an additional assumption that $d_i > 0$, for each i. We assure the reader that this is not at all indispensable and can be easily done away with. From the underlying stipulation that information on $\mathcal{O}(i)$ is available from each sample unit i, it is easy to see that the value of $\theta(i,j)$ will be known for each **pair** of units (i,j) in the sample. Now, in case of sample size $n = 2$, the unbiased estimator that the (SRSWOR, sample mean) strategy proposes boils down to

$$\hat{\bar{\theta}}_1(s) = \theta(i,j), \quad \text{if } s = \{i,j\}. \tag{15}$$

What we propose in our alternative strategy is the estimator

$$\hat{\bar{\theta}}_2(s) = \frac{d_i}{N-1}\theta(i,j), \quad \text{if } s \text{ is the } \textbf{ORDERED} \text{ sample } (i,j). \tag{16}$$

It is fairly easy to simplify $E\Big(\hat{\bar{\theta}}_2\Big) = \sum\limits_{\substack{i,j\in\Omega \\ i\neq j}}\sum P\Big(s=(i,j)\Big)\frac{d_i}{N-1}\theta(i,j)$ and hence show that $\hat{\bar{\theta}}_2$ is unbiased. What is really important is that one can do a little computation to get the second moments of the unbiased estimators (15) and (16) and deduce that

$$E\Big(\hat{\bar{\theta}}_1^2\Big) - E\Big(\hat{\bar{\theta}}_2^2\Big) = \frac{1}{N(N-1)}\sum_{\substack{i,j\in\Omega \\ i\neq j}}\sum \Big(1 - \frac{d_i}{N-1}\Big)\,\theta(i,j) \geq 0. \tag{17}$$

An immediate consequence of (17) is that $\hat{\bar{\theta}}_2$ performs uniformly better than $\hat{\bar{\theta}}_1$, in the sense of reducing variance. Further, the right-hand-side of (17) also shows that $\hat{\bar{\theta}}_2$ has **strictly smaller variance** except in the extreme case when $d_i = N-1$ for **all** i with $\sum_j \theta(i,j) > 0$. Of course, noting that our proposed estimator $\hat{\bar{\theta}}_2$ is "order dependent", an initiated reader will immediately see an opportunity of further improving on it by using the classical idea of what is widely known as "*Blackwellization*". By "averaging over order", one gets an even more improved estimator given by

$$\hat{\bar{\theta}}_3(s) = \frac{2d_id_j}{(N-1)(d_i+d_j)}\,\theta(i,j), \quad \text{if } s = \{i,j\}. \tag{18}$$

Having thus described our improved strategy in detail for the case $n = 2$, it is quite natural now to try and extend this idea for a general sample size n. This (and much more) was indeed done and reported in detail in Goswami, Sinha and Sengupta (1990). In particular, we were able to exhibit a sequential strategy (p_0, e_0), for any sample size n, that performs uniformly better than a size n (SRSWOR, sample mean) strategy. An initiated reader would surely recall that a sampling stategy consists of a pair (sampling scheme, estimator). Our proposed sequential strategy (p_0, e_0) is described below, from which it will be clear to the reader that it is a generalization of what was done for $n = 2$.

Sequential strategy (p_0, e_0): For distinct population units $i_1, \cdots, i_k$, let us denote

$$\mathcal{O}(i_1, \cdots, i_k) = \Big(\mathcal{O}(i_1) \cup \cdots \cup \mathcal{O}(i_k)\Big) \setminus \{i_1, \cdots, i_k\} \text{ and } d(i_1, \cdots, i_k) = |\mathcal{O}(i_1, \cdots, i_k)|.$$

- p_0 : First draw a SRSWOR of size $(n-1)$, say, $\{i_1, \cdots, i_{n-1}\}$
- p_0 : Now draw a random unit, say, i_n from $\mathcal{O}(i_1, \cdots, i_{n-1})$
- e_0 : For $s = \{i_1, \cdots, i_{n-1}; i_n\}$ define

$$e_0(s) = \frac{1}{n(n-1)} \sum_{\substack{1 \le k,l \le n-1 \\ k \ne l}} \sum \theta(i_k, i_l) + \frac{2d(i_1, \cdots, i_{n-1})}{n(n-1)(N-n+1)} \sum_{k=1}^{n-1} \theta(i_k, i_n) \qquad (19)$$

The following theorem captures one of our main results in the context of this investigation.

Theorem 4: For every $n \ge 2$, the sequential sampling strategy (p_0, e_0) of sample size n perfoms better than the size n (SRSWOR, sample mean) strategy uniformly, in the sense of having smaller variance.

Did we bump into this sequential strategy by some magic or chance? The answer to that is an emphatic 'No'. In fact, what we were able to do is to describe in detail the string of main ideas that essentially leads one to not just this particular sequential strategy (p_0, e_0), but to a whole class of possible sequential strategies, each of which beats the (SRSWOR, sample mean) strategy uniformly. The strategy (p_0, e_0) is just a special case. The story behind the closed doors is that we did not arrive at our sequential strategies at one go. We did it in two steps. To briefly describe it, let us fix $n > 2$ and denote the (SRSWOR, sample mean) strategy of size n by (p, e). In our first step, we construct a variable sample size strategy (p^*, e^*), with sample size varying between n and $n-1$, which is equivalent to (p, e), in the sense that $E_{p^*}(e^*) = E_p(e)$ and $E_{p^*}(e^{*2}) = E_p(e^2)$. Then, in our next step, we construct a sequential strategy (p^{**}, e^{**}) of sample size n and show that it performs uniformly better than (p^*, e^*) and hence uniformly better than (p, e). The important point is that in this last step, we actually prescribe not just one sequential strategy but **a whole class** of possible sequential strategies (p^{**}, e^{**}) of sample size n, each of which performs uniformly better than (p^*, e^*) (and hence, better than (p, e)). It will be too much to give the complete descriptions of all of these and the corresponding proofs here. For complete details, we refer to Goswami, Sinha and Sengupta (1990).

4. Some concluding remarks

Section 2.1 :

It is well-known that an arbitrary $\mathbf{p}$ vector is "majorized" by the vector $\mathbf{p}_0$ and hence the results on Schur concave functions will directly apply, provided one can establish Schur concavity of $\Phi(t, m, \mathbf{p})$ as a function of $\mathbf{p}$. This is worth exploring (see Marshall and Olkin (1979)). Another question that occured to the author while taking a fresh relook at the paper just before the talk, is that $\mathbf{p}_0$ is known to have the maximum Shannon entropy among all probability vectors $\mathbf{p}$ of size m. It is worth investigating whether that has any role to play. Cracking this may lead to formulating a large number of more general problems and getting interesting answers to those.

Section 2.2 :

Our proposed sequential strategy (p_0, e_0) (and, more generally, (p^{**}, e^{**})) leaves several questions unanswered. Firstly, the estimator e_0 is order dependent and so, it leads one to possibly think of improving it further by using "Blackwellization" (that is, symmetrization). We tried it, but the variance of the symmetrized version seemed rather intractable. Another more significant limitation with (p_0, e_0) is that the sampling design p_0 differs very little from SRSWOR, in the sense that the sequential nature of p_0 appears only while drawing the last unit. This would mean that the improvent wouldn't mean much when n is large (as is evidenced by the expression for variance). So, a natural question is whether we can devise a strategy that brings in the sequential nature much earlier, thereby hoping to make more significant improvement over (p, e). Some efforts in this direction were undertaken for sample size $n = 3$. The results we obtained are reported in Goswami, Sinha and Sengupta (1990). However, the issue remains wide open for general n and is certainly worth pursuing.

Acknowledgements

I am indeed grateful to the Society of Statistics, Computer and Applications (SSCA) and, in particular, to Professor V.K.Gupta, President (SSCA), for inviting me to deliver the prestigeous Bikas K. Sinha Endowment Lecture during the 27th annual conference of the SSCA held in February 2025 at NEHU, Shillong. Many thanks to my teacher and friend '*Bimal-da*' (Professor Bimal K. Sinha) for proposing and endorsing my name for delivering this 4th lecture in the series. I would also like to thank the organizers, specially Professor Bishal Gurung and my good friend '*Tapan*' (Professor T. K. Chakrabarty), for their wonderful hospitality and efforts in ensuring smooth organization of this big event. Finally, I must again thank Professor V.K.Gupta, for encouraging me to submit this article based on my lecture, for possible inclusion in the Special Proceedings of the 27th Conference.

References

Gore, A. P. and Paranjape, S. A. (1997). Effort needed to measure biodiversity. *International Journal of Ecology and Environmental Sciences*, **23**, 173-183.

Gore, A. P. and Paranjape, S. A. (2001). *A Course in Mathematical and Statistical Ecology.* Kluwer, Budapest.

Goswami, A., Sinha, Bikas K., and Sengupta, S. (1990). Optimal strategies in sampling from a social network. *Sequential Analysis*, **9**, 1-18.

Goswami, A. and Sinha, Bikas K. (2006). Some probabilistic aspects in the discovery of species. *Sequential Analysis*, **25**, 103-115.

Marshall, A. W. and Olkin, I. (1979). *Inequalities: Theory of Majorization with Applications.* Academic Press, New York.

Rao, A. R. and Bandyopadhyay, S. (1987). Measures of reciprocity in a social network. *Sankhya Series A*, **49**, 141-188.

Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025; pp 43-50

On Exploring Tails via Tail Equivalence

Sreenivasan Ravi
Department of Studies in Statistics, University of Mysore, Manasagangotri, Mysore 570 006, India

Received: 18 May 2025; Revised: 24 June 2025; Accepted: 26 June 2025

Abstract

Tail equivalence between two distribution functions was introduced in Resnick, S.I. (1971). Tail equivalence and its applications, Journal of Applied Probability, 8(1), 136-156. After clarifying a few properties and giving examples of classes of tail equivalent distributions, this article looks briefly at some interesting applications of tail equivalence in establishing tail behaviours of mixtures and order statistics, in particular, of limit laws of normalised k-th upper order statistics from a random sample, for fixed integer k. The tail behaviours of such limit laws have been studied via tail equivalence. It turns out that tail equivalence simplifies much of the apparent difficulty in handling the tails of such limit laws. A consequence is a method of generating random observations from regularly varying tails having different exponents of regular variation.

Key words: Extreme value theory; Limit laws; Mixtures; Partial maximum; Tail equivalence; Upper order statistics.

AMS Subject Classifications: 60F05, 60G70, 62G30.

1. Introduction

Resnick (1971) introduces the concept of tail equivalence between two distribution functions (dfs) on the real line $\mathbb{R}$. Here tail refers to right tail and we confine to right tail in this article. Similar results for left tail can be derived from the results discussed here. Tail equivalence divides the class of all dfs on the real line into equivalence classes. In this article, after giving known definitions of heaviness of tail, illustrations of the use of tail equivalence to study the tail behaviour of limit laws of normalised mixtures and k-th upper order statistics from a random sample for fixed integer k are given, under fixed and random sample sizes. These results were derived by the author and co-workers in several articles.

1.1. Tail equivalence

Definition (Resnick, 1971): Two dfs F and G on $\mathbb{R}$ are said to be tail equivalent, denoted by $F \stackrel{T}{=} G$, if

$$\lim_{x\to\infty} \frac{1-F(x)}{1-G(x)} = A, 0 < A < \infty. \tag{1}$$

Sreenivasan Ravi
ravi@statistics.uni-mysore.ac.in

We refer to Resnick (1971) for applications of tail equivalence in extreme value theory. The following are easy consequences of the definition:

- If $F \stackrel{T}{=} G$, and $r(F) = \sup\{x \in \mathbb{R} : F(x) < 1\}$ denotes the right extremity of F, then $r(F) = r(G)$, finite or infinite. This is because, otherwise, A in (1) will be 0 or ∞ according as $r(F) < r(G) \leq \infty$ or $r(G) < r(F) \leq \infty$, respectively.
- Since $F \stackrel{T}{=} F$ with $A = 1$ in (1), the relation $\stackrel{T}{=}$ is reflexive.
- If $F \stackrel{T}{=} G$ with the limit in (1) as A, then $G \stackrel{T}{=} F$ with the limit in (1) as $1/A$, so that the relation $\stackrel{T}{=}$ is symmetric.
- If $F \stackrel{T}{=} G$ with the limit in (1) as A, and $G \stackrel{T}{=} H$ with the limit in (1) as B, then $F \stackrel{T}{=} H$ with the limit in (1) as AB, so that the relation $\stackrel{T}{=}$ is transitive, proving that the relation is an equivalence relation.

Now we give some examples of tail equivalent families of dfs on $\mathbb{R}$.

Examples of classes of tail equivalent dfs:

- Family of exponential distributions with different location parameters:
  If $F(x;\mu) = 1 - e^{x-\mu}, x > \mu$, and 0 elsewhere, with $\mu \in \mathbb{R}$ as a location parameter, then $\lim_{x\to\infty} \frac{1-F(x;\mu_1)}{1-F(x;\mu_2)} = e^{-(\mu_1-\mu_2)}$. However, note that family of exponential distributions with different scale parameters, is not a tail equivalent class.
- Family of Pareto distributions with location and scale parameters:
  If $F(x;\mu,\sigma) = 1 - \frac{1}{\left(\frac{x-\mu}{\sigma}\right)}, x > \mu+\sigma$, and 0 elsewhere, with $\mu \in \mathbb{R}$ as a location parameter and $\sigma > 0$ as a scale parameter, then $\lim_{x\to\infty} \frac{1-F(x;\mu_1,\sigma_1)}{1-F(x;\mu_2,\sigma_2)} = \frac{\sigma_2}{\sigma_1}$.
- Family of log-Pareto distributions with scale and shape parameters:
  If $F(x;\mu,\sigma) = 1 - \frac{1}{\ln\left(\frac{x}{\mu}\right)^\sigma}, x > \mu e^{1/\sigma}$, and 0 elsewhere, with $\mu > 0$ as a scale parameter and $\sigma > 0$ as a shape parameter, then $\lim_{x\to\infty} \frac{1-F(x;\mu_1,\sigma_1)}{1-F(x;\mu_2,\sigma_2)} = \frac{\sigma_2}{\sigma_1}$.

1.2. Heavy tails

We refer to Praveena and Ravi (2023, 2025) and Nair *et al.* (2023) for definitions and results mentioned below and some recent work. We give some definitions now, followed by some examples.

Definitions:

- A df F on $\mathbb{R}$ is heavy tailed if $\limsup_{x\to\infty} \dfrac{1-F(x)}{e^{-x}} = \infty$.
- If not, F is said to be light tailed.
- A df F on $\mathbb{R}$ is super-heavy tailed to the right if $\limsup_{x\to\infty} \dfrac{1-F(x)}{x^{-\alpha}} = \infty$, for all $\alpha > 0$.

Examples:

- Pareto df $F(x) = 1 - \dfrac{1}{x^\alpha}, x > 0, \alpha > 0$, Weibull df with shape parameter greater than 1, are examples of heavy tailed dfs.
- Normal, exponential dfs are examples of light tailed dfs.
- Cauchy, Fréchet, Burr dfs are super-heavy tailed distributions.
- But there can be heavier tails like dfs log-Pareto, log-log-Pareto, *etc.*

1.3. Extremes and upper order statistics

The extreme value laws: If $X_1, X_2, \ldots,$ are independent and identically distributed random variables with common df F, $M_n = \max\{X_1, \ldots, X_n\}$, and $\lim_{n\to\infty} P(M_n \le a_n x + b_n) = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x), \quad x \in \mathcal{C}(G)$, the set of all continuity points of the limit df G, then we denote this as $F \in \mathcal{D}_l(G)$. It is known that G is a type of the extreme value laws, given by:

- Fréchet law: $\Phi_\alpha(x) = \exp(-x^{-\alpha})$, $0 \le x$,
- Weibull law: $\Psi_\alpha(x) = \exp(-|x|^\alpha)$, $x < 0$,
- Gumbel law, $\Lambda(x) = \exp(-\exp(-x))$, $x \in \mathbb{R}$; where $\alpha > 0$ a parameter.

Max stability: The extreme value laws satisfy the following stability property:

$$\Phi_\alpha^n\left(n^{1/\alpha}x\right) = \Phi_\alpha(x), \quad \Psi_\alpha^n\left(n^{-1/\alpha}x\right) = \Psi_\alpha(x), \quad \Lambda^n(x + \log n) = \Lambda(x), x \in \mathbb{R}.$$

1.4. Order statistics and k-th extremes

We denote the order statistics of $\{X_1, \ldots, X_n\}$ by $X_{1:n} \le \ldots \le X_{n:n}$ and assume that $F \in \mathcal{D}_l(G)$ for some G. The df of the k-th upper order statistic $X_{n-k+1:n}$, for a fixed positive integer k is given by

$$F_{k:n}(x) = P\left(X_{n-k+1:n} \le x\right) = \sum_{i=0}^{k-1} \binom{n}{i} F^{n-i}(x)(1 - F(x))^i, x \in \mathbb{R}.$$

The limit $G_k(x) = \lim_{n\to\infty} F_{k:n}(a_n x + b_n)$ is given by

$$G_k(x) = G(x) \sum_{i=0}^{k-1} \frac{(-\log G(x))^i}{i!}, \quad x \in \{y : G(y) > 0\}.$$

2. Applications of tail equivalence to tail behaviour

The following questions on tails of dfs were answered by using tail equivalence.

2.1. Questions and motivation

- If $F(.) = \alpha F_1(.) + (1-\alpha)F_2(.)$ is a mixture df with component dfs F_1, F_2, how are the tails of F related to those of F_1, F_2? Or, how is $F \in \mathcal{D}_l(.)$ related to $F_i \in \mathcal{D}_l(.), i = 1, 2$?
- What is the tail of G_k like? Or, $G_k \in \mathcal{D}_l(.)$?

2.2. On mixtures

The following discussion is from Praveena *et al.* (2019).

- If F is the mixture df, $r(F) = \max\{r(F_1), r(F_2)\}$.
- If $F \in \mathcal{D}_l(.)$ with some norming constants and $r(F_1) < r(F_2)$, then $F_2 \in \mathcal{D}_l(.)$ with the same norming constants. This is because $F \stackrel{T}{=} F_2$.
- If $F \in \mathcal{D}_l(.)$ with some norming constants and $r(F_1) = r(F_2)$, then nothing can be said about the max domains to which F_1, F_2 may belong to. Examples have been given.
- If $F_1 \stackrel{T}{=} F_2$ and one of them belong to $\mathcal{D}_l(.)$ with some norming constants, then $F \in \mathcal{D}_l(.)$ with the same norming constants.

2.3. On k-th extremes via tail equivalence

The discussion here is from Ravi and Manohar (2018).

A recurrence relation: For any df F, fixed integer $k \geq 1$, define $F_k(x) = F(x) \sum_{i=0}^{k-1} \frac{(-\ln F(x))^i}{i!}$, $x \in \{y : F(y) > 0\}$. The df F_k satisfies the recurrence relation

$$F_{k+1}(x) = F_k(x) + \frac{F(x)}{k!}(-\ln F(x))^k, \quad k \geq 1, \ x \in \{y : F(y) > 0\}.$$

The pdf of F_{k+1} is

$$f_{k+1}(x) = \frac{f(x)}{k!}(-\ln F(x))^k, \quad k \geq 1, \ x \in \{y : F(y) > 0\}.$$

A result for fixed sample size: If F is a df with pdf f, then for every positive integer k, $(1-F(x))^k$ is tail of the df $H_k(x) = 1 - (1-F(x))^k$, $x \in \mathbb{R}$, and H_k is also absolutely continuous with pdf $H_k'(x) = k\{1-F(x)\}^{k-1} f(x), x \in \mathbb{R}$. Further, the following are true:

- If $F \in D_l(\Phi_\alpha)$, then $r(H_k) = r(F) = \infty$, and $H_k \in D_l(\Phi_{k\alpha})$ with $a_n = F^-(1-(1/n)^{1/k}), b_n = 0$.
- If $F \in D_l(\Psi_\alpha)$ then $r(H_k) = r(F) < \infty$, and $H_k \in D_l(\Psi_{k\alpha})$ with $a_n = r(F) - F^-(1-(1/n)^{1/k}), b_n = r(F)$.

- If $F \in D_l(\Lambda)$, $a_n = v(b_n)$ and $b_n = F^-\left(1 - \frac{1}{n}\right)$ then $r(H_k) = r(F)$, and $H_k \in D_l(\Lambda)$ with $a_n = \dfrac{v(b_n)}{k}, b_n = H_k^-(1 - 1/n)$.

Another result for fixed sample size: Let rv X have absolutely continuous df F with pdf f and k be a positive integer. Then for $F_k(x) = F(x)\sum_{i=0}^{k-1}\dfrac{(-\ln F(x))^i}{i!}$, $x \in \{y : F(y) > 0\}$, the following results are true:

- F_k is a df with $r(F_k) = r(F)$, pdf $f_k(x) = \dfrac{f(x)}{(k-1)!}(-\ln F(x))^{k-1}$, $x \in \{y \in \mathbb{R} : F(y) > 0\}$; and $\lim_{x\to r(F)}\dfrac{1 - F_k(x)}{(1 - F(x))^k} = \dfrac{1}{k!}$, so that $F_k \overset{TE}{=} H_k$.
- If $F \in D_l(\Phi_\alpha)$, then $r(F_k) = r(F) = \infty$, and $F_k \in D_l(\Phi_{k\alpha})$ with $a_n = F^-(1 - (k!/n)^{1/k}), b_n = 0$.
- If $F \in D_l(\Psi_\alpha)$ then $r(F_k) = r(F) < \infty$, and $F_k \in D_l(\Psi_{k\alpha})$ with $a_n = r(F) - F^-(1 - (k!/n)^{1/k}), b_n = r(F)$.
- If $F \in D_l(\Lambda)$, $a_n = v(b_n)$ and $b_n = F^-\left(1 - \frac{1}{n}\right)$ then $r(F_k) = r(F)$, and $F_k \in D_l(\Lambda)$ with $a_n = \dfrac{v(b_n)}{k}, b_n = F_k^-(1 - 1/n)$.

2.3.1. Results for random sample size

Uniform k-th extremes: Suppose that n in the previous section is replaced by a discrete uniform rv N_n with $P(N_n = r) = \frac{1}{n}$, $r = m+1, m+2, \ldots, m+n$, N_n independent of the iid rvs $X_1, X_2, \ldots$, $m \geq 1$ a fixed integer. We look at the tail behaviour of the limit of linearly normalized $X_{N_n-k+1:N_n}$. Observe that $X_{N_n-k+1:N_n}$ is well defined for $1 \leq k \leq m$. We have $F_{k:N_n}(x) = P(X_{N_n-k+1:N_n} \leq x) = \sum_{r=m}^{\infty} P(X_{N_n-k+1:N_n} \leq x, N_n = r)$
$= \sum_{r=m}^{\infty}\sum_{i=0}^{k-1}\binom{r}{i}F^{r-i}(x)(1 - F(x))^i P(N_n = r)$, $x \in \mathbb{R}$. The following results are true:

- If $F \in D_l(G)$ for some max stable df G then $\lim_{n\to\infty} F_{k:N_n}(a_n x + b_n)$ is equal to $U_{k,G}(x) = k\{\frac{1-G(x)}{-\ln G(x)}\} - G(x)\sum_{l=1}^{k-1}(k-l)\frac{(-\ln G(x))^{l-1}}{l!}$, $x \in \{y \in \mathbb{R} : G(y) > 0\}$, $G = \Phi_\alpha$ or Ψ_α or Λ.
- For any df F, and fixed integer $k \geq 1$, let $U_{k,F}(x) = k\{\frac{1-F(x)}{-\ln F(x)}\} - F(x)\sum_{l=1}^{k-1}(k - l)\frac{(-\ln F(x))^{l-1}}{l!}$, $x \in \{y : F(y) > 0\}$. If X has df F, pdf f, k is a fixed positive integer, and $U_{1,F}(x) = \dfrac{1 - F(x)}{-\ln F(x)}$, $x \in \{y : F(y) > 0\}$, then $U_{1,F}$ is a df with $r(U_{1,F}) = r(F)$, pdf $u_{1,F}(x) = \dfrac{f(x)}{F(x)}\dfrac{U_{1,F}(x) - F(x)}{-\ln F(x)} = \dfrac{f(x)}{F(x)}\left\{\dfrac{1 - F(x) + F(x)\ln F(x)}{(-\ln F(x))^2}\right\}$, $x \in \{y \in \mathbb{R} : F(y) > 0\}$; and $\lim_{x\to r(F)}\dfrac{1 - U_{1,F}(x)}{1 - F(x)} = \dfrac{1}{2}$ so that $U_{1,F} \overset{T}{=} F$.

For the family of dfs $U_{k,F}$ and $H_k(x) = 1 - (1 - F(x))^k, x \in \mathbb{R}$, the following results are true.

- $U_{k,F}$ is a df with $r(U_{k,F}) = r(F)$, pdf $u_{k,F}(x) = \dfrac{kf(x)}{(-\ln F(x))^2}\left\{\dfrac{1}{F(x)} - \sum_{l=0}^{k}\dfrac{(-\ln F(x))^l}{l!}\right\}$, $x \in \{y \in \mathbb{R} : F(y) > 0\}$, $\lim_{x\to r(F)}\dfrac{1 - U_{k,F}(x)}{(1 - F(x))^k} = \dfrac{1}{(k+1)!}$, and $U_{k,F} \stackrel{T}{=} H_k$.
- If $F \in D_l(\Phi_\alpha)$, then $r(U_{k,F}) = r(F) = \infty$, $U_{k,F} \in D_l(\Phi_{k\alpha})$ with $a_n = F^-(1 - ((k+1)!/n)^{1/k}), b_n = 0$.
- If $F \in D_l(\Psi_\alpha)$ then $r(U_{k,F}) = r(F) < \infty$, $U_{k,F} \in D_l(\Psi_{k\alpha})$ with $a_n = r(F) - F^-(1 - ((k+1)!/n)^{1/k}), b_n = r(F)$.
- If $F \in D_l(\Lambda)$, $a_n = v(b_n)$ with $b_n = F^-\left(1 - \frac{1}{n}\right)$ then $r(U_{k,F}) = r(F)$, $U_{k,F} \in D_l(\Lambda)$ with $F = U_{k,F}, G = \Lambda$, $a_n = \dfrac{v(b_n)}{k}, b_n = U^-_{k,F}(1 - 1/n)$.
- The df $U_{k,F}$ satisfies the recurrence relation

$$U_{k+1}(x) = U_{k,F}(x) + U_{1,F}(x) - F(x)\sum_{l=1}^{k}\frac{(-\ln F(x))^{l-1}}{l!}.$$

Geometric k-th extremes: Let N_n be a shifted geometric rv with pmf $P(N_n = r) = p_n q_n^{r-m}, r = m, m+1, m+2, \ldots$, $0 < p_n < 1, q_n = 1 - p_n$ and $\lim_{n\to\infty} np_n = 1$.

- If $F \in D_l(G)$ for some max stable law G, then for fixed integer $k, 1 \le k \le m$, $\lim_{n\to\infty} F_{k:N_n}(a_n x + b_n)$ is equal to

$$R_{k,G}(x) = 1 - \left(\frac{-\ln G(x)}{1 - \ln G(x)}\right)^k,\ x \in \{y \in \mathbb{R} : G(y) > 0\}, \quad \text{with}$$

$$R_{k,G}(x) = \begin{cases} 1 - \left(\frac{1}{1+x^\alpha}\right)^k & \text{if } G(x) = \Phi_\alpha(x), \\ 1 - \left(\dfrac{(-x)^\alpha}{1 + (-x)^\alpha}\right)^k & \text{if } G(x) = \Psi_\alpha(x), \\ 1 - \left(\frac{e^{-x}}{1+e^{-x}}\right)^k & \text{if } G(x) = \Lambda(x). \end{cases}$$

  The first two are Burr distributions of XII kind (Burr, 1942) and the last is the logistic distribution.

- If X has df F, pdf f, k a positive integer and $R_{k,F}$ is as defined above, then the following are true:

- $R_{k,F}$ is a df with pdf $r_{k,F}(x) = \dfrac{kf(x)(-\ln F(x))^{k-1}}{F(x)(1 - \ln F(x))^{k+1}}$, $x \in \{y \in \mathbb{R} : F(y) > 0\}$, $r(R_{k,F}) = r(F)$, and $\lim_{x\to r(F)}\dfrac{1 - R_{k,F}(x)}{(1 - F(x))^k} = 1$, and $R_{k,F} \stackrel{T}{=} H_k$.

- If $F \in D_l(\Phi_\alpha)$, then $r(R_{k,F}) = r(F) = \infty$, and $R_{k,F} \in D_l(\Phi_{k\alpha})$ with $a_n = F^-(1-(1/n)^{1/k}), b_n = 0$.
- If $F \in D_l(\Psi_\alpha)$ then $r(R_{k,F}) = r(F) < \infty$, and $R_{k,F} \in D_l(\Psi_{k\alpha})$ with $F = R_{k,F}, G = \Psi_{k\alpha}$, $a_n = r(F) - F^-(1-(1/n)^{1/k}), b_n = r(F)$.
- If $F \in D_l(\Lambda)$, $a_n = v(b_n)$ and $b_n = F^-\left(1-\frac{1}{n}\right)$ then $r(R_{k,F}) = r(F)$, and $R_{k,F} \in D_l(\Lambda)$ with $F = R_{k,F}, G = \Lambda$, $a_n = \frac{v(b_n)}{k}, b_n = R^-_{k,F}(1-1/n)$.

The Burr connection: Burr (1942) proposed twelve explicit forms of dfs which have since come to be known as the Burr system of distributions. A number of well-known distributions such as the uniform, Rayleigh, logistic, and log-logistic are special cases of Burr dfs.

A df W is said to belong to the Burr family if it satisfies the differential equation $\frac{dW(x)}{dx} = W(x)(1-W(x))h(x,W(x))$, where $h(x,W(x))$ is a non-negative function for x for which the function is increasing, $h(x,W(x))$ could be $h(x,W(x)) = \frac{h_1(x)}{W(x)}$ where $h_1(x) \geq 0$. Then $\frac{dW(x)}{dx} = (1-W(x))h_1(x)$.

The dfs $R_{k,F}$ belong to the Burr family.

Negative Binomial k-th extremes: Let N_n be a shifted negative binomial rv with $P(N_n = l) = \binom{l-m+r-1}{l-m} p_n^r q_n^{l-m}, r = m, m+1, m+2, \ldots,$ where $0 < p_n < 1, q_n = 1-p_n$ and $\lim_{n\to\infty} np_n = 1$.

If $F \in D_l(G)$ for some G, then for fixed integer $k, 1 \leq k \leq m$, $\lim_{n\to\infty} F_{k:N_n}(a_n x + b_n)$ is equal to

$$T_{k,G}(x) = \sum_{l=0}^{k-1} \binom{l+r-1}{l} \frac{(-\ln G(x))^l}{(1-\ln G(x))^{r+l}},\ x \in \{y \in \mathbb{R} : G(y) > 0\}.$$

The df $T_{k,F}$ satisfies the recurrence relation

$$T_{k+1,F}(x) = T_{k,F}(x) + \binom{k+r-1}{k} \frac{(-\ln F(x))^k}{(1-\ln F(x))^{k+r}},\quad k \geq 1, x \in \mathbb{R}.$$

Its pdf is $t_{k+1,F}(x) = \frac{1}{B(r,k+1)} \frac{f(x)}{F(x)} \frac{(-\ln F(x))^k}{(1-\ln F(x))^{r+k+1}},\quad k \geq 1, x \in \mathbb{R}.$

Let rv X have df F with pdf f and k be a fixed positive integer. Then for $T_{k,F}$, the following results are true.

- $T_{k,F}$ is a df with pdf $t_{k,F}(x) = \frac{1}{B(r,k)} \frac{f(x)}{F(x)} \frac{(-\ln F(x))^{k-1}}{(1-\ln F(x))^{r+k}},\ x \in \{y \in \mathbb{R} : F(y) > 0\}$, right extremity $r(T_{k,F}) = r(F)$, and $\lim_{x\to r(F)} \frac{1-T_{k,F}(x)}{(1-F(x))^k} = \frac{k}{B(r,k)}$.

- If $F \in D_l(\Phi_\alpha)$, then $r(T_{k,F}) = r(F) = \infty$, and $T_{k,F} \in D_l(\Phi_{k\alpha})$ with $a_n = F^-(1-(1/n)^{1/k}), b_n = 0$.
- If $F \in D_l(\Psi_\alpha)$ then $r(T_{k,F}) = r(F) < \infty$, and $T_{k,F} \in D_l(\Psi_{k\alpha})$ with $a_n = r(F) - F^-(1-(1/n)^{1/k}), b_n = r(F)$.
- If $F \in D_l(\Lambda)$, $a_n = v(b_n)$ and $b_n = F^-\left(1-\frac{1}{n}\right)$ then $r(T_{k,F}) = r(F)$, and $T_{k,F} \in D_l(\Lambda)$ with $a_n = \frac{v(b_n)}{k}, b_n = T^-_{k,F}(1-1/n)$.

3. Conclusion

In this article, tail behaviour of several interesting tails are explored through the concept of tail equivalence which simplifies several proofs. After recalling the definition of tail equivalence and clarifying some simple properties of tail equivalence, the article explores tail behaviour of mixtures of dfs and the limit laws of linearly normalised k upper order statistics of a random sample of size n, when n is fixed and n is replaced by Uniform, Geometric and Negative Binomial random sample sizes. Several results stated here can be used to simulate random observations from a variety of tails.

Acknowledgements

I sincerely thank Professor Vinod Kumar Gupta, Chair Editor, for inviting me to contribute to this Proceedings of the Annual Conference.

Conflict of interest

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

References

Burr, I. W. (1942). Cumulative frequency distributions. *Annals of Mathematical Statistics*, **13**, 215-232.

Nair, J., Wierman, A., and Zwart, B. (2023). *The Fundamentals of Heavy Tails.* Cambridge University Press.

Praveena, A. S. and Ravi, S. (2023). On the exponential max-domain of the standard log-Frechet distribution and subexponentiality. *Sankhya Series A,* **85**, 1607-1622.

Praveena, A. S. and Ravi, S. (2025). On standard log-Frechet and log-log-Frechet laws. *Mathematical Methods of Statistics,* **34**, 55-67.

Praveena, A. S., Srinath, S., and Ravi, S. (2019). On mixture distributions and their max domains. *Journal of the Indian Society for Probability and Statistics*, **20**, 173-183.

Ravi, S. and Manohar, M. C. (2018). On tail behaviour of k-th upper order statistics under fixed and random sample sizes via tail equivalence. *Statistics,* **52**, 156-176.

Resnick, S. I. (1971). Tail equivalence and its applications. *Journal of Applied Probability*, **8**, 136-156.

Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025; pp 51-66

Advances in Stepped Wedge Design: A Comprehensive Review

Soumadeb Pain and Satya Prakash Singh

Department of Mathematics and Statistics
Indian Institute of Technology Kanpur, 208016, India

Received: 12 June 2025; Revised: 23 June 2025; Accepted: 26 June 2025

Abstract

Stepped wedge designs (SWDs) are increasingly gaining popularity in cluster randomized trials. This review provides a comprehensive overview of stepped wedge cluster randomised trials (SW–CRTs), beginning with their historical development and rationale in the Introduction. We classify and compare different types of stepped wedge designs, highlighting their relative advantages and practical considerations. We then examine the primary statistical models used for analysis, key approaches to sample size determination, and the impact of various intracluster and temporal correlation structures on trial inference. Special attention is given to trials with unequal cluster sizes, addressing design adaptations and efficiency implications. We review recent advances in Bayesian optimal design strategies for SW–CRTs and extend the discussion to include adaptations for non-normal outcome data. As an alternative framework, we explore the staircase design, comparing its logistical and analytical features with those of traditional stepped wedge trials.

Key words: Staircase design; Stepped wedge trials; Cluster randomized trials; Optimal design; Bayesian design.

1. Introduction

Cluster Randomized Trials (CRTs) are pivotal in evaluating interventions where groups, rather than individuals, are randomized. Statistical methods for CRTs have been the focus of extensive research over the past several decades and are well-documented in various methodological reviews [Donner and Klar (2000), Turner *et al.* (2017a), and Turner *et al.* (2017b)]. Among CRT designs, the stepped wedge cluster randomized trial (SW-CRT) has garnered increasing attention, alongside traditional parallel and crossover CRTs [Brown and Lilford (2006), Mdege *et al.* (2011), Hemming *et al.* (2015)]. While parallel designs randomize clusters to fixed intervention or control arms and crossover designs alternate clusters between arms over time, the SW-CRT employs a unidirectional roll-out of all the clusters from control to intervention in sequential "steps". A SW–CRT comparing control and intervention conditions is illustrated in Table 1. The order in which the different individuals

Corresponding Author: Satya Prakash Singh
Email: singhsp@iitk.ac.in

or clusters receive the intervention is random. The study continues until all the clusters are assigned to the intervention and outcome data will be collected from each cluster. The SW design offers a versatile framework for CRTs, particularly in public health and service delivery. SW–CRTs offer several distinctive advantages including:

- **Evaluation During Rollout:** SW–CRTs are particularly useful for assessing the community level effectiveness of an intervention while it is being gradually rolled out across clusters.

- **Acceptability (Social, Political, Ethical):** Since all clusters eventually receive the intervention, this design is often more acceptable to stakeholders, especially in contexts where withholding a potentially beneficial intervention may be controversial.

- **Logistical and Financial Feasibility:** In many cases, it is not practical due to resource, personnel, or policy constraints to implement the intervention across all units simultaneously. SW–CRTs accommodate such staged implementation.

- **Statistical Efficiency:** Because each cluster serves as its own control at different time points, this design can increase statistical efficiency and may require fewer clusters compared to parallel-arm trials.

While its ethical and practical benefits are significant, researchers must address analytical complexities and potential biases. Hemming and Taljaard (2020) discussed several key factors that should be considered when implementing a SW-CRT.

The Gambia Hepatitis Intervention Study (Hall *et al.* (1987)) is the first ever reported stepped wedge trial which is also the longest running. The study was set up in 1986 to investigate whether vaccination against hepatitis B in infancy could reduce the risk of liver cancer over the next 30 to 40 years of life. The usefulness of SW-CRTs was recognized only later, but they are now highly regarded and widely used in medical research, as demonstrated by a recent review by Varghese *et al.* (2025), which examines studies published in high-impact journals. Some previous reviews on SW–CRTs include Brown and Lilford (2006) and Beard *et al.* (2015).

This paper focuses on reviewing recent statistical developments related to the design of SW-CRT trials. In doing so, we also introduce several variants of SW-CRTs that are widely used in current practice. The organization of this article is as follows: Section 2 reviews types of stepped wedge designs; Section 3 covers models, sample size determination and correlation structures; Section 4 examines unequal cluster sizes; Section 5 presents Bayesian optimal SW designs; Section 6 addresses non-normal outcome data; Section 7 introduces the staircase design as an alternative to SWD; and Section 8 provides discussion and conclusions.

2. Types of stepped wedge designs

Following (Copas *et al.*, 2015), SW–CRTs can be broadly classified in three categories: (i) Cohort SW–CRTs; (ii) Cross–sectional; and (iii) Continuous recruitment SW–CRTs.

2.1. Cohort SW–CRTs

When researchers track a group of individuals over time and assess their health outcomes at regular intervals, this is referred to as a cohort study design. A cohort design further classified in two categories: (a) Closed cohort SW–CRTs and (b) Open cohort SW–CRTs. In a stepped wedge trial, where the same clusters are revisited at different time points, it is common to measure outcomes for some or all of the same individuals on multiple occasions some before and some after the intervention is introduced. When the measurements are taken from same individuals at each time point in a cluster, the design is known as a closed cohort stepped wedge trial (Li *et al.* (2018a), Li (2020), Gasparini *et al.* (2025)). While this is a reasonably appropriate design, it may not always be realistic. In practice, it is more likely that participants may enter or leave the study over time. So, when some individuals are the same and others differ across measurement periods within a cluster, the design is referred to as an open cohort stepped wedge trial (Copas *et al.* (2015), Kasza *et al.* (2020)). Here, the term "open cohort" reflects the natural flow of participants in and out of the study population.

2.2. Cross–sectional SW–CRT

Sometimes clusters that are very large or densely populated, where researchers do not try to measure outcomes for everyone, but instead select a small, random sample at each visit. In that case, the chance of observing the same person twice is minimal. At this point, no individual is followed longitudinally across multiple time points (steps). Instead, each sample taken from a cluster represents a snapshot or cross-section of the population at that specific time. This type of design is known as a repeated cross-section stepped wedge trial (Hussey and Hughes (2007), Martin *et al.* (2019), Thompson *et al.* (2017)). It is relatively logistically simpler than cohort SW, since we do not need to track the same individuals over months, or in other words, we do not have to deal with individual's auto correlation.

2.3. Continuous recruitment SW–CRTs

In aforementioned SW–CRTs, we have described scenarios where extending the duration of a cluster randomized trial involves returning to the same clusters multiple times to collect outcome data. But what if, instead, participants enter and exit the trial in a continuous flow like an ongoing stream of eligible individuals? In this case, extending the trial simply means recruiting over a longer period of time, allowing more people from that continuous stream to be included. This type of design is known as a continuous recruitment stepped wedge trial (Hooper and Copas (2019), Hooper *et al.* (2020)). An example of a this type of stepped wedge trial is the Gambia Hepatitis Intervention Study discussed in the Introduction section. In this study, new eligible participants (newborn infants) arrived at a fairly steady rate, as is natural. In any continuous one year period of recruitment the researchers expected to recruit around 30,000 children into the study, and by scheduling the trial over a total of four years they hoped to see 120,000.

Table 1: A T–period stepped wedge design with $T-1$ sequences, comparing the control (in grey) and treatment (in blue).

Sequence	Time					
	0	**1**	**2**	**3**	**4**	**5**
1						
2						
3						
4						
5						

Cluster unexposed to intervention Cluster exposed to intervention

3. Models, sample size determination and correlation structures in SW–CRTs

This section reviews key developments in the statistical methodology of SW-CRTs. We begin by exploring various modeling structures, followed by a review of sample size determination methods and the different forms of correlation, a key aspect of SW–CRTs.

3.1. Models

Broadly, SW–CRTs are analyzed using conditional (cluster/subject/time specific) and marginal (population average) models. Conditional models, commonly implemented via linear mixed effects models (LMMs) or generalized linear mixed-effects models (GLMMs), account for clustering through random effects and estimate intervention effects conditional on these latent cluster/subject/time–level factors. Conditional models are utilized by Hussey and Hughes (2007), Hughes *et al.* (2015), Hooper *et al.* (2016), Girling and Hemming (2016), Kasza *et al.* (2019), Kasza and Forbes (2019), and Hemming *et al.* (2018).

Marginal models, typically fitted via generalized estimating equations (Liang and Zeger (1986)), directly target the population average treatment effect, offering robust inference even under correlation structure misspecification. Within each framework, a variety of correlation structures such as exchangeable, nested exchangeable, and exponential decay have been proposed to capture within and between period intraclass correlations (ICCs). Marginal models are used by Hussey and Hughes (2007), Li *et al.* (2018b), Ford and Westgate (2020), Li (2020), Thompson *et al.* (2021), and Li *et al.* (2022).

The choice between conditional and marginal approaches affects both interpretation and efficiency. There has existed controversy about the use of marginal and conditional models. Lee and Nelder (2004) discussed the advantages of conditional models over marginal models and regarded the conditional model as fundamental, from which marginal predictions can be made. Various models employed in SW design are thoroughly discussed in the review paper by Li and Wang (2022).

3.2. Sample size determination

3.2.1. Foundational work

A critical aspect of SWDs is sample size calculation, which has seen significant methodological advancements. The seminal paper by Hussey and Hughes (2007) introduced analytical formulas for power and sample size calculations in SWDs. The model proposed by Hussey and Hughes (2007), though foundational for sample size estimation in SWDs, has notable limitations. First, it assumes a cross-sectional design with no repeated measurements on individuals, rendering it unsuitable for cohort-based studies where individual correlation must be accounted for (Hooper *et al.*, 2016). Second, it presumes constant cluster sizes and a simplistic intracluster correlation (ICC) structure, ignoring variability in cluster sizes (Matthews, 2020) or more complex correlation patterns (*e.g.*, decaying correlations over time). Third, the model assumes fixed time effects and a constant intervention effect, failing to accommodate time-varying treatment effects or interactions between time and intervention exposure (Kenny *et al.*, 2022). Finally, it is restricted to continuous outcomes and does not generalize readily to binary, count, or survival data without modification (Zhou *et al.*, 2020).

3.2.2. Sample size calculations based on design effects

The design effect quantifies the increase in variance of an estimator due to deviations from a simple random sampling design. In cluster-based studies, it accounts for correlations within clusters, which reduces the effective sample size. The standard approach to calculating sample size in parallel group CRTs begins with estimating the required sample size under individual randomization, denoted as N_u. This unadjusted sample size is then scaled by the design effect $[1+(n-1)\rho]$ to account for clustering, where n is the number of individuals per cluster and ρ is the intracluster correlation coefficient (Donner and Klar, 2000). To adopt a similar framework, Woertman *et al.* (2013) derived the following design effect for SWDs:

$$DE_{sw} = \frac{1+\rho(ktn+bn-1)}{1+\rho\left(\frac{1}{2}ktn+bn-1\right)}\frac{3(1-\rho)}{2t\left(k-\frac{1}{k}\right)}.$$

Here, k represents the number of steps, b is the number of baseline measurements, and t is the number of measurements after each step. Thus, each cluster is measured $(b+kt)$ times. This design effect appropriately adjusts for both clustering and the stepped wedge structure and the required sample size for a stepped wedge trial is $N_{sw} = N_u DE_{sw}$.

The design effect DE_{sw} is influenced by three key parameters: the number of post–step measurements t, the number of baseline measurements b, and the number of steps k. Increasing any of these reduces the design effect and, consequently, the required sample size. In contrast, increasing the cluster size n slightly increases the design effect. Additionally, DE_{sw} depends on the intracluster correlation coefficient (ICC), ρ, which reflects variability between clusters. While ρ is contextd ependent and not under direct control, it should be estimated using prior studies, pilot data, or domain knowledge. As ρ increases, the design effect initially rises and then begins to decline. Woertman *et al.* (2013) have shown that increasing the number of steps improves efficiency in terms of sample size and also the gain is substantially larger when increasing from 2 to 3 steps than from 6 to 12 steps.

3.2.3. Simulation–based sample size calculations

Analytical sample size formulas, while computationally efficient, are often constrained by simplifying assumptions that limit their applicability in real world SW–CRTs. For example, Hussey and Hughes (2007), Woertman *et al.* (2013) *etc* assumed balanced design and intervention effect is modeled as constant across clusters. Also analytical formulas work well for continuous outcomes but struggle with binary or count outcomes Xia *et al.* (2021) or when repeated measures are taken on the same individuals over time, due to the additional level of correlation implied in this case. Simulation–based sample size calculation has emerged as a robust and flexible approach for designing SW-CRT, particularly when analytical formulas are insufficient due to the complexity of the design or outcome types. Simulation-based sample size calculation typically follows these steps:

- Define the Data-Generating Model: Specify fixed effects (*e.g.*, intervention effect, time trends) and random effects (*e.g.*, cluster-level or time-level variability).
- Simulate Datasets: Generate repeated datasets under the assumed model, incorporating design parameters (for example number of clusters, steps, and observations per cluster).
- Analyze Simulated Data: Apply the planned statistical method (*e.g.*, mixed-effects regression) to each dataset, estimate the intervention effect and its standard error and p-value and then record the proportion of simulations where the intervention effect is statistically significant (empirical power).
- Iterate Until Target Power is Achieved: Adjust parameters (*e.g.*, cluster size or number of steps) and repeat simulations until the desired power is reached

But at the same time there are challenges in simulation-based sample size calculation. Simulations require significant computational resources, especially for large trials or complex models. Analyzing thousands of datasets with complex models can be slow. Parallel computing is often essential and also complex models may fail to converge in some simulations. Further, results depend on accurate pre–specification of nuisance parameters (*e.g.*, ICC), which may be uncertain in practice.

3.3. Correlation structures

Stepped wedge designs inherently involve longitudinal and clustered data, leading to multiple correlation structures that complicate statistical analysis. These structures arise from repeated measurements within clusters over time, participant-level dependencies in cohort designs, and temporal trends.

3.3.1. Correlation parameters in SW–CRTs

Hemming *et al.* (2015) incorporated both within-period and between-period ICCs in their sample size calculation for cross-sectional designs. Hooper *et al.* (2016) and Li *et al.* (2018b) extended this by considering a three correlation structure that also accounts for within individual repeated measurements in closed cohort designs.

- Within period intracluster correlation (wp–ICC): Measures similarity of outcomes within the same cluster and time period.

- Between period intracluster correlation (bp–ICC): Captures correlation between outcomes from the same cluster across different periods.

- Individual level autocorrelation: Relevant in closed cohort designs where the same participants are measured repeatedly.

For example, in cross sectional designs (new participants each period), wp–ICC and bp–ICC dominate, while closed cohort designs require an additional parameter for individual autocorrelation. Ignoring these distinctions can lead to biased variance estimates and underpowered studies (Girling and Hemming, 2016).

Table 2: Different types of correlation structures in SW design ($0 < r < 1$ is any constant value)

(a) Constant ICC over time: within period ICC = between period ICC
Used in Hussey and Hughes (2007) − no decay

	Period 1	Period 2	Period 3	Period 4	Period 5
Period 1	ρ	ρ	ρ	ρ	ρ
Period 2		ρ	ρ	ρ	ρ
Period 3			ρ	ρ	ρ
Period 4				ρ	ρ
Period 5					ρ

(b) Fixed between period ICC and within period ICC > between period ICC
Used in Hooper *et al.* (2016)− no decay

	Period 1	Period 2	Period 3	Period 4	Period 5
Period 1	ρ	$r\rho$	$r\rho$	$r\rho$	$r\rho$
Period 2		ρ	$r\rho$	$r\rho$	$r\rho$
Period 3			ρ	$r\rho$	$r\rho$
Period 4				ρ	$r\rho$
Period 5					ρ

(c) Between ICCs decay exponentially and within period ICC > between period ICC
Used in Kasza and Forbes (2019)− allows decay

	Period 1	Period 2	Period 3	Period 4	Period 5
Period 1	ρ	$r\rho$	$r^2\rho$	$r^3\rho$	$r^4\rho$
Period 2		ρ	$r\rho$	$r^2\rho$	$r^3\rho$
Period 3			ρ	$r\rho$	$r^2\rho$
Period 4				ρ	$r\rho$
Period 5					ρ

3.3.2. Modeling decaying correlation structures

In longitudinal studies, "decay in correlation over time" refers to the phenomenon where the correlation between two measurements decreases as the time interval between them increases. In other words, observations made closer together in time tend to be more similar (more correlated) than observations that are farther apart.

The assumption of constant ICC, as used by Hooper *et al.* (2016) and Li *et al.* (2018b), may not reflect real world data structures. Therefore, alternative design and analysis strategies that account for temporal correlation decay are essential in stepped wedge trials. In cross sectional designs, where different individuals are observed in each period, some studies (Hemming *et al.*, 2015) allowed between period ICC to differ from within period ICC but assumed constancy across time. Kasza *et al.* (2019), Kasza and Forbes (2019) introduced a nonuniform correlation model incorporating exponential decay, improving sample size estimation. Grantham *et al.* (2019) extended this to continuous time correlation decay in multiple periods CRTs with continuous recruitment. Ignoring correlation decay, as shown by Kasza *et al.* (2019), can lead to misestimate intervention effects and incorrect sample size calculations.

4. SW–CRTs with unequal cluster size

Methods for calculating power and sample size in SW-CRTs assuming equal cluster sizes have been extensively discussed in the literature see for example Hussey and Hughes (2007), Woertman *et al.* (2013), Baio *et al.* (2015), and Hemming and Taljaard (2016). In many studies, such as observational studies, unequal cluster sizes are a common occurrence. This presents significant challenges in the design and analysis of SW–CRTs. A comprehensive methodological review addressing unequal cluster sizes in cluster randomized trials, including SW-CRTs, is provided in Zhan *et al.* (2021b). The impact of cluster size imbalance on the power is discussed in Ouyang *et al.* (2020). Martin *et al.* (2019) examined how randomly allocating clusters of varying sizes to sequences impacts different aspects of the analysis. They investigated cluster–balanced stepped wedge designs (SWDs) with unequal cluster sizes and observed that, when the total number of individuals is fixed, such designs can be more efficient than those with equal cluster sizes. This finding contrasts with traditional cluster balanced designs, where equal sized clusters are typically considered optimal. Girling (2018) investigate the impact of unequal cluster size and found the expressions for the relative efficiency (RE) of the treatment effect estimate relative to that for the equal cluster design with the same total number of observations. Matthews (2020) proposed near optimal designs for unequal cluster size. Kristunas *et al.* (2017) proposed corrections to the design effect(DE) for SWD with unequal cluster sizes. Girling (2018) investigate the impact of unequal cluster size and found the expressions for the relative efficiency (RE) of the treatment effect estimate relative to that for the equal cluster design with the same total number of observations. Using simulations Martin *et al.* (2019) showed that the while the average power reduction in SW-CRTs is smaller than in parallel designs, the variance in power across allocations is higher, particularly with fewer clusters.

Typically, larger clusters are assigned to the extreme sequences. However, this pattern may not hold in closed-cohort stepped wedge designs (SWDs), where optimal allocation depends on various correlation parameters. In a working paper, we observed that an efficient

design tends to allocate an equal number of clusters to sequences i and $T-1-i$ for $i = 1, \ldots, T-1$. A similar symmetry is also observed in the allocation of total cluster size across these sequences. The determination of optimal cluster–to–sequence proportions in the context of unequal cluster sizes remains an area that requires further detailed investigation.

5. Bayesian optimal SWD

An optimal design is obtained by optimizing a specific criterion. For example, by minimizing the variance of the estimated treatment effect or by maximizing the study's power or precision. Lawrie *et al.* (2015) found out optimal allocation of clusters into sequences under the linear mixed effect model given by Hussey and Hughes (2007) by minimizing the variance of the treatment effect. They demonstrated that when cluster sizes are equal, the extreme sequences (first and last) receive the same level of allocation, while all intermediate sequences receive an equal but smaller allocation compared to the extremes. This work is then extended to closed cohort SW–CRT designs with repeated measures per subject by Li *et al.* (2018a). Thompson *et al.* (2017) examined the optimal structure of stepped wedge cluster randomized trial (SW–CRT) designs under the assumption of normally distributed data and equal allocation of clusters across sequences. In contrast, Zhan *et al.* (2018) explored optimal designs where some clusters may not be sampled during certain stages of the trial. Optimal design, thus obtained, is called locally optimized design as they are sensitive to the choice of different correlation parameters. More recently, to obtain a robust optimal design a Bayesian approach is adopted. Zhan *et al.* (2021a) demonstrated that incorporating prior information on time effects through a Bayesian approach can significantly reduce the required sample size. However, due to the risk of bias from mis-specified prior distributions, they do not recommend this as the default method for sample size calculation. Nevertheless, when it is difficult to recruit enough clusters or participants, using external information on time effects with a Bayesian approach can help assess if a smaller sample size would still be sufficient, making it easier to decide whether the trial can go ahead. Singh (2024) proposed a Bayesian optimal SWD by placing priors on the ICC and demonstrating robustness against ICC misspecification compared to locally optimal designs. Under a marginal (GEE) model with either exchangeable or exponential-decay working correlation, Etfer *et al.* (2024) developed a framework for finding Bayesian D-optimal SW designs for binary outcomes.

Bayesian designs for stepped wedge trials remain a significantly underdeveloped area of research with considerable potential. For instance, in closed–cohort studies, the presence of multiple correlation parameters introduces substantial uncertainty in the design process. This challenge can be effectively addressed by adopting a Bayesian framework. Moreover, in the case of non–normal responses where the optimal design criteria depend on unknown model parameters, a Bayesian approach can offer substantial advantages.

6. SW design for non–normal data

In recent years, a substantial body of work has extended the SW-CRT framework to accommodate non-normal outcomes, most notably binary and count data through a variety of methodological and practical innovations. Stepped-wedge trials with non-normal outcomes (counts or binary) extend the usual mixed-effects framework by replacing the linear mixed model with a generalized linear mixed model (GLMM) or generalized estimating equations (GEE). Broadly, for binary outcomes one uses logistic-link GLMMs or marginal GEE,

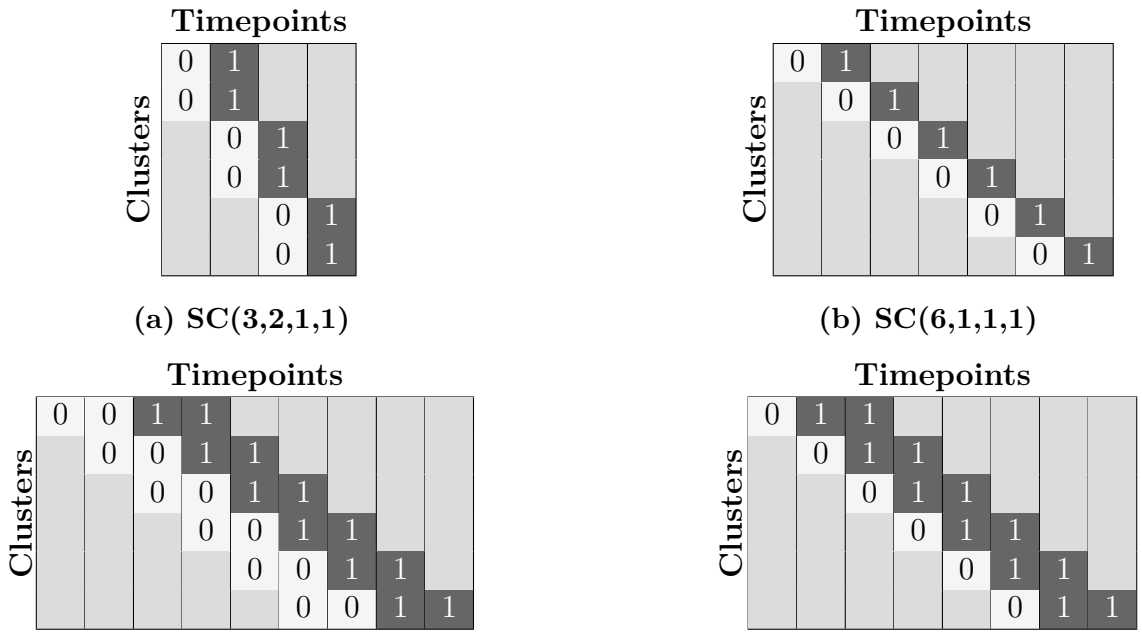
whereas for counts one adopts Poisson (or negative-binomial) GLMMs. Zhou *et al.* (2020) developed a numerical method for the power analysis for stepped-wedge cluster randomized trials (SW-CRTs) with binary outcomes, utilizing a maximum-likelihood estimation framework. Their approach allows researchers to assess the statistical power of complex SW-CRT designs without relying on simplified analytical approximations, making it particularly useful for settings with unequal cluster sizes or varying intraclass correlations.Wang *et al.* (2021) found out a sample size and power calculation method using GEE that can be broadly applied to both closed-cohort and cross-sectional SW-CRTs with binary outcomes. Also, they introduced a correction method to address the problem of underestimated variance in the GEE approach when the number of clusters is small in SW-CRTs. Building on the Laplace approximation of Breslow and Clayton (1993), Xia *et al.* (2021) have derived an analytical variance formula for the intervention effect estimator using GLMM, encompassing both normal (identity link) and non-normal (*e.g.*, logistic, Poisson) outcomes. Lastly as mentioned in the previous section, Etfer *et al.* (2024) develop a Bayesian D-optimal design framework for stepped-wedge cluster randomized trials with binary outcomes by combining generalized estimating equations and approximate design theory under both exchangeable and exponential-decay correlation structures.

7. Staircase design: An alternative to SWD

Stepped wedge designs require clusters to collect data across all trial periods, leading to high logistical and financial burdens. A staircase design is an "incomplete" variant of the stepped-wedge, in which each cluster contributes data only for a small number of periods immediately before and after its switch from control to intervention. The staircase design was first formalized by Grantham *et al.* (2024), who noted that the most informative observations in a stepped-wedge lie along its main diagonal (the "zigzag" of switches) and proposed focusing data collection there only. Like a stepped-wedge, all clusters eventually receive the intervention and the rollout is staggered; unlike a complete stepped-wedge, clusters do not collect data in every period, reducing burden and potentially attrition.

The general class of staircase designs is denoted by $SC(S, K, R_0, R_1)$, where S and R_0 denote the number of distinct treatment sequences and the number of clusters per sequence, R_0 is the number of control periods before the switch to intervention, and R_1 is the number of intervention periods after the switch. Different types of staircase designs can be achieved depending on the values of R_0 and R_1 (see Figure 1). In total, the design includes SK clusters, and the trial spans $S + R_0 + R_1 - 1$ periods. Clusters in sequence s are observed from period s through $s + R_0 + R_1 - 1$. A balanced staircase design has equal numbers of control and intervention periods in each sequence (*i.e.* $R_0 = R_1$). In contrast, an imbalanced staircase design allows for different numbers of pre and post switch periods ($R_0 \neq R_1$).

Grantham *et al.* (2024) have derived explicit expressions for the variance of the generalized least squares estimator of treatment effect for the basic staircase design under the assumption that the observed periods in each sequence follow the same schedule of control and intervention periods. This expression can be used to calculate sample size and power for staircase designs. Grantham *et al.* (2025) examined the relative efficiency of the stepped wedge design compared to various forms of the basic staircase design, where each sequence consists of one control period followed by one intervention period. Their analysis began with a basic staircase design embedded within a stepped wedge framework, and extends to



(a) SC(3,2,1,1) **(b) SC(6,1,1,1)**

(c) SC(6,1,2,2) **(d) SC(6,1,1,2)**

Figure 1: Design schematics for several staircase designs with 6 clusters: a basic staircase with two clusters assigned to each of three unique sequences (top left), a basic staircase with one cluster assigned to each of six unique sequences (top right), a balanced staircase with two control periods followed by two intervention periods in each sequence and one cluster assigned to each of six unique sequences (bottom left), and an imbalanced staircase with one control period followed by two intervention periods in each sequence and one cluster assigned to each of six unique sequences (bottom right).

versions with either more clusters or larger cluster-period sizes–some maintaining the same total number of participants as the stepped wedge design, and others using fewer participants overall. The relative efficiency of these designs is influenced by the intracluster correlation structure, correlation parameters, and trial configuration, including the number of sequences and the size of each cluster–period. They concluded that basic staircase design is a particularly lean and potentially powerful alternative to the stepped wedge design as across a broad range of realistic trial scenarios, the basic staircase design often provides greater statistical power than the stepped wedge design, even when using the same or even fewer total participants. A comprehensive analysis of staircase design including optimal cluster proportion to the sequences, appropriate cluster sizes, and Bayesian design strategies should be thoroughly explored.

8. Discussion and conclusion

In this review we have traced the evolution of stepped wedge designs (SWDs) from the foundational Hussey–Hughes random-intercept model through modern Bayesian and "staircase" alternatives. A recurring theme is the trade–off between analytical simplicity and realistic correlation structure. Early formulas for power and sample size assume constant ICCs, cross-sectional sampling, and equal cluster sizes; these yield closed-form design effects

but can mislead when within- and between-period correlations differ or clusters vary in size. Extensions to cohort and open-cohort SWDs introduced additional ICC parameters and decay models, but at the cost of analytical intractability. Simulation-based approaches remedy this at the expense of computational burden and reliance on assumed nuisance parameters. Our survey of optimal-allocation methods highlights how design efficiency depends critically on the allocation of clusters to sequences. Bayesian D-optimal frameworks then add robustness by placing priors on ICC or time-effect parameters, reducing required sample size when external information is reliable but risking bias under prior misspecification. Lastly, the staircase design represents a pragmatic compromise: by sampling only around each cluster's switch point, it retains most information on treatment contrasts while cutting data-collection burden. Across a broad range of ICC scenarios, basic staircase trials can even outperform full SWDs in power per participant.

Despite these advances, several gaps remain. First, most methods target continuous outcomes; extensions to binary, count or time–to–event endpoints require further development. Second, while correlation-decay models are conceptually appealing, real world validation via intensive pilot data or retrospective re-analysis of completed SWDs remains scarce. Third, the increasing complexity of hybrid designs (*e.g.* unequal cluster sizes, open cohorts, Bayesian priors) calls for user friendly software that integrates power, sample size, and optimal allocation routines under a unified interface. Finally, practical considerations such as staggered enrollment logistics, missing data, and secular trends–deserve more attention in design-stage simulations.

In sum, the stepped wedge framework has matured from simple cross sectional formulas to a rich design space encompassing complex correlation structures, Bayesian robustness, and lean staircase variants. The choice among these should be driven by the substantive context–outcome type, anticipated ICC patterns, logistical constraints and cluster sizes. In this review paper, we not only mention a few relevant works in various field of SW design but also explain fundamental terminologies related to this design in a concise manner, aiming to assist readers who are encountering these concepts for the very first time. We hope this introductory yet informative overview provides a solid foundation for further exploration into the field of SW design.

Acknowledgements

The work of Soumadeb Pain is funded through an IIT Kanpur assistantship, funded by the Ministry of Education(MoE), Govt. of India. The work of Satya Prakash Singh is funded by Science and Engineering Research Board, Grant/Award Number: MTR/2022/000627. We are indeed grateful to the Editors for their guidance and counsel. We also sincerely appreciate the reviewer's valuable comments, insightful suggestions, and the generous inclusion of numerous 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

Baio, G., Copas, A., Ambler, G., Hargreaves, J., Beard, E., and Omar, R. Z. (2015). Sample size calculation for a stepped wedge trial. *Trials*, **16**.

Beard, E., Lewis, J. J., Copas, A., Davey, C., Osrin, D., and Baio, G. e. a. (2015). Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. *Trials*, **16**.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Brown, C. A. and Lilford, R. J. (2006). The stepped wedge trial design: a systematic review. *BMC Medical Research Methodology*, **6**.

Copas, A. J., Lewis, J. J., Thompson, J. A., Davey, C., Baio, G., and Hargreaves, J. R. (2015). Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. *Trials*, **16**, 1–12.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomised Trials in Health Research*. Arnold, London.

Etfer, L., Wason, J., and Grayling, M. J. (2024). Optimal Bayesian stepped-wedge cluster randomised trial designs for binary outcome data. *arXiv preprint arXiv:2402.09938*, **abs/2402.09938**, 1–15.

Ford, W. P. and Westgate, P. M. (2020). Maintaining the validity of inference in small-sample stepped wedge cluster randomized trials with binary outcomes when using generalized estimating equations. *Statistics in Medicine*, **39**, 2779–2792.

Gasparini, A., Crowther, M. J., Hoogendijk, E. O., Li, F., and Harhay, M. O. (2025). Analysis of cohort stepped wedge cluster-randomized trials with nonignorable dropout via joint modeling. *Statistics in Medicine*, **44**, e10347.

Girling, A. J. (2018). Relative efficiency of unequal cluster sizes in stepped wedge and other trial designs under longitudinal or cross-sectional sampling. *Statistics in Medicine*, **37**, 4652–4664.

Girling, A. J. and Hemming, K. (2016). Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Statistics in Medicine*, **35**, 2149–2166.

Grantham, K. L., Forbes, A. B., Hooper, R., and Kasza, J. (2024). The staircase cluster randomised trial design: a pragmatic alternative to the stepped wedge. *Statistical Methods in Medical Research*, **33**, 24–41.

Grantham, K. L., Forbes, A. B., Hooper, R., and Kasza, J. (2025). The relative efficiency of staircase and stepped wedge cluster randomised trial designs. *Statistical Methods in Medical Research*, **34**, 701–716.

Grantham, K. L., Kasza, J., Heritier, S., Hemming, K., and Forbes, A. B. (2019). Accounting for a decaying correlation structure in cluster randomized trials with continuous recruitment. *Statistics in Medicine*, **38**, 1918–1934.

Hall, A. J., Inskip, H. M., and Loik, F. e. a. (1987). The gambia hepatitis intervention study. *Cancer Research*, **47**, 5782–5787.

Hemming, K., Haines, T. P., Chilton, P. J., Girling, A. J., and Lilford, R. J. (2015). The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ*, **350**.

Hemming, K. and Taljaard, M. (2016). Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. *Journal of Clinical Epidemiology*, **69**, 137–46.

Hemming, K. and Taljaard, M. (2020). Reflection on modern methods: when is a stepped-wedge cluster randomized trial a good study design choice? *International Journal of Epidemiology*, **49**, 1043–1052.

Hemming, K., Taljaard, M., and Forbes, A. (2018). Modeling clustering and treatment effect heterogeneity in parallel and stepped-wedge cluster randomized trials. *Statistics in Medicine*, **37**, 883–898.

Hooper, R. and Copas, A. (2019). Stepped wedge trials with continuous recruitment require new ways of thinking. *Journal of Clinical Epidemiology*, **116**, 161–166.

Hooper, R., Kasza, J., and Forbes, A. (2020). The hunt for efficient, incomplete designs for stepped wedge trials with continuous recruitment and continuous outcome measures. *BMC Medical Research Methodology*, **20**, 1–9.

Hooper, R., Teerenstra, S., de Hoop, E., and Eldridge, S. (2016). Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Statistics in Medicine*, **35**, 4718–4728.

Hughes, J. P., Granston, T. S., and Heagerty, P. J. (2015). Current issues in the design and analysis of stepped wedge trials. *Contemporary Clinical Trials*, **45**, 55–60.

Hussey, M. A. and Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, **28**, 182–191.

Kasza, J. and Forbes, A. B. (2019). Inference for the treatment effect in multiple-period cluster randomised trials when random effect correlation structure is misspecified. *Statistical Methods in Medical Research*, **28**, 3112–3122.

Kasza, J., Hemming, K., Hooper, R., Matthews, J., and Forbes, A. (2019). Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. *Statistical Methods in Medical Research*, **28**, 703–716.

Kasza, J., Hooper, R., Copas, A., and Forbes, A. B. (2020). Sample size and power calculations for open cohort longitudinal cluster randomized trials. *Statistics in Medicine*, **39**, 1871–1883.

Kenny, A., Voldal, E. C., Xia, F., Heagerty, P. J., and Hughes, J. P. (2022). Analysis of stepped wedge cluster randomized trials in the presence of a time-varying treatment effect. *Statistics in Medicine*, **41**, 4311–4339.

Kristunas, C. A., Smith, K. L., and Gray, L. J. (2017). An imbalance in cluster sizes does not lead to notable loss of power in cross-sectional, stepped-wedge cluster randomised trials with a continuous outcome. *Trials*, **18**, 1–11.

Lawrie, J., Carlin, J. B., and Forbes, A. B. (2015). Optimal stepped wedge designs. *Statistics and Probability Letters*, **99**, 210–214.

Lee, Y. and Nelder, J. A. (2004). Conditional and marginal models: another view. *Statistical Science*, **19**, 219–238.

Li, F. (2020). Design and analysis considerations for cohort stepped wedge cluster randomized trials with a decay correlation structure. *Statistics in Medicine*, **39**, 438–455.

Li, F., Turner, E. L., and Preisser, J. S. (2018a). Optimal allocation of clusters in cohort stepped wedge designs. *Statistics and Probability Letters*, **137**, 257–263.

Li, F., Turner, E. L., and Preisser, J. S. (2018b). Sample size determination for gee analyses of stepped wedge cluster randomized trials. *Biometrics*, **74**, 1450–1458.

Li, F. and Wang, R. (2022). Stepped wedge cluster randomized trials: a methodological overview. *World Neurosurgery*, **161**, 323–330.

Li, F., Yu, H., Rathouz, P. J., Turner, E. L., and Preisser, J. S. (2022). Marginal modeling of cluster-period means and intraclass correlations in stepped wedge designs with binary outcomes. *Biostatistics*, **23**, 772–788.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Martin, J. T., Hemming, K., and Girling, A. (2019). The impact of varying cluster size in cross–sectional stepped-wedge cluster randomised trials. *BMC Medical Research Methodology*, **19**.

Matthews, J. N. S. (2020). Highly efficient stepped wedge designs for clusters of unequal size. *Biometrics*, **76**, 1167–1176.

Mdege, N. D., Man, M.-S., Taylor, C. A., and Torgerson, D. J. (2011). Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *Journal of Clinical Epidemiology*, **64**, 936–948.

Ouyang, Y., Karim, M. E., Gustafson, P., Field, T. S., and Wong, H. (2020). Explaining the variation in the attained power of a stepped-wedge trial with unequal cluster sizes. *BMC Medical Research Methodology*, **20**.

Singh, S. P. (2024). Bayesian optimal stepped wedge design. *Biometrical Journal*, **66**, 2300168.

Thompson, J., Hemming, K., Forbes, A., Fielding, K., and Hayes, R. (2021). Comparison of small sample standard error corrections for generalised estimating equations in stepped wedge cluster randomised trials with a binary outcome: a simulation study. *Statistical Methods in Medical Research*, **30**, 425–439.

Thompson, J. A., Fielding, K., Hargreaves, J., and Copas, A. (2017). The optimal design of stepped wedge trials with equal allocation to sequences and a comparison to other trial designs. *Clinical Trials*, **14**, 639–647.

Turner, E. L., Li, F., Gallis, J. A., Prague, M., and Murray, D. M. (2017a). Review of recent methodological developments in group-randomized trials: part 1—design. *American Journal of Public Health*, **107**, 907–915.

Turner, E. L., Prague, M., Gallis, J. A., Li, F., and Murray, D. M. (2017b). Review of recent methodological developments in group-randomized trials: part 2—analysis. *American Journal of Public Health*, **107**, 1078–1086.

Varghese, E., Briola, A., Kennel, T., Pooley, A., and Parker, R. A. (2025). A systematic review of stepped wedge cluster randomized trials in high impact journals: assessing the design, rationale, and analysis. *Journal of Clinical Epidemiology*, **178**, 111622.

Wang, J., Cao, J., Zhang, S., and Ahn, C. (2021). Sample size and power analysis for stepped wedge cluster randomised trials with binary outcomes. *Statistical Theory and Related Fields*, **5**, 162–169.

Woertman, W., de Hoop, E., Moerbeek, M., Zuidema, S. U., Gerritsen, D. L., and Teerenstra, S. (2013). Stepped wedge designs could reduce the required sample size in cluster randomized trials. *Journal of Clinical Epidemiology*, **66**, 752–758.

Xia, F., Hughes, J. P., Voldal, E. C., and Heagerty, P. J. (2021). Power and sample size calculation for stepped-wedge designs with discrete outcomes. *Trials*, **22**, 1–10.

Zhan, D., Ouyang, Y., Xu, L., and Wong, H. (2021a). Improving efficiency in the stepped–wedge trial design via Bayesian modeling with an informative prior for the time effects. *Clinical Trials*, **18**, 295–302.

Zhan, D., Xu, L., Ouyang, Y., Sawatzky, R., and Wong, H. (2021b). Methods for dealing with unequal cluster sizes in cluster randomized trials: A scoping review. *PLoS ONE*, **16**, e0255389.

Zhan, Z., de Bock, G. H., and van den Heuvel, E. R. (2018). Optimal unidirectional switch designs. *Statistics in Medicine*, **37**, 3573–3588.

Zhou, X., Liao, X., Kunz, L. M., Normand, S.-L. T., Wang, M., and Spiegelman, D. (2020). A maximum likelihood approach to power calculations for stepped wedge designs of binary outcomes. *Biostatistics*, **21**, 102–121.

Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025; pp 67-75

Challenges in Flexible and Scalable Modeling of Point-referenced Spatial Data

Suman Guha
Department of Statistics, Presidency University, Kolkata

Received: 15 June 2025; Revised: 25 June 2025; Accepted: 27 June 2025

Abstract

Statistical analysis of point-referenced spatial/geostatistical data generally considers a multivariate Gaussian distribution as the underlying probability model. That way, the related statistical inference boils down to estimating the mean vector and the covariance matrix of some multivariate normal distribution. While a fully general specification of the covariance matrix yields a flexible model for the data, it introduces too many parameters for consideration, thereby rendering statistical inference impossible. Alternatively, one can use a parametric covariance function that aligns with the underlying data. This covariance function is then used to form the elements of the covariance matrix under consideration. Parametric covariance functions often rely on the assumption of isotropy, or if not so, at least assume stationarity. However, stationary covariance functions are inadequate for explaining the complex dependence structure of spatial data arising out of environmental applications. In this article, we review prominent approaches for the construction of non-stationary covariance functions. Once a suitable covariance function is selected, the next challenge that one faces is to carry out computation using that covariance function. Non-stationary covariance functions although flexibly capture the spatial dependence structure, model fitting with them requires $O(n^3)$ computation, which is impossible to commence if n is massive. Basis function-based construction of non-stationary covariance functions can reduce the computational cost by a large margin. Recently, the Vecchia approximation-based nearest-neighbour Gaussian process has gained popularity among applied researchers. In this article, we review these approaches and some more for the construction of scalable spatial covariance functions for point-referenced spatial data.

Key words: Geostatistical data; Non-stationary covariance function; Vecchia approximation; Scalable spatial models.

AMS Subject Classifications: 62M30.

1. Introduction

Point-referenced spatial/geostatistical data arises when observations are made at n spatial/geographical locations $\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n$. They are routinely encountered in a broad range

Corresponding Author: Suman Guha
Email: suman.stat@presiuniv.ac.in

of areas like environmental, meteorological, ecological, and economic studies. The observations $y(\mathbf{s}_1), y(\mathbf{s}_2), \cdots, y(\mathbf{s}_n)$ are generally scalar-valued, and can signify temperature, rainfall, ground-level ozone concentration, house price, *etc.* Often the observations are recorded from satellite-based images and as a result, n can be as large as of the order ~ 100000. In addition to the dependent nature of the background random variables $Y(\mathbf{s}_1), Y(\mathbf{s}_2), \cdots, Y(\mathbf{s}_n)$, a large value of n renders the statistical modeling to be doubly difficult. In fact, a large n implies that the number of parameters p required for model specification may also be large. So, it is a high dimension high sample size (HDHSS) problem that couples high dimension (large p) with big data (large n).

Statistical inference is often carried out assuming that $(Y(\mathbf{s}_1), Y(\mathbf{s}_2), \cdots, Y(\mathbf{s}_n))$ is distributed according to a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. An exception to Gaussianity is noted for spatial data associated with extreme events like daily maximum windspeed, daily maximum temperature (Huser and Wadsworth (2022)), extreme snow depth (Blanchet and Davison (2011)), *etc.*, which requires modeling with multivariate extreme value distribution. Apart from that, spatial data that are positive-valued, skewed (Ayalew *et al.* (2024)) with a possible heavy tail, are also modeled by multivariate skewed distributions (Hazra *et al.* (2020)). Nevertheless, for the majority of cases, the inference boils down to estimating the mean vector and the covariance matrix of a multivariate normal distribution. Generally, the mean vector $\boldsymbol{\mu} = (\mu(\mathbf{s}_1), \mu(\mathbf{s}_2), \cdots, \mu(\mathbf{s}_n))$ is assumed to be a function of spatial locations; for example, $\mu(\mathbf{s}_i) = \beta_0 + \beta_1 \mathbf{s}_{i,1} + \beta_2 \mathbf{s}_{i,2}$. Geostatistical data $y(\mathbf{s}_1), y(\mathbf{s}_2), \cdots, y(\mathbf{s}_n)$ is often accompanied by measurements on other spatially indexed covariates, and those covariates are included in the model by extending the formula of mean to $\mu(\mathbf{s}_i) = \beta_0 + \beta_1 \mathbf{s}_{i,1} + \beta_2 \mathbf{s}_{i,2} + \gamma_1 x_1(\mathbf{s}_i) + \gamma_2 x_2(\mathbf{s}_i)$. Unlike $\boldsymbol{\mu}$, the specification of $\boldsymbol{\Sigma}$ requires additional care. A fully general specification of $\boldsymbol{\Sigma}$ as $\begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ & & \vdots & \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}$ elicits a flexible covariance structure for the data, although it brings $\frac{n(n+1)}{2}$ parameters under consideration, thereby making statistical inference impossible. Note that, the number of data points is n, and hence, any meaningful specification of $\boldsymbol{\Sigma}$ must not exceed n parameters. One way of achieving that is to consider a parametric covariance function $c_Y(\mathbf{s}, \mathbf{s}')$ and use it to specify the spatial covariance matrix as $\boldsymbol{\Sigma} = \begin{pmatrix} c_Y(\mathbf{s}_1, \mathbf{s}_1) & c_Y(\mathbf{s}_1, \mathbf{s}_2) & \cdots & c_Y(\mathbf{s}_1, \mathbf{s}_n) \\ c_Y(\mathbf{s}_2, \mathbf{s}_1) & c_Y(\mathbf{s}_2, \mathbf{s}_2) & \cdots & c_Y(\mathbf{s}_2, \mathbf{s}_n) \\ & & \vdots & \\ c_Y(\mathbf{s}_n, \mathbf{s}_1) & c_Y(\mathbf{s}_n, \mathbf{s}_2) & \cdots & c_Y(\mathbf{s}_n, \mathbf{s}_n) \end{pmatrix}$. In that case, the estimation of $\boldsymbol{\Sigma}$ translates to the estimation of only a few unknown parameters associated with $c_Y(\mathbf{s}, \mathbf{s}')$.

A well-known and much-used parametric covariance function is the exponential covariance function defined as $c_Y(\mathbf{s}, \mathbf{s}') := \sigma^2 e^{-\phi \|\mathbf{s} - \mathbf{s}'\|_2}$. The two parameters σ^2 and $\phi > 0$ are used to specify the shape of the covariance function. σ^2 specifies the variance of the underlying spatial process $\{Y(\mathbf{s})\}$ and ϕ, which is the decay parameter, decides how strong the spatial correlation is between $Y(\mathbf{s})$ and $Y(\mathbf{s}')$. Sometimes, an additional τ^2 parameter is brought in to define a squared exponential covariance function with nugget as

$$c_Y(\mathbf{s},\mathbf{s}') := \begin{cases} \sigma^2+\tau^2 & \text{if } 0 = \|\mathbf{s}-\mathbf{s}'\|_2 \\ \sigma^2 e^{-\phi\|\mathbf{s}-\mathbf{s}'\|_2} & \text{if } 0 < \|\mathbf{s}-\mathbf{s}'\|_2. \end{cases}$$

τ^2 is referred to as the nugget variance and it quantifies the variability of the microscale spatial components. The microscale spatial components are those parts of $\{Y(\mathbf{s})\}$ which are uncorrelated at even the minutest spatial resolution, and hence practically behave like an iid process. Besides the exponential covariance function, there are many other parametric covariance functions that are used to model geostatistical data. Some of them are presented in the following table.

Table 1: Useful parametric covariance functions for geostatistical modeling

Covariance function	Formula
Spherical	$c_Y(\mathbf{s},\mathbf{s}') := \begin{cases} \sigma^2+\tau^2 \text{if } 0 = \|\mathbf{s}-\mathbf{s}'\|_2 \\ \sigma^2\left[1-\frac{3}{2}\phi\|\mathbf{s}-\mathbf{s}'\|_2+\frac{1}{2}\phi^3\|\mathbf{s}-\mathbf{s}'\|_2^3\right] \text{if } 0 < \|\mathbf{s}-\mathbf{s}'\|_2 \end{cases}$
Exponential	$c_Y(\mathbf{s},\mathbf{s}') := \begin{cases} \sigma^2+\tau^2 \text{if } 0 = \|\mathbf{s}-\mathbf{s}'\|_2 \\ \sigma^2 e^{-\phi\|\mathbf{s}-\mathbf{s}'\|_2} \text{if } 0 < \|\mathbf{s}-\mathbf{s}'\|_2 \end{cases}$
Squared exponential	$c_Y(\mathbf{s},\mathbf{s}') := \begin{cases} \sigma^2+\tau^2 \text{if } 0 = \|\mathbf{s}-\mathbf{s}'\|_2 \\ \sigma^2 e^{-\phi\|\mathbf{s}-\mathbf{s}'\|_2^2} \text{if } 0 < \|\mathbf{s}-\mathbf{s}'\|_2 \end{cases}$
Powered exponential	$c_Y(\mathbf{s},\mathbf{s}') := \begin{cases} \sigma^2+\tau^2 \text{if } 0 = \|\mathbf{s}-\mathbf{s}'\|_2 \\ \sigma^2 e^{-\phi\|\mathbf{s}-\mathbf{s}'\|_2^\alpha} \text{if } 0 < \|\mathbf{s}-\mathbf{s}'\|_2; 0<\alpha\le 2 \end{cases}$
Rational quadratic	$c_Y(\mathbf{s},\mathbf{s}') := \begin{cases} \sigma^2+\tau^2 \text{if } 0 = \|\mathbf{s}-\mathbf{s}'\|_2 \\ \sigma^2\left(1-\frac{\|\mathbf{s}-\mathbf{s}'\|_2^2}{\phi+\|\mathbf{s}-\mathbf{s}'\|_2^2}\right) \text{if } 0 < \|\mathbf{s}-\mathbf{s}'\|_2 \end{cases}$
Matérn ($\nu>0$)	$c_Y(\mathbf{s},\mathbf{s}') := \begin{cases} \sigma^2+\tau^2 \text{if } 0 = \|\mathbf{s}-\mathbf{s}'\|_2 \\ \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}\left(\sqrt{2\nu}\phi\|\mathbf{s}-\mathbf{s}'\|_2\right)^\nu K_\nu\left(\sqrt{2\nu}\phi\|\mathbf{s}-\mathbf{s}'\|_2\right) \text{if } 0 < \|\mathbf{s}-\mathbf{s}'\|_2 \end{cases}$

The squared exponential covariance function resembles the exponential covariance function but suffers from the limitation that the associated spatial process $\{Y(\mathbf{s})\}$ is infinitely many times differentiable. The spherical covariance function on the other hand has compact support and hence is useful in creating a sparse $\mathbf{\Sigma}$. The Matérn covariance function is attractive in the sense that the associated spatial process $\{Y(\mathbf{s})\}$ has controllable smoothness with $\lceil\nu\rceil - 1$ times differentiability. However, all these parametric covariance functions depend only on the distance $\|\mathbf{s}-\mathbf{s}'\|_2$ disregarding the direction along which $\mathbf{s}'$ is separated from $\mathbf{s}$. This property is known as isotropy. Isotropic covariance functions are not suitable for modeling environmental datasets that are under the influence of wind flow. For such datasets, the observations separated along the direction of wind flow typically display stronger dependence compared to the ones separated along other directions. Banerjee *et al.* (2003) analyzed a scallop catch dataset where the dependence structure along different directions varied substantially thereby necessitating the use of anisotropic covariance functions.

Unlike the isotropic covariance functions, whose covariance contours are circles, the covariance contours of the anisotropic covariance functions can take the shape of arbitrary closed curves. Different notions of anisotropy have been introduced by different researchers. Zimmerman (1993) systematically studied them and classified them roughly into three different broad categories. In order to understand them we first need to define the variogram function $\gamma_Y(\mathbf{s},\mathbf{s}')$ associated with a spatial process $\{Y(\mathbf{s})\}$. It is defined by the formula

$\gamma(\mathbf{s},\mathbf{s}') := \frac{1}{2}var(Y(\mathbf{s}) - Y(\mathbf{s}'))$. In an alternative approach to geostatistical modeling, the whole theory that can be developed using the covariance function $c_Y(\mathbf{s},\mathbf{s}')$, has been developed parallel using the variogram function $\gamma(\mathbf{s},\mathbf{s}')$. Zimmerman (1993) defined three classes of anisotropy as follows. The differential dependence along the different directions is referred to as sill anisotropy if $\lim_{a\to\infty}\gamma(\frac{a\mathbf{h}}{\|\mathbf{h}\|})$ depends not on $\|\mathbf{h}\|$ but on $\mathbf{h}$. Here $\mathbf{h} = (\mathbf{s}-\mathbf{s}')$ is the lag between two spatial locations $\mathbf{s}$ and $\mathbf{s}'$. When $\lim_{a\to 0}\gamma(\frac{a\mathbf{h}}{\|\mathbf{h}\|})$ depends on $\mathbf{h}$, one says the process shows nugget anisotropy. The third and last type of anisotropy occurs when the decay parameter ϕ depends on $\mathbf{h}$. It is referred to as range anisotropy. A particularly interesting subclass of the range anisotropy is the geometric anisotropy. Geometrically anisotropic spatial process $\{Y(\mathbf{s})\}$ has a covariance function with elliptical covariance contours. A simple recipe for creating a geometrically anisotropic covariance function is to replace $\|\mathbf{s}-\mathbf{s}'\|_2$ by $\sqrt{(\mathbf{s}-\mathbf{s}')'\mathbf{A}(\mathbf{s}-\mathbf{s}')}$ in the formula of a parametric isotropic covariance function $c_Y(\mathbf{s},\mathbf{s}')$. The 2×2 matrix $\mathbf{A}$ is a pd matrix with 3 unknown parameters, that control the shape and the alignment of the elliptical covariance contours.

Although useful, the anisotropic covariance functions are not the best choice for modeling the complex dependence structure associated with geostatistical data arising out of environmental applications. The reason is that such a covariance function $c_Y(\mathbf{s},\mathbf{s}')$, although invokes differential dependence structure along different directions, is still a function of the lag $\mathbf{h}$ between two spatial locations $\mathbf{s}$ and $\mathbf{s}'$. This property is known as stationarity. Stationarity implies that the covariance between $Y(\mathbf{s})$ and $Y(\mathbf{s}')$ remains unchanged if both the spatial locations are shifted by the same lag $\mathbf{h}$, *i.e.*, $C_Y(\mathbf{s},\mathbf{s}') = C_Y(\mathbf{s}+\mathbf{h},\mathbf{s}'+\mathbf{h})$. Efforts have been made to create non-stationary covariance functions $C_Y(\mathbf{s},\mathbf{s}')$ which depend on both $\mathbf{s}$ and $\mathbf{s}'$.

2. Towards non-stationary covariance functions

Over the years different strategies to create non-stationary covariance functions have been proposed. Here we discuss a few prominent ones.

Approach 1 : Direct construction The simplest approach is to propose a formula of $C_Y(\mathbf{s},\mathbf{s}')$ that involves both $\mathbf{s}$ and $\mathbf{s}'$ and then subsequently show that $C_Y(\mathbf{s},\mathbf{s}')$ is a valid covariance function. However, guessing such functions and then showing them to be valid covariance functions can be difficult.

Approach 2 : Transformation of the original process Alternatively, one can start with a spatial process $\{Y(\mathbf{s})\}$ that has an isotropic covariance function and then take a transformation of $\{Y(\mathbf{s})\}$ to define a new process $\{Y^*(\mathbf{s})\}$ which has an anisotropic covariance function. The transformations used are generally elementary in nature. One such transformation $Y^*(\mathbf{s}) = \sigma(\mathbf{s})Y(\mathbf{s})$ gives rise to the non-stationary covariance function of the form $C_{Y^*}(\mathbf{s},\mathbf{s}') = \sigma(\mathbf{s})\sigma(\mathbf{s}')C_Y(\mathbf{s},\mathbf{s}') = \sigma(\mathbf{s})\sigma(\mathbf{s}')f(\|\mathbf{s}-\mathbf{s}'\|_2)$. $\sigma(\mathbf{s})$ is a geographically varying positive function that enforces the departure from stationarity in a multiplicative manner. In another transformation, one can propose $Y^*(\mathbf{s}) = Y(\mathbf{s}) + \delta(\mathbf{s})Z$, where Z is a random variable with mean 0 and variance σ_Z^2 and $\delta(\mathbf{s})$ is a positive function of $\mathbf{s}$. The transformed process has the covariance function $C_Y(\mathbf{s},\mathbf{s}') + \delta(\mathbf{s})\delta(\mathbf{s}')\sigma_Z^2 = f(\|\mathbf{s}-\mathbf{s}'\|_2) + \delta(\mathbf{s})\delta(\mathbf{s}')\sigma_Z^2$. In this case, the departure from stationarity takes place in an additive manner. To combine both, one can define $Y^*(\mathbf{s}) = \sigma(\mathbf{s})Y(\mathbf{s}) + \delta(\mathbf{s})Z$ leading to a non-stationary covariance func-

tion of the form $\sigma(\mathbf{s})\sigma(\mathbf{s}')f(\|\mathbf{s}-\mathbf{s}'\|_2)+\delta(\mathbf{s})\delta(\mathbf{s}')\sigma_Z^2$. However, the class of non-stationary covariance functions that can be generated by the transformation approach are very limited.

Approach 3 : Deformation approach Richer class of non-stationary covariance functions can be created by the deformation approach. In a seminal paper, Sampson and Guttorp (1992) first came up with the idea of deformation $g(\cdot)$ of the original geographical space so that the observed spatial process $\{Y(\mathbf{s})\}$ is stationary with respect to to the deformed geographical locations $g(\mathbf{s}_1), g(\mathbf{s}_2), \cdots, g(\mathbf{s}_n)$. Hence, the covariance of $Y(\mathbf{s})$ and $Y(\mathbf{s}')$ is of the form $f(\|g(\mathbf{s})-g(\mathbf{s}')\|_2)$. When considered in terms of the original geographical space, the covariance function is not a function of $\|\mathbf{s}-\mathbf{s}'\|_2$, hence non-stationary. This brilliant idea however suffers from the serious shortcoming that estimating the deformation function $g(\cdot)$ from the data is a highly non-linear optimization problem that can be numerically very challenging. Moreover, the estimated $g(\cdot)$ can sometimes fold over its domain leading to a meaningless covariance function, and also, the estimation process as proposed by Sampson and Guttorp (1992) requires replicated samples at the original geographical locations $\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n$.

Approach 4 : Process convolution approach In time series analysis, starting with a white noise process $\{Z_t\}$ that has the simplest covariance function, one can create a moving average process $\{X_t\}$ by taking a linear combination of Z_t as $X_t := Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$. The process $\{X_t\}$ has a substantially improved covariance function compared to the original process $\{Z_t\}$. Much to the same spirit, starting with a spatial white noise process $\{Z(\mathbf{s})\}$ with a simple spatial covariance function one can create a new process $\{Y(\mathbf{s})\}$ by the following process convolution

$$Y(\mathbf{s}) := \int_{\mathbb{R}^2} K(\mathbf{s}-\mathbf{s}')Z(\mathbf{s}')d\mathbf{s}'. \tag{1}$$

Strictly speaking, the above integral is not defined and should be interpreted as $Y(\mathbf{s}) := \int_{\mathbb{R}^2} K(\mathbf{s}-\mathbf{s}')dB(\mathbf{s}')$ where $B(\mathbf{s})$ denotes a two-dimensional Brownian motion on $\mathbb{R}^2$. When interpreted as above, the process $\{Y(\mathbf{s})\}$ has a stationary covariance function given by the formula $c_Y(\mathbf{s},\mathbf{s}') = \sigma^2 \int_{\mathbb{R}^2} K(\mathbf{s}-\mathbf{t})K(\mathbf{s}'-\mathbf{t})d\mathbf{t}$. Higdon (1998) used spatially varying kernel functions in the above formula to generate a non-stationary covariance function. In that case, $Y(\mathbf{s}) := \int_{\mathbb{R}^2} K_{\mathbf{s}}(\mathbf{s}-\mathbf{s}')dB(\mathbf{s}')$, where $K_{\mathbf{s}}(\mathbf{s}-\mathbf{t})$ is a non-negative real-valued integrable function (bivariate kernel function); the associated covariance function is non-stationary, and is given by the formula

$$c_Y(\mathbf{s},\mathbf{s}') = \sigma^2 \int_{\mathbb{R}^2} K_{\mathbf{s}}(\mathbf{s}-\mathbf{t})K_{\mathbf{s}'}(\mathbf{s}'-\mathbf{t})d\mathbf{t}. \tag{2}$$

The convolution approach was later extended by Paciorek and Schervish (2006) to produce a flexible non-stationary Matérn covariance function with nugget as

$$c_{Y,NS}(\mathbf{s},\mathbf{s}') := \begin{cases} \tau^2+\sigma^2 & \text{if } \mathbf{s}=\mathbf{s}' \\ \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} |\Sigma(\mathbf{s})|^{\frac{1}{4}} |\Sigma(\mathbf{s}')|^{\frac{1}{4}} \left|\frac{\Sigma(\mathbf{s})+\Sigma(\mathbf{s}')}{2}\right|^{-\frac{1}{2}} & \\ \times \left(2\sqrt{\nu}\phi\sqrt{Q(\mathbf{s},\mathbf{s}')}\right)^{\nu} K_\nu\left(2\sqrt{\nu}\phi\sqrt{Q(\mathbf{s},\mathbf{s}')}\right) & \text{if } \mathbf{s}\neq\mathbf{s}'. \end{cases} \tag{3}$$

Here $\Sigma(\mathbf{s}) = \begin{pmatrix}\cos(\theta(\mathbf{s})) & -\sin(\theta(\mathbf{s}))\\ \sin(\theta(\mathbf{s})) & \cos(\theta(\mathbf{s}))\end{pmatrix}\begin{pmatrix}\lambda_1(\mathbf{s}) & 0\\ 0 & \lambda_2(\mathbf{s})\end{pmatrix}\begin{pmatrix}\cos(\theta(\mathbf{s})) & \sin(\theta(\mathbf{s}))\\ -\sin(\theta(\mathbf{s})) & \cos(\theta(\mathbf{s}))\end{pmatrix}$ and $Q(\mathbf{s},\mathbf{s}') = (\mathbf{s}-\mathbf{s}')'\left(\frac{\Sigma(\mathbf{s})+\Sigma(\mathbf{s}')}{2}\right)^{-1}(\mathbf{s}-\mathbf{s}')$. Paciorek and Schervish (2003) also used the $c_{Y,NS}(\mathbf{s},\mathbf{s}')$ as

the covariance function of a Gaussian process prior for a Bayesian non-parametric regression problem. While $\lambda_1(\mathbf{s})$ and $\lambda_2(\mathbf{s})$ determine the length of the major and minor axis of the elliptical covariance contours at location $\mathbf{s}$, the $\theta(\mathbf{s})$ determines the alignment of the contours. $c_{Y,NS}(\mathbf{s},\mathbf{s}')$ is non-stationary since the shape of the covariance contours vary with respect to $\mathbf{s}$.

3. Scalable covariance functions for massive geostatistical data

Once an appropriate non-stationary covariance function is selected, the next step is to estimate the unknown parameters associated with the mean and the covariance function. Assume that the vector of those unknown parameters is denoted by $\boldsymbol{\theta}$. Hence, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are functions of $\boldsymbol{\theta}$ and are better represented as $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Under the classical paradigm, the estimation is mostly carried out by maximizing the Gaussian likelihood function $L(\boldsymbol{\theta}|\mathbf{y}) := \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{|\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))}$. The likelihood function is highly non-linear in $\boldsymbol{\theta}$ and hence requires numerical algorithms for finding the global maximum. On the other hand, if the Bayesian path is chosen, one needs to find the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{|\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))}\pi(\boldsymbol{\theta})$. The posterior generally does not appear in the form of nice well-known distributions, and hence its exploration requires an MCMC method. Regardless of the classical or Bayesian approach being adopted, one needs to evaluate the terms $|\boldsymbol{\Sigma}(\boldsymbol{\theta})|$ and $(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))$ repeatedly. Each evaluation of $|\boldsymbol{\Sigma}(\boldsymbol{\theta})|$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$ requires $O(n^3)$ operations and n being a large number of the order ~ 100000, the computational cost jumps to a staggering $O(10^{15})$ operations making it impossible to implement. Suitable strategies have been developed to bring the computational cost down to a manageable level. Below we discuss some such strategies. Most of these approaches are based on replacing the terms $|\boldsymbol{\Sigma}(\boldsymbol{\theta})|$ and $(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))$ by some approximation and then carrying out the computation. Their success depends on whether the approximation to the original term is good and itself is easily computable.

3.1. Fixed rank Kriging

In one of the earliest works in this direction Cressie and Johannesson (2008), while fitting a centered Gaussian process $Y(\mathbf{s}) \sim GP(0, c_Y(\mathbf{s},\mathbf{s}'))$ to the observed geostatistical data, approximated $\{Y(\mathbf{s})\}$ by a new process defined as $\tilde{Y}(\mathbf{s}) = \sum_{r=1}^{R}\sum_{k=1}^{K_r}\theta_{r,k}\varphi_{r,k}(\mathbf{s}) + \epsilon(\mathbf{s})v(\mathbf{s})$. $\varphi_{r,k}(\mathbf{s})$ are basis functions of resolution r and $\theta_{r,k}$ are dependent Gaussian random variables with covariance matrix $K(\boldsymbol{\phi})$. So, $\tilde{Y}(\mathbf{s}) \sim GP(0, c_{\tilde{Y}}(\mathbf{s},\mathbf{s}'))$ and its covariance function $c_{\tilde{Y}}(\mathbf{s},\mathbf{s}')$ approximates $c_Y(\mathbf{s},\mathbf{s}')$. Consequently, the covariance matrix $\boldsymbol{\Sigma}$ is also being approximated by the covariance matrix $\boldsymbol{\varphi}\mathbf{K}(\boldsymbol{\phi})\boldsymbol{\varphi}' + \tau^2\mathbf{V}$. For the approximating covariance matrix $\boldsymbol{\varphi}\mathbf{K}(\boldsymbol{\phi})\boldsymbol{\varphi}' + \tau^2\mathbf{V}$ calculating the determinant and inverse it takes $O(n)$ operations only. Thus they approximated the original likelihood $L(\boldsymbol{\theta}|\mathbf{y})$ by a new likelihood $L(\boldsymbol{\phi},\tau^2|\mathbf{y})$ where $(\boldsymbol{\phi},\tau^2)$ is the vector comprising the new parameters.

3.2. Gaussian predictive process models

In another approach more geared towards the Bayesian paradigm, Banerjee *et al.* (2008) considered a centered Gaussian process $Y(\mathbf{s}) \sim GP(0, c_Y(\mathbf{s},\mathbf{s}'))$ and approximated it by a predictive process $[Y(\mathbf{s})|Y(\mathbf{s}_1^*), Y(\mathbf{s}_2^*), \cdots, Y(\mathbf{s}_k^*)] + \epsilon(\mathbf{s})$. So, the predictive process

can be expressed as $\tilde{Y}(\mathbf{s}) = [c_Y(\mathbf{s}_1, \mathbf{s}^*), c_Y(\mathbf{s}_2, \mathbf{s}^*), \cdots, c_Y(\mathbf{s}_n, \mathbf{s}^*)]{\Sigma_Y^*}^{-1} \begin{pmatrix} Y(\mathbf{s}_1^*) \\ Y(\mathbf{s}_2^*) \\ \vdots Y(\mathbf{s}_k^*) \end{pmatrix} + \epsilon(\mathbf{s}) = \mathbf{S}(\boldsymbol{\theta}) \begin{pmatrix} Y(\mathbf{s}_1^*) \\ Y(\mathbf{s}_2^*) \\ \vdots Y(\mathbf{s}_k^*) \end{pmatrix} + \epsilon(\mathbf{s})$. $\mathbf{s}_1^*, \mathbf{s}_2^*, \cdots, \mathbf{s}_k^*$ are some knot points on the geographical plane. So, the approximating covariance matrix $\Sigma_{\tilde{Y}}$ is of the form $\mathbf{S}(\boldsymbol{\theta})\mathbf{K}(\boldsymbol{\theta})\mathbf{S}'(\boldsymbol{\theta}) + \tau^2\mathbf{I}$. Calculating the determinant and inverting $\mathbf{S}(\boldsymbol{\theta})\mathbf{K}(\boldsymbol{\theta})\mathbf{S}'(\boldsymbol{\theta}) + \tau^2\mathbf{I}$ takes only $O(n)$ operations. For fixed rank Kriging, the number of basis functions $n = K_1 + K_2 + \cdots + K_R$ determines the the quality of approximation, and it should be chosen judiciously to trade quality of approximation for computational cost. In the case of the Gaussian predictive process model, the number of the knot points n plays the same role.

3.3. Covariance tapering

While the last two approaches were based on approximating the original Gaussian process $\{Y(\mathbf{s})\}$ by a new Gaussian process $\{\tilde{Y}(\mathbf{s})\}$ with which the computational cost reduces significantly to $O(n)$, other approaches directly approximate $\boldsymbol{\Sigma}$ by a new covariance matrix $\tilde{\boldsymbol{\Sigma}}$. In the covariance tapering approach, instead of approximating $\boldsymbol{\Sigma}$ by a new covariance matrix $\tilde{\boldsymbol{\Sigma}}$, one transforms $\boldsymbol{\Sigma}$ to convert it to a sparse matrix. With that, the likelihood can be rewritten as $\left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}tr((\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}))} = \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}tr((\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1})}$, and can be approximated by $\left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}tr((\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1})}$. Here $\boldsymbol{\Sigma}$ is approximated by $\boldsymbol{\Sigma} \odot \mathbf{T}$. The transformation $\mathbf{T}$ is referred to as the one-taper transform and it transforms $\boldsymbol{\Sigma}$ to $\boldsymbol{\Sigma} \odot \mathbf{T}$. $\mathbf{T}$ is a covariance matrix formed by a compactly supported covariance function (Kaufman *et al.* (2008)). The tapered matrix $\boldsymbol{\Sigma} \odot \mathbf{T}$ is also a covariance matrix and it is sparse, thereby making the approximating likelihood scalable to massive n. A variation of the one-taper transform is called a two-taper transform that transforms $\boldsymbol{\Sigma}$ to $\boldsymbol{\Sigma} \odot \mathbf{T}$ as well as the empirical covariance matrix $(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'$ to $(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})' \odot \mathbf{T}$. That way, both the model covariance matrix and the empirical covariance matrix become sparse.

3.4. Vecchia approximation and nearest neighbour Gaussian process (NNGP)

Any likelihood function can be expressed as products of conditional distributions as follows

$$[Y(\mathbf{s}_1), Y(\mathbf{s}_2), \cdots, Y(\mathbf{s}_n)] = [Y(\mathbf{s}_n) \mid Y(\mathbf{s}_{n-1}), \cdots, Y(\mathbf{s}_1)] \times [Y(\mathbf{s}_{n-1}) \mid Y(\mathbf{s}_{n-2}), \cdots, Y(\mathbf{s}_1)] \times \cdots \times [Y(\mathbf{s}_2) \mid Y(\mathbf{s}_1)] \times [Y(\mathbf{s}_1)].$$

Based on this representation Vecchia (1988) in an early work figured out how to reduce the computational cost of evaluating a likelihood. He demonstrated that the above expression is

$$\approx [Y(\mathbf{s}_n) \mid Y(\mathbf{s})_{s \in N_n}] \times [Y(\mathbf{s}_{n-1}) \mid Y(\mathbf{s})_{s \in N_{n-1}}] \times \cdots \times [Y(\mathbf{s}_2) \mid Y(\mathbf{s})_{s \in N_2}] \times [Y(\mathbf{s}_1)].$$

where N_i denotes of neighbourhood set of $\mathbf{s}_i$ that contains atmost k spatial locations. So under the traditional Gaussian setup evaluating $[Y(\mathbf{s}_i) \mid Y(\mathbf{s})_{s \in N_i}]$ requires calculating the determinant and inverse of at most a $k \times k$ covariance matrix. The computational cost

is atmost $O(k^3)$. As there are $n-1$ such products, the overall computational cost for calculating the approximating likelihood $[Y(\mathbf{s}_n) \mid Y(\mathbf{s})_{s\in N_n}] \times [Y(\mathbf{s}_{n-1}) \mid Y(\mathbf{s})_{s\in N_{n-1}}] \times \cdots \times [Y(\mathbf{s}_2) \mid Y(\mathbf{s})_{s\in N_2}] \times [Y(\mathbf{s}_1)]$ sum up to $O(nk^3)$. Although the idea was first presented by Vecchia (1988), it became familiar when Datta *et al.* (2016) applied it successfully to model a massive forest inventory dataset. Moreover, it is to the credit of Datta *et al.* (2016) who showed that the approximating likelihood can be associated with another Gaussian process, which they referred to as the nearest neighbour Gaussian process (NNGP). Although, the NNGP produced promising result, the method's dependence on the number of neighbours, and the set of neighbouring locations requires further investigation. Another issue is that the decomposition of the likelihood function as products of conditional distributions is not unique, and hence the success of the Vecchia approximation and the NNGP depend on the particular version one uses.

4. Concluding remarks

In this article, we have briefly touched upon different approaches for the creation of non-stationary covariance functions. The list is ever growing and many of them are not discussed here. For example, Fuentes (2002) considered the convolution of stationary processes and created locally stationary covariance functions. Then we have seen that the problem does not just end with the selection of an appropriate non-stationary covariance function. The advent of GIS-based data collection system coupled with advancement in data storage capacity, allows us to gather data at millions. Directly working with a non-stationary covariance function for such massive dataset leads to $O(n^3)$ computations making the task impossible to commence. In this regard, we have discussed different methods of scalable modeling of massive geostatistical data. Among them, the Vecchia approximation has recently gained popularity with the work of Datta *et al.* (2016). In a recent work, Zheng *et al.* (2023) extended the idea to a non-Gaussian spatial process. Besides the approaches discussed here, the multiresolution analysis proposed by Katzfuss (2017) is also useful in modeling massive geostatistical data. There are many more methods for scalable modeling of geostatistical data and a comparative analysis of them have been carried out in Heaton *et al.* (2019). The field is growing rapidly. For a more comprehensive review of the Bayesian methods for massive geostatistical data, one can consider the recent articles by Banerjee and Fuentes (2012) and Banerjee (2017).

Conflict of interest

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

References

Ayalew, K. A., Manda, S., and Cai, B. (2024). Multivariate skew-normal distribution for modelling skewed spatial data. *Spatial and Spatio-temporal Epidemiology*, **51**, 100692.

Banerjee, S. (2017). High-dimensional Bayesian geostatistics. *Bayesian Analysis*, **12**, 583.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2003). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC.

Banerjee, S. and Fuentes, M. (2012). Bayesian modeling for large spatial datasets. *Wiley Interdisciplinary Reviews: Computational Statistics*, **4**, 59–66.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **70**, 825–848.

Blanchet, J. and Davison, A. C. (2011). Spatial modeling of extreme snow depth. *The Annals of Applied Statistics*, **5**, 1699–1725.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **70**, 209–226.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, **111**, 800–812.

Fuentes, M. (2002). Spectral methods for nonstationary spatial processes. *Biometrika*, **89**, 197–210.

Hazra, A., Reich, B. J., and Staicu, A.-M. (2020). A multivariate spatial skew-t process for joint modeling of extreme precipitation indexes. *Environmetrics*, **31**, e2602.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., et al. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, **24**, 398–425.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environmental and Ecological Statistics*, **5**, 173–190.

Huser, R. and Wadsworth, J. L. (2022). Advances in statistical modeling of spatial extremes. *Wiley Interdisciplinary Reviews: Computational Statistics*, **14**, e1537.

Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, **112**, 201–214.

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, **103**, 1545–1555.

Paciorek, C. and Schervish, M. (2003). Nonstationary covariance functions for gaussian process regression. *Advances in Neural Information Processing Systems*, **16**.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, **17**, 483–506.

Sampson, P. D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, **87**, 108–119.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **50**, 297–312.

Zheng, X., Kottas, A., and Sansó, B. (2023). Nearest-neighbor mixture models for non-gaussian spatial processes. *Bayesian Analysis*, **18**, 1191–1222.

Zimmerman, D. L. (1993). Another look at anisotropy in geostatistics. *Mathematical Geology*, **25**, 453–470.

Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025; pp 77-86

Use of Designs for Statistical Experiments in Constructing Cryptographic Schemes

Mausumi Bose

Indian Statistical Institute and St. Xavier's College, Kolkata

Received: 23 June 2025; Revised 08 July 2025; Accepted 10 July 2025

Abstract

In this article we highlight how the combinatorial properties of statistical designs of experiments have been used by many researchers for constructing various types of cryptographic schemes. In particular, we discuss key predistribution schemes for distributed sensor networks in some detail and show through examples, how useful schemes can be constructed from the duals of certain block designs.

Key words: Balanced incomplete block designs; Partially balanced incomplete block designs; Steiner's triple systems; Distributed sensor networks; Resilience.

AMS Subject Classifications: 05B05, 94A6.

1. Introduction

Combinatorial structures of different kinds have been extensively studied over the years by mathematicians, for example, Hadamard Matrices, orthogonal arrays, Latin squares, Steiner's triple systems, *etc.* The construction and existence of these structures have been well-developed and a considerable literature is available on such structures.

Later, statisticians found that many of these structures are useful in the field of Design of Experiments. Subsequently, optimality properties of the designs based on these structures, were also proved. For example, it was found that Hadamard matrices were useful in obtaining optimal weighing designs using the chemical balance, Steiner's triple systems were useful as incomplete block designs for one-way elimination of heterogeneity, Latin squares and mutually orthogonal Latin squares were useful as optimal designs for eliminating heterogeneity in two or three directions, orthogonal arrays were useful in obtaining fractional factorial designs, and the list goes on. For a comprehensive discussion on these designs, their combinatorial properties and construction, and their statistical optimality aspects, we refer to Raghavarao (1971), Street and Street (1987), Shah and Sinha (1989) and Hedayat, Stufken and Sloane (1999).

Much later, cryptographers found that many of these statistical designs of experiments

Corresponding Author: Mausumi Bose
Email: mausumi.bose@gmail.com

based on combinatorial structures can also be used to generate good cryptographic schemes. For some details of such use, we refer to Stinson (2004) and Stinson and Patterson (2023).

Cryptography is the practice of scrambling communications so that only the intended recipient can access them. In modern times, cryptography is used to protect confidentiality of sensitive information and protect it from hackers and other cyber criminals. It can be used to obscure various forms of digital communication, including text, images, video, or audio; protect confidentiality and integrity in communication. *e.g.*, computer passwords, email, online transactions, transmitting confidential information, *etc.* For a historical perspective of the development of the subject since ancient to recent times, we refer to Kahn (1996) and Bauer (2021). For a technical perspective of some schemes, we refer to Stinson and Patterson (2023).

In this paper we mention some cryptographic schemes which can be obtained from combinatorial structures. In Section 1, we give a brief description of two such schemes and mention the combinatorial structures which lead to these schemes. In Section 2 we focus on distributed sensor networks and describe how they can be obtained from statistical designs. References are given for all these results and the reader may obtain the details from these references.

2. Some cryptographic schemes and related combinatorial structures

In this section we mention two cryptographic schemes, error-correcting codes and visual cryptographic schemes, and mention the designs that may be used to construct these schemes. Our objective here is to only give a flavor of the versatility of the application of designs to cryptography. There are many other schemes which are not mentioned here for the sake of brevity.

2.1. Error correcting codes and Hadamard matrices

Error-correcting codes are used to detect and correct errors that can occur when transmitting data over noisy channels. They add extra bits, *i.e.*, redundant information, to the original data in such a way that the recipient of the data can compare the received data with the redundancies and identify the errors which arise due to noise or other factors. Each code is a collection of codewords, or k-tuples, say, with symbols from a set of symbols or an alphabet.

It is well known that optimal weighing designs are given by Hadamard matrices, *e.g.*, to optimally weigh 8 objects using 8 weighings with a chemical balance, the optimal design matrix will be given by a Hadamard matrix of order 8. In the cryptography context, the rows of this same Hadamard matrix, after replacing -1 by 1 and 1 by 0, will give an error-correcting code for transmitting a binary message of 3 bits as a message of 8 bits, and it can correct one error. More generally, using Hadamard matrices, one can construct the first-order Reed-Muller code over the binary alphabet which is useful in transmitting messages over noisy channels. For some applications in this context, we refer to Serberry, Wysocky and Wysocki (2005) and Yarlagadda and Hershey (1997).

2.2. Error-correcting codes and orthogonal arrays

Orthogonal arrays are well-known structures which are useful in statistics for obtaining suitable fractions of factorial experiments for experimentation. These orthogonal arrays also give useful error-correcting codes, namely MDS (Maximum distance separable) codes, the Reed Solomon Codes, Hamming codes, *etc.* These codes have optimal properties of various kinds.

2.3. Visual cryptographic schemes and BIBD, PBIBD

In a (k, n) visual cryptography scheme, a secret image (or text) is encoded to form n *'shares'* and each share is printed on a transparency sheet. There are n participants, each of whom gets one share. The encryption is such that only when $k(\geq 2)$ participants get together and stack their sheets one above another, the secret image is revealed, no set of $k-1$ or fewer participants can decode the secret image. This scheme is useful as decoding can be done simply by the human eye without the need of any computers or equipment. More details can be obtained from Naor and Shamir (1994) and Kang, Arce and Lee (2011), Ibrahim, Teh and Abdullah (2021) and Climato, Prisco and Santis (2005).

It has been shown in Blundo, Santis and Stinson (1999) that balanced incomplete block designs (BIBDs) are useful in encoding the secret image and forming the shares. Adhikari and Bose (2004) and Adhikari, Bose, Kumar and Roy (2007) showed that partially balanced incomplete block Designs (PBIBDs) lead to schemes where the sharpness of the recovered image is better for certain set of participants. Bose and Mukerjee (2006, 2010) showed that various other incomplete block designs like regular graph designs, symmetrical unequal block designs may also be used to obtain schemes with many desirable properties.

There are several other schemes in the literature which have been developed from designs of experiments and which have not been mentioned here, *e.g.*, general threshold access structures, anti-collusion digital fingerprinting, *etc.* Some references on these are Kang, Sinha and Lee (2006), Yagi, Matsushima and Hirasawa (2007), Bose and Mukerkee (2013, 2014). Moreover, there could be many other possibilities of using designs to construct useful cryptographic schemes of various types in future.

3. Distributed sensor networks

Distributed sensor networks (DSNs) are used in a wide range of applications. Some examples of their use are in air quality monitoring, water quality monitoring, wildlife tracking, seismic activity detection *etc.* These are also used in military applications such as battlefield surveillance, target tracking, perimeter security, reconnaissance missions, *etc.* Another interesting use of this system is in smart cities where they prove useful in traffic management, congestion monitoring, parking availability detection, street lighting control, *etc.*

This wide applicability of these network schemes is due to the ability of DSNs to collect real-time data from geographically dispersed sensors, enabling comprehensive monitoring and analysis of various physical phenomenon across large areas.

We now discuss key-predistribution schemes for DSNs in some detail, based on results from Bose, Dey and Mukerjee (2013); more references may be found therein.

3.1. Key predistribution schemes(KPS) for DSNs

We begin with an example of a situation where sensor nodes are pre-distributed in several locations. Suppose in a military operation, several sensor nodes, each with some secret keys installed in them, are randomly scattered over a sensitive area. The keys in each node are taken from a large set of keys. Each node can send or receive signals only over a certain wireless communication range or neighbourhood. Once deployed, these nodes have to communicate with each other through secure keys in order to gather and relay information.

In this context, some metrics of the KPS are important:

1. Network size, *i.e.*, the number of nodes deployed, say, n.
2. Key storage, *i.e.*, the number of keys stored per node, say, k.
3. Intersection Threshold *i.e.*, the number of keys common between 2 nodes, say, q.
4. Communication rule *i.e.*, if two nodes are within each other's neighbourhood, they can communicate with each other
   (a) directly, if they have $q \geq 1$ common keys, or
   (b) *via* one hop if there is a third node within the intersection of their neighbourhoods which shares q common keys with each of them. If needed, multiple secure links can also be used if there is a sequence of nodes connecting them such that every pair of successive nodes in this sequence share $q(\geq 1)$ common keys.

Now, after deployment, some nodes may be captured in an attack. In that case, all the keys in these captured nodes are considered to be lost and cannot be used for communication by the other nodes. However, if the remaining nodes can still communicate using their remaining keys as per 4 (a) or 4 (b) above, then the KPS is said to be 'resilient'. Resilience is a desirable property of a KPS.

3.2. Correspondence with block designs

Now we introduce a correspondence between some terms used in the context of block designs with the terms used in the context of the KPS as introduced in section 3.1.

The set of all keys of the KPS corresponds to the set of all treatments in a block design.

The sensor nodes of the KPS correspond to the blocks in a block design. Here, since we would like a large number of nodes in the system, we need a large number of blocks in the designs, as opposed to fewer blocks preferred in designs of experiments.

The key storage of a KPS corresponds to the block size of a design.

With the above correspondence, it is clear that the 'intersection threshold' of a KPS corresponds to the number of treatments that are common to two blocks. This means that the block intersection number of a block design becomes important. We will consider the duals of block designs where the roles of treatment and block in the original block design are reversed, and so, the incidence between the treatments and blocks is also reversed.

We give examples of two designs, one PBIBD(d_1) wih 2 associate classes, and one BIBD(d_2), and their corresponding dual designs d_1^* and d_2^*, as shown below. These duals will be used subsequently to construct KPS. For the design d_i, let v_i, b_i, r_i and k_i denote the number of treatments, number of blocks, replication number, and block size, respectively, $i = 1, 2$. Then, from the properties of BIBD and PBIBD it may be noted that these duals $d_i^*,\ i = 1, 2$, are such that:

1. every symbol occurs at most once in any block
2. every symbol occurs in k_i blocks, $2 \leq k_i < b_i$
3. every block contains r_i symbols, $v_i > r_i \geq 2$, and
4. there is an association scheme with 2 associate classes on the sets of blocks of d_i^*, $i = 1, 2$. Any 2 distinct blocks will either have no symbol in common (then we call these blocks 1st associates of each other) or they will have exactly one symbol in common (then we call these blocks 2nd associates of each other). Each block is called the 0th associate of itself. Clearly, any 2 distinct blocks of d_2^* will be 2nd associates, while any two distinct blocks of d_1^* may be either 1st or 2nd associates.

Example 1: PBIB design with GD scheme $d_1(v_1 = 6, b_1 = 9, r_1 = 3, k_1 = 2, \lambda_1 = 0, \lambda_2 = 1)$, blocks shown as columns labeled $1, \ldots, 9$.

	1	2	3	4	5	6	7	8	9
d_1 :	1	1	1	2	2	2	3	3	3
	4	5	6	4	5	6	4	5	6

Dual of d_1: $d_1^*(v_1^* = 9, b_1^* = 6, r_1^* = 2, k_1^* = 3)$, blocks shown as columns labeled $B_1, \ldots, B_6$.

	B_1	B_2	B_3	B_4	B_5	B_6
d_i^* :	1	4	7	1	2	3
	2	5	8	4	5	6
	3	6	9	7	8	9

Example 2: BIB design d_2 $(v_2 = 9, b_2 = 12, r_2 = 4, k_2 = 3, \lambda = 1)$, blocks shown as columns labeled $1, \ldots, 12$.

	1	2	3	4	5	6	7	8	9	10	11	12
d_2 :	4	7	1	5	8	2	6	9	3	1	4	7
	7	1	4	8	2	5	9	3	6	2	5	8
	2	5	8	3	6	9	1	4	7	3	6	9

Dual of d_2: $d_2^*(v_2^* = 12, b_2^* = 9, r_2^* = 3, k_2^* = 4)$, blocks shown as columns labeled $C_1, \ldots, C_9$.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
d_2^* :	2	1	4	1	2	5	1	3	6
	3	5	8	3	4	7	2	4	7
	7	6	9	8	6	9	9	5	8
	10	10	10	11	11	11	12	12	12

3.3. Use of block designs

We can build useful key predistribution schemes based on block designs because using the combinatorial structures of the designs we can

1. study the connectivity property of the scheme
2. study the resilience property of the scheme, and
3. carry out shared-key discovery and path-key establishment in a structured manner.

Schemes are evaluated on the basis of their connectivity and resilience using the measures Pr_1, Pr_2 and $fail(s)$ as proposed by Lee and Stinson (2004) and defined below:

For any 2 randomly chosen nodes in each other's neighbourhood, let Pr_1 be the probability that the 2 nodes can securely communicate directly with each other, *i.e.*, they have q keys in common.

Again, for any 2 randomly chosen nodes in each other's neighbourhood, let Pr_2 be the probability that these 2 nodes do not share q common keys but there is a third key in the neighbourhood of both of them which shares q common keys with both these nodes. So these 2 nodes can communicate securely via this third node.

Finally, $Pr_1 + Pr_2$ is used to study the connectivity of a KPS, either through a secure direct path, or through a secure path *via* a third node. The larger the value of $Pr_1 + Pr_2$, the better is the connectivity of the KPS.

In the event of an attack a number of nodes are compromised and the keys in the compromised nodes are rendered unusable for communication. Let A and B be 2 uncompromised nodes which share q common keys. Then, the resilience of the KPS is measured by $fail(s)$ which is equal to the conditional probability that the link between A and B will fail, when out of the other $n-2$ nodes, s randomly chosen nodes are compromised. A smaller $fail(s)$ means a larger resilience property for the KPS.

Several researchers have studied this problem. Lee and Stinson (2004) considered KPS with $q = 1$ and $q = 2$ and used transversal designs for their construction. Bose, Dey and Mukerjee (2013) studied KPS for general q and used various types of designs for their construction, *e.g.*, BIBD, PBIBD based on GD, LS and triangular association schemes, and suitable duals of these designs, for general q.

3.4. An illustration of the construction of KPS for $q = 2$

We now illustrate how duals of some suitable block designs can be used in the construction of the schemes. For our illustration, we use the designs shown in Section 3.2. For more examples, details and theoretical justifications, we refer to Bose, Dey and Mukerjee (2013). We only consider the case where $q = 2$; the case with $q = 1$ is easier and omitted here.

We can onstruct a KPS with $q = 2$ as follows:

(1) We start with 2 designs, each being either a PBIB design with $\lambda_1 = 0, \lambda_2 = 1$, or a BIB design with $\lambda = 1$, and then we consider their dual designs. *e.g.*, we start with d_1 and d_2 shown in Section 3.2 and take their duals d_1^* and d_2^*.

(2) We identify the symbols of d_1^* and d_2^* as the keys. So, the number of possible keys is $v_1^* + v_2^* = b_1 + b_2$, which equals to $9 + 12 = 21$ keys in our example.

(3) We take all possible selections of a block from each of d_1^* and d_2^*, and consider their union as a node. So, any node in our example is of the form: $B_i \cup C_j,\ i = 1, \ldots, 6,\ j = 1, \ldots, 9$. Thus we get the number of nodes as $n = b_1^* \times b_2^* = v_1 \times v_2$, which equals $6 \times 9 = 54$ nodes in our example. Each of these nodes have $k_1^* + k_2^* = r_1 + r_2$ keys, which equals $3+4 = 7$ keys in our example.

We can check the properties of the KPS from the properties of the constituent designs.

For example, by taking the union of block B_1 from d_1^* and block C_1 from d_2^*, and writing the symbols of d_2^* in italics to differentiate them from the symbols of d_1^*, we get the node as

$$B_1 \cup C_1 = 1\ 2,\ \ 3,\ \ \textit{2},\ \ \textit{3},\ \ \textit{7},\ \ \textit{10}$$

Similarly, taking union of block B_3 from d_1^* and block C_4 from d_2^*, and writing the symbols of d_2^* in italics, we get the node as

$$B_3 \cup C_4 = 7,\ \ 8,\ \ 9,\ \ \textit{1},\ \ \textit{3},\ \ \textit{8},\ \ \textit{11}$$

Note that B_1 and B_3 have no symbol in common and hence these blocks are 1st associates of each other. Again, blocks C_1 and C_4 have 1 symbol in common and hence these blocks are 2nd associates of each other. So we will say that the 2 nodes given by $B_1 \cup C_1$ and $B_3 \cup B_4$ are *12*th associates of each other.

Now, since d_1 is a PBIB design with 2 associate classes, blocks of d_1^* can be either 0, 1, or 2 associates. Again, as d_2 is a BIB design, blocks of d_2^* can be either 0, or 2 associates. So the association relationship between any 2 *distinct* nodes $B_{i_1} \cup C_{j_1}$ and $B_{i_2} \cup Cj_2$ in this KPS will be given by the set

$$\{02, 10, 12, 20, 22\}$$

Using this association structure between two nodes, we can deduce which two nodes can directly communicate with each other and which two nodes need a path via a third node to communicate.

It can be shown that with $q = 2$, all pairs of nodes except those which are *12* associates of each other can communicate directly with one another.

In this example, it can be checked that the number of *12* associates of any node in the KPS is 16. So the remaining $54 - 16 = 38$ nodes can directly communicate with each other.

Algebraic expressions for Pr_1, Pr_2 and $fail(s)$ can also be obtained using the combinatorial properties of the component designs. We omit the details here.

3.5. Evaluating Local connectivity and resilience for the above KPS

For the KPS obtained from the designs d_1^* and d_2^*, it may be shown that:

$$Pr_1 = 0.6981, \quad Pr_2 = \frac{16}{53}[1 - (1 - \frac{29}{52})^{\eta}]$$

where the intersection of the neighbourhoods of nodes A and B contain η nodes, excluding A and B themselves. So for $q = 2$ and for some choices of η, the probability that any 2 randomly chosen nodes in the KPS can communicate with each other is equal to

η	1	2	3	4	5	10	15	20
$Pr_1 + Pr_2$	0.867	0.941	0.974	0.988	0.995	0.9999	1.000	1.000

The above table shows that this KPS has quite high local connectivity. Different choices of the constituent designs will lead to different KPS and their metrics can be computed.

This idea of construction for $q = 2$ can be extended to general $q(\geq 2)$ where we start with q suitable initial designs, take their duals, and then form KPS as in steps (1), (2) and (3) in Section 3.3. Each time, n is multiplicative in the b_i^* while k is additive in k_i^*. Thus, this method gives schemes with many nodes but small key storage. The properties of such KPS can be similarly ascertained from the properties of the designs.

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

Adhikari, A. and Bose, M. (2004). Construction of new visual threshold schemes using combinatorial designs. *IEICE Transactions*, **E87-A**, 1198-1202.

Adhikari, A., Bose, M., Kumar, D., and Roy, B. (2007). Applications of partially balanced incomplete block designs in developing (2, n) Visual cryptographic Schemes. *IEICE Transactions*, **E90-A**, 949-951.

Bauer, C. P. (2013). *Secret History: The Story of Cryptology.* Chapman and Hall/CRC.

Blundo, C., De Santis, A., and Stinson, D. R. (1999). On the contrast in visual cryptography schemes. *Journal of Cryptology*, **12**, 261-289.

Bose, M, Dey, A., and Mukerjee, R. (2013). Key predistribution schemes distributed sensor networks via block designs. *Design, Codes and Cryptography*, **467**, 111-136.

Bose, M. and Mukerjee, R. (2006). Optimal $(2, n)$ visual cryptographic schemes. *Design, Codes and Cryptography*, **40**, 255-267.

Bose, M. and Mukerjee, R. (2010). Optimal (k, n) visual cryptographic schemes for general k. *Designs, Codes and Cryptography*, **55**, 19-35.

Bose, M. and Mukerjee, R. (2013). Union distinct families of sets, with an application to cryptography. *Ars Combinatoria,* **110**, 179-192.

Bose, M. and Mukerjee, R. (2014). An unequal probability scheme for improving anonymity in shared key operations. *Journal of Statistical Theory and Practice*, **8**, 100-112.

Bose, R. C. and Shrikhande, S. S. (1959). A note on result in the theory of code construction. *Information and Control*, **2**, 183–194.

Bush, K. A. (1952). Orthogonal arrays of index unity. *Annals of Mathematical Statistics*, **23**, 416-434.

Clatworthy, W. H. (1973). *Tables of Two-associate Partially Balanced Designs*. National Bureau of Standards, Applied Maths, series no **63**, Washington D.C.

Climato, S., Prisco, R. D., and Santis, A. D. (2005). Optimal coloured threshold visual cryptography schemes. *Designs Codes and Cryptography*, **35**, 311-335.

Hedayat, A. S., Stufken, J., and Sloane, N. J. A. (1999). *Orthogonal Arrays: Theory and Applications*. Springer-Verlag, New York.

Ibrahim, D. R., Teh, J. S., and Abdullah, R. (2021). An overview of visual cryptography techniques. *Multimedia Tools and Applications*, **80**, 31927-31952.

Kahn, D. (1996). *The Codebreakers: The Comprehensive History of Secret Communication from Ancient Times to the Internet*. Simon and Schuster.

Kang, I., Arce, G., and Lee, H. K. (2011). Color extended visual cryptography using error diffusion. *IEEE Transactions on Image Processing*, **20**, 132-145.

Kang, I., Sinha, K., and Lee, H. K. (2006). New digital fingerprint code scheme using group-divisible design. *IEICE Transactions on Fundamentals*, **E89-A**, 3732-3735.

Lee, J. and Stinson, D. R. (2004). Deterministic key predistribution schemes for distributed sensor networks. *SAC 2004 Proceedings, Lecture notes in Computer Science*, **3357**, 294-307.

Mukerjee, R. and Wu, C. F. J. (2006) *A Modern Theory of Factorial Design.* Springer, New York.

Naor, M. and Shamir, A. (1994). Visual cryptography. Advances in Cryptology, Eurocrypt'94. Lecture Notes in Computer Science, **950**, 1-12, Springer-Verlag.

Plotkin, M. (1960). Binary codes with specified minimum distance. *IRE Transactions*, **IT-6** 445–450.

Raghavarao, D. (1971). *Construction and Combinatorial Problems in Design of Experiments*. New York, Wiley.

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

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

Serberry, J., Wysocky, B. J., and Wysocki T. A. (2005). On some applications of Hadamard matrices. *Metrika*, **62**, 221-239.

Shah, K. and Sinha, B. K. (1989). *Theory of Optimal Designs*. Springer-Verlag, New York.

Stinson, D. R. (2004). *Combinatorial Designs / Constructions and Analysis*. Springer, ISBN 978-0-387-95487-5.

Stinson, D. R. and Paterson, M. B. (2023). *Cryptography: Theory and Practice*. 4th ed. CRC Press.

Street, A. P. and Street, D. J. (1987). *Combinatorics of Experimental Design.* Oxford. Clarendon Press.

Takeuchi, K. (1962). A table of difference sets generating balanced incomplete block designs. *Review of International Statistical Institute*, **30**, 361-366.

Trappe, W., Wu, M., Wang, Z. J., and Liu, K. J. R. (2003). Anti-collusion finger-printing for multimedia. *IEEE Transactions on Signal Processing*, **51**, 1069-1087.

Yagi, H., Matsushima, T., and Hirasawa, S. (2007). Improved collusion-secure codes for digital fingerprinting based on finite geometries. *IEEE International Conference on System, Man and Cybernetics*, 948-953.

Yarlagadda, R. K. and Hershey, J. E. (1997). *Hadamard Matrix Analysis and Synthesis: with Applications to Communications and Signal/Image Processing*. Kluwer.

Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025; pp 87-100

Large Language Models in Practice: Training Paradigms, Knowledge Systems, and Production-Scale Deployments

Utkarsh Tripathi
Solventum Health Information Systems, Pittsburgh, PA

Received: 26 July 2025; Revised: 03 August 2025; Accepted: 06 August 2025

Abstract

This survey presents a comprehensive overview of current methodologies and challenges in the development of large language models (LLMs), focusing on training processes, knowledge integration techniques, and evaluation frameworks. The review examines both traditional and innovative approaches, including the DeepSeek methodology, and discusses critical challenges such as static knowledge limitations, hallucinations, and the need for robust guardrails. The analysis covers the full spectrum from foundational training to production deployment, providing insights into the evolving landscape of LLM systems and their practical applications.

Key words: Large language models; Knowledge bases; RLHF18.

1. Introduction

Large Language Models (LLM) have emerged as transformative technologies in artificial intelligence, demonstrating remarkable capabilities across diverse natural language processing tasks. However, their development, deployment, and evaluation present complex challenges that require sophisticated frameworks and methodologies. This survey synthesizes current approaches to LLM training, knowledge integration, and evaluation, drawing from recent advances in the field and practical implementation experiences.

The rapid evolution of LLMs necessitates a comprehensive understanding of their underlying mechanisms, from initial training processes to production-ready systems. This review addresses key challenges including knowledge cutoff limitations (Chen *et al.*, 2023), hallucination mitigation (Zhang *et al.*, 2023), and the development of robust evaluation metrics that ensure both performance and safety.

2. Large language model training frameworks

2.1. Traditional training pipeline

The conventional LLM training process follows a structured approach involving several critical stages, each presenting unique technical challenges and optimization opportuni-

Corresponding Author: Utkarsh Tripathi
Email: utkarshbitsp@gmail.com

ties.

2.2. Data collection and preprocessing

The foundation of any LLM involves gathering vast amounts of text data from diverse sources and preparing it for training. This stage encompasses several processes: (1) **Web crawling and curation**, where large-scale internet scraping operations collect terabytes of textual data from websites, forums, and digital repositories, requiring advanced filtering mechanisms to ensure quality and remove duplicates (OpenAI, 2023); (2) **Multilingual corpus construction**, involving the careful balance of languages to prevent model bias toward dominant languages while ensuring adequate representation of low-resource languages; (3) **Quality assessment algorithms**, implementing perplexity-based filtering, n-gram overlap detection, and semantic coherence scoring to eliminate low-quality content; and (4) **Tokenization strategies**, employing subword tokenization methods like Byte-Pair Encoding (BPE) or SentencePiece to handle out-of-vocabulary words and optimize vocabulary size for computational efficiency (Vaswani *et al.*, 2017).

2.3. Self-supervised learning

Models are trained to predict missing words in sequences through attention mechanisms that enhance language understanding *via* pattern recognition and contextual learning (Vaswani *et al.*, 2017). This phase implements the transformer architecture's core innovation: multi-head self-attention, where the model computes attention weights $A_{ij} = \frac{\exp(Q_i K_j^T/\sqrt{d_k})}{\sum_{k=1}^{n} \exp(Q_i K_k^T/\sqrt{d_k})}$, allowing each position to attend to all positions in the input sequence. The self-supervised objective maximizes the likelihood $\mathcal{L} = \sum_{t=1}^{T} \log P(x_t|x_{<t})$, where the model learns to predict token x_t given all previous tokens. This approach builds fundamental language capabilities through masked language modeling (MLM) and next sentence prediction (NSP) tasks, establishing the semantic and syntactic understanding necessary for more complex reasoning tasks.

2.4. Supervised learning and fine-tuning

The transition from self-supervised pre-training to supervised fine-tuning adapts the model for specific tasks using curated instruction datasets. This process uses gradient-based optimization, updating parameters as $\theta_{t+1} = \theta_t - \alpha \nabla_\theta \mathcal{L}(\theta)$, where $\mathcal{L}(\theta)$ is the task-specific loss. The supervised phase employs: (1) **Instruction tuning**, where models learn to follow human instructions via prompt-response datasets; (2) **Task-specific adaptation**, involving fine-tuning on datasets such as SQuAD for question-answering or WMT for translation; and (3) **Multi-task learning**, where models simultaneously optimize multiple objectives to boost generalization.

2.5. Distributed training infrastructure

The computational intensity of LLM training requires parallel computing architectures that utilize multiple GPUs in distributed systems (Shoeybi *et al.*, 2020). Modern training implementations employ several parallelization strategies: (1) **Data parallelism**, where different GPU nodes process separate batches of data while maintaining synchro-

nized model parameters through all-reduce operations; (2) **Model parallelism**, splitting the model architecture across multiple devices, particularly useful for models exceeding single-GPU memory capacity; (3) **Pipeline parallelism**, dividing the model into sequential stages across different devices, enabling concurrent processing of different micro-batches; and (4) **Tensor parallelism**, partitioning individual tensor operations across multiple devices to handle extremely large parameter matrices.

2.6. DeepSeek methodology: an alternative paradigm

The DeepSeek training framework (DeepSeek Team, 2024) follows established LLM training practices with architectural innovations, implementing a mixture-of-experts (MoE) architecture with 671B total parameters and 37B activated parameters. DeepSeek-R1 specifically uses reinforcement learning without supervised fine-tuning to develop reasoning capabilities.

2.7. Architectural innovations

The DeepSeek architecture integrates several advanced components: (1) **Multi-head latent attention mechanisms**, extending traditional attention by incorporating latent variable modeling where attention weights are computed through a latent space z: $A_{ij} = \text{softmax}(f(Q_i, K_j, z))$, allowing for more flexible attention patterns; (2) **Chain-of-Thought integration** (Wei *et al.*, 2022), embedding reasoning pathways directly into the model architecture through specialized attention heads that track logical dependencies; (3) **Mixture of Experts (MoE) architectures** (Fedus *et al.*, 2022), implementing sparse activation patterns where only a subset of parameters are active for any given input, defined by the gating function $G(x) = \text{softmax}(W_g \cdot x)$ that routes inputs to appropriate expert networks.

2.8. Training methodology distinctions

The DeepSeek approach differs fundamentally from standard training in several key aspects:

Data Usage Philosophy: While conventional approaches require extensive human-labeled datasets often exceeding billions of examples, DeepSeek employs a cold-start methodology with minimal initial supervision, typically requiring only thousands of high-quality seed examples. The system then implements iterative synthetic data generation through rejection sampling, where candidate responses are generated and filtered based on quality metrics $Q(r) = \alpha \cdot \text{coherence}(r) + \beta \cdot \text{relevance}(r) + \gamma \cdot \text{factuality}(r)$.

Reinforcement Learning Integration: Traditional RLHF (Ouyang *et al.*, 2022) applies reinforcement learning as a post-processing step, whereas DeepSeek integrates RL throughout the training process. The system alternates between supervised fine-tuning phases and pure reinforcement learning episodes, implementing policy gradient methods where the policy $\pi_\theta(a|s)$ is updated according to $\nabla_\theta J(\theta) = \mathbb{E}[\nabla_\theta \log \pi_\theta(a|s) A(s,a)]$, where $A(s,a)$ represents the advantage function estimating the quality of action a in state s.

3. Foundational challenges in LLM knowledge systems

3.1. Static knowledge limitations and temporal boundaries

Large language models face fundamental epistemological challenges related to knowledge representation and temporal validity that significantly impact their practical deployment and reliability.

3.2. The knowledge freeze problem

LLMs experience "knowledge freeze" at their training cutoff dates, creating a temporal boundary beyond which the model lacks awareness of events, discoveries, or factual updates (Chen *et al.*, 2023). Scientific knowledge continuously evolves, with research showing that various domains experience different rates of knowledge obsolescence, though specific quantification varies significantly across fields. One possible mathematical representation of knowledge decay could follow an exponential decay model $K(t) = K_0 \cdot e^{-\lambda t}$, though empirical validation of such models remains an area of active research.

The implications extend beyond simple factual updates to encompass: (1) **Causal relationship evolution**, where the relationships between entities change over time, requiring dynamic graph structures to represent evolving knowledge networks; (2) **Semantic drift**, where word meanings and contextual associations shift, particularly in rapidly evolving domains like technology and social media; and (3) **Emerging concept integration**, where entirely new concepts, terminologies, or frameworks arise that require knowledge incorporation mechanisms.

3.3. Parametric versus non-parametric knowledge trade-offs

The tension between internal (parametric) knowledge storage and external (non-parametric) knowledge retrieval presents complex optimization challenges. Parametric knowledge, encoded within model weights, offers rapid access but suffers from staleness and limited update mechanisms. The storage capacity can be estimated as $C = \frac{N \cdot \log_2(Q)}{B}$ bits, where N is the number of parameters, Q is quantization levels, and B is bits per parameter.

Non-parametric knowledge systems, while offering currency and updateability, introduce latency and consistency challenges. The trade-off can be formalized as an optimization problem: $\min_\alpha \alpha \cdot \text{Latency(retrieval)} + (1 - \alpha) \cdot \text{Staleness(parametric)}$, where α balances between retrieval overhead and knowledge currency.

3.4. The hallucination frontier

Hallucinations represent a critical failure mode where models generate seemingly plausible but factually incorrect information (Zhang *et al.*, 2023; Ji *et al.*, 2023; Manakul *et al.*, 2023). The phenomenon occurs primarily when models encounter queries that exceed their knowledge boundaries, leading to confabulation based on statistical patterns rather than factual grounding.

Research has identified several hallucination triggers: (1) **Knowledge boundary proximity**, where queries approach the limits of training data coverage; (2) **Confidence**

calibration failures, where models express high confidence in incorrect information; (3) **Context insufficient disambiguation**, where ambiguous queries lead to incorrect assumption propagation; and (4) **Training data biases**, where systematic errors in training corpora propagate to model outputs.

Mitigation strategies include: (1) **Uncertainty quantification**, implementing Bayesian approaches to estimate prediction confidence: $P(y|x) = \int P(y|x,\theta)P(\theta|D)d\theta$; (2) **Attention mechanism analysis**, monitoring attention patterns to detect when models rely on weak or irrelevant context; and (3) **Consistency checking**, validating responses through multiple generation paths and cross-referencing.

3.5. The paradox of re-use and training data ecosystem

An emerging concern involves the "paradox of re-use," where increased LLM adoption potentially degrades the quality of future training data through feedback loops. As LLMs generate increasing amounts of web content, subsequent training iterations may incorporate model-generated text, leading to potential quality degradation through recursive training effects.

This phenomenon can be modeled as a Markov chain where each generation G_n of models trains on data that includes outputs from previous generations: $D_{n+1} = (1 - \rho)D_{\text{human}} + \rho \sum_{i=1}^{n} \alpha_i O_{G_i}$, where ρ represents the proportion of synthetic content, and α_i weights the contribution of generation i outputs.

4. Knowledge editing methodologies

4.1. Retrieval-augmented generation

RAG systems (Lewis *et al.*, 2020) implement sophisticated information retrieval pipelines that dynamically incorporate external knowledge during generation. The architecture comprises several interconnected components operating in a coordinated fashion. The **embedding subsystem** converts both queries and document collections into high-dimensional vector representations using transformer-based encoders. Query embedding $q = \text{Encoder}_q(x)$ and document embeddings $d_i = \text{Encoder}_d(doc_i)$ are typically generated using models like BERT or specialized sentence transformers, producing dense vectors in $\mathbb{R}^d$ where d commonly ranges from 384 to 1024 dimensions.

The **retrieval mechanism** implements similarity search through vector databases (Anderson *et al.*, 2023) that support efficient approximate nearest neighbor queries. The similarity function, typically cosine similarity $\text{sim}(q, d_i) = \frac{q \cdot d_i}{||q|| \cdot ||d_i||}$, ranks documents by relevance. Advanced implementations employ learned sparse retrieval methods combining dense embeddings with traditional term-frequency approaches.

The **context integration module** aggregates the retrieved information with the original query through prompt engineering, where the selected documents are formatted according to task-specific templates that optimize information utilization while respecting context length constraints.

4.2. Multi-level RAG complexity framework

A comprehensive survey by researchers (Lewis *et al.*, 2020) categorizes RAG tasks into four levels based on external data requirements and reasoning complexity:

Level 1: Explicit Fact Queries implement direct factual lookup mechanisms where queries map to specific knowledge entries. The retrieval function operates as $R(q) = \arg\max_{d \in D} \text{match}(q, d)$, where exact or near-exact matches suffice for response generation. This level handles queries like "What is the capital of France?" through straightforward entity-attribute lookups.

Level 2: Implicit Fact Queries require multi-hop reasoning across connected knowledge pieces. The system must identify relevant fact chains $\{f_1, f_2, \ldots, f_n\}$ where each fact f_i provides context for subsequent facts. The reasoning process implements graph traversal algorithms over knowledge graphs, where edges represent relationships and nodes represent entities or concepts.

Level 3: Interpretable Rationale Queries extend beyond factual retrieval to incorporate logical reasoning patterns from external sources. The system must identify and apply reasoning templates that provide step-by-step solution methodologies. This involves template matching where query patterns $P(q)$ are matched against reasoning frameworks $R(t)$ to generate structured response sequences.

Level 4: Hidden Rationale Queries requires discovery of implicit reasoning strategies not explicitly present in retrieved documents. The system must synthesize reasoning approaches from multiple sources, implementing meta-learning mechanisms that identify optimal problem-solving strategies for novel query types.

4.3. Hypernetwork-based knowledge updates

Hypernetworks (Ha *et al.*, 2016) provide a mechanism for targeted knowledge modification without full model retraining. These auxiliary networks generate weight modifications for the primary model, implementing the transformation W'=W+H(c), where W represents original weights, H is the hypernetwork function, and c is the conditioning context representing the knowledge update requirement.

The hypernetwork architecture typically employs a multi-layer perceptron that takes knowledge update specifications as input and produces delta weights for specific model components. The training objective minimizes $\mathcal{L} = \mathbb{E}_{(x,y,c)}[||f(x; W + H(c)) - y||^2]$, where f represents the primary model, and (x, y, c) are input-output-context triples representing desired knowledge updates.

Advanced implementations employ attention mechanisms within hypernetworks to selectively modify relevant parameter subsets, reducing computational overhead and minimizing interference with existing knowledge. The attention-weighted modification becomes W prime equals W plus the sum over i of alpha sub i times H sub i of c, where alpha sub i represents attention weights determining the relevance of each hypernetwork component.

4.4. Localized knowledge neuron editing

Recent research has identified specific neural pathways responsible for the storage of factual knowledge within transformer architectures (Meng *et al.*, 2021). These "knowledge neurons" can be precisely targeted for updates without affecting broader model capabilities.

The identification process employs gradient-based attribution methods, computing $\nabla_{h_i} \log P(y|x)$ for each hidden unit h_i to determine its contribution to specific factual predictions. Neurons with high attribution scores for particular facts become candidates for targeted modification.

The editing process implements constrained optimization where knowledge neuron activations are modified to reflect updated information while preserving surrounding model behavior. The objective function balances update accuracy with behavioral consistency: $\min_{\Delta W} ||f(x_{edit}; W + \Delta W) - y_{new}||^2 + \lambda \sum_{x \in X_{preserve}} ||f(x; W + \Delta W) - f(x; W)||^2$.

4.5. Continual learning integration

Continual learning approaches (Parisi *et al.*, 2019) enable incremental knowledge updates while mitigating catastrophic forgetting. These methods implement memory systems and regularization techniques to maintain previously acquired knowledge during updates.

Elastic Weight Consolidation (EWC) computes parameter importance scores based on Fisher Information Matrix diagonal elements: $F_i = \mathbb{E}[(\frac{\partial \log P(y|x)}{\partial \theta_i})^2]$. The regularization term $\lambda \sum_i F_i(\theta_i - \theta_i^*)^2$ prevents important parameters from deviating significantly during updates.

Progressive Neural Networks implement modular architectures where new knowledge modules are added while preserving existing ones. The architecture employs lateral connections $h_i^{(k)} = f(W^{(k)} h_{i-1}^{(k)} + \sum_{j<k} U_i^{(k:j)} h_{i-1}^{(j)})$, where knowledge from previous modules j influences current module k processing.

Memory-Augmented Networks maintain explicit episodic memories of previous learning experiences, implementing retrieval mechanisms that recall relevant examples during new learning episodes. The memory update process balances between adding new experiences and maintaining diverse historical knowledge.

5. Evaluation frameworks for LLM systems

5.1. Perplexity-based assessment

Perplexity serves as a fundamental intrinsic evaluation metric measuring model uncertainty in predicting text sequences (Brown *et al.*, 2023). Mathematically defined as $\text{PPL}(X) = \exp\left(-\frac{1}{N} \sum_{i=1}^{N} \log P(x_i|x_{<i})\right)$, perplexity quantifies the model's predictive confidence, with lower values indicating superior language modeling capabilities.

Advanced perplexity analysis employs domain-specific decomposition, computing separate scores for different text types: $\text{PPL}_{\text{domain}} = \exp\left(-\frac{1}{N_{\text{domain}}} \sum_{x \in D_{\text{domain}}} \log P(x)\right)$. This approach reveals model strengths and weaknesses across different knowledge domains and

text genres.

Conditional perplexity measurements evaluate model performance given specific contexts or constraints, implementing $\text{PPL}(X|C) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(x_i|x_{<i}, c)\right)$, where c represents conditioning information. This metric proves particularly valuable for assessing context utilization in RAG systems and domain adaptation effectiveness.

5.2. Reference-based similarity metrics

BLEU Score Implementation (Papineni *et al.*, 2002) computes n-gram overlap between generated and reference texts through the geometric mean of precision scores: $\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$, where p_n represents n-gram precision and BP is the brevity penalty addressing length disparities.

The metric implements modified precision calculations to prevent repetition: p_n = (sum over C in Candidates of sum over n-gram in C of $Count_{clip}(n-gram)$ / (sum over C' in Candidates of sum over n-gram' in C' of Count(n-gram')), where $Count_{clip}$ limits n-gram counts to reference frequencies.

ROUGE Metrics (Lin, 2004) implement recall-oriented evaluation through various formulations: ROUGE-N computes n-gram recall, ROUGE-L employs longest common subsequence matching, and ROUGE-S utilizes skip-bigram co-occurrence. The ROUGE-L formulation $R_{lcs} = \frac{LCS(X,Y)}{m}$ and $P_{lcs} = \frac{LCS(X,Y)}{n}$ compute recall and precision based on longest common subsequences, providing robust similarity assessment for variable-length outputs.

BLEU Score Implementation (Papineni *et al.*, 2002) computes n-gram overlap between generated and reference texts through the geometric mean of precision scores: $\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$, where p_n represents n-gram precision and BP is the brevity penalty addressing length disparities.

ROUGE Metrics (Lin, 2004) implement recall-oriented evaluation through various formulations: ROUGE-N computes n-gram recall, ROUGE-L employs longest common subsequence matching, and ROUGE-S utilizes skip-bigram co-occurrence. The ROUGE-L formulation $R_{lcs} = \frac{LCS(X,Y)}{m}$ and $P_{lcs} = \frac{LCS(X,Y)}{n}$ compute recall and precision based on longest common subsequences, providing robust similarity assessment for variable-length outputs.

5.3. Advanced evaluation methodologies: multi-dimensional human assessment

Human evaluation protocols implement structured assessment frameworks encompassing multiple quality dimensions. Evaluators assess responses across: (1) **Fluency**, measuring grammatical correctness and natural language flow; (2) **Coherence**, evaluating logical consistency and thematic unity; (3) **Relevance**, assessing response appropriateness to query context; (4) **Informativeness**, measuring content richness and factual density; and (5) **Truthfulness**, verifying factual accuracy and consistency with reliable sources.

Calibration techniques align LLM judge scores with human evaluations through regression models or distribution matching. The calibration function $f : S_{\text{LLM}} \to S_{\text{human}}$ learns

mappings from LLM scores to human-equivalent scores, improving evaluation validity.

5.4. Benchmark dataset assessment

Standardized benchmarks provide systematic performance comparison across models and methodologies. **GLUE and SuperGLUE** implement comprehensive evaluation suites covering diverse NLP tasks including sentiment analysis, textual entailment, and question answering. Performance aggregation employs weighted averages accounting for task difficulty and dataset size.

Domain-specific benchmarks evaluate specialized capabilities such as mathematical reasoning (GSM8K), coding proficiency (HumanEval), and scientific knowledge (SciBench). These benchmarks implement rigorous evaluation protocols with automated scoring systems and comprehensive test suites covering edge cases and challenging scenarios.

5.5. Adversarial robustness testing

Adversarial evaluation assesses model robustness through deliberately challenging inputs designed to expose failure modes. Techniques include: (1) **Prompt injection attacks**, testing resistance to malicious instruction manipulation; (2) **Context manipulation**, evaluating performance degradation under misleading or contradictory context; (3) **Semantic perturbations**, testing sensitivity to paraphrasing and synonym substitution; and (4) **Out-of-distribution queries**, assessing behavior on inputs significantly different from training data.

The evaluation protocol implements systematic perturbation generation through automated techniques and human-crafted challenging examples. Robustness metrics quantify performance degradation: $R = 1 - \frac{\text{Performance}_{\text{adversarial}}}{\text{Performance}_{\text{clean}}}$, where lower values indicate better robustness.

6. LLM guardrails and safety frameworks

6.1. Input validation and preprocessing

Modern LLM deployment requires comprehensive safety frameworks implementing defense-in-depth strategies across multiple system layers (Johnson *et al.*, 2024).

The first line of defense implements input analysis to detect potentially harmful or manipulative queries. **Prompt injection detection** employs trained classifiers that identify attempts to override system instructions or extract sensitive information. The detection system analyzes query patterns using features such as instruction keywords, context breaks, and linguistic anomalies.

The classifier implements a multi-stage approach: (1) **Syntactic analysis** identifying structural patterns common in injection attempts; (2) **Semantic analysis** using embedding similarity to detect attempts to mimic system prompts; and (3) **Contextual analysis** evaluating query appropriateness given conversation history and system role.

Content sanitization processes inputs to remove or neutralize potentially harmful

elements while preserving legitimate query intent. This involves entity recognition for sensitive information, toxicity scoring using specialized models, and context-aware filtering that considers domain-specific content policies.

6.2. Response generation controls

During the generation process, multiple safeguards ensure output quality and safety. **Real-time monitoring** tracks model attention patterns and internal states to detect potential safety violations before completion. The monitoring system implements threshold-based intervention where concerning patterns trigger alternative generation paths or safety responses.

Content filtering pipelines evaluate generated text across multiple dimensions: (1) **Toxicity detection** using specialized classifiers trained on harmful content datasets; (2) **Bias assessment** measuring unfair treatment of protected groups through demographic parity metrics; (3) **Factuality verification** cross-referencing claims against reliable knowledge bases; and (4) **Coherence validation** ensuring logical consistency and topical relevance.

6.3. Post-processing and output validation

The final safety layer implements comprehensive output validation before response delivery. **Multi-model consensus** employs multiple independent models to evaluate response quality and safety, implementing voting mechanisms where responses require majority approval for release.

Dynamic policy enforcement applies context-sensitive rules based on user profiles, conversation history, and application domain. The rule engine implements conditional logic trees evaluating multiple safety criteria simultaneously.

Audit trail generation maintains comprehensive logs of all safety interventions, enabling continuous improvement of safety systems through analysis of edge cases and system failures.

7. Production-scale LLM infrastructure

7.1. Data pipeline infrastructure

Production LLM systems require integrated data processing pipelines handling diverse input types and sources. **Stream processing systems** like Apache Kafka and Apache Pulsar manage real-time data ingestion with low latency and high throughput requirements. The architecture implements pub-sub patterns enabling scalable data distribution across processing components.

ETL frameworks such as Apache Airflow orchestrate complex data transformation workflows, implementing DAG-based scheduling with dependency management and error recovery mechanisms. These systems handle: (1) Data ingestion from multiple sources including APIs, databases, and file systems; (2) Transformation and normalization ensuring consistent data formats; (3) Quality validation through automated testing and anomaly detection; and (4) Loading into downstream systems with appropriate partitioning and indexing

strategies.

Vector database systems (Anderson *et al.*, 2023) provide specialized storage and retrieval for high-dimensional embeddings. Production implementations employ distributed architectures with horizontal scaling capabilities, implementing approximate nearest neighbor algorithms like HNSW or IVF for efficient similarity search. The query processing pipeline optimizes for both accuracy and latency through techniques such as query caching, index warming, and adaptive batching.

7.2. Model serving and orchestration

Model serving infrastructure implements request routing and load balancing across multiple model instances. The architecture employs containerized deployments using technologies like Docker and Kubernetes, enabling dynamic scaling based on demand patterns. Advanced implementations utilize model parallelism across multiple GPUs or nodes, implementing tensor sharding strategies that distribute computational load while maintaining response coherence.

Orchestration frameworks coordinate complex workflows involving multiple models, retrieval systems, and validation components. Systems like LangChain and LlamaIndex provide abstraction layers enabling composable AI workflows, implementing retry mechanisms, timeout handling, and fallback strategies for robust production operation.

Caching systems optimize performance through multi-level caching strategies: (1) Response caching storing complete answers for frequently asked questions; (2) Embedding caching maintaining computed vector representations; (3) Context caching preserving processed conversation history; and (4) Model state caching reducing initialization overhead for dynamically loaded models.

7.3. Performance monitoring systems

Production LLM systems require comprehensive monitoring across multiple performance dimensions. **Latency tracking** measures end-to-end response times with percentile-based analysis identifying performance outliers and degradation patterns. The monitoring system tracks: (1) Model inference time including tokenization and generation phases; (2) Retrieval system latency for RAG implementations; (3) Network communication overhead; and (4) Queue waiting times during high-load periods.

Throughput monitoring tracks request processing rates with capacity planning metrics. The system implements predictive scaling based on traffic patterns and resource utilization trends, automatically adjusting compute resources to maintain target performance levels.

Resource utilization tracking monitors GPU memory usage, CPU consumption, and network bandwidth to identify bottlenecks and optimization opportunities. Advanced implementations employ machine learning models to predict resource requirements and detect anomalous usage patterns indicating potential issues.

7.4. Quality assurance and safety monitoring

Response quality tracking implements automated assessment of output quality using multiple evaluation metrics. The system continuously monitors response relevance, coherence, and factual accuracy, alerting operators to quality degradation that may indicate model drift or system issues.

Safety violation detection tracks incidents where safety guardrails activate, analyzing patterns to identify potential attack vectors or system vulnerabilities. The monitoring system implements real-time alerting for serious safety violations while maintaining comprehensive audit logs for forensic analysis.

User satisfaction metrics collect implicit and explicit feedback signals, implementing sentiment analysis on user interactions and conversion rate tracking for task completion metrics. These signals provide early warning of system degradation and guide improvement efforts.

8. Conclusion and future directions

The landscape of LLM development encompasses sophisticated technical challenges requiring integrated solutions across training methodologies, knowledge management, evaluation frameworks, and production systems. The survey has examined the evolution from traditional training approaches to innovative methodologies like DeepSeek, highlighting the trade-offs between human supervision and automated optimization.

Key technical advances include the development of multi-level RAG systems that enable reasoning capabilities, the implementation of precise knowledge editing techniques targeting specific neural pathways, and the deployment of comprehensive safety frameworks addressing the complex challenges of production AI systems.

Future research directions encompass several critical areas: (1) Development of more efficient training paradigms that reduce computational requirements while maintaining or improving model capabilities; (2) Advanced knowledge integration techniques that enable real-time updates without catastrophic forgetting; (3) End to end evaluation frameworks that better capture real-world performance and safety characteristics; and (4) Scalable production architectures that can handle the increasing demands of widespread LLM deployment.

The convergence of these technical advances represents a pathway toward more capable, efficient, and trustworthy AI systems that can effectively serve diverse human needs while maintaining appropriate safety standards and ethical considerations. As the field continues to evolve rapidly, the integration of these comprehensive frameworks will be essential for realizing the full potential of large language models in practical applications.

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 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 work included in this article.

References

Anderson, R., Lee, J., and Martinez, C. (2023). Vector databases for large-scale similarity search. *ACM Computing Surveys*, **56**, 1–35.

Brown, S., Davis, M., and Wilson, J. (2023). Perplexity and its applications in language model evaluation. *Computational Linguistics*, **49**, 345–378.

Chen, W., Liu, M., and Zhang, Y. (2023). Temporal knowledge boundaries in large language models. *Nature Machine Intelligence*, **8**, 234–249.

DeepSeek Team (2024). Deepseek-r1: Incentivizing reasoning capability in llms *via* reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Fedus, W., Zoph, B., and Shazeer, N. (2022). Switch transformer: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, **23**, 1–39.

Ha, D., Dai, A., and Le, Q. V. (2016). Hypernetworks. *arXiv preprint arXiv:1609.09106*.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., and Fung, P. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Johnson, A., Smith, S., and Brown, D. (2024). Implementing safety guardrails for large language models in production. *AI Safety Review*, **12**, 78–95.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., *et al.*. (2020). Retrieval augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, **33**, 9459–9474.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81.

Manakul, P., Liusie, A., and Gales, M. J. (2023). Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9004–9017.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2021). Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, **34**, 17359–17372.

OpenAI (2023). Gpt-4 technical report. *https://arxiv.org/abs/2303.08774*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., *et al.*. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, **35**, 27730–27744.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, **113**, 54–71.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. (2020). Efficient large-scale language model training on gpu clusters using megatron-lm. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–15.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, **30**.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., *et al.*. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, **35**, 24824–24837.

Zhang, W., Chen, L., and Wang, M. (2023). Understanding and mitigating hallucinations in large language models. *Machine Learning Research*, **24**, 1234–1256.

Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025; pp 101-120

A Comparative Study of Estimation Methods for Nakagami Distribution in Reliability Analysis

Rahul Gupta[1] **and Bhagwati Devi**[2]

[1]*Department of Statistics, University of Jammu, Jammu, India*
[2]*Department of Statistics, Central University of Jharkhand, Ranchi, India*

Received: 20 May 2025; Revised: 08 August 2025; Accepted: 14 August 2025

Abstract

In this study, the Nakagami distribution is examined in the context of reliability analysis, focusing on key reliability measures. Various estimation techniques for the distribution parameters are explored and compared. A novel approach for deriving different estimators is introduced. Asymptotic confidence intervals for the parameters are constructed based on both MLE and log-MLE methods. In addition, hypothesis testing procedures are developed for different scenarios. The performance of the proposed estimation methods is assessed through a comprehensive Monte Carlo simulation study. Finally, the applicability of these methods is demonstrated using a real data set, providing clarity and practical insight into the estimation process.

Key words: Nakagami distribution; Reliability; Classical methods; Markov chain Monte Carlo.

AMS Subject Classifications: 62N05, 62E20 , 62F03 , 62F05, 62F10.

1. Introduction and preliminaries

Reliability function is the probability that a system performs its intended function without any failure at time t under the prescribed conditions. So, if we suppose that lifetime of an item or any system is denoted by the random variate X then the reliability is $R(t) = Pr(X > t)$. The other important measure of reliability is $P = Pr(X > Y)$ and this represents reliability of X (random strength) subject to Y (random stress). This is known as the reliability of an item under the stress strength set up. This measure is very useful to find the reliability of an item in no time like in case we want to test the reliability of an electric wire. Various authors have conducted estimation and testing of the reliability measure $R(t)$ and P considering different distributions. For literature, one can refer Pugh (1963), Basu (1964), Tong (1974, 1975), Johnson (1975), Sathe and Shah (1981), Chao (1982). Chaturvedi and Surinder (1999) developed the inferential procedures for testing these reliability measures of exponential distribution. Awad and Gharraf (1986) estimated P in case of Burr distribution. Tyagi and Bhattacharya (1989) and, Chaturvedi and Rani (1998) done esti-

Corresponding Author: Rahul Gupta
Email: rahulgupta68@yahoo.com

mation related to Maxwell and generalized Maxwell distributions respectively. Chaturvedi and Pathak (2012) derived inferential procedures for exponentiated Weibull and Lomax distributions. Chaturvedi and Rani (1997) and Chaturvedi and Tomer (2003) draw inferences for $R(t)$ and P for the families of lifetime distributions which are very useful as they cover many distributions as particular cases. Chaturvedi and Vyas (2018) have done estimation of $R(t)$ and P for three parameter Burr distribution under different censoring schemes.
In the present communication, a very important distribution known as the Nakagami distribution is taken into consideration which is most useful in communication engineering. It was Nakagami (1960) who proposed this distribution which models the fading of radio signals by name as Nakagami-m distribution with shape parameter m.
If a random variable (rv) X follows the Nakagami distribution with shape parameter $\alpha \geq 0.50$ and scale parameter $\lambda > 0$ then its probability density function (P.d.f) is as follows

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

Hereafter, we denote Nakagami distribution by $ND(\alpha,\lambda)$, where shape parameter α is known and scale parameter λ is unknown. The corresponding cumulative distribution function (cdf) of $ND(\alpha,\lambda)$ is given by,

$$F(x) = \frac{1}{\Gamma\alpha}\Gamma\left(\frac{\alpha}{\lambda}x^2,\alpha\right); \quad x>0, \alpha \geq 0.5, \lambda > 0. \tag{2}$$

where $\Gamma(x,a) = \int_0^x t^{a-1}e^{-t}dt$ is the lower incomplete gamma function.
The reliability function of $ND(\alpha,\lambda)$ is

$$R(t) = 1 - \frac{1}{\Gamma\alpha}\Gamma\left(\frac{\alpha}{\lambda}t^2,\alpha\right); \quad t>0, \alpha \geq 0.5, \lambda > 0 \tag{3}$$

The failure rate of $ND(\alpha,\lambda)$ is

$$h(t) = \frac{\frac{2}{\Gamma\alpha}\left(\frac{\alpha}{\lambda}\right)^{\alpha} t^{2\alpha-1} exp\left(-\frac{\alpha}{\lambda}t^2\right)}{1-\frac{1}{\Gamma\alpha}\Gamma\left(\frac{\alpha}{\lambda}t^2,\alpha\right)}; \quad t>0, \alpha \geq 0.5, \lambda > 0. \tag{4}$$

1.1. Relations with other distribution

1. For $\alpha = 0.5$, $ND(\alpha,\lambda)$ reduces to Half Normal distribution.
2. With $\alpha = 1, ND(\alpha,\lambda)$ becomes Rayleigh distribution.
3. If rv Y is distributed as $Gamma(k,\lambda)$ with shape k and scale λ then $\sqrt{Y}$ follows $ND(k,k\lambda)$.
4. If Z follows chi-square with parameter 2α and 2α is integer-valued then $\sqrt{\frac{\lambda}{2\alpha}Z}$ is $ND(\alpha,\lambda)$ variate.

This distribution has found applications in various disciplines such as in hydrology, multimedia data traffic over networks, medical imaging studies, in modeling of seismogram envelope of high frequency etc. For review, one may see Schwartz *et al.* (2013). Using the Monte Carlo simulation technique, Abdi and Kaveh (2000) made comparison of three different estimators of Nakagami-m distribution. Cheng and Beaulieu (2001) estimated the distribution using Maximum Likelihood method. Schwartz *et al.* (2013) discussed the estimation of the shape parameter using improved maximum likelihood estimation and also gave some distributional properties.

The main aim of this paper is to develop point estimation and hypotheses testing procedures for two measures of reliability *viz.*, $R(t)$ and P. In Section 2, we present point estimation when shape parameter is known but scale parameter is unknown. Uniformly Minimum Variance Unbiased Estimators (U.M.V.U.Es), Maximum Likelihood Estimators(M.L.Es) and Moment estimators have been found in this section. Section 3 comprises of point estimation when both scale and shape parameters are unknown. Asymptotic confidence intervals are developed for the parameters in Section 4. Section 5 is devoted for developing testing procedures for testing different hypotheses. In Section 6, we present the simulation study using Monte Carlo techniques with Section 6.1 devoted for the case when shape parameter α is known and scale parameter λ is unknown, Section 6.2 for the case when both α and λ are unknown and Section 6.3 for hypotheses testing . In Section 7, a real data study is performed and finally the paper is concluded in Section 8.

2. Point estimation when shape parameter is known

Let us take a random sample $X_1, X_2, \ldots, X_n$ from the model (1) having size n. Taking α to be known, the likelihood function of the parameter λ given the sample observations $\underline{x}$ comes out to be

$$L(\lambda|\underline{x}) = \prod_{i=1}^{n} f(x_i, \lambda) = \left(\frac{2\alpha^{\alpha}}{\Gamma\alpha}\right)^{n} \frac{1}{\lambda^{\alpha n}} \prod_{i=1}^{n} {x_i}^{2\alpha-1} exp\left(-\frac{\alpha}{\lambda}\sum_{i}^{n} {x_i}^{2}\right) \tag{5}$$

Theorem 1: For $q \in (-\infty, \infty)$, $q \neq 0$, U.M.V.U.E of λ^q is

$$\tilde{\lambda^q} = \begin{cases} \left\{\frac{\Gamma(n\alpha - q)}{\Gamma(n\alpha)}\right\} S^q & ; n\alpha > q \\ 0 & ; \text{Otherwise} \end{cases} \tag{6}$$

Proof: From the likelihood (5) and factorization theorem Rohtagi and Saleh (2012, pp.361) it can be easily obtained that $S = \sum_{i=1}^{n} {x_i}^2$ is a sufficient statistic for λ and the $P.d.f$ of S is

$$f_s(S|\lambda) = \frac{S^{n\alpha-1}}{\Gamma(n\alpha)\lambda^{n\alpha}} exp\left(-\frac{S}{\lambda}\right) \tag{7}$$

From (7), since the distribution of S belongs to the exponential family, it is also complete Rohtagi and Saleh (2012, pp.367).

Now, from (7), we have

$$E[S^{-q}] = \frac{1}{\Gamma(n\alpha)\lambda^{n\alpha}} \int_0^\infty S^{n\alpha-q-1} exp\left(- \frac{S}{\lambda}\right) dS$$
$$= \left\{ \frac{\Gamma(n\alpha - q)}{\Gamma(n\alpha)} \right\} \frac{1}{\lambda^q}$$

and the theorem holds on using Lehmann-Scheffe theorem Rohtagi (1976, pp.357). □

Theorem 2: The U.M.V.U.E of the reliability function is

$$\tilde{R}(t) = \begin{cases} 1 - I_{\frac{t^2}{S}}[\alpha, (n-1)\alpha] & ; t^2 < \frac{S}{\alpha} \\ 0 & ; \text{Otherwise} \end{cases} \tag{8}$$

where $I_x(p,q) = \frac{1}{\beta(p,q)} \int_0^x y^{p-1}(1-y)^{q-1}\, dy; 0 \le y \le 1, x < 1, p, q > 0$ is the incomplete beta function.

Proof: Let us define a random variable as

$$V = \begin{cases} 1, & X_1 > t \\ 0, & \text{Otherwise} \end{cases} \tag{9}$$

which is based on a single observation and is an unbiased estimator of $R(t)$. Using Rao-Blackwellization and (9), we have

$$\tilde{R}(t) = E(V|S)$$
$$= P\left(X_1 > t|S\right)$$
$$= P\left(v_1 > \frac{t^2}{S}\right), \; say; \tag{10}$$

where $v_1 = \frac{{X_1}^2}{S}$. From (7), we see that v_1 follows beta distribution of first kind with parameters $[\alpha, (n-1)\alpha]$. Applying Basu's theorem, from (10), we have

$$\tilde{R}(t) = 1 - P\left(v_1 \le \frac{t^2}{S}\right)$$
$$= 1 - \frac{\beta[\frac{t^2}{S}; \alpha, (n-1)\alpha]}{\beta[\alpha, (n-1)\alpha]} \tag{11}$$

and the theorem holds. □

Corollary 2.1: The Reliability estimate of the distribution for which $\alpha = 1$ is

$$\tilde{R}(t) = \begin{cases} \left(1 - \frac{t^2}{S}\right)^{n-1} & ; t^2 < S \\ 0 & ; \text{Otherwise} \end{cases} \tag{12}$$

Corollary 2.2: The U.M.V.U.E of sampled *P.d.f* (1) at a specified point x is :

$$\tilde{f}(x;\lambda)=\begin{cases}\left(\frac{\alpha}{S}\right)^{\alpha}\frac{x^{2\alpha-1}}{\beta[\alpha,(n-1)\alpha]}\left(1-\frac{\alpha x^2}{S}\right)^{(n-1)\alpha-1} & ;x^2<\frac{S}{\alpha_1}\\ 0 & ;\text{Otherwise}\end{cases}\tag{13}$$

Let us take two independent random variables X and Y with P.d.fs $f(x,\alpha_1,\lambda_1)$ and $f(y,\alpha_2,\lambda_2)$ respectively, where

$$f(x;\alpha_1,\lambda_1)=\frac{2}{\Gamma\alpha_1}\left(\frac{\alpha_1}{\lambda_1}\right)^{\alpha_1}x^{2\alpha_1-1}exp\left(-\frac{\alpha_1}{\lambda_1}x^2\right);\quad x>0,\alpha_1\geq 0.5,\lambda_1>0.\tag{14}$$

and

$$f(y;\alpha_2,\lambda_2)=\frac{2}{\Gamma\alpha_2}\left(\frac{\alpha_2}{\lambda_2}\right)^{\alpha_2}x^{2\alpha_2-1}exp\left(-\frac{\alpha_2}{\lambda_2}y^2\right);\quad y>0,\alpha_2\geq 0.5,\lambda_2>0.\tag{15}$$

Now draw a random sample $X_1,X_2,\ldots,X_n$ from $f(x;\alpha_1,\lambda_1)$ and random sample $Y_1,Y_2,\ldots,Y_m$ from $f(y;\alpha_2,\lambda_2)$. Denote $S=\sum_{i=1}^{n}{x_i}^2$ and $T=\sum_{i=1}^{m}{y_i}^2$.

Theorem 3: The U.M.V.U.E of P is

$$\tilde{P}=\begin{cases}\frac{1}{2\beta[\alpha_2,(m-1)\alpha_2]}\int_0^{\frac{\alpha_2 S}{\alpha_1 T}}\{1-I_{\frac{Tz}{\alpha_2 S}}[\alpha_1,(n-1)\alpha_1]\}z^{\alpha_2-1}(1-z)^{(m-1)\alpha_2-1}\,dz\\ \qquad ;\left(\frac{S}{\alpha_1}\right)^{\frac{1}{2}}\leq\left(\frac{T}{\alpha_2}\right)^{\frac{1}{2}}\\ \frac{1}{2\beta[\alpha_2,(m-1)\alpha_2]}\int_0^{1}\{1-I_{\frac{Tz}{\alpha_2 S}}[\alpha_1,(n-1)\alpha_1]\}z^{\alpha_2-1}(1-z)^{(m-1)\alpha_2-1}\,dz\\ \qquad ;\left(\frac{S}{\alpha_1}\right)^{\frac{1}{2}}>\left(\frac{T}{\alpha_2}\right)^{\frac{1}{2}}\end{cases}$$

Proof:

Proceeding as in case of proving Corollary 2, we can rewrite U.M.V.U.E of P in terms of $\tilde{R}(y,\lambda_1)$ as follows

$$\begin{aligned}\tilde{P}&=\int_{y=0}^{\infty}\int_{x=y}^{\infty}\tilde{f}(x;\lambda_1)\tilde{f}(y;\lambda_2)\,dx\,dy\\&=\int_{y=0}^{\infty}\tilde{R}(y;\lambda_1)\tilde{f}(y;\lambda_2)\,dy\end{aligned}$$

Now, using Theorem 2, we have

$$\begin{aligned}\tilde{P}=&\int_0^{min\left[\left(\frac{S}{\alpha_1}\right)^{\frac{1}{2}},\left(\frac{T}{\alpha_2}\right)^{\frac{1}{2}}\right]}\left[1-I_{\frac{y^2}{S}}(\alpha_1,(n-1)\alpha_1)\right]\\&\left(\frac{\alpha_2}{T}\right)^{\alpha_2}\frac{y^{2\alpha_2-1}}{\beta[\alpha_2,(m-1)\alpha_2]}\left(1-\frac{\alpha_2 y^2}{T}\right)^{(m-1)\alpha_2-1}dy\\=&\frac{\left(\frac{\alpha_2}{T}\right)^{\alpha_2}}{\beta[\alpha_2,(m-1)\alpha_2]}\int_0^{min\left[\left(\frac{S}{\alpha_1}\right)^{\frac{1}{2}},\left(\frac{T}{\alpha_2}\right)^{\frac{1}{2}}\right]}\\&\left[1-I_{\frac{y^2}{S}}(\alpha_1,(n-1)\alpha_1)\right]y^{2\alpha_2-1}\left(1-\frac{\alpha_2 y^2}{T}\right)^{(m-1)\alpha_2-1}dy\end{aligned}\tag{16}$$

Now from (16), when $\left(\frac{S}{\alpha_1}\right)^{\frac{1}{2}} \leq \left(\frac{T}{\alpha_2}\right)^{\frac{1}{2}}$

$$\tilde{P} = \frac{1}{2\beta[\alpha_2,(m-1)\alpha_2]} \int_0^{\frac{\alpha_2 S}{\alpha_1 T}} \{1 - I_{\frac{Tz}{\alpha_2 S}}[\alpha_1,(n-1)\alpha_1]\} z^{\alpha_2-1} (1-z)^{(m-1)\alpha_2-1}\, dz$$

and we have the first assertion.

Furthermore, for $\left(\frac{S}{\alpha_1}\right)^{\frac{1}{2}} > \left(\frac{T}{\alpha_2}\right)^{\frac{1}{2}}$,

$$\tilde{P} = \frac{1}{2\beta[\alpha_2,(m-1)\alpha_2]} \int_0^{1} \{1 - I_{\frac{Tz}{\alpha_2 S}}[\alpha_1,(n-1)\alpha_1]\} z^{\alpha_2-1} (1-z)^{(m-1)\alpha_2-1}$$

and this proves the second assertion. □

Corollary 3.1: U.M.V.U.E of P when $\alpha_1 = \alpha_2 = 1$ is given by

$$\tilde{P} = \begin{cases} \frac{1}{2\beta[1,m-1]} \int_0^{\frac{S}{T}} \{1 - I_{\frac{Tz}{S}}[1,(n-1)]\}(1-z)^{(m-2}\, dz; & S^{\frac{1}{2}} \leq T^{\frac{1}{2}} \\ \frac{1}{2\beta[1,m-1]} \int_0^{1} \{1 - I_{\frac{Tz}{S}}[1,(n-1)]\}(1-z)^{(m-2)}\, dz; & S^{\frac{1}{2}} > T^{\frac{1}{2}} \end{cases}$$

We provide M.L.E. of λ^q, $R(t)$ and P under the assumption that α is known in the following given theorems.
From (5), M.L.E of λ is

$$\hat{\lambda} = \frac{S}{n} \tag{17}$$

Theorem 4: The M.L.E. of λ^q is

$$\hat{\lambda}^q = \left(\frac{S}{n}\right)^q \tag{18}$$

Theorem 5: The M.L.E. of $R(t)$ is given by

$$\hat{R}(t) = 1 - \frac{1}{\Gamma\alpha}\Gamma\left(\frac{n\alpha t^2}{S}, \alpha\right) \tag{19}$$

We obtain M.L.E. of sampled $P.d.f$ with the help of Theorem 5 in the following corollary. This will be used to obtain M.L.E. of P.

Corollary 5.1: The M.L.E. of $f(x;\lambda)$ at a specified point x is

$$\hat{f}(x;\lambda) = \frac{2}{\Gamma\alpha}\left(\frac{n\alpha}{S}\right)^{\alpha} x^{2\alpha-1} exp\left(-\frac{n\alpha x^2}{S}\right) \tag{20}$$

Theorem 6: The M.L.E. of P is

$$\hat{P} = 1 - \frac{1}{\Gamma\alpha_1\Gamma\alpha_2} \int_{z=0}^{\infty} z^{\alpha_2-1} e^{-z}\, \Gamma\left(\frac{n\alpha_1 zT}{m\alpha_2 S}, \alpha_1\right) dz \tag{21}$$

Proof: We know that

$$\hat{P} = \int_{y=0}^{\infty} \int_{x=y}^{\infty} \hat{f}(x;\lambda_1)\hat{f}(y;\lambda_2)\,dx\,dy$$
$$= \int_{y=0}^{\infty} \hat{R}(y;\lambda_1)\hat{f}(y;\lambda_2)\,dy$$

Now using (20) and Theorem 5, we have

$$\hat{P} = \int_{y=0}^{\infty} \left[1 - \frac{1}{\Gamma\alpha_1}\Gamma\left(\frac{n\alpha_1 y^2}{S}, \alpha_1\right)\right] \frac{2}{\Gamma\alpha_2}\left(\frac{m\alpha_2}{T}\right)^{\alpha_2}$$
$$.y^{2\alpha_2-1} exp\left(-\frac{m\alpha_2 y^2}{T}\right) dy$$

Substituting $\frac{m\alpha_2 y^2}{T} = z$ and solving for the above integral, we get the desired result. □
Next, we provide moment estimators for the parameters. For this, below given theorem provides the r^{th} moment generating function of the distribution.

Theorem 7: For $r = 1, 2, 3 \ldots$, the moment generating function r^{th} is given by

$$u_r = E(X^r) = \frac{\Gamma(\alpha + \frac{r}{2})}{\Gamma(\alpha)} \left(\frac{\lambda}{\alpha}\right)^{\frac{r}{2}} \tag{22}$$

From (22), we have

$$u_1 = \frac{\Gamma(\alpha + \frac{1}{2})}{\Gamma(\alpha)} \left(\frac{\lambda}{\alpha}\right)^{\frac{1}{2}}$$

and

$$u_2 = \frac{\Gamma(\alpha + \frac{2}{2})}{\Gamma(\alpha)} \left(\frac{\lambda}{\alpha}\right)^{\frac{2}{2}} = \lambda$$

Equating the population moments with the sample moments, we have

$$\hat{\lambda}_m = \frac{S}{n} \tag{23}$$

Using (23), the moment estimator $\hat{\alpha}_m$ of α is obtained by the solution of

$$\bar{X} - \frac{\Gamma(\alpha + \frac{1}{2})}{\Gamma(\alpha)} \sqrt{\left(\frac{\hat{\lambda}_m}{\alpha}\right)} = 0 \tag{24}$$

uniroot function in R-software is used for finding the roots of the above equation.

3. Point estimation when shape parameter is unknown

Now we discuss the case when both the parameters are unknown. The log-likelihood function of the parameters α and λ given the sample observations $\underline{x}$ is:

$$l(\lambda|\underline{x}) = nlog(2\alpha^n) - nlog(\Gamma\alpha) - n\alpha log(\lambda) + \sum_{i=1}^{n} log({x_i}^{2\alpha-1}) - \frac{\alpha}{\lambda}\sum_{i}^{n} {x_i}^2$$

The M.L.E of α is given by the solution of the following equation

$$\frac{\partial l}{\partial \alpha} = \frac{n^2}{\alpha} - n\Psi_0(\alpha) - nlog(\lambda) + 2\sum_{i=1}^{n} log(x_i) - \frac{1}{\lambda}\sum_{i=0}^{n} {x_i}^2 = 0 \tag{25}$$

where Ψ_0 is a polygamma function of order zero and $\Psi_0(\alpha) = \frac{\partial log\Gamma(\alpha)}{\partial \alpha}$ is diagamma function.

$$\frac{\partial l}{\partial \lambda} = -\frac{n\alpha}{\lambda} + \frac{\alpha}{\lambda^2}\sum_{i=1}^{n} {x_i}^2 = 0 \tag{26}$$

$$\implies \hat{\lambda} = \frac{S}{n}$$

Since (25) do not have a closed form solution, therefore any iterative procedure have to be use to compute M.L.E.

Theorem 8: The M.L.E. of $R(t)$ is given by:

$$\hat{R}(t) = 1 - \frac{1}{\Gamma\hat{\alpha}}\Gamma\left(\frac{n\hat{\alpha}t^2}{S}, \hat{\alpha}\right) \tag{27}$$

Corollary 8.1: The M.L.E. of $f(x;\alpha,\lambda)$ at a specified point x is

$$\hat{f}(x;\alpha,\lambda) = \frac{2}{\Gamma\hat{\alpha}}\left(\frac{n\hat{\alpha}}{S}\right)^{\hat{\alpha}} x^{2\hat{\alpha}-1} exp\left(-\frac{n\hat{\alpha}x^2}{S}\right) \tag{28}$$

Theorem 9: The M.L.E. of P is

$$\hat{P} = 1 - \frac{1}{\Gamma\hat{\alpha_1}\Gamma\hat{\alpha_2}}\int_{z=0}^{\infty} z^{\hat{\alpha_2}-1}e^{-z}\,\Gamma\left(\frac{n\hat{\alpha_1}zT}{m\hat{\alpha_2}S}, \hat{\alpha_1}\right)dz \tag{29}$$

4. Asymptotic confidence intervals

The Confidence Intervals (C.I) can be obtained by using the variance-covariance matrix of the M.L.Es of the parameters. The asymptotic variance-covariance matrix of $\hat{\eta} = (\hat{\alpha}, \hat{\lambda})$ is the inverse of the following Fisher Information matrix

$$I(\eta) = -E\begin{bmatrix} \frac{\partial^2 l}{\partial^2 \alpha} & \frac{\partial^2 l}{\partial \alpha \partial \lambda} \\ \frac{\partial^2 l}{\partial \lambda \partial \alpha} & \frac{\partial^2 l}{\partial^2 \lambda} \end{bmatrix}$$

This is very cumbersome to obtain the exact distributions of the M.L.Es and the alternative is to use the observed Fisher information matrix which is

$$I(\hat{\eta}) = \begin{bmatrix} -\frac{\partial^2 l}{\partial^2 \alpha} & -\frac{\partial^2 l}{\partial \alpha \partial \lambda} \\ -\frac{\partial^2 l}{\partial \lambda \partial \alpha} & -\frac{\partial^2 l}{\partial^2 \lambda} \end{bmatrix}$$

Thus, we have observed variance-covariance matrix as

$$I^{-1}(\hat{\eta}) = \begin{bmatrix} \hat{V}ar(\hat{\alpha}) & \hat{C}ov(\hat{\alpha}, \hat{\lambda}) \\ \hat{C}ov(\hat{\lambda}, \hat{\alpha}) & \hat{V}ar(\hat{\lambda}) \end{bmatrix}$$

Assuming asymptotic normality of the M.L.Es, confidence intervals for α and λ are constructed. Let $\hat{\sigma}^2(\hat{\alpha})$ and $\hat{\sigma}^2(\hat{\lambda})$ be the estimated variances of α and λ. Then the two sided equal tail asymptotic $100(1-\delta)\%$ confidence intervals for the parameters α and λ are $\left(\hat{\alpha} \pm Z_{\frac{\delta}{2}}\hat{\sigma}(\hat{\alpha})\right)$ and $\left(\hat{\lambda} \pm Z_{\frac{\delta}{2}}\hat{\sigma}(\hat{\lambda})\right)$, respectively, where $Z_{\frac{\delta}{2}}$ is the $\left(\frac{\delta}{2}\right)^{th}$ percentile of the standard normal distribution. The coverage probabilities (CP) are given as,

$$CP_{\alpha} = P\left[\left|\frac{\hat{\alpha}-\alpha}{\hat{\sigma}(\hat{\alpha})}\right| \leq Z_{\frac{\delta}{2}}\right]$$

and

$$CP_{\lambda} = P\left[\left|\frac{\hat{\lambda}-\lambda}{\hat{\sigma}(\hat{\lambda})}\right| \leq Z_{\frac{\delta}{2}}\right]$$

The asymptotic C.I based on log(M.L.E) has better coverage probability as reported by Meeker and Escober (1998). An approximate $100(1-\delta)\%$ C.I for $log(\alpha)$ and $log(\lambda)$ are

$$\left\{log(\hat{\alpha}) \pm Z_{\frac{\delta}{2}}\hat{\sigma}[log(\hat{\alpha})]\right\} \text{ and } \left\{log(\hat{\lambda}) \pm Z_{\frac{\delta}{2}}\hat{\sigma}[log(\hat{\lambda})]\right\},$$

where $\hat{\sigma}^2[log(\hat{\alpha})]$ and $\hat{\sigma}^2[log(\hat{\lambda})]$ are the estimated variance of $log(\alpha)$ and $log(\lambda)$ respectively, and are approximated by

$$\hat{\sigma}^2[log(\hat{\alpha})] = \frac{\hat{\sigma}^2(\hat{\alpha})}{\hat{\alpha}^2} \text{ and } \hat{\sigma}^2[log(\hat{\lambda})] = \frac{\hat{\sigma}^2(\hat{\lambda})}{\hat{\lambda}^2}.$$

Hence, approximate $100(1-\delta)\%$ C.I for α and λ are

$$\left(\hat{\alpha}e^{\pm Z_{\frac{\delta}{2}}\frac{\hat{\sigma}(\hat{\alpha})}{\hat{\alpha}}}\right) \text{ and } \left(\hat{\lambda}e^{\pm Z_{\frac{\delta}{2}}\frac{\hat{\sigma}(\hat{\lambda})}{\hat{\lambda}}}\right).$$

5. Testing of statistical hypotheses

Under this section, we consider the following three cases of hypothesis testing.
1. $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$, when α is known.
2. $H_0 : \lambda \leq \lambda_0$ versus $H_1 : \lambda > \lambda_0$, when α is known.
3. $H_0 : P = P_0$ versus $H_1 : P \neq P_0$, when $\alpha_1 = \alpha_1$ is known.

Testing $H_0 : \lambda = \lambda_0$ against $H_1 : \lambda \neq \lambda_0$ is considered to be the most important. From (5), we can have the likelihood of observing λ given the sample observations $\underline{x}$ as

$$L(\lambda|\underline{x}) = \left(\frac{2\alpha^{\alpha}}{\Gamma\alpha}\right)^{n} \frac{1}{\lambda^{\alpha n}} \prod_{i=1}^{n} {x_i}^{2\alpha-1} exp\left(-\frac{\alpha}{\lambda}\sum_{i}^{n} {x_i}^{2}\right) \tag{30}$$

Under H_0, we have

$$\sup_{\Theta_0} L(\lambda; \underline{x}, \alpha) = \left(\frac{2\alpha^{\alpha}}{\Gamma\alpha}\right)^{n} \frac{1}{\lambda_0^{\alpha n}} \prod_{i=1}^{n} {x_i}^{2\alpha-1} exp\left(-\frac{\alpha}{\lambda_0}\sum_{i}^{n} {x_i}^{2}\right); \quad \Theta_0 = \{\lambda : \lambda = \lambda_0\} \tag{31}$$

and

$$\sup_{\Theta} L(\lambda; \underline{x}, \alpha) = \left(\frac{2\alpha^{\alpha}}{\Gamma\alpha}\right)^{n} \left(\frac{n}{S}\right)^{\alpha n} \prod_{i=1}^{n} {x_i}^{2\alpha-1} exp\left(-\frac{n\alpha}{S}\sum_{i}^{n} {x_i}^{2}\right); \quad \Theta_0 = \{\lambda : \lambda > 0\} \tag{32}$$

Therefore, the Likelihood Ratio (L.R) is given by

$$\begin{aligned}\phi(\underline{x}) &= \frac{\sup_{\Theta_0} L(\lambda; \underline{x}, \alpha)}{\sup_{\Theta} L(\lambda; \underline{x}, \alpha)} \\ &= \left(\frac{S}{n\lambda_0}\right)^{n\alpha} exp\left[-\alpha\left(\frac{S}{\lambda_0} - n\right)\right]\end{aligned} \tag{33}$$

From Right Hand Side (R.H.S) of above equation, it is clear that first term is an increasing whereas the second term is monotonically decreasing function in S. As $2\frac{S}{\lambda_0} \sim \chi^2_{(2n)}$, where $\chi^2_{(2n)}$ is the Chi-Square statistics with $2n$ degrees of freedom, the critical region is given by

$$\{0 < S < \gamma_0\} \cup \{\gamma_0^i < S < \infty\},$$

where the constants γ_0 and γ_0^i are obtained such that

$$P\left[\chi^2_{(2n)} < 2\frac{\gamma_0}{\lambda_0} \quad or \quad 2\frac{\gamma_0^i}{\lambda_0} < \chi^2_{(2n)}\right] = \varepsilon$$

Thus,

$$\gamma_0 = \frac{\lambda_0 \chi^2_{(2n)}\left(1 - \frac{\varepsilon}{2}\right)}{2}$$

and

$$\gamma_0^i = \frac{\lambda_0 \chi^2_{(2n)}\left(\frac{\varepsilon}{2}\right)}{2}$$

where ε is the probability of type I error.

The second important hypothesis is $H_0 : \lambda \leq \lambda_0$ versus $H_1 : \lambda > \lambda_0$. For $\lambda_1 > \lambda_2$, we have from (5)

$$\frac{L(\lambda_1|\underline{x})}{L(\lambda_2|\underline{x})} = \left(\frac{\lambda_2}{\lambda_1}\right)^{n\alpha} exp\left[-S\left(\frac{\alpha}{\lambda_1} - \frac{\alpha}{\lambda_2}\right)\right] \tag{34}$$

From (34), we can see $L(\lambda, \underline{x})$ has Monotone Likelihood Ratio (M.L.R) in S. Thus, the Uniformly Most Powerful Critical Region (U.M.P.C.R) for testing $H_0 : \lambda \leq \lambda_0$ against $H_1 : \lambda > \lambda_0$ is given as Lehmann (1959, pp.88)

$$\phi(\underline{x}) = \begin{cases} 1, & S \leq \gamma_0^{ii} \\ 0, & \text{Otherwise.} \end{cases}$$

where, γ_0^{ii} is obtained such that $P\left[\chi^2_{(2n)} < 2\frac{\gamma_0^{ii}}{\lambda_0}\right] = \varepsilon$. Therefore,

$$\gamma_0^{ii} = \frac{\lambda_0 \chi^2_{(2n)}\left(1 - \varepsilon\right)}{2}$$

It can be seen that for two independent random variables X and Y with $\alpha_1 = \alpha_2 = 1$,

$$P = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

Let us test $H_0 : P = P_0$ against $H_1 : P \neq P_0$. Thus, H_0 is equivalent to $\lambda_1 = k\lambda_2$, where $k = \frac{P_0}{1-P_0}$. Therefore, $H_0 : \lambda_1 = k\lambda_2$ and $H_1 : \lambda_1 \neq k\lambda_2$.
Under H_0, we can have

$$\hat{\lambda}_1 = \frac{S + Tk}{(n+m)}$$

and

$$\hat{\lambda}_2 = \frac{S + Tk}{k(n+m)}$$

Thus, for C (a generic constant), we have the likelihood of observing λ_1 and λ_2 as

$$L(\lambda_1, \lambda_2 | \underline{x}, \underline{y}) = \frac{C}{{\lambda_1}^n {\lambda_2}^m} exp\left[-\left(\frac{S}{\lambda_1} + \frac{T}{\lambda_2}\right)\right] \tag{35}$$

Thus,

$$\sup_{\Theta_0} L(\lambda_1, \lambda_2 | \underline{x}, \underline{y}) = C\left[\frac{k(n+m)}{S+Tk}\right]^{n+m} exp\left[-(n+m)\right]; \quad \Theta_0 = \{\lambda_1, \lambda_2 : \lambda_1 = k\lambda_2\} \tag{36}$$

and

$$\sup_{\Theta} L(\lambda_1, \lambda_2 | \underline{x}, \underline{y}) = C\left(\frac{n}{S}\right)^n \left(\frac{m}{T}\right)^m exp\left[-(n+m)\right]; \quad \Theta = \{\lambda_1, \lambda_2 : \lambda_1 > 0, \lambda_2 > 0\} \tag{37}$$

From (36) and (37), the Likelihood ratio criterion is

$$\phi(\underline{x}, \underline{y}) = \frac{C\left(\frac{S}{T}\right)^n}{\left(1 + \frac{S}{Tk}\right)^{n+m}} \tag{38}$$

Let us denote the F- statistic with (a, b) degrees of freedom by $F_{a,b}(.)$. As

$$\frac{S}{T} \sim \frac{n\lambda_1}{m\lambda_2} F_{(2n,2m)},$$

the critical region is

$$\left\{\frac{S}{T} < \gamma_2 \, or \, \frac{S}{T} > {\gamma_2}^i\right\},$$

where γ_2 and ${\gamma_2}^i$ are obtained such that

$$P\left\{\frac{nk\gamma_2}{m} < F_{2n,2m} \cup \frac{nk{\gamma_2}^i}{m} > F_{2n,2m}\right\} = \varepsilon$$

Thus, we have $\gamma_2 = \frac{nk}{m} F_{2n,2m}\left(1 - \frac{\varepsilon}{2}\right)$ and ${\gamma_2}^i = \frac{nk}{m} F_{2n,2m}\left(\frac{\varepsilon}{2}\right)$.

6. Simulation study

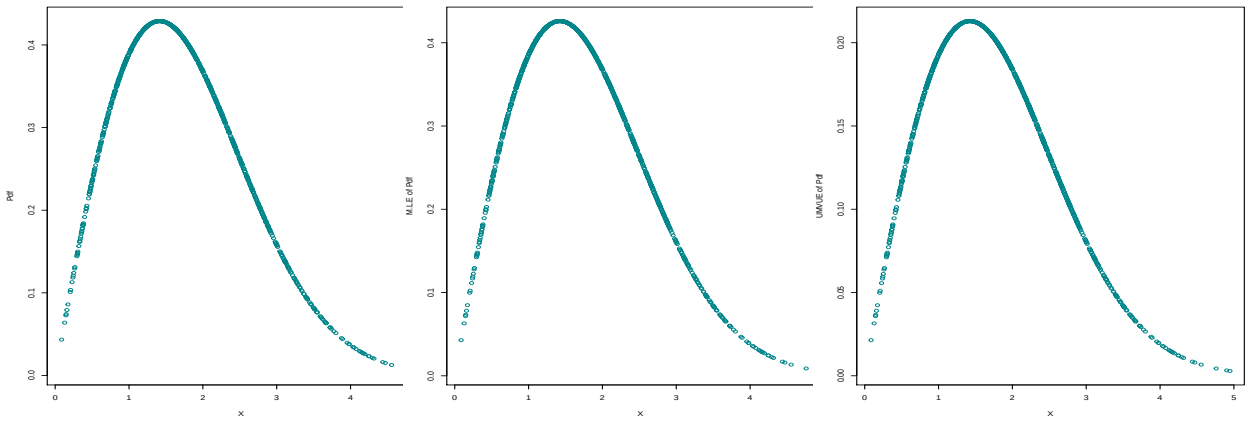
For validating the results obtained theoretically in Section 2 and Section 4, we, firstly present results which are based on Monte Carlo simulation technique. We have computed Mean Square Error (M.S.E) for comparison purpose. All the analyses have been done using R 3.4.3 software R Core Team (2013).

6.1. When shape parameter is known

For acquiring the performance of $\tilde{\lambda}^q$ and $\hat{\lambda}^q$, we have generated 1000 random samples from (1) of different sizes $n = (20, 30, 40, 60)$ with $\alpha = (0.8, 0.9, 1.0)$. We have computed average $\tilde{\lambda}^q$, $\hat{\lambda}^q$, corresponding average biases and M.S.E, and approximate $(1-\delta)100\%$, where $\delta = 0.05$, confidence intervals for λ^q. As $q \in (-\infty, \infty)$, $q \neq 0$, we choose a negative and a positive power of q to have better look into the performance of the estimators. For $q = -1$ and $q = 1$, results are given in Table **??**. The $1^{st}, 2^{nd}, 3^{rd}$ row represents average estimates, average bias, M.S.E and 4^{th} row represents the confidence interval. From Table **??**, we can infer that for negative values of q, U.M.V.U.E is performing better than the M.L.E but for positive value of q M.L.E is performing better than U.M.V.U.E. It can be seen that as the value of sample size is increasing M.S.E is decreasing for both the estimators. The length of the confidence interval is shorter for both estimators in all cases which means it is more informative. So, U.M.V.U.E should be preferred if we want to estimate the negative power of λ and for positive power, we should opt for M.L.E. It is interesting to note here that for $\alpha = 1$ and $q = -1$ both estimators are yielding the same results for all values of sample sizes.

Now, to acquire and compare the performance of the two estimators of $R(t)$, 1000 random samples are generated from (1) of different sizes $n = (10, 20, 30, 40, 60)$ with $\alpha = 3$ and $\lambda = 0.5$. Taking values of $t = (0.10, 0.15, 0.20, 0.25, 0.35)$, $\tilde{R}(t)$ and $\hat{R}(t)$, corresponding biases, M.S.E and approximate $(1 - \delta)100\%$ C.I have been calculated. The obtained results are presented in Table 2 where $I^{st}, 2^{nd}, 3^{rd}$ row represents average estimates, average bias, M.S.E and 4^{th} row represents the C.I.
Looking at M.S.E values in Table 2, we can say that performance of M.L.E of $R(t)$ is better than that of U.M.V.U.E of $R(t)$. Performance of estimators is decreasing with the increase in time t as the M.S.E values are increasing. Estimators tends to perform better in case of large sample sizes. Table 3 presents Moment estimators $\hat{\alpha}_m$ and $\hat{\lambda}_m$ of the parameters α and λ are given for different values of n=(500,1000,1500) and different set of the parameters $(\alpha, \lambda) = (0.6, 0.8), (1.5, 0.8)$ and $(1.5, 1.0)$. The moment estimator and M.L.E of λ are equal and both the estimators are the functions of the sufficient statistics. So, both M.L.E and Moment estimators are equally efficient and works good.

In order to investigate how well estimators of P performs, 1000 random samples are generated from (14) and (15) of sizes $(n, m) = (5, 10), (10, 5)$ and $(10, 10)$ with $\alpha_1 = 0.5$ and $\alpha_2 = 10$, and $(\lambda_1, \lambda_2) = (3, 5), (3, 6), (4, 5), (4, 6)$. The obtained results are presented in Table 4 where $I^{st}, 2^{nd}, 3^{rd}$ row represents average estimates, average bias, M.S.E and 4^{th} row represents the confidence interval. Data in table 4 reveals that M.L.E of P gives better estimates than U.M.V.U.E of P for all combinations of (λ_1, λ_2) and (n, m).



(a) P.d.f **(b) M.L.E** **(c) U.M.V.U.E**

Figure 1: Estimates of Sampled Probability Density Function

The estimates of P.d.f obtained in Section 2 are plotted in Figure 1. From the figure, we can see that the estimates of P.d.f fits well to the actual model.

6.2. When both scale and shape parameters are unknown

For obtaining the estimate of $R(t)$ when both scale and shape parameters are unknown, we have generated first random sample of size $n = 15$ from (1) with $\alpha_1 = 5$ and $\lambda_1 = 4$. Let it be X- population or random strength X given as

X- Population: 2.049508, 1.697911, 2.057258, 2.093914, 1.830376, 2.299230, 1.030369, 1.352851, 1.910835, 2.518206, 1.717836, 2.318662, 1.932082, 1.436255, 2.776821.

The M.L.Es of α and λ comes out to be $\hat{\alpha_1} = 5.512470$ and $\hat{\lambda}_1 = 3.937005$. For $t = 0.2$, actual $R(t) = 0.9999869$ and $\hat{R}(t) = 0.9999869$.

Now, for estimation of P, we have generated another random sample say Y population or random stress Y from (1) of size $m = 10$ with $\alpha_2 = 4$ and $\lambda_2 = 3$. The sample is

Y Population: 1.263758, 1.816875, 1.346044, 2.083317, 2.489531, 1.119266, 1.714329, 1.912815, 1.371682, 2.496376. The M.L.Es of α and λ comes out to be $\hat{\alpha_2} = 4.118107$ and $\hat{\lambda}_2 = 3.321188$. For $t = 0.2$, actual $R(t) = 0.9964545$ and $\hat{R}(t) = 0.9994916$. The actual $P = 0.4363503$ and the M.L.E of P comes out to be $\hat{P} = 0.4696917$. All the estimates can be seen validating the theoretical results obtained.

6.3. Hypothesis testing

This section comprises of checking the validity of the hypothesis testing procedures developed in section 5. Firstly, we test the hypothesis $H_0 : \lambda = \lambda_0 = 4$ against $H_1 : \lambda = \lambda_0 \neq 4$. For this we have generated a random sample of size 50 from (1) with $(\lambda = 4, \alpha = 5)$, given by

Sample 1 : 1.8715040, 2.4957160, 1.3041026, 1.0625339, 1.9552509, 1.8412767, 1.4635787, 1.6677863, 2.1402472, 1.6651901, 1.4523474, 2.4088220, 1.6413565, 2.2162550, 1.6001383, 2.0236934, 2.0894237, 1.7744711, 2.0995504, 2.9366243, 2.3269415, 1.6324515, 1.5328350, 0.9560068, 2.4759661, 2.0723630, 2.2769360, 1.3536968, 2.0298724, 2.4644942, 2.0113171, 1.6845441, 1.8919575, 2.5608773, 1.9408668, 1.8201857, 2.3742209, 2.1374813, 2.5166206,

2.4151387, 1.4238698, 1.4754821, 1.6192035, 2.1958351, 1.7966403, 2.2790533, 2.0138617, 1.4063136, 1.8715380, 1.6387806.
Now, using chi-square table at $\varepsilon = 5\%$ Level of Significance (LOS), we obtained $\gamma_0 = 259.12$ and $\gamma_0^i = 148.44$. From the sample we have $S = 192.5144$. Here, it can be seen that the value of S is not lying in the critical region. So, we do not have enough evidence to reject the null hypothesis at 5% LOS.

Consider the above sample 1 again for testing $H_0 : \lambda = \lambda_0 \leq 4$ against $H_1 : \lambda = \lambda_0 > 4$ at 5% LOS, we obtained $\gamma_0^{ii} = 148.68$. As $S = 192.5144$ is not lying in the critical region so we do not have sufficient evidence in support of alternate hypothesis. Thus, we do not reject the null hypothesis.

Now, to test $H_0 : P = P_0 = 0.5$ against $H_1 : P = P_0 \neq 0.5$, we have generated two random samples X_i and Y_i of sizes $n = 12$ and $m = 10$ from the distribution with $\lambda_1 = \lambda_2 = 4$ and $\alpha_1 = \alpha_2 = 1$ given by X= 0.6960666, 1.9268595, 2.1383461, 1.4266733, 1.8846088, 2.1335468, 2.0400911, 0.2361899, 3.8670944, 1.0444884, 1.5116124, 1.0438313 and Y=0.843277, 2.642273, 2.342635, 1.710249, 2.558351, 2.145001, 1.933631, 2.861752, 3.867026, 1.158481. From the two samples, we get $\frac{S}{T} = 0.766934$. Using F-table at 5% LOS, we computed $\gamma_2 = 2.7924$ and $\gamma_2^i = 0.498$. Thus, we do not reject the null hypothesis on the basis above information.

7. Real data analysis

Now we present two real data set to understand and illustrate the procedures discussed in the previous sections broadly.

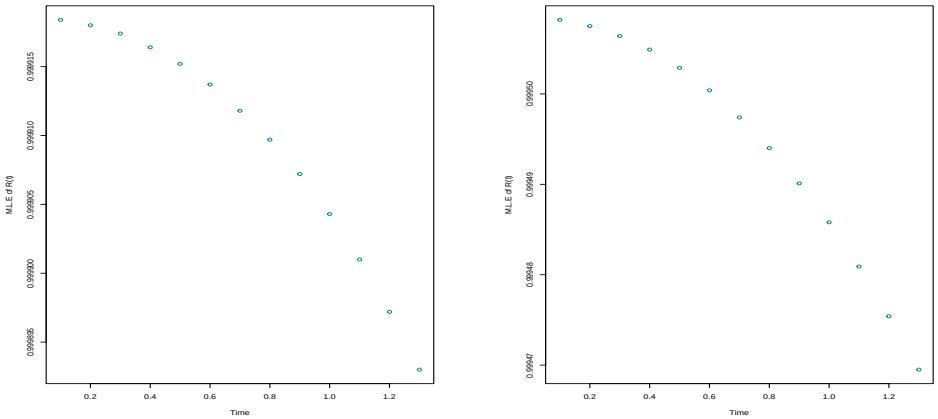
7.1. First data set

The data set has been taken from Lawless (2003, pp.267). This was originally reported by Schafft *et al.* (1987). This data represents the hours to failure of 59 conductors of 400-micrometer length. The specimens are put on a test with same temperature and current density and they all ran to failure at a certain high temperature with current density.
X-Population: 6.545, 9.289, 7.543, 6.956, 6.492, 5.459, 8.120, 4.706, 8.687, 2.997, 8.591, 6.129, 11.038, 5.381, 6.958, 4.288, 6.522, 4.137, 7.459, 7.495, 6.573, 6.538, 5.589, 6.087, 5.807, 6.725, 8.532, 9.663, 6.369, 7.024, 8.336, 9.218, 7.945, 6.869, 6.352, 4.700, 6.948, 9.254, 5.009, 7.489, 7.398, 6.033, 10.092, 7.496,4.531, 7.974, 8.799, 7.683, 7.224, 7.365, 6.923, 5.640, 5.434, 7.937, 6.515, 6.476, 6.071, 10.491, 5.923.
Kumar *et al.* (2017) used this data set and found that Nakagami distribution fits well to the data with M.L.Es as $\hat{\alpha} = 4.8336$ and $\hat{\lambda} = 51.2823$. For $t = (0.1, 0.2, 0.3, 0.4, 0.5)$ we have computed $R(t) = (0.9985447, 0.9985376, 0.9985258, 0.998509, 0.9984872)$ and their M.L.Es are $\hat{R}(t) = (0.9999184, 0.999918, 0.9999174, 0.9999164, 0.9999152)$.

7.2. Second data set

The second data set given below is taken from Murthy *et al.* (2004, pp.180)(2004, pp.180). This data represents 50 items that are put on use at t=0 and failure times are in recorded (in weeks). The data set is
Y-Population: 0.013, 0.065, 0.111, 0.111, 0.163, 0.309, 0.426, 0.535, 0.684, 0.747, 0.997,

1.284, 1.304, 1.647, 1.829, 2.336, 2.838, 3.269, 3.977, 3.981, 4.520, 4.789, 4.849, 5.202, 5.291, 5.349, 5.911, 6.018, 6.427, 6.456, 6.572, 7.023, 7.291, 7.087, 7.787, 8.596, 9.388, 10.261, 10.713, 11.658, 13.006, 13.388, 13.842, 17.152, 17.283, 19.418, 23.471, 24.777, 32.795, 48.105. Mudasir and Ahmed (2017) used this data set for analysis and comparison purpose in case of weighted Nakagami distribution. The M.L.Es of α and λ came out to be $\hat{\alpha} = 4.1$ and $\hat{\lambda} = 144.2292$. For $t = (0.1, 0.2, 0.3, 0.4, 0.5)$, $R(t) = (0.9966496, 0.9966451, 0.9966375, 0.9966269, 0.9966132)$ and their M.L.Es are $\hat{R}(t) = (0.9995082, 0.9995075, 0.9995064, 0.9995049, 0.9995029)$. The MLE estimate of $R(t)$ for both data sets is plotted in Figure 2. From the figure, it can



(a) Population X **(b) Population Y**

Figure 2: M.L.Es of R(t)

be seen that in both cases the survival is very high at initial time but as the time increases survival probability goes on decreasing.
To evaluate M.L.E of P, first data set is taken as X population and second set as Y population. Actual P came out to be $P = 0.7377018$ and its M.L.E is $\hat{P} = 0.703271$.

8. Conclusion

This paper presents estimation and testing procedures for the reliability functions of the Nakagami distribution. A new, simpler technique for obtaining Uniformly Minimum Variance Unbiased Estimators (UMVUEs) and Maximum Likelihood Estimators (MLEs) of $R(t)$ and P is introduced, requiring no explicit forms of the parametric functions. In addition to these estimators, moment estimators for the parameters are derived. The efficiency of MLEs and moment estimators is compared through simulations, showing similar performance as both are functions of the sufficient statistic. Hypothesis testing is also performed, with real data analysis on strength (X) and stress (Y) datasets.

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 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

Abdi, A. and Kaveh, M. (2000). Performance comparison of three different estimators for the nakagami-m parameter using monte carlo simulation. *IEEE Communication Letters*, **4**, 119–121.

Awad, A. and Gharraf, M. (1986). Estimation of p(y<x) in the Burr case: A comparative study. *Communication in Statistics- Simulation and Computation*, **15**, 389–403.

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(x>y) in the exponential case. *IEEE Transactions on Reliability*, **R-26**, 389–392.

Chaturvedi, A. and Pathak, A. (2012). Estimation of the reliability functions for exponentiated Weibull distribution. *Journal of Statistics and Applications*, **7**, 1–8.

Chaturvedi, A. and Rani, U. (1997). Estimation procedures for a family of density functions representing various life-testing models. *Metrika*, **46**, 213–219.

Chaturvedi, A. and Rani, U. (1998). The sampling distribution of an estimate arising in life testing. *Journal of Statistical Research*, **32**, 113–120.

Chaturvedi, A. and Surinder, K. (1999). Further remarks on estimating the reliability function of exponential distribution under type i and type ii censorings. *Brazilian Journal of Probability and Statistics*, **13**, 29–39.

Chaturvedi, A. and Tomer, S. (2003). UMVU estimation of the reliability function of the generalized life distributions. *Statistical Papers*, **44**, 301–313.

Chaturvedi, A. and Vyas, S. (2018). Estimation and testing procedures for the reliability function of three parameter Burr distribution under censoring. *Statistica*, **77**, 207–235.

Cheng, J. and Beaulieu, N. (2001). Maximum-likelihood based estimation of the nakagami m parameter. *Technometrics*, **5**, 101–103.

Johnson, N. (1975). Letter to the editor. *Technometrics*, **17**, 393.

Kumar, K., Garg, R., and Krishna, H. (2017). Nakagami distribution as a reliability model under progressive censoring. *International Journal of System Assurance Engineering and Management*, **8**, 109–122.

Lawless, J. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley, New York.

Lehmann, E. (1959). *Testing Statistical Hypotheses*. John Wiley and Sons, New York.

Meeker, W. and Escober, L. (1998). *Statistical Methods for Reliability Data*. John Wiley and Sons, New York.

Mudasir, S. and Ahmed, S. (2017). Weighted Nakagami distribution with applications using r- software. In *International Conference on Recent Innovations in Science, Agriculture, Engineering and Management*. University College of Computer Applications, Guru Kashi University, Bathinda, India.

Murthy, D., Xie, M., and Jiang, R. (2004). *Weibull Models*. John Wiley and Sons, New York.

Nakagami, M. (1960). The m-distribution—a general formula of intensity distribution of rapid fading. In *Statistical Methods in Radio Wave Propagation*, pages 3–36. Elsevier.

Pugh, E. (1963). The best estimate of reliability in the exponential case. *Operations Research*, **11**, 57–61.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rohtagi, V. (1976). *An Introduction to Probability Theory and Mathematical Statistics*. John Wiley and Sons, New York.

Rohtagi, V. and Saleh, A. (2012). *An Introduction to Probability and Statistics*. Wiley, New York.

Sathe, Y. and Shah, S. (1981). On estimating p(x<y) for the exponential distribution. *Communication in Statististics-Theory and Methods*, **A10**, 39–47.

Schafft, H., Staton, T., Mandel, J., and Shott, J. (1987). Reproducibility of electro-migration measurements. *IEEE Transactions on Electron Devices*, **34**, 673—-681.

Schwartz, J., Godwin, R., and Giles, D. (2013). Improved maximum likelihood estimation of the shape parameter in the Nakagami distribution. *Journal of Statistical Computation and Simulation*, **83**, 434–445.

Tong, H. (1974). A note on the estimation of p(y<x) in the exponential case. *Technometrics*, **16**, 625.

Tong, H. (1975). Letter to the editor. *Technometrics*, **17**, 393.

Tyagi, R. and Bhattacharya, S. (1989). A note on the MVU estimation of reliability for the maxwell failure distribution. *Estadistica*, **41**, 73–79.

ANNEXURE

Table 1: U.M.V.U.Es and M.L.Es of λ^q for Different Values of α

$n \rightarrow$		20		30		40		60	
$\alpha \downarrow$	$q \downarrow$	$\tilde{\lambda}^q$	$\hat{\lambda}^q$	$\tilde{\lambda}^q$	$\hat{\lambda}^q$	$\tilde{\lambda}^q$	$\hat{\lambda}^q$	$\tilde{\lambda}^q$	$\hat{\lambda}^q$
0.8	-1	0.5671 -0.0996 **0.0346** (0.557,0.577)	0.7089 0.0422 **0.0403** (0.697,0.721)	0.5568 -0.1098 **0.0215** (0.55,0.564)	0.696 0.0294 **0.0221** (0.687,0.705)	0.5486 - 0.1181 **0.0168** (0.542,0.555)	0.6857 0.019 **0.0221** (0.678,0.694)	0.5436 -0.123 **0.0218** (0.539,0.549)	0.6796 0.0129 **0.0106** (0.673,0.686)
	1	1.9853 0.4853 **0.4617** (1.956,2.015)	1.4889 -0.0111 **0.1274** (1.467,1.511)	1.9615 0.4615 **0.3693** (1.937,1.986)	1.5038 0.0038 **0.0919** (1.485,1.523)	1.9284 0.4284 **0.2966** (1.908,1.949)	1.4945 -0.0055 **0.0679** (1.478,1.511)	1.9187 0.4187 **0.259** (1.901,1.937)	1.503 0.003 **0.0513** (1.489,1.517)
0.9	-1	0.6256 -0.041 **0.0277** (0.616,0.636)	0.6951 0.0285 **0.0329** (0.684,0.706)	0.6266 -0.0401 **0.0184** (0.619,0.635)	0.6962 0.0295 **0.0216** (0.687,0.705)	0.6113 -0.0554 **0.0142** (0.605,0.618)	0.6792 0.0126 **0.0143** (0.672,0.687)	0.6143 -0.0523 **0.0098** (0.609,0.62)	0.6826 0.0159 **0.0089** (0.677,0.688)
	1	1.7631 0.2631 **0.2198** (1.739,1.787)	1.4986 -0.0014 **0.1088** (1.478,1.519)	1.712 0.212 **0.1491** (1.692,1.732)	1.4838 -0.0162 **0.0785** (1.466,1.501)	1.6971 0.1971 **0.114** (1.68,1.714)	1.485 -0.015 **0.0578** (1.47,1.5)	1.6868 0.1868 **0.0888** (1.672,1.701)	1.49 -0.01 **0.0422** (1.477,1.503)
1.0	-1	0.7082 0.0415 **0.0307** (0.698,0.719)	0.7082 0.0415 **0.0307** (0.698,0.719)	0.6926 0.0259 **0.0169** (0.685,0.7)	0.6926 0.0259 **0.0169** (0.685,0.7)	0.6898 0.0231 **0.0131** (0.683,0.697)	0.6898 0.0231 **0.0131** (0.683,0.697)	0.6804 0.0137 **0.0086** (0.675,0.686)	0.6804 0.0137 **0.0086** (0.675,0.686)
	1	1.5692 0.0692 **0.1272** (1.548,1.591)	1.4908 -0.0092 **0.1106** (1.47,1.511)	1.5413 0.0413 **0.0804** (1.524,1.559)	1.4899 -0.0101 **0.0736** (1.473,1.507)	1.5452 0.0452 **0.0623** (1.53,1.56)	1.5066 0.0066 **0.0574** (1.492,1.521)	1.5269 0.0269 **0.04** (1.515,1.539)	1.5015 0.0015 **0.0379** (1.489,1.514)

Table 2: U.M.V.U.Es and M.L.Es of $R(t)$

$n \rightarrow$		20		40		60	
$t \downarrow$	$R(t) \downarrow$	$\tilde{R}(t)$	$\hat{R}(t)$	$\tilde{R}(t)$	$\hat{R}(t)$	$\tilde{R}(t)$	$\hat{R}(t)$
0.1	0.99294	0.999966 0.007028 **4.9e − 05** (0.999965,0.999966)	0.992928 -1e-05 **5.7e − 09** (0.992923,0.992932)	0.999966 0.007028 **4.9e − 05** (0.999965,0.999967)	0.992935 -3e-06 **2.5e − 09** (0.992932,0.992938)	0.999966 0.007028 **4.9e − 05** (0.999965,0.999966)	0.992935 -3e-06 **1.7e − 09** (0.992932,0.992937)
0.15	0.9922	0.999621 0.007419 **5.5e − 05** (0.999611,0.99963)	0.992166 -3.6e-05 **4.02e − 08** (0.992154,0.992178)	0.999627 0.007426 **5.5e − 05** (0.999621,0.999633)	0.992185 -1.6e-05 **1.67e − 08** (0.992178,0.992193)	0.999627 0.007425 **5.5e − 05** (0.999621,0.999632)	0.992189 -1.2e-056 **1.17e − 08** (0.992183,0.992196)
0.2	0.99104	0.998047 0.007011 **5e − 05** (0.998001,0.998093)	0.990963 -7.3e-05 **1.527e − 07** (0.99094,0.990987)	0.998042 0.007006 **4.9e − 05** (0.998009,0.998075)	0.990992 -4.4e-05 **7.65e − 08** (0.990975,0.991009)	0.998074 0.007038 **5e − 05** (0.998048,0.998101)	0.991019 -1.7e-05 **5.09e − 08** (0.991005,0.991033)
0.25	0.98927	0.993315 0.004043 **2.3e − 05** (0.993161,0.993469)	0.989146 -0.000126 **5.689e − 07** (0.9891,0.989192)	0.993407 0.004135 **2e − 05** (0.993305,0.99351)	0.989232 -4e-05 **2.493e − 07** (0.989201,0.989263)	0.993363 0.004091 **1.9e − 05** (0.993276,0.993449)	0.989239 -3.3e-05 **1.709e − 07** (0.989213,0.989264)
0.35	0.98261	0.960736 -0.021879 **0.000664** (0.959891,0.96158)	0.982016 -0.000599 **7.1786e − 06** (0.981854,0.982178)	0.961043 -0.021572 **0.000547** (0.960482,0.961604)	0.982313 -0.000301 **2.9177e − 06** (0.982209,0.982418)	0.961205 -0.02141 **0.000512** (0.960752,0.961657)	0.98242 -0.000195 **1.8391e − 06** (0.982336,0.982503)

Table 3: Moment estimators of α and λ

$n \rightarrow$		500		1000		1500	
$\alpha \downarrow$	$\lambda \downarrow$	$\hat{\alpha}_m$	$\hat{\lambda}_m$	$\hat{\alpha}_m$	$\hat{\lambda}_m$	$\hat{\alpha}_m$	$\hat{\lambda}_m$
0.6	0.8	0.5576	0.8025	00.5988	0.8130	0.6001	0.8164
1.5	0.8	1.4235	0.7710	1.4477	0.7796	01.4935	0.8195
1.5	1.0	1.4925	1.0162	1.4928	1.0079	1.50041	1.0017

Table 4: U.M.V.U.Es and M.L.Es of P

$\lambda_1 \rightarrow$	3		3		4		4	
$\lambda_2 \rightarrow$	5		6		5		6	
$P \rightarrow$	0.625		0.6666667		0.5555556		0.6	
$(n,m) \downarrow$	$\tilde{P}$	$\hat{P}$	$\tilde{P}$	$\hat{P}$	$\tilde{P}$	$\hat{P}$	$\tilde{P}$	$\hat{P}$
(5,10)	0.415 -0.21 **0.0447** (0.413,0.416)	0.6487 0.0237 **8e − 04** (0.6478,0.6496)	0.4075 -0.2592 **0.0679** (0.406,0.409)	0.6443 -0.0224 **7e − 04** (0.6433,0.6453)	0.4348 -0.1207 **0.0149** (0.434,0.436)	0.6605 0.105 **0.0111** (0.6599,0.6612)	0.4283 -0.1717 **0.0299** (0.427,0.43)	0.6567 0.0567 **0.0034** (0.6559,0.6574)
(10,5)	0.4104 -0.2146 **0.0463** (0.409,0.411)	0.6505 0.0255 **7e − 04** (0.6499,0.6511)	0.403 -0.2636 **0.0698** (0.402,0.404)	0.6465 -0.0202 **5e − 04** (0.6459,0.6471)	0.4324 -0.1231 **0.0153** (0.432,0.433)	0.6626 0.1071 **0.0115** (0.6622,0.663)	0.4269 -0.1731 **0.0301** (0.426,0.428)	0.6596 0.0596 **0.0036** (0.6592,0.66)
(10,10)	0.4111 -0.2139 **0.046** (0.41,0.412)	0.6509 0.0259 **7e − 04** (0.6504,0.6514)	0.4026 -0.0204 **5e − 04** (0.402,0.404)	0.6462 0.0259 **7e − 04** (0.6457,0.6468)	0.4325 -0.1231 **0.0153** (0.432,0.433)	0.6626 0.1071 **0.0115** (0.6623,0.663)	0.4256 -0.1744 **0.0306** (0.425,0.426)	0.6589 0.0588 **0.0035** (0.6584,0.6593)

Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025; pp 121-134

Predicting Stock Market Crash with Bayesian Generalised Pareto Regression

Sourish Das
Chennai Mathematical Institute

Received: 21 June 2025; Revised: 14 August 2025; Accepted: 16 August 2025

Abstract

This paper develops a Bayesian Generalised Pareto Regression (GPR) model to forecast extreme losses in Indian equity markets, with a focus on the Nifty 50 index. Extreme negative returns, though rare, can cause significant financial disruption, and accurate modelling of such events is essential for effective risk management. Traditional Generalised Pareto Distribution (GPD) models often ignore market conditions; in contrast, our framework links the scale parameter to covariates using a log-linear function, allowing tail risk to respond dynamically to market volatility. We examine four prior choices for Bayesian regularisation of regression coefficients: Cauchy, Lasso (Laplace), Ridge (Gaussian), and Zellner's g-prior. Simulation results suggest that the Cauchy prior delivers the best trade-off between predictive accuracy and model simplicity, achieving the lowest RMSE, AIC, and BIC values. Empirically, we apply the model to large negative returns (exceeding 5%) in the Nifty 50 index. Volatility measures from the Nifty 50, S&P 500, and gold are used as covariates to capture both domestic and global risk drivers. Our findings show that tail risk increases significantly with higher market volatility. In particular, both S&P 500 and gold volatilities contribute meaningfully to crash prediction, highlighting global spillover and flight-to-safety effects. The proposed GPR model offers a robust and interpretable approach for tail risk forecasting in emerging markets. It improves upon traditional EVT-based models by incorporating real-time financial indicators, making it useful for practitioners, policymakers, and financial regulators concerned with systemic risk and stress testing.

All codes and datasets for this paper are available in the following GitHub repository: `https://github.com/sourish-cmi/quant/tree/main/Predicting_Stock_Market_Crash`

Key words: Bayesian regularisation; Generalised Pareto regression; Financial tail risk; Market crash prediction; Volatility spillover.

AMS Subject Classifications: 62K05, 05B05.

Corresponding Author: Sourish Das
Email: sourish@cmi.ac.in

1. Introduction

Stock market crashes have significant economic implications, often resulting in large wealth erosion, investor panic, and long-lasting financial instability, see Liu *et al.* (2021); Song *et al.* (2022). Understanding and anticipating such extreme events is vital for effective risk management, financial regulation, and economic forecasting, see Dai *et al.* (2021). During the COVID-19 crash, for instance, Giglio *et al.* (2020) document a sharp decline in short-term investor expectations and a surge in disagreement about future market performance. Mazur *et al.* (2021) further highlight that while sectors like healthcare and software showed resilience, others such as hospitality and energy experienced extreme negative returns and asymmetric volatility, accompanied by varied corporate responses.

Extreme Value Theory (EVT) provides a principled framework for modelling the statistical behaviour of rare but severe events, particularly through the Generalised Pareto Distribution (GPD), which is commonly used for modelling exceedances over a high threshold, see Smith (1985). In the context of financial crashes, Fry (2008) presents a comprehensive investigation into bubbles, volatility, and contagion, deriving a GPD-based model for market drawdowns to analyse tail risk dynamics.

Several studies have applied the GPD to model extreme financial phenomena. Malevergne *et al.* (2006) and Das and Halder (2016) use EVT to characterise financial extremes, but their models lack covariate inputs, limiting their capacity to account for underlying market conditions such as volatility or liquidity. Liu (2011) advances the field by detecting structural breaks in tail behaviour using a transformed GPD and fluctuation tests, highlighting the need for flexible models when estimating extreme quantiles under changing market regimes.

More recent work by Rai *et al.* (2022) analyses the statistical behaviour of aftershocks in stock market crashes and finds that tail patterns and inter-occurrence times vary depending on the nature of the crash. This supports the hypothesis that covariate information, such as macroeconomic signals, volatility, or liquidity, can influence the shape and scale of extreme return distributions.

Motivated by these findings, we extend the standard GPD framework by modelling the scale parameter as a function of covariates through a log-linear link. This Generalised Pareto Regression (GPR) model enhances both interpretability and predictive capacity by incorporating market-relevant features directly into the tail distribution.

Earlier examples of GPR models include Das *et al.* (2010), who model extreme alcohol consumption events using a covariate-linked scale parameter within a GPD framework. In the operational risk domain, Hambuckers *et al.* (2018) employ a regularised GPD regression where both the scale and shape parameters vary with covariates. While flexible, such models risk overparameterisation and identifiability issues. To maintain parsimony and model explainability, we assume a common shape parameter across observations and model only the scale parameter as a function of predictors.

While it is possible to model the shape parameter ξ as a function of covariates, we keep it constant across observations to ensure model parsimony and stability. Varying ξ with covariates substantially increases model complexity and introduces identifiability challenges, especially with limited tail data. Our preliminary experiments indicated that such extensions

did not significantly improve predictive performance but led to unstable estimates. Hence, we opt for a simpler model with a shared shape parameter across all observations.

Next in Section 2 we present the data description and exploratory data analysis. In Section 3 we present the Bayesian methodology for Generalised Pareto Regression. In Section 4, we present a thorough simulation study to evaluate the performance of different methodologies. In Section 5, we present the Bayesian analysis of Indian market crash events. Section 6 conclude the paper.

2. Data insight

We analyse the Nifty 50 index over the period from 17 September 2007 to 13 June 2025. Figure 1(a) shows the daily closing prices of the Nifty 50, and Figure 1(b) displays the corresponding log-returns. The blue dashed line marks the -2% threshold and the red dashed line indicates the -5% threshold in returns. During this period, the index experienced a few sharp declines exceeding 5%, most notably during the 2008 global financial crisis and the COVID-19 pandemic. Given that the daily volatility of log-returns is approximately 1.32%, a drop of more than 2% is typically regarded as a significant shock. The choice of 2% as the lower threshold aligns with commonly used multiples of standard deviation in financial risk literature, indicating a statistically significant deviation.

The 5% threshold corresponds to historically observed crash magnitudes, such as those during the 2008 financial crisis and COVID-19 market panic, and serves as a benchmark for severe market stress. These thresholds were also supported by exploratory diagnostics that revealed natural separation in the distribution of extreme returns. A decline exceeding 5% in a single day can trigger widespread panic, often resulting in substantial margin calls in the derivatives market, particularly in futures and options, which may lead to forced liquidations and amplified market instability.

Figure 1(c) displays the empirical volatility of daily returns. This is computed using an exponentially weighted moving average (EWMA) method that considers past return variance over a rolling window of $k = 21$ trading days. Specifically, the conditional variance at time t is calculated as

$$\sigma_t^2 = \alpha \cdot s_{t-1}^2 + (1 - \alpha) \cdot r_t^2,$$

where r_t is the daily return, s_{t-1}^2 is the sample variance of returns over the previous $k - 1$ days, and $\alpha = 0.9$ is the smoothing parameter; see Sen and Das (2023). The resulting daily volatility is annualised by multiplying by $\sqrt{250}$, reflecting the approximate number of trading days in a year. This approach yields a more adaptive and responsive measure of recent market volatility compared to traditional historical estimates.

Figure 1(d) presents the Garman-Klass estimate of intraday volatility. This estimator offers a more efficient measure of daily volatility by incorporating high, low, opening, and closing prices, see Garman and Klass (1980). It reduces estimation variance by using price range data instead of just closing prices. The Garman-Klass volatility is defined as

$$\sigma_{\text{GK}}^2 = \frac{1}{2}\left(\log\left(\frac{H}{L}\right)\right)^2 - (2\log(2) - 1)\left(\log\left(\frac{C}{O}\right)\right)^2,$$

where O, H, L, and C denote the opening, high, low, and closing prices, respectively. This

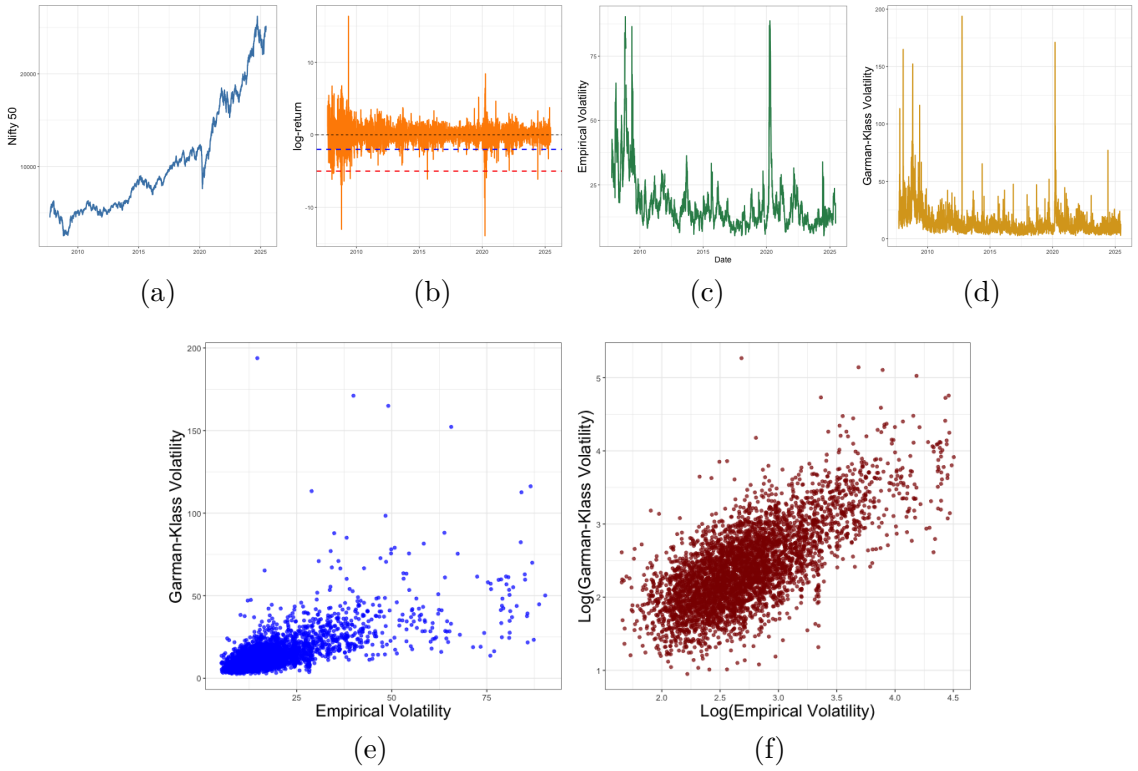
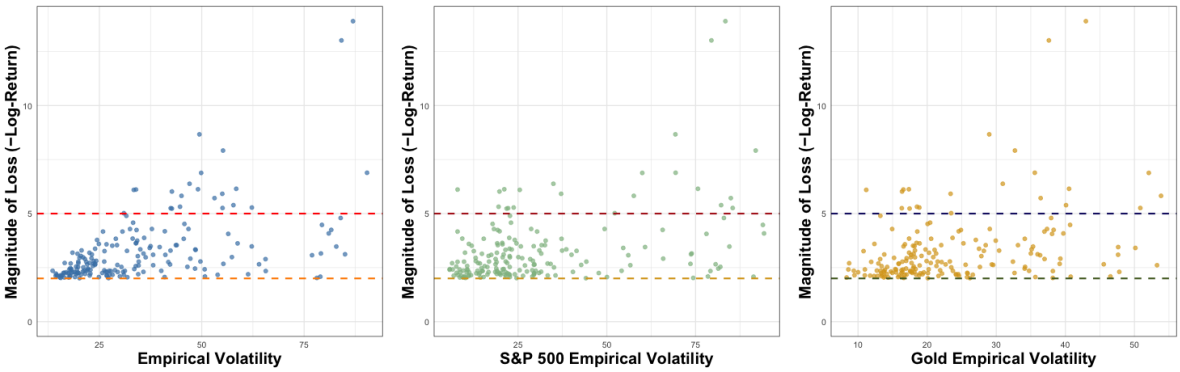


Figure 1: Visual summary of Nifty50 prices and volatility measures: (a) closing price, (b) log-return, (c) empirical volatility, (d) Garman-Klass volatility, (e) empirical *vs* GK volatility, and (f) log-transformed comparison of empirical *vs* GK volatility.

formula assumes zero drift and the absence of overnight jumps, making it well suited for capturing intraday risk characteristics.

Figure 1(e) shows the scatter plot of empirical volatility against Garman-Klass volatility. Figure 1(f) presents the same comparison on a logarithmic scale. The correlation between the logarithms of empirical and Garman-Klass intra-day volatility is approximately 0.71, suggesting a strong association in the log scale, and highlighting consistency between the two methods of volatility estimation.

In addition, we incorporate the S&P 500 index and gold prices into our analysis to capture global market volatility and assess its influence on the Indian stock market. To study the tail behaviour of returns, we focus on large daily losses in the Nifty 50 index, specifically, those drops, exceeding 2% level, and examine their relationship with daily empirical volatility. Figure 2 visualises this association across three dimensions of volatility: panel (a) presents the Nifty's own empirical volatility, panel (b) shows S&P 500 volatility, and panel (c) depicts gold volatility. The plots reveal that large negative returns often coincide with elevated volatility, both domestically and globally. The strong alignment with S&P 500 volatility suggests the presence of spillover effects from global equity markets, while the relationship with gold volatility may reflect investor flight-to-safety behaviour during periods of financial stress.



(a) (b) (c)

Figure 2: Relationship between large losses (negative returns) in the Nifty 50 and empirical volatility measures: (a) own volatility of Nifty 50, (b) S&P 500 volatility, and (c) gold volatility. Dashed lines indicate thresholds at 2% and 5% loss levels.

3. Methodology

Suppose $\mathcal{D} = \{(y_i, \mathbf{x}_i) \mid i = 1, 2, \cdots, n\}$ is the observed dataset, where y_i follows a Generalised Pareto Distribution (GPD), *i.e.*,

$$y_i \sim \text{GPD}(\mu, \sigma_i, \xi), \quad i = 1, 2, \cdots, n,$$

with probability density function:

$$f(y_i \mid \sigma_i, \xi) = \frac{1}{\sigma_i}\left(1 + \xi\frac{y_i - \mu}{\sigma_i}\right)^{-1/\xi - 1},$$

and $\mathbf{x}_i = \{x_{i1}, x_{i2}, \cdots, x_{ip}\}$ are the covariates associated with the i^{th} observation. The threshold μ is assumed known; the shape parameter ξ is same across observations, and the scale parameter σ_i is modelled as a log-linear function of covariates:

$$\log(\sigma_i) = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

The support of Y_i is (μ, ∞) if $\xi < 0$, and $(\mu, \mu + \sigma_i/\xi)$ if $\xi > 0$. The special case $\xi = 0$ corresponds to the exponential distribution with mean σ_i, interpreted as the limit as $\xi \to 0$. The survival function of the GPD regression model is:

$$\begin{aligned}\mathbb{P}(Y > y_0 \mid Y > \mu) &= \left(1 - \frac{\xi(y_0 - \mu)}{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}\right)^{1/\xi} \\ &= \left(1 + \frac{\xi(y_i - \mu)}{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}\right)^{-1/\xi}.\end{aligned}$$

In financial applications, if Y represents the daily negative return (in percentage), and a market crash is defined as a daily drop exceeding $y_0 = 5\%$, then this survival function

quantifies the conditional probability of a crash, given that a drop of, say, $\mu = 2\%$ has already occurred. The conditional expectation of Y_i given the covariates is:

$$\mathbb{E}(Y_i \mid Y_i \geq \mu) = \begin{cases} \infty, & \xi \geq 1, \\ \mu + \dfrac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1-\xi}, & \xi < 1. \end{cases}$$

The variance exists only when $\xi < 0.5$ and is given by:

$$\mathbb{V}ar(Y_i \mid Y_i \geq \mu) = \frac{\exp(2\mathbf{x}_i^\top \boldsymbol{\beta})}{(1-\xi)^2(1-2\xi)}.$$

This characterisation highlights the influence of covariates on the first two moments of the tail distribution and emphasises the model's ability to capture heavy-tailed behaviour.

3.1. Log-likelihood function

The log-likelihood function of the GPD regression model is:

$$\log L(\xi, \boldsymbol{\beta}) = -\sum_{i=1}^{n} \mathbf{x}_i^\top \boldsymbol{\beta} + \left(\frac{1}{\xi} - 1\right) \sum_{i=1}^{n} \log\left[1 + \xi \frac{y_i - \mu}{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}\right].$$

This formulation treats the exceedances $y_i > \mu$ as GPD-distributed with covariate-dependent scale. The shape parameter ξ governs the heaviness of the tail, while the covariates modulate the scale of exceedance. In particular, the conditional mean:

$$\mathbb{E}(Y_i \mid Y_i \geq \mu) = \mu + \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1-\xi}$$

exists only for $\xi < 1$, and the variance exists only if $\xi < 0.5$. This structure provides a coherent framework for modelling the magnitude of extreme outcomes while adjusting for covariate information. As shown in Smith (1985), the maximum likelihood estimator (MLE) exists asymptotically when $\xi < 1$ and is consistent, asymptotically normal, and efficient provided $\xi < 0.5$.

3.2. Truncated Cauchy prior on ξ

The `truncated Cauchy(0,1)` prior truncated from above at 1, *i.e.*, the support is $\xi < 1$. The *probability density function* of the standard Cauchy distribution is:

$$p(\xi) = \frac{1}{\pi(1+\xi^2)}, \quad \xi \in (-\infty, \infty).$$

We define the *truncated density*:

$$p_{\text{trunc}}(\xi) = \frac{1}{Z} \cdot \frac{1}{\pi(1+\xi^2)}, \quad \text{for } \xi < 1$$

where Z is the normalising constant, *i.e.* the total probability over the truncated domain:

$$\begin{aligned} Z &= \int_{-\infty}^{1} \frac{1}{\pi(1+\xi^2)}\, d\xi = \frac{1}{\pi} \cdot \Big[\tan^{-1}(\xi)\Big]_{-\infty}^{1} = \frac{1}{\pi} \cdot \Big(\tan^{-1}(1) - \tan^{-1}(-\infty)\Big) \\ &= \frac{1}{\pi}\left(\frac{\pi}{4} - (-\frac{\pi}{2})\right) = \frac{1}{\pi} \cdot \frac{3\pi}{4} = \frac{3}{4} \end{aligned} \tag{1}$$

Now plug in $Z = \frac{3}{4}$:

$$p_{\text{trunc}}(x) = \frac{4}{3\pi(1+\xi^2)}, \quad \xi < 1.$$

The pdf of `truncated Cauchy(0,1)` prior density, with upper truncation at $\xi = 1$.

$$p(\xi) = \frac{4}{3\pi(1+\xi^2)}, \quad \text{for } \xi < 1$$

3.3. Cauchy prior on $\boldsymbol{\beta}$

In our Bayesian regression framework, we assign a standard Cauchy prior to the regression coefficients $\boldsymbol{\beta}$, see Gelman *et al.* (2008). The Cauchy distribution is symmetric and heavy-tailed, making it well-suited for scenarios where large effect sizes are plausible or when robust shrinkage is needed. The standard Cauchy prior has the following probability density function:

$$p(\beta) = \frac{1}{\pi(1+\beta^2)}, \quad \beta \in \mathbb{R}.$$

A key property of the Cauchy distribution is that it lacks a finite mean, variance, and higher moments. This feature makes it a particularly attractive choice as a weakly informative prior, see Berger (1985). It imposes minimal structure on the parameter space, allowing the data to dominate the posterior inference, while still penalising extremely large values less severely than priors with finite variance, such as the Gaussian. Consequently, the Cauchy prior provides a balance between regularisation and flexibility, making it suitable for sparse or high-dimensional regression models where robustness and parsimony are desired.

3.4. Lasso prior on $\boldsymbol{\beta}$

An alternative to the Cauchy prior is the Lasso prior, which corresponds to a Laplace (double-exponential) distribution on the regression coefficients $\boldsymbol{\beta}$, see Hastie *et al.* (2009), Park and Casella (2008). The Lasso prior encourages sparsity in the estimated coefficients by applying stronger shrinkage towards zero, making it particularly useful for high-dimensional models or when variable selection is of interest. The probability density function of the Laplace distribution with scale parameter $\lambda > 0$ is given by:

$$p(\beta) = \frac{\lambda}{2}\exp(-\lambda|\beta|), \quad \beta \in \mathbb{R}.$$

This prior has a sharp peak at zero and heavier tails than the Gaussian, enabling it to shrink small coefficients more aggressively while allowing larger coefficients to remain relatively unaffected. The result is a sparse posterior mode, with many coefficients estimated as exactly zero, aligning with the behaviour of the classical Lasso estimator. Unlike the Cauchy prior, the Laplace distribution has finite mean and variance, making it more suitable when a moderate degree of regularisation is desired without the extreme heavy tails of the Cauchy prior. Its effectiveness in automatic variable selection and computational tractability makes the Lasso prior a popular choice in Bayesian sparse regression models.

3.5. Ridge prior on $\boldsymbol{\beta}$

The Ridge prior assumes a Gaussian (normal) distribution on the regression coefficients $\boldsymbol{\beta}$, offering smooth shrinkage without inducing sparsity, Hastie *et al.* (2009). It is commonly used when multicollinearity is present or when all predictors are believed to have small, non-zero effects. The probability density function of the Ridge prior with precision parameter τ is:

$$p(\beta) = \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{\tau}{2}\beta^2\right), \quad \beta \in \mathbb{R}.$$

This is equivalent to placing an independent $\mathcal{N}(0, \tau^{-1})$ prior on each coefficient β. The Ridge prior results in posterior estimates that are biased toward zero, but unlike the Lasso prior, it does not set any coefficients exactly to zero. Consequently, it is well-suited for settings where all predictors contribute weakly to the response.

In the Bayesian framework, the Ridge prior leads to a conjugate posterior when combined with a Gaussian likelihood, allowing for closed-form posterior inference in linear models. Its computational simplicity and regularising effect make it a widely used choice in models where interpretability through sparsity is not essential but stabilisation of estimates is desired.

3.6. Zellner's g-prior on $\boldsymbol{\beta}$

Zellner's g-prior is a popular conjugate prior used in Bayesian linear regression, particularly when the design matrix $\mathbf{X}$ is fixed and known, see Zellner (1986),Sabanés Bové and Held (2011). It imposes a multivariate normal prior on the regression coefficients $\boldsymbol{\beta}$, with the covariance structure informed by the design matrix:

$$\boldsymbol{\beta} \sim \mathcal{N}\left(\mathbf{0},\ g(\mathbf{X}^\top \mathbf{X})^{-1}\right),$$

where $g > 0$ is a scalar hyperparameter controlling the strength of the prior relative to the likelihood. This prior has several attractive properties:

- It is invariant under linear transformations of the design matrix.
- The posterior mean shrinks toward zero as $g \to 0$, while as $g \to \infty$, the prior becomes non-informative.
- The use of $(\mathbf{X}^\top \mathbf{X})^{-1}$ ensures the prior reflects the geometry of the predictors.

Zellner's g-prior is especially useful for Bayesian model comparison and variable selection, as it facilitates closed-form expressions for marginal likelihoods and Bayes factors. However, care must be taken in the choice of g, as it significantly influences the inference. Empirical Bayes, fixed g, or hierarchical models placing a hyperprior on g are common approaches to address this.

3.7. MAP estimation for efficient Bayesian model exploration

Maximum a Posteriori (MAP) estimation provides a computationally efficient alternative to full Bayesian inference, especially when exploring different model specifications or

prior structures. Unlike posterior summaries such as the posterior mean or credible intervals that rely on Markov Chain Monte Carlo (MCMC) methods, MAP estimation avoids the computational burden of sampling by framing inference as an optimisation problem. Formally, the MAP estimate is defined as:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \, p(\theta \mid \mathcal{D}) = \arg\max_{\theta} \left[\log p(\mathcal{D} \mid \theta) + \log p(\theta)\right],$$

where $\mathcal{D}$ denotes the observed data, $p(\mathcal{D} \mid \theta)$ is the likelihood function, and $p(\theta)$ is the prior distribution. In our case $\theta = (\boldsymbol{\beta}, \xi)$. This above expression highlights that MAP estimation balances the fit to the data and the influence of prior information, analogous to penalised likelihood methods in the frequentist setting.

MAP estimation is often preferred in high-dimensional problems or when computational resources are limited, as it offers faster and more stable inference than MCMC, which can suffer from slow convergence and poor mixing. Moreover, it enables quick evaluation of multiple models or regularisation schemes, such as Lasso or Ridge, where priors play the role of sparsity-inducing penalties.

Importantly, MAP estimation is well-suited for the model exploration phase. Once a final model is selected, based on predictive performance or interpretability, one may then employ MCMC on that specific model to obtain a full characterisation of the posterior distribution. In this way, MAP provides an efficient screening tool, while full Bayesian inference is reserved for the final model of interest.

4. Simulation study of frequentist properties

To evaluate the performance of different regularised Generalised Pareto Distribution (GPD) regression models, we conduct a simulation study comparing four Bayesian prior structures on the regression coefficients: the **Cauchy prior**, **Lasso** (ℓ_1 penalty), **Ridge** (ℓ_2 penalty), and **Zellner's g-prior**. Our aim is to assess and compare their predictive accuracy, estimation error, model complexity (via AIC and BIC), and computational efficiency.

We simulate $N = 100$ datasets, each consisting of $n = 100$ observations. For each dataset, the covariate matrix $X \in \mathbb{R}^{n \times p}$ is generated from a multivariate normal distribution with identity covariance, where the number of predictors is $p = 5$. The true regression coefficients β are sampled from a standard normal distribution, and the GPD shape parameter ξ is drawn from a uniform distribution on the interval $(-0.5, 0.5)$. The response variable y is generated from a GPD model with a fixed location parameter $\mu = 2$ and a scale parameter $\sigma = \exp(X\beta)$. Each dataset is then split into a training set (80%) and a testing set (20%).

Each of the four models is estimated using maximum a posteriori (MAP) estimation via the `BFGS` optimisation algorithm. The Cauchy prior is specified as a weakly informative heavy-tailed prior on the coefficients β, along with a truncated Cauchy prior on the shape parameter ξ. The Lasso model imposes an ℓ_1 penalty on the regression coefficients, while the Ridge model uses an ℓ_2 penalty; in both cases, the regularisation parameter is chosen using 5-fold cross-validation. The Zellner's g-prior assumes a prior of the form $\beta \sim \mathcal{N}(0, g(X^\top X)^{-1})$, and the hyperparameter g is selected through cross-validation.

We assess model performance using several metrics. These include the root mean squared error (RMSE) of predictions on the test data, as well as the RMSE for recovering

the true coefficient vector β and the shape parameter ξ. We also compute the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), both based on the negative log-likelihood (excluding the contribution of the prior). In addition, we record the computational time taken to fit each model, and express it as a multiple relative to the time taken for the Cauchy prior model, which serves as the baseline.

Table 1: Simulation comparison of GPD regression models with different priors

Metric	Cauchy	Lasso	Ridge	g-prior
RMSE (y)	11.36	10.37	10.65	10.43
RMSE (β)	0.10	0.16	0.16	0.18
RMSE (ξ)	0.08	0.08	0.09	0.11
AIC	162.67	177.15	175.74	180.94
BIC	176.97	191.44	190.03	195.24
Time (sec)	0.00	0.60	0.53	8.45
Time (relative)	1.00	175.64	155.69	2476.86

Finally, the results across the 100 simulations are aggregated using the median to ensure robustness. The outcome is summarised in the Table 1, displaying predictive accuracy, parameter estimation error, model complexity, and computational cost for each prior, facilitating a comprehensive evaluation of their relative performance.

5. Empirical study of market crashes

In this section, we investigate the application of Generalised Pareto Distribution (GPD) regression for modelling extreme negative returns in the Indian equity market. Our analysis is based on daily returns of the NSE Nifty 50 index, augmented with realised volatility measures derived from domestic and global financial assets. The central objective is to assess the predictive performance and interpretability of GPD regression models under a range of prior assumptions on the regression coefficients, within a Bayesian framework for financial tail risk modelling.

All codes and datasets for this paper are available in the following GitHub repository: `https://github.com/sourish-cmi/quant/tree/main/Predicting_Stock_Market_Crash`

Data description and preprocessing

We work with a cleaned dataset comprising daily log-returns of the NSE Nifty 50 index, with all missing entries removed. Denoting the daily return at time t as r_t, we extract a subset of observations where $r_t < -2\%$, corresponding to the left tail of the return distribution. This subset constitutes approximately 4.6% of all trading days. Among these tail events, 11.6% show losses exceeding 5%, indicating the presence of disproportionately large market crashes within the extreme left tail.

To facilitate interpretation and estimation, negative returns are transformed into

absolute values, *i.e.*, $y_t = |r_t|$ for $r_t < -2\%$. We construct covariates based on (i) empirical volatility and (ii) Garman–Klass (GK) volatility for the Nifty 50, S&P 500, and gold. These volatility variables are standardised to have mean zero and unit variance. The baseline model includes only the intercept and Nifty empirical volatility, while richer covariate sets are examined in subsequent analyses.

Predictive evaluation and model comparison

We randomly split the dataset into training (80%) and test (20%) sets. Out-of-sample predictive performance is assessed via Root Mean Squared Error (RMSE), and in-sample fit is evaluated using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), both computed from the negative log-likelihood (excluding prior terms). For the Lasso model, degrees of freedom are adjusted to reflect the number of non-zero coefficients.

Table 2: Model evaluation using RMSE, AIC, and BIC.

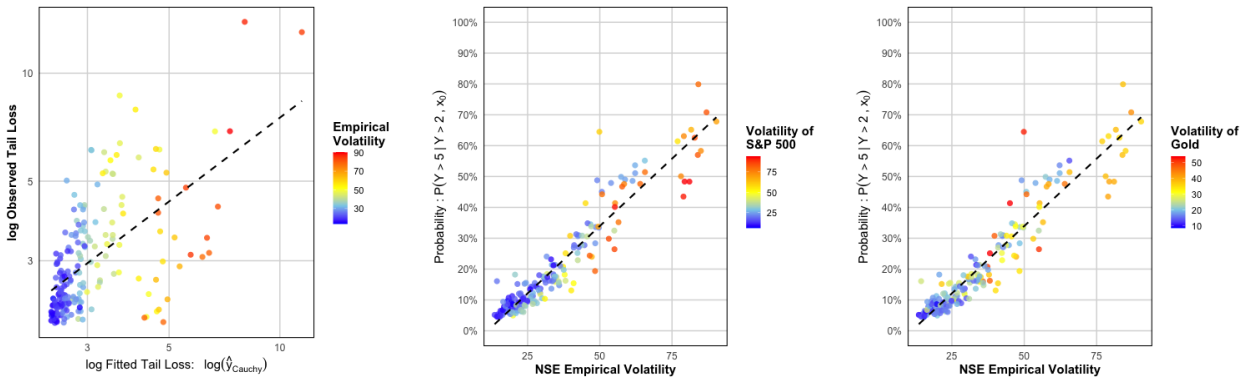
Prior Models	RMSE	AIC	BIC
Cauchy	1.58	298.06	322.20
Lasso	1.80	311.57	335.71
Ridge	1.73	302.07	326.20
Zellner's g-prior	1.94	308.12	332.25

Among all models, the Cauchy prior delivers the best fit, achieving the lowest AIC and BIC values, along with the smallest RMSE, suggesting strong predictive accuracy and robust in-sample performance. The Ridge and g-prior models provide competitive alternatives, whereas the Lasso model performs relatively poorly, possibly due to over-shrinkage.

To understand the economic implications of volatility on tail risk, we compute both the conditional expectation $\mathbb{E}[Y \mid Y > \mu, x]$ and the conditional exceedance probability $\mathbb{P}(Y > 5 \mid Y > 2, x)$ under the fitted Cauchy model. Results show a nonlinear increase in both quantities with rising volatility. At the lower end (*e.g.*, 10th percentile of volatility), expected losses are around 2.4 with tail event probabilities below 5%. At the upper end (90th percentile), expected losses exceed 7 with crash probabilities above 65%.

Figure 3 presents three visual diagnostics. Panel (a) plots the fitted *vs.* observed tail losses on the log scale, with colour indicating empirical Nifty volatility. The upward trend and colour gradient affirm that higher volatility coincides with more severe losses, though dispersion increases in the tail. Panels (b) and (c) display the fitted probability $\mathbb{P}(Y > 5 \mid Y > 2, x_0)$ as a function of Nifty volatility, coloured by the S&P 500 and gold volatilities, respectively. *These plots reveal clear global spillover effects: both S&P 500 and gold volatilities enhance the likelihood of extreme losses in the Indian market, the former via market correlation and the latter via flight-to-safety dynamics.*

This empirical exercise highlights that the tail behaviour of Indian equity returns is jointly shaped by local volatility conditions and global financial signals. The GPD model with a Cauchy prior not only provides superior predictive performance but also captures the nonlinear and interactive effects of volatility on extreme losses. These findings underscore



(a) (b) (c)

Figure 3: Relationship between extreme tail events in the Nifty 50 index and market volatility. Panel (a) compares the log-fitted tail losses from the GPD model under a Cauchy prior with actual observed losses. Panels (b) and (c) show the model-based conditional tail probabilities as functions of Nifty volatility, coloured by S&P 500 and gold volatilities, respectively.

the importance of robust regularisation and the inclusion of international covariates and global spill-over effect in tail risk modelling of Indian stock market.

6. Conclusion

In this study, we developed and applied a Bayesian Generalised Pareto Regression (GPR) framework to model and predict extreme negative returns in the Indian equity market, particularly focusing on the Nifty 50 index. The core innovation lies in modelling the scale parameter of the Generalised Pareto Distribution as a log-linear function of market volatility indicators, while keeping the shape parameter same for all data points. This allows the model to adapt to changing volatility regimes while preserving interpretability and parsimony.

Our theoretical development was supported by a thorough simulation study comparing different regularisation strategies, including Cauchy, Lasso, Ridge, and Zellner's g-prior; within a MAP estimation framework. Among these, the Cauchy prior emerged as the most effective in balancing predictive accuracy and model complexity. This observation was further validated in the empirical analysis of actual market crash events, where the Cauchy prior achieved the lowest AIC, BIC, and RMSE values.

Empirical findings highlight a strong nonlinear association between volatility and the likelihood and magnitude of extreme losses. We demonstrated that both domestic (Nifty) and global (S&P 500, gold) volatility measures significantly influence the tail risk. The inclusion of global covariates like S&P 500 and gold volatilities proved crucial in capturing spillover and flight-to-safety effects during periods of financial stress.

Visual diagnostics revealed that in high-volatility regimes, the conditional probability of a crash; defined as a loss exceeding 5% given a 2% drop; can exceed 60%, with expected losses rising sharply. These insights underscore the importance of incorporating volatility-

sensitive covariates and flexible modelling strategies when forecasting rare but impactful financial events.

Overall, our study demonstrates that Bayesian GPD regression, particularly under heavy-tailed priors like the Cauchy, provides a powerful and interpretable tool for tail risk modelling in financial markets. It holds promise for future applications in systemic risk monitoring, stress testing, and portfolio tail risk management. Future research could explore dynamic extensions, integrate macroeconomic signals, and investigate hierarchical Bayesian formulations that allow for time-varying shape parameters or latent volatility drivers.

Acknowledgements

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

Conflict of interest

The author does not have any financial or non-financial conflicts of interest to declare for the research work included in this article.

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer, New York, 2nd edition.

Dai, P.-F., Xiong, X., Liu, Z., Huynh, T. L. D., and Sun, J. (2021). Preventing crash in stock market: The role of economic policy uncertainty during covid-19. *Financial Innovation*, **7**, 1–15.

Das, K. P. and Halder, S. C. (2016). Understanding extreme stock trading volume by generalized Pareto distribution. *The North Carolina Journal of Mathematics and Statistics*, **2**, 45–60.

Das, S., Harel, O., Dey, D. K., Covault, J., and Kranzler, H. R. (2010). Analysis of extreme drinking in patients with alcohol dependence using Pareto regression. *Statistics in Medicine*, **29**, 1057–1065.

Fry, J. M. (2008). *Statistical Modelling of Financial Crashes.* PhD thesis, University of Sheffield.

Garman, M. B. and Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *Journal of Business*, **53**, 67–78. JSTOR.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, **2**, 1360–1383.

Giglio, S., Maggiori, M., Stroebel, J., and Utkus, S. (2020). Inside the mind of a stock market crash. Technical report, National Bureau of Economic Research.

Hambuckers, J., Groll, A., and Kneib, T. (2018). Understanding the economic determinants of the severity of operational losses: A regularized generalized Pareto regression approach. *Journal of Applied Econometrics*, **33**, 898–935.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2nd edition.

Liu, W. (2011). Detecting structural breaks in tail behaviour – from the perspective of fitting the generalized Pareto distribution. *Applied Economics*, **45**, 1273–1286.

Liu, Z., Huynh, T. L. D., and Dai, P.-F. (2021). The impact of covid-19 on the stock market crash risk in China. *Research in international Business and Finance*, **57**, 101419.

Malevergne, Y., Pisarenko, V., and Sornette, D. (2006). On the power of generalized extreme value (gev) and generalized Pareto distribution (gpd) estimators for empirical distributions of stock returns. *Applied Financial Economics*, **16**, 271–289.

Mazur, M., Dang, M., and Vega, M. (2021). Covid-19 and the march 2020 stock market crash. evidence from S&P1500. *Finance Research Letters*, **38**, 101690.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, **103**, 681–686.

Rai, A., Mahata, A., Nurujjaman, M., and Prakash, O. (2022). Statistical properties of the aftershocks of stock market crashes revisited: Analysis based on the 1987 crash, financial-crisis-2008 and covid-19 pandemic. *International Journal of Modern Physics C*, **33**, 2250019.

Sabanés Bové, D. and Held, L. (2011). Hyper-g priors for generalized linear models. *Bayesian Analysis*, **6**, 387–410.

Sen, R. and Das, S. (2023). *Computational Finance with R*. Indian Statistical Institute Series. Springer Singapore. Published: 18 May 2023 (Hardcover), 16 May 2023 (eBook).

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, **72**, 67–90.

Song, R., Shu, M., and Zhu, W. (2022). The 2020 global stock market crash: Endogenous or exogenous? *Physica A: Statistical Mechanics and Its Applications*, **585**, 126425.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In Goel, P. K. and Zellner, A., editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. North-Holland, Amsterdam.

Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025; pp 135-152

Unraveling Biological Complexity: AI and Statistical Approaches to Multi-Omics Data Integration

D. C. Mishra[1], Shesh Nath Rai[2], Mamatha Y. S.[1], K. K. Chaturvedi[1], Sudhir Srivastava[1], Neeraj Budhlakoti[1] and Girish Kumar Jha[1]

[1]*Division of Agricultural Bioinformatics*
ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110012
[2]*Cancer Data Science Center*
University of Cincinnati, College of Medicine, Cincinnati, OH, USA.

Received: 16 June 2025; Revised: 14 August 2025; Accepted: 17 August 2025

Abstract

In the era of precision medicine, understanding the intricate biological mechanisms underlying diseases requires a comprehensive analysis of multi-omics data, including genomics, transcriptomics, proteomics and metabolomics. The sheer volume and complexity of these datasets present significant challenges in deciphering the interactions and regulatory networks that govern cellular functions. This paper will explore how cutting-edge artificial intelligence (AI) and statistical methodologies, including deep learning approaches like Variational Autoencoder (VAE) and Graph Neural Networks (GNNs), are transforming the integration of multi-omics data, enabling new insights into biological complexity. We will discuss advanced statistical models, such as Bayesian Networks, Canonical Correlation Analysis (CCA) and Multi-Omics Factor Analysis (MOFA), that facilitate the integration of diverse data types, revealing deeper layers of biological information that are often obscured in traditional analyses. From identifying biomarkers for early disease detection to uncovering therapeutic targets, the combination of AI, deep learning and statistical approaches holds great promise in advancing our understanding of health and disease.

Key words: Multiomics; Data integration; MOFA; Deep learning; Network based approach.

AMS Subject Classifications: 62K05, 05B05.

1. Introduction

The central dogma of molecular biology, which describes the flow of genetic information from DNA to RNA to protein, has long served as a cornerstone of biological understanding. However, a comprehensive understanding of biological systems requires the integration of data from multiple 'omics' layers. Genomics, transcriptomics, proteomics and

Corresponding Author: D.C. MISHRA
Email: dwij.mishra@gmail.com

metabolomics each offer a unique perspective, capturing different aspects of cellular function and regulation. The advent of high-throughput technologies, such as next-generation sequencing and mass spectrometry, has led to an explosion of omics data, creating both opportunities and challenges for systems biology, see Misra (2018).

While each omics layer provides valuable information, studying them in isolation offers an incomplete and potentially misleading picture. For instance, changes in mRNA transcript levels do not always directly correlate with corresponding protein abundances due to post-transcriptional regulation, protein turnover and other factors. Multi-omics integration seeks to address these limitations by combining data from multiple sources to provide a more holistic and accurate representation of biological systems, see Subramanian *et al.* (2020).

In this paper, we explore a range of statistical and AI-based methods for multi-omics data integration, with a focus on Canonical correlation analysis, Network modeling, Bayesian inference and Deep learning strategies like Variational autoencoders. We review existing tools such as mixOmics, RGCCA, and PINSPlus, which leverage these methods for practical applications in agricultural and biomedical research.

2. Statistical approaches to multi-omics data integration

Statistical methods play a crucial role in managing the high-dimensional, heterogeneous nature of multi-omics data. Several widely used methods for integrating multi-omics data are given below, see Naserkheil *et al.* (2022).

2.1. Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis is a statistical method designed to identify and quantify the linear relationships between two multidimensional datasets. In the context of multi-omics data integration, CCA helps in discovering correlated patterns across different omics layers—such as transcriptomics and proteomics—thus uncovering shared biological signals, see Wróbel *et al.* (2024).

Let $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times q}$ be two centered datasets representing two omics layers, where n is the number of samples, and p and q are the number of variables in each omics type. CCA seeks linear combinations of the variables in each dataset such that the correlation between these combinations is maximized. We aim to find vectors $\mathbf{a} \in \mathbb{R}^p$ and $\mathbf{b} \in \mathbb{R}^q$ such that the correlation between the canonical variates $X\mathbf{a}$ and $Y\mathbf{b}$ is maximized:

$$\max_{\mathbf{a},\mathbf{b}} \rho = \frac{\mathbf{a^T C_{XY} b}}{\sqrt{\mathbf{a^T C_{XX} a}}\sqrt{\mathbf{b^T C_{YY} b}}} \tag{1}$$

where:

- $\mathbf{C}_{XX} = \frac{1}{n-1}X^TX$ is the covariance matrix of $\mathbf{X}$.
- $\mathbf{C}_{YY} = \frac{1}{n-1}Y^TY$ is the covariance matrix of $\mathbf{Y}$.
- $\mathbf{C}_{XY} = \frac{1}{n-1}X^TY$ is the cross covariance matrix of $\mathbf{XY}$.

This leads to the generalized eigenvalue problem:

$$\mathbf{C}_{XY}\,\mathbf{C}_{YY}^{-1}\,\mathbf{C}_{YX}\,\mathbf{a}=\lambda\,\mathbf{C}_{XX}\,\mathbf{a}$$

$$\mathbf{C}_{YX}\mathbf{C}_{XX}^{-1}\mathbf{C}_{XY}\mathbf{b}=\lambda\mathbf{C}_{YY}\mathbf{b}$$

The first pair $(\mathbf{a}_1, \mathbf{b}_1)$ gives the directions of maximal correlation. Subsequent canonical directions are obtained by enforcing orthogonality constraints with previous variates.

In high-dimensional multi-omics data (where p or q is much larger than n), classical CCA may become ill-posed. In such cases, regularized or sparse variants are used.

2.1.1. Regularized CCA

Regularized CCA adds penalties to the denominator to stabilize the solution, see Parkhomenko *et al.* (2009).

$$\max_{\mathbf{a},\mathbf{b}} \rho=\frac{\mathbf{a}^T\mathbf{C}_{XY}\mathbf{b}}{\sqrt{\mathbf{a}^T(\mathbf{C}_{XX}+\kappa_x\mathbf{I})\mathbf{a}}\,\sqrt{\mathbf{b}^T(\mathbf{C}_{YY}+\kappa_y\mathbf{I})\mathbf{b}}} \tag{2}$$

where κ_x and κ_y are regularization parameters.

2.1.2. Sparse CCA (sCCA)

Sparse CCA (sCCA) imposes sparsity constraints on $\mathbf{a}$ and $\mathbf{b}$, leading to feature selection and interpretability:

$$\begin{aligned} \max_{\mathbf{a},\mathbf{b}} \quad & \mathbf{a}^T\mathbf{C}_{XY}\mathbf{b} \\ \text{subject to} \quad & \|\mathbf{a}\|_2 \le 1, \quad \|\mathbf{b}\|_2 \le 1 \\ & \|\mathbf{a}\|_1 \le c_1, \quad \|\mathbf{b}\|_1 \le c_2 \end{aligned} \tag{3}$$

These constraints $\|\cdot\|_1$ enforce sparsity, making sCCA particularly useful in the context of omics data where many variables are irrelevant or noisy, see Witten and Tibshirani (2009).

2.1.3. Advantages and limitations

CCA is a powerful tool for identifying relationships between multi-omics datasets. It can handle high-dimensional data and identify complex dependencies. However, CCA is sensitive to outliers and assumes a linear relationship between the variables. In cases where the relationship is non-linear, other methods, such as kernel CCA, may be more appropriate.

2.1.4. Tools implementing CCA for multi-omics data integration

a. mixOmics R package with multivariate methods (including CCA) for exploring and integrating omics datasets, see Rohart *et al.* (2017).

b. RGCCA R package offering generalized CCA for integrating multiple datasets.

c. BLOCCS R package for Block Sparse CCA, estimating multiple canonical directions for enhanced interpretability.

2.2. Similarity-based approaches

Similarity-based approaches represent a powerful class of methods in multi-omics data integration. These methods focus on quantifying the similarity or distance between samples within each omics layer and then combining these relationships to gain a unified understanding of biological patterns, such as disease subtypes, cellular states, or treatment responses.

Unlike direct feature-level integration, which merges raw data matrices, similarity-based methods operate by first computing sample-sample similarity matrices independently for each omics type (*e.g.*, transcriptomics, proteomics, metabolomics). These matrices reflect the relationship between samples based on their respective omics profiles.

Let us consider K different omics datasets $\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(K)}\}$, each with n samples and their respective similarity matrices $\{\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, \ldots, \mathbf{S}^{(K)}\}$, where each $\mathbf{S}^{(k)} \in \mathbb{R}^{n \times n}$.

The key idea is to integrate these K similarity matrices into a single consensus matrix $\mathbf{S}_{\text{integrated}}$, which captures the shared structure across all data types.

2.2.1. Similarity Network Fusion (SNF)

One of the most popular similarity-based methods is Similarity Network Fusion, which iteratively updates each similarity matrix using neighborhood information from other omics layers, see Wang *et al.* (2014).

The SNF algorithm involves the following steps:

1. **Compute sample similarity matrices** $\mathbf{S}^{(k)}$ for each omics data type using a distance metric (*e.g.*, Euclidean distance or Gaussian kernel similarity)
2. **Normalize** the matrices to maintain comparability.
3. **Iteratively update** each matrix by combining it with others through a message-passing mechanism:

$$\mathbf{W}_{t+1}^{(k)} = \alpha \mathbf{P}^{(k)} . \left(\frac{1}{K-1} \sum_{l \neq k} \mathbf{W}_t^{(l)} \right) \mathbf{P}^{(k)T} + (1-\alpha) \mathbf{W}_t^{(k)} \tag{4}$$

   where $\mathbf{P}^{(k)}$ is the transition probability matrix of $\mathbf{S}^{(k)}$, and α is a regularization parameter (typically 0.5).

4. **Fuse the final networks** after convergence:

$$\mathbf{S}_{\text{integrated}} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{W}_T^{(k)} \tag{5}$$

The resulting integrated similarity matrix is then used for downstream tasks such as spectral clustering, dimensionality reduction, or classification.

2.2.2. Other tools and methods

a. PINSPlus An extension of perturbation clustering that performs multiple clustering runs on each omics dataset and integrates the results using co-clustering frequencies.

b. NEMO (Neighborhood-based multi-omics clustering) Designed for partial datasets with missing omics layers, it builds local sample neighborhoods and combines them across modalities.

c. iClusterPlus Although fundamentally a latent variable model, it also aligns sample similarities and can be categorized under similarity-based frameworks.

2.2.3. Advantages and limitations

Similarity based integration methods offer several advantages and challenges. Among the advantages, they are robust to missing data, as similarity matrices can still be computed even when some features are absent. They also allow flexible integration, effectively handling heterogeneous omics types without requiring normalization across different data scales. Additionally, these methods enhance interpretability by providing integrated similarity networks that visually and intuitively represent relationships among samples. However, there are notable challenges as well. The choice of similarity metric is critical, as different distance measures can produce significantly different outcomes. Computational complexity is another concern, especially with large datasets, as calculating pairwise similarities can be both memory and time-intensive. Lastly, parameter tuning is essential for algorithms like Similarity Network Fusion, which rely on parameters such as the number of neighbors and kernel width, requiring careful adjustment to ensure reliable results.

2.3. Bayesian models

Bayesian models offer a powerful and principled framework for multi-omics data integration by treating uncertainty explicitly and allowing incorporation of prior biological knowledge. These models are particularly useful in handling heterogeneous, high-dimensional and often noisy datasets typical in multi-omics studies, such as genomics, transcriptomics, epigenomics and proteomics, see Kirk *et al.* (2012).

2.3.1. Bayesian clustering models

These models assign samples to latent clusters using probability distributions, rather than hard assignments. A popular non parametric Bayesian clustering method is the Dirichlet Process Mixture Model (DPMM).

In multi-omics integration, each omics dataset contributes to the clustering through its own likelihood component. For instance, assuming omics data $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(K)}$ share

a common clustering structure **Z**:

$$P\left(Z, \theta^{(1)}, \ldots, \theta^{(K)} \mid \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(K)}\right) \propto P(Z) \prod_{k=1}^{K} P\left(\mathbf{X}^{(k)} \mid Z, \theta^{(k)}\right) P\left(\theta^{(k)}\right) \tag{6}$$

where: θ^k is cluster specific parameter.

Tools and methods include:

a. MDI (Multiple Dataset Integration) A joint Bayesian model that performs clustering on multiple omics layers and identifies consensus clusters.

b. BCC (Bayesian Consensus Clustering) Estimates shared cluster structure while allowing for data-specific variations.

c. LRAcluster (Low-Rank Approximation Clustering) Incorporates low-rank approximations to simplify the Bayesian model for high-dimensional omics data.

2.3.2. Bayesian networks

Bayesian networks are graphical models that represent conditional dependencies among random variables. In multi-omics integration, they are used to model causal relationships between genes, proteins, and metabolites.

A Bayesian network is a directed acyclic graph (DAG), where nodes represent variables (*e.g.*, gene expression, protein levels), and edges encode conditional dependencies. The joint distribution is factorized as:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid \text{Parents}(X_i)\right) \tag{7}$$

This formulation enables modeling of regulatory pathways or signaling cascades across omics layers. Examples:

a. PARADIGM (Pathway Recognition Algorithm using Data Integration on Genomic Models) Integrates copy number and gene expression data to infer pathway activity, see Vaske *et al.* (2021).

b. CONEXIC (COpy Number and EXpression In Cancer) Uses Bayesian networks to identify driver genes by integrating copy number alterations and expression profiles, see Akavia *et al.* (2010).

2.3.3. Advantages and limitations

Bayesian models offer several compelling advantages and face notable challenges. On the positive side, they excel at uncertainty modeling by providing full posterior distributions, which yield credible intervals and enhance confidence in predictions. They also allow the incorporation of prior knowledge, such as known biological pathways or disease associations,

directly into the model. Thanks to modern techniques like variational inference and Markov Chain Monte Carlo (MCMC) sampling, Bayesian methods have become scalable to large datasets. Additionally, they handle missing data naturally as part of the inference process, eliminating the need for imputation. However, these benefits come with challenges. Bayesian inference can be computationally expensive, particularly when dealing with multiple omics layers or a high number of variables. The complexity of designing and validating hierarchical models or directed acyclic graphs (DAGs) demands significant expertise and domain knowledge. Moreover, the results can be sensitive to the choice of priors—poorly chosen or inadequate priors may bias outcomes or impede model convergence.

2.4. Multivariate methods

Multivariate methods are essential tools in multi-omics data integration, offering the capability to jointly analyze multiple variables from different omics layers. Unlike univariate methods that treat each variable independently, multivariate approaches capture correlations, co-variations, and shared structures across datasets, making them ideal for discovering hidden biological relationships and reducing dimensionality in high-throughput omics data. These methods are particularly valuable when integrating datasets from genomics, transcriptomics, proteomics, metabolomics, and other omics types, where the number of variables far exceeds the number of observations, and variables often interact in complex, non-linear ways.

2.4.1. Principal Component Analysis (PCA)

PCA is one of the most widely used unsupervised multivariate techniques for dimensionality reduction. It identifies orthogonal directions (principal components) that capture the maximum variance in the data. When applied to multi-omics datasets either jointly or separately, PCA can reveal dominant variation patterns, batch effects, and clustering structures, see Jolliffe and Cadima (2016).

Given a centered data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, PCA solves the eigenvalue problem:

$$\mathbf{X}^T \mathbf{X} \mathbf{v} = \lambda \mathbf{v} \tag{8}$$

where $\mathbf{v}$ is the eigenvector corresponding to the principal component, and λ is its associated eigenvalue.

2.4.2. Partial Least Squares (PLS)

PLS is a supervised multivariate method that models relationships between predictor and response datasets, see Tenenhaus (1998). In multi-omics, PLS is useful for integrating two or more omics layers (*e.g.*, gene expression and metabolite levels) and relating them to phenotypic outcomes, see Lê C. *et al.* (2008).

PLS finds weight vectors $\mathbf{w}_X$ and $\mathbf{w}_Y$ such that the covariance between the projections $\mathbf{X}\mathbf{w}_X$ and $\mathbf{Y}\mathbf{w}_Y$ is maximized:

$$\max_{\mathbf{w}_X, \mathbf{w}_Y} \operatorname{Cov}(\mathbf{X}\mathbf{w}_X, \mathbf{Y}\mathbf{w}_Y) \tag{9}$$

Variants like sparse PLS introduce regularization to enable feature selection.

2.4.3. Multi-Omics Factor Analysis (MOFA)

MOFA is a latent variable model specifically developed for the integration of multi-omics data. It decomposes each omics dataset into shared and data-specific factors, which correspond to biological or technical sources of variation, see Argelaguet *et al.* (2018).

Given K omics matrices $\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(K)}\}$, MOFA models each as:

$$\mathbf{X}^{(k)} = \mathbf{Z}\mathbf{W}^{(k)} + \mathbf{E}^{(k)} \tag{10}$$

where:

- $\mathbf{Z} \in \mathbb{R}^{n \times d}$ is a matrix of latent factors shared across datasets,
- $\mathbf{W}^{(k)} \in \mathbb{R}^{d \times p_k}$ are weights for dataset k,
- $\mathbf{E}^{(k)}$ is residual noise.

MOFA is probabilistic and handles missing data naturally. It enables unsupervised clustering, dimensionality reduction, and exploration of latent drivers in biological systems, see Vahabi and Michailidis (2022).

2.4.4. Sparse Multi-Block PLS (sMBPLS)

sMBPLS extends PLS to more than two data blocks and incorporates sparsity to identify the most informative features across all omics layers, see Li *et al.* (2012). It is especially suited for studies where multiple omics are related to a common response (*e.g.*, disease status or treatment outcome).

This method builds a global latent structure and optimizes for interpretability, making it useful in complex systems biology studies.

2.4.5. Gene-wise weights and feature selection

In some multivariate frameworks, gene-wise weights are assigned to different omics variables to evaluate their contribution to observed variance or phenotype association. These weights help rank and select biologically relevant features from high-dimensional data.

One example is the CNAmet model, which integrates copy number, methylation, and expression data using correlation structures and statistical weighting.

2.4.6. Advantages and limitations

Multivariate methods offer a range of advantages and face several challenges in the analysis of complex datasets. They enable joint analysis by accounting for co-variation and correlations among variables, which enhances the understanding of interdependencies in the data. These methods also facilitate dimensionality reduction, making high-dimensional omics data more tractable and interpretable. Additionally, they are powerful tools for discovering

latent factors that may represent hidden biological drivers of variation. Their flexibility allows them to be applied in both supervised and unsupervised learning contexts. However, multivariate methods can be computationally intensive, especially when applied to large-scale omics datasets, necessitating efficient algorithmic implementations. They are also prone to overfitting, particularly in scenarios with small sample sizes, which requires the use of regularization techniques. Furthermore, while these methods can uncover latent components, interpreting these components in terms of clear biological processes can be challenging.

3. AI and machine learning approaches

3.1. Variational Autoencoders (VAEs) in multi-omics data integration

Variational Autoencoders are a class of generative models that have gained popularity in multi-omics data integration due to their ability to model complex, non-linear relationships and uncover latent representations of high-dimensional biological data, see Kingma and Welling (2013). VAEs are especially well-suited for handling the noise, sparsity, and heterogeneity commonly found in multi-omics datasets, see Simidjievski *et al.* (2019).

3.1.1. Theoretical foundations of VAEs

VAEs belong to the family of probabilistic generative models and extend classical autoencoders by introducing a probabilistic framework. Instead of encoding an input x into a deterministic latent vector, VAEs encode it into a distribution over latent variable z. The goal is to learn the parameters of the generative model $p_\theta(x \mid z)$, and the inference model $q_\phi(z \mid x)$, typically with neural networks.

The VAE objective is to maximize the evidence lower bound (ELBO):

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] - D_{\mathrm{KL}} \left(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z})\right) \tag{11}$$

where:

- $\mathbb{E}[\log p_\theta(x \mid z)]$ is the reconstruction loss,
- D_{KL} is the Kullback–Leibler divergence between the approximate posterior and the prior $p(z)$, typically $\mathcal{N}(0, I)$.

This formulation ensures that the latent space z is both continuous and regularized, which enables smooth sampling and interpolation—useful for capturing underlying biological variation.

3.1.2. Application in multi-omics integration

In multi-omics studies, VAEs can be used to learn shared or modality-specific latent representations that capture the biological signal common across omics layers while accounting for layer-specific variation.

3.1.2.1. Integration strategies

a. Early integration (Full Fusion) Concatenate all omics datasets as input to a single VAE model.

b. Intermediate integration Each omics layer has a separate encoder, but a shared latent space is learned.

c. Late integration Separate VAEs are trained for each omics dataset, and their latent embeddings are later combined for downstream tasks (*e.g.*, clustering, classification).

These approaches support modularity, scalability, and flexibility in integrating omics with different feature spaces and distributions.

3.1.3. Tools

a. scVI A VAE model for single-cell RNA-seq data, modeling gene expression while correcting batch effects.

b. Multi-omics VAE Custom-built frameworks where omics-specific encoders feed into a joint decoder, enabling integrative modeling of transcriptomics, proteomics, and epigenomics, see Xin *et al.* (2024)

3.1.4. Advantages and limitations

Variational Autoencoders offer several benefits in biological research, particularly in the analysis of complex omics data. They enable dimensionality reduction by compressing high-dimensional data into low-dimensional latent factors that capture key biological variation. Their probabilistic framework enhances robustness to noise and batch effects, making them well-suited for real-world biological datasets. VAEs also handle missing data naturally by modeling the underlying data distribution, allowing for effective imputation. The latent space learned by VAEs often reveals meaningful clusters that correspond to phenotypes or disease subtypes, aiding in visualization and interpretation. Furthermore, VAEs support biomarker discovery by identifying important features that contribute to latent factors, which can be biologically interpreted. However, VAEs also come with challenges. The interpretability of latent dimensions can be limited, as they may not directly map to known biological processes. Training complexity is another issue, requiring careful tuning of the model architecture and learning parameters. Additionally, data scaling is crucial, as different omics types must be normalized to prevent bias in the latent space. Lastly, over-regularization due to the KL divergence term can overly constrain the latent space, potentially leading to underfitting and loss of important biological signals.

3.2. Graph-based learning in multi-omics data integration

Graph-based learning has emerged as a powerful strategy for integrating multi-omics data, particularly because biological systems are naturally structured as networks—whether they be gene regulatory networks, protein–protein interaction (PPI) networks, metabolic pathways, or cell–cell communication maps. Graph-based methods model the relationships

between entities (*e.g.*, genes, proteins, samples) as edges in a graph, enabling the analysis of topological structure, dependency, and contextual interactions across multiple omics layers.

In traditional machine learning, samples are often treated as independent and identically distributed. However, in multi-omics analysis, samples or features often exhibit nonlinear dependencies and interconnected behaviors that are better captured by graphs. For example: 1. Genes may co-express or be co-regulated, 2. Proteins interact physically or functionally, 3. Samples (patients) may be similar based on integrated omics profiles. Graph-based learning encodes this structure using nodes (*e.g.*, genes, proteins, samples) and edges (*e.g.*, co-expression, similarity, interaction), and applies machine learning techniques tailored for graphs, see Bengio *et al.* (2013).

3.2.1. Types of graph-based approaches

a. Similarity networks In this approach, each omics dataset is used to construct a similarity matrix between samples, which is then converted into a graph. These graphs are fused to form a unified network using methods such as Similarity Network Fusion. The final network can be analyzed using spectral clustering or community detection to identify subgroups (*e.g.*, disease subtypes).

b. Graph Neural Networks (GNNs) GNNs are deep learning models designed to operate on graph-structured data. They aggregate information from neighboring nodes and learn node embeddings that capture structural and feature information, see Kipf and Welling (2017). For multi-omics, nodes may represent genes with features from multiple omics. Edges may encode gene–gene relationships or pathway links. The GNN learns to predict phenotypes or latent node properties using neighborhood context, see Velickovic *et al.* (2017).

A common formulation in a GNN layer is:

$$\mathbf{h_v^{(l+1)}} = \sigma\left(\sum_{u\in\mathcal{N}(v)} \frac{1}{c_{vu}} \mathbf{W^{(l}h_u^{(l)}}\right) \tag{12}$$

where:

- $\mathbf{h_v^{(l)}}$ is the representation of node v at layer l,
- $\mathcal{N}(v)$ is the set of neighbors of node v,
- c_{vu} is a normalization constant,
- $\mathbf{W^{(l)}}$ is the learnable weight matrix, and
- σ is a non-linear activation function.

c. Network propagation and diffusion These algorithms propagate information (*e.g.*, expression signals, mutation scores) over a network to prioritize relevant nodes, see Köhler *et al.* (2008). Examples include:

- **Random Walk with Restart (RWR)** A random walker starts at a node and probabilistically explores the network, returning to the start with probability **r**. This helps rank nodes based on their proximity to known disease genes.

$$p_{t+1} = (1 - r)Wp_t + rp_0 \tag{13}$$

where: p_t is the probability vector at time t, W is the transition matrix, p_0 is the initial distribution.

- **NetICS, TieDIE** Used for integrating mutation data with expression or pathway data using directed propagation, see Paull *et al.* (2013).

d. Probabilistic graphic models These include Bayesian Networks and Markov Random Fields (MRFs) that model conditional dependencies among variables (genes, proteins, *etc.*). For example, PARADIGM infers pathway activities by combining multiple omics layers within a Bayesian graphical model framework.

3.2.2. Advantages and limitations

Graph-based learning has emerged as a powerful approach in multi-omics analysis due to its ability to model complex, structured biological relationships. It has been applied in various domains such as cancer subtype classification, where methods like Graph Neural Networks and Similarity Network Fusion cluster patients based on integrated omics profiles; biomarker discovery, where network diffusion identifies genes or proteins functionally related to known disease markers; pathway activity inference, with tools like PARADIGM integrating gene expression and copy number data to predict pathway status; and feature selection, where GNN attention mechanisms highlight informative nodes for downstream analysis. The advantages of graph-based methods include their natural representation of biological systems using existing knowledge like gene networks, flexibility in handling non-Euclidean and structured data, context-aware learning through neighborhood-informed node embeddings, and scalability enabled by recent computational advances. However, challenges remain, such as the need for careful data preprocessing to construct reliable graphs, limited interpretability of deep graph models, complexity in integrating heterogeneous omics layers without introducing bias or losing specificity, and the high computational demands of training large-scale graph models.

4. Conclusion

Multi-omics data integration is at the forefront of systems biology, enabling a holistic view of cellular function by combining genomic, transcriptomic, proteomic, metabolomic, and other omics data types. Each method explored—statistical, machine learning, and network-based—offers unique strengths in addressing the challenges of high-dimensionality, heterogeneity, and noise inherent in biological data. Statistical approaches, particularly Canonical Correlation Analysis and its variants (sparse and regularized CCA), provide interpretable linear models for discovering cross-domain correlations between omics layers. These models are well-suited for moderate-dimensional data and are often used as a first step in integrative analysis. Similarity-based methods, such as Similarity Network Fusion, excel in clustering

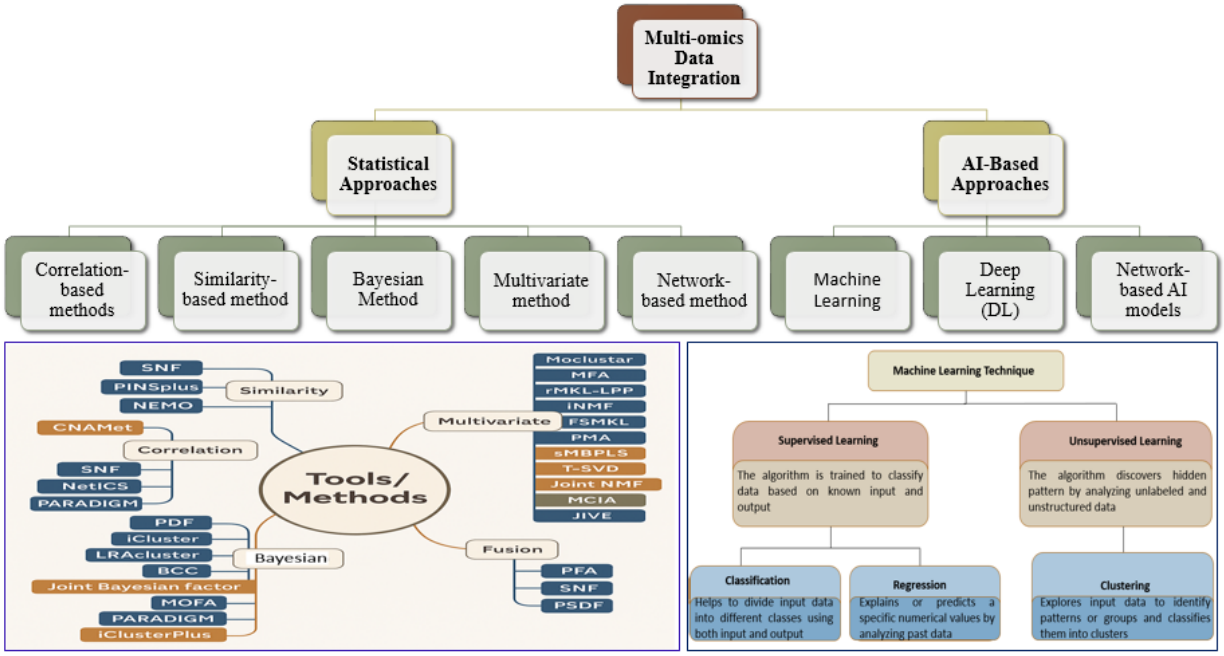


Figure 1: Workflow diagram of AI and statistical methods of multi-omics data integration

Table 1: Multi-omics public datasets and compatible methods

Dataset / Resource	Multi-omics layers	Compatible methods
TCGA (*via* GDC portal)	mRNA, miRNA, methylation, CNV, proteomics	PCA, PLS, SNF, BCC, PARADIGM, MOFA, *etc.*
ICGC	Genomics, transcriptomics, epigenomics	Same as TCGA, broader diversity
CMOB benchmark (TCGA-based)	Processed multi-cancer data	All listed ML/stat methods
MixOmics example sets	mRNA, proteome, metabolome	PCA, PLS, sMBPLS, CCA
BioGRID interactions + TCGA	Network /expression/proteomics	GNN, RWR

and patient stratification by leveraging sample-level relationships across different datasets. These methods are robust to missing features and offer flexible data-type integration through graph-based fusion strategies. Bayesian models introduce a probabilistic framework that explicitly handles uncertainty and allows for the incorporation of prior biological knowledge. They are particularly effective in unsupervised clustering, causal inference, and modeling hidden structures in multi-omics data, though often computationally demanding. Multivariate methods, including PCA, PLS, MOFA, and sMBPLS, help in reducing dimensionality and uncovering latent variables that drive shared or specific biological variation across omics layers. These techniques are scalable and interpretable, making them widely adopted in both research and clinical settings.

Variational Autoencoders represent a more recent advancement, leveraging deep learning to capture complex, non-linear patterns and generate latent representations. Their flexibility in integration strategies (early, intermediate, late) and natural handling of missing data make them highly promising for large, noisy, and heterogeneous datasets. Graph-based learning, including Graph Neural Networks and network propagation methods, allows integration of biological interaction networks with omics data. These methods encode structural dependencies, enhance biological interpretability, and enable feature prioritization based on contextual relevance within the network. However, no single method is universally superior; instead, the choice depends on the specific research question, data type, sample size, and computational resources. As computational methods advance and multi-omics datasets expand, integrative approaches will continue to unlock new insights into complex diseases, biological pathways, and precision medicine.

Acknowledgements

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.

Table 2: Multi-omics data integration methods

Method	Function	Advantages	Limitations
Canonical Correlation Analysis (CCA)	Finds linear combinations of features in two datasets that are maximally correlated.	Simple and interpretable; suitable for moderate-dimensional data.	Assumes linearity; unstable when number of variables exceeds samples; sensitive to noise.
Sparse/Regularized CCA	Extends CCA with sparsity (L1) or regularization to improve feature selection or stability.	Feature selection; better suited for high-dimensional omics data.	Parameter tuning required; interpretability can decrease with complexity.
Similarity Network Fusion (SNF)	Constructs sample-sample similarity networks from each omics and fuses them iteratively.	Handles heterogeneous data; robust to missing features; good for clustering.	Sensitive to similarity metric choice; requires careful normalization and parameter tuning.
Bayesian Clustering (MDI, BCC)	Uses probabilistic models to assign samples to latent clusters across datasets.	Models uncertainty; incorporates prior knowledge; captures hidden structure.	Computationally intensive; may require strong assumptions or priors.
Bayesian Networks (*e.g.*, PARADIGM)	Models conditional dependencies among omics variables *via* DAGs.	Captures causal relationships; integrates multiple data types with biological priors.	Complex to construct; inference can be slow and sensitive to data quality.
Principal Component Analysis (PCA)	Reduces dimensionality by capturing directions of maximum variance.	Simple, fast, and unsupervised; good for visualization and variance exploration.	Assumes linearity; may overlook class-specific patterns; not tailored to response variables.
Partial Least Squares (PLS)	Projects data onto latent variables that correlate with outcomes.	Supervised; identifies correlated features across data types.	May overfit with small sample sizes; assumes linear relationships.
Multi-Omics Factor Analysis (MOFA)	Learns shared and specific latent factors across omics layers.	Probabilistic; handles missing data; interpretable latent structure.	Assumes Gaussian distributions; requires tuning of latent dimensionality.
sMBPLS (Sparse Multi-Block PLS)	Integrates multiple omics datasets with sparsity constraints.	Simultaneous integration and feature selection; interpretable loadings.	Computationally demanding; sensitive to sparsity level selection.
Variational Autoencoders	Learns probabilistic latent representations; used for denoising, imputation, clustering.	Captures nonlinear patterns; handles missing data; flexible integration strategies.	Requires deep learning expertise; difficult to interpret latent variables biologically.
Graph Neural Networks (GNNs)	Learns on graph-structured data to capture node-level and graph-level representations.	Exploits interaction networks; context-aware; scalable with recent advances.	Graph construction can be noisy; hard to interpret; requires large labeled datasets.
Network Propagation (*e.g.*, RWR)	Spreads signals across biological networks to prioritize genes or features.	Integrates prior knowledge; useful for ranking and feature prioritization	Performance depends on quality of network; propagation may dilute weak but important signals.

Table 3: Comparison of multi-omics integration methods

Method	Description	Software / Package	Platform	Link / Notes
CCA (Canonical Correlation Analysis)	Identifies linear relationships between two data matrices	mixOmics::rcc / PMA::CCA	R	mixOmics, PMA
SNF (Similarity Network Fusion)	Constructs sample similarity networks and fuses them	SNFtool / SNFpy	R / Python	SNFtool (R)
BCC (Bayesian Consensus Clustering)	Unsupervised clustering across multiple data types	BayesCC	R	GitHub - BayesCC
PARADIGM	Integrates multi-omics using pathway information	Java tool, also in UCSC Cancer Genomics Browser	Java / Web	PARADIGM GitHub, UCSC site
PCA (Principal Component Analysis)	Linear dimensionality reduction	Base R/prcomp, sklearn .decomposition. PCA	R / Python	scikit-learn PCA
PLS (Partial Least Squares)	Projects predictor and response variables to a new space	mixOmics::pls / sklearn.cross $_{d}ecomposition.PLSRegression$	R / Python	mixOmics, scikit-learn PLS
MOFA (Multi-Omics Factor Analysis)	Probabilistic latent variable model for multiple omics	MOFA2	R / Python	MOFA2 GitHub, Documentation
sMBPLS (Sparse Multi-block PLS)	PLS extension for multi-block data, sparse variant	mixOmics::block .spls	R	mixOmics - block.spls
VAE (Variational Autoencoder)	Deep learning model to learn latent representations	TensorFlow, PyTorch, scVI	Python	scVI, PyTorch VAE example
GNN (Graph Neural Networks)	Deep models on graph-structured omics data	PyTorch Geometric, DGL, Spektral	Python	PyTorch Geometric, DGL, Spektral
RWR (Random Walk with Restart)	Graph-based propagation for gene prioritization	Custom or igraph, NetWalker	R / Python / Java	NetWalker, igraph

References

Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., Pochanard, P., Mozes, E., Garraway, L. A., and Pe'er, D. (2010). An integrated approach to uncover drivers of cancer. *Cell*, **143**, 1005–1017.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, **14**, e8124.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 1798–1828.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, **374**, 531–547.

Kingma, D. P. and Welling, M. (2013). *Auto-Encoding Variational Bayes*, **1312**, Banff, Canada.

Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z., and Wild, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, **28**, 3290–3297.

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, **82**, 949–958.

Lê C., Kim, A., Rossouw, D., Robert, G., and Besse, P. (2008). A sparse pls for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, **7**, p.35.

Li, W., Zhang, S., Liu, C.-C., and Zhou, X. J. (2012). Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, **28**, 2458–2466.

Misra, B. B. (2018). New tools and resources in metabolomics: 2016–2017. *Electrophoresis*, **39**, 909–923.

Naserkheil, M., Ghafouri, F., Zakizadeh, S., Pirany, N., Manzari, Z., Ghorbani, S., Banabazi, M. H., Bakhtiarizadeh, M. R., Huq, M. A., and Park, M. N. (2022). Multi-omics integration and network analysis reveal potential hub genes and genetic mechanisms regulating bovine mastitis. *Current Issues in Molecular Biology*, **44**, 309–328.

Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, **8**, 1–34.

Paull, E. O., Carlin, D. E., Niepel, M., Sorger, P. K., Haussler, D., and Stuart, J. M. (2013). Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (tiedie). *Bioinformatics*, **29**, 2757–2764.

Rohart, F., Gautier, B., Singh, A., and Lê Cao, K.-A. (2017). mixomics: An r package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, **13**, e1005752.

Simidjievski, N., Bodnar, C., Tariq, I., Scherer, P., Andres Terre, H., Shams, Z., Jamnik, M., and Liò, P. (2019). Variational autoencoders for cancer data integration: design principles and computational practice. *Frontiers in Genetics*, **10**, 1205.

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights*, **14**, 1177932219899051.

Tenenhaus, M. (1998). *La régression PLS: théorie et pratique*. Editions technip.

Vahabi, N. and Michailidis, G. (2022). Unsupervised multi-omics data integration methods: a comprehensive review. *Frontiers in Genetics*, **13**, 854752.

Vaske, C. J., Benz, S. C., Stuart, J. M., and Haussler, D. (2021). Pathway recognition algorithm using data integration on genomic models (paradigm). **10**. US Patent 991,448.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. *Stat*, **1050**, 10–48550.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, **11**, 333–337.

Witten, D. and Tibshirani, R. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, **8**, 1–27.

Wróbel, S., Turek, C., Stepień, E., and Piwowar, M. (2024). Data integration through canonical correlation analysis and its application to omics research. *Journal of Biomedical Informatics*, **151**, 104575.

Xin, L., Huang, C., Li, H., Huang, S., Feng, Y., Kong, Z., Liu, Z., Li, S., Yu, C., and Shen, F. (2024). Artificial intelligence for central dogma-centric multi-omics: Challenges and breakthroughs. *arXiv preprint arXiv:2412.12668*, **.**

Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025; pp 153-166

An Overview of Bayesian Semiparametric Approaches for Genetic Association Studies

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: 01 July 2025; Revised: 16 August 2025; Accepted: 19 August 2025

Abstract

In human genetics, Bayesian semiparametric approaches have proven especially effective in disease gene association studies, where genetic heterogeneity and complex interactions are common. They are particularly advantageous in stratified subpopulation settings with an unknown number of subgroups. Unlike traditional parametric models that require prespecifying the number of subpopulations, nonparametric methods such as Dirichlet Process mixture models allow the number and structure of subpopulations to be learned from the data. This flexibility enables more accurate detection of disease-associated variants while accounting for population structure, which are key challenges in complex trait analysis and precision medicine. This work provides an overview of how Dirichlet Process based mixture models can be used to flexibly model gene-gene and gene-environment interactions and identify disease-associated variants in complex, stratified populations with unknown heterogeneity.

Key words: Dirichlet process; Genetic association studies; Mixture model; Parallel computing; TMCMC.

AMS Subject Classifications: 62K05, 05B05.

1. Introduction

1.1. Gene-gene and gene-environment interaction

With recent technological advances, it is now possible to assay millions of loci in an individual's genomic DNA to identify disease-associated genes. While this capability has revolutionized genetic research, it has also introduced substantial analytical challenges, particularly in managing the massive volume of data generated. Addressing these challenges requires the development of sophisticated statistical models that integrate current biological and biochemical knowledge of disease mechanisms. Such models not only facilitate efficient

Corresponding Author: Durba Bhattacharya
Email: durba@sxccal.edu

computation but also enable deeper insights into the complex pathways underlying multifactorial diseases.

Genome-wide association studies (GWAS) have identified numerous single nucleotide polymorphisms (SNPs) associated with complex diseases, yet they explain only a small fraction of heritable genetic variation; see Larson and Schaid (2013). A growing body of research indicates that genes often function through intricate interaction networks, which significantly shape the genetic basis of complex traits Bonetta (2010). The limited explanatory power of GWAS may stem from the absence of models that incorporate gene-gene interactions into genomic analysis Cordell (2009), thereby overlooking important biological mechanisms Yi (2010).

A major obstacle in studying genetic interactions lies in the lack of a clear definition of epistasis. Phillips (2008) distinguishes between functional and compositional biological epistasis, both of which differ from the classical statistical definition proposed by Fisher (1918) and extended by Kempthorne (1954). While VanderWeele (2009) identifies conditions for alignment between statistical and biological definitions, most statistical tests fail to reflect the biological complexity of interactions. Still, statistical models are essential for quantifying these effects Cordell (2002), Wang *et al.* (2010).

SNP-SNP interactions are often used to model gene-gene interactions in case-control studies Yi *et al.* (2011). However, SNP-level models are computationally intensive due to the large number of interaction terms required, whereas gene-level models offer dimensionality reduction at the expense of finer detail Larson and Schaid (2013), Musameh *et al.* (2015). Moreover, additive linear models can oversimplify interaction mechanisms and obscure interpretability, especially when principal components are used for reduction Wang *et al.* (2010).

These challenges are compounded by the frequent neglect of population substructure. Genetic effects can vary across subpopulations, and ignoring such heterogeneity can lead to biased inference and inflated false positives Bhattacharjee *et al.* (2010). Since the number and structure of subgroups are usually unknown, flexible models that can infer latent structure are critical.

Beyond genetic interactions, the interplay between genes and environmental factors is critical to understanding disease etiology. Although most diseases arise from a combination of genetic and environmental influences, only a small subset are purely monogenic. Environmental exposures can alter genetic risk Mapp (2003), Khouri (2005), and in certain cases, disease manifestation occurs only beyond specific environmental thresholds. Hunter (2005) emphasize that neglecting such interactions can lead to misestimation of the population-level disease burden. These interactions are particularly salient in pharmacogenetics, where treatment efficacy and safety may vary by genotype Scott (2011). Mechanistically, gene–environment interactions may act through pathways such as epigenetic modification and transcriptional regulation Purcell (2002), Ottman (2010). However, existing statistical approaches, particularly linear and log-linear models, often fail to adequately capture these complex dependencies Mukherjee *et al.* (2008), Mukherjee and Chatterjee (2008), Mukherjee *et al.* (2010), Mukherjee *et al.* (2012), Sanchez *et al.* (2012), Ahn *et al.* (2013), Ko *et al.* (2013).

These limitations point to the need for more general, data-adaptive approaches.

Bayesian nonparametric methods based on Dirichlet Process mixture models offer the flexibility to address gene-gene and gene-environment interactions while accounting for population stratification. This article surveys recent developments in this direction, building on the framework proposed in Bhattacharya and Bhattacharya (2018) and Bhattacharya and Bhattacharya (2024).

2. An overview of our Bayesian nonparametric ideas

This paper presents a Bayesian nonparametric/semiparametric framework for analyzing gene–gene interactions, fundamentally differing from traditional logistic regression approaches. Rather than modeling disease status conditional on genotype, we model genotype distributions conditional on disease status. To account for hidden population substructure, Dirichlet process-based finite mixtures (Bhattacharya, 2008) are embedded within a hierarchical model that captures interactions at both gene and SNP levels via matrix-normal priors. The framework extends naturally to gene–environment interactions through covariate-dependent priors, enabling the assessment of how environmental factors influence genetic associations.

Our Bayesian approach addresses multiple sources of uncertainty and moves beyond binary presence–absence tests by modeling the magnitude and structure of interaction effects using correlation-based measures. Disease-predisposing loci (DPLs) are detected through novel posterior-clustering-based hypothesis testing. For computational efficiency in high-dimensional settings, Transformation-based Markov Chain Monte Carlo (TMCMC) (Dutta and Bhattacharya, 2014) is employed, facilitating block updates with high acceptance rates. Combined with parallel Gibbs sampling tailored for Dirichlet process mixtures, the method achieves substantial computational gains.

We validate the methodology through simulations and apply it to a myocardial infarction case–control SNP dataset. The results corroborate known associations and reveal novel gene–gene and gene–environment interactions, illustrating the flexibility and inferential power of the proposed framework.

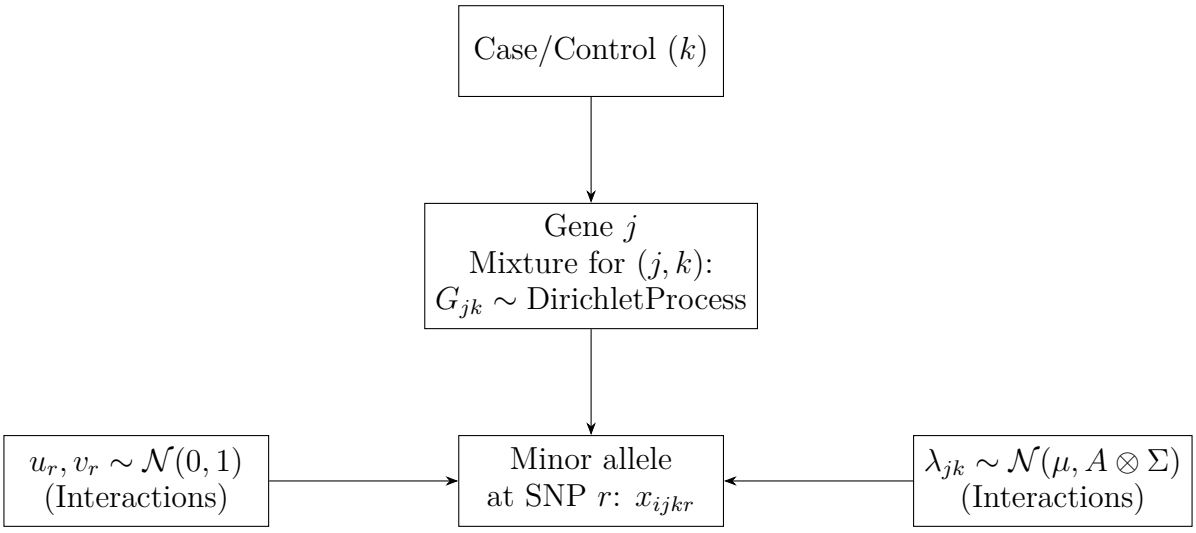


Figure 1: Schematic representation of the Bayesian model for gene–gene interactions

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 transformation-based MCMC (TMCMC).

2.1. Case-control type genotype data

Humans have 22 pairs of autosomes and one pair of sex chromosomes in the nuclear genome. Chromosomes are composed of tightly coiled DNA, containing genes that may exist in alternative forms, known as *alleles*, at the same genetic locus. Variation in alleles can lead to phenotypic differences, and the specific allelic combination at a locus defines an individual's *genotype*. The most common genetic variation is the Single Nucleotide Polymorphism (SNP), a single base change in the DNA sequence. This study analyses SNP data from case and control cohorts in relation to a specific disease.

Let $s = 1, 2$ represent the two chromosomes. For an individual indexed by i, gene j, group k, and locus r, define $x^s_{ijkr} = 1$ if the minor allele is present, and $x^s_{ijkr} = 0$ otherwise. The indices range as follows: $i = 1, \ldots, N_k$; $j = 1, \ldots, J$; $k = 0, 1$, where $k = 1$ corresponds to the case group; and $r = 1, \ldots, L_j$. Given any (j, k), let $\boldsymbol{x}_{ijkr} = (x^1_{ijkr}, x^2_{ijkr})$, and $\boldsymbol{X}_{ijk} = (\boldsymbol{x}_{ijk1}, \boldsymbol{x}_{ijk2}, \ldots, \boldsymbol{x}_{ijkL_j})$.

2.2. Gene-gene interaction based mixture models driven by Dirichlet processes

We assume that for every triplet (i, j, k), $\boldsymbol{X}_{ijk}$ are independently distributed with mixture probability mass function with a *maximum* of M components, given by

$$[\boldsymbol{X}_{ijk}] = \sum_{m=1}^{M} \pi_{mjk} \prod_{r=1}^{L_j} f\left(\boldsymbol{x}_{ijkr} | p_{mjkr}\right), \tag{1}$$

where $f\left(\cdot | p_{mjkr}\right)$ is the probability mass function of independent Bernoulli distributions, given by

$$f\left(\boldsymbol{x}_{ijkr} | p_{mjkr}\right) = \left\{p_{mjkr}\right\}^{x^1_{ijkr}+x^2_{ijkr}} \left\{1 - p_{mjkr}\right\}^{2-(x^1_{ijkr}+x^2_{ijkr})}. \tag{2}$$

Using allocation variables z_{ijk}, with probability distribution

$$[z_{ijk} = m] = \pi_{mjk}, \tag{3}$$

for $i = 1, \ldots, N_k$ and $m = 1, \ldots, M$, (1) can be represented as

$$[\boldsymbol{X}_{ijk} | z_{ijk}] = \prod_{r=1}^{L_j} f\left(\boldsymbol{x}_{ijkr} | p_{z_{ijk}jkr}\right). \tag{4}$$

We may assume appropriate Dirichlet distribution priors on $(\pi_{1jk}, \ldots, \pi_{Mjk})$ for $j = 1, \ldots, J$; $k = 0, 1$. Following Mukhopadhyay and Bhattacharya (2021), we set $\pi_{mjk} = 1/M$, for $m = 1, \ldots, M$, and for all (j, k).

Letting $\boldsymbol{p}_{mjk} = \left(p_{mjk1}, p_{mjk2}, \ldots, p_{mjkL_j}\right)$, we further assume that

$$\boldsymbol{p}_{1jk}, \boldsymbol{p}_{2jk}, \ldots, \boldsymbol{p}_{Mjk} \overset{iid}{\sim} \boldsymbol{G}_{jk}; \tag{5}$$

$$\boldsymbol{G}_{jk} \sim \mathrm{DP}\left(\alpha_{jk}\boldsymbol{G}_{0,jk}\right), \tag{6}$$

where $\mathrm{DP}\left(\alpha_{jk}\boldsymbol{G}_{0,jk}\right)$ stands for Dirichlet process with expected probability measure $\boldsymbol{G}_{0,jk}$ having precision parameter α_{jk}. We assume that under $\boldsymbol{G}_{0,jk}$, for $m = 1, \ldots, M$ and $r = 1, \ldots, L_j$,

$$p_{mjkr} \overset{iid}{\sim} \mathrm{Beta}\left(\nu_{1jkr}, \nu_{2jkr}\right). \tag{7}$$

Discreteness of Dirichlet processes causes coincidences among the parameter vectors of $\boldsymbol{P}_{Mjk} = \{\boldsymbol{p}_{1jk}, \boldsymbol{p}_{2jk}, \ldots, \boldsymbol{p}_{Mjk}\}$ with positive probability, so that, with positive probability, the actual number of mixture components in (1) falls below M, the maximum number of components, the mixing probabilities taking the form M^*/M, where $1 \leq M^* \leq M$. The property of coincidences among the parameter vectors is clearly preserved by the Polya urn scheme. Notationally, we shall denote the number of distinct elements of $\boldsymbol{P}_{Mjk} = \{\boldsymbol{p}_{1jk}, \boldsymbol{p}_{2jk}, \ldots, \boldsymbol{p}_{Mjk}\}$ by τ_{jk} and that of $\boldsymbol{P}_{Mjk} \backslash \{\boldsymbol{p}_{mjk}\}$ by $\tau_{jk}^{(m)}$.

Conditioned on $\boldsymbol{G}_{jk}$, our fixed-M mixture model mimics an infinite-dimensional Dirichlet process mixture despite the non-iid nature of the data (Mukhopadhyay and Bhattacharya (2021)). The number of distinct components in $\boldsymbol{P}_{Mjk}$ can vary across (j, k) due to random duplication. This flexibility aligns with biological expectations, as genotype distributions often differ between cases and controls under genetic influence. Such heterogeneity is naturally accommodated within our framework.

2.3. Gene-gene, SNP-SNP interactions and parallel processing

To incorporate the SNP-SNP dependence, which may exist within each gene and also among the genes, The Beta parameters are modelled as ν_{1jkr} and ν_{2jkr} of (7) as follows:

For $r = 1, \ldots, L$, where $L = \max\{L_j;\ j = 1, \ldots, J\}$, and for every (j, k),

$$\nu_{1jkr} = \exp\left(u_r + \lambda_{jk}\right); \tag{8}$$

$$\nu_{2jkr} = \exp\left(v_r + \lambda_{jk}\right). \tag{9}$$

We further assume that for $r = 1, \ldots, L$,

$$u_r \overset{iid}{\sim} N(0, 1); \tag{10}$$

$$v_r \overset{iid}{\sim} N(0, 1). \tag{11}$$

The Gaussian priors on u_r and v_r with other means and variances did not yield significantly different results, establishing the prior robustness in our modeling strategy.

Subsequently, the SNP-wise dependence in a gene is modelled using matrix-normal distribution

$$\boldsymbol{\lambda} = \{\lambda_{jk};\ j = 1, \ldots, J,\ k = 0, 1\} \sim N(\boldsymbol{\mu}, \boldsymbol{A} \otimes \boldsymbol{\Sigma}),$$

as a prior for $\boldsymbol{\lambda}$ ($\boldsymbol{\Lambda}$ in matrix form) with appropriate inverse-Wishart priors on $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$. Furthermore, the matrix-normal prior induces dependence among genes, which in turn creates dependencies among the SNPs belonging to different genes.

Given that the mixture distributions for each gene $j \in 1, \ldots, J$ and case-control group $k \in 0, 1$ are conditionally independent when the interaction parameters are known, we take advantage of this structure for efficient computation. Mixture components are updated simultaneously across multiple processors, while the interaction parameters, which govern the dependencies, are updated afterward on a single processor using a specialized TMCMC approach.

This separation in the update steps enables the method to handle large-scale data effectively while preserving the ability to capture complex gene-gene and SNP-SNP relationships.

2.4. Summary of analysis of the MI dataset

In our analysis of the real Myocardial Infarction (MI) dataset, we focused on a total of 1251 SNPs, out of which only 33 had prior evidence suggesting a possible link to the disease. The remaining 1218 SNPs had no documented association with MI and were largely considered unlikely candidates for influencing disease risk. In fact, apart from a few among the 33 literature-supported SNPs, most of the others were included not because of prior biological relevance, but to test the robustness of our model in distinguishing meaningful signals from noise. Interestingly, in several instances, the disease-predisposing loci (DPL) identified by our Bayesian approach matched those already highlighted in the literature as relevant to MI. Notable examples include SNP $rs7395662$ in gene $OR4A48P$, SNP $rs964184$ in $AP006216.10$, SNP $rs4420638$ in $APOC1$, SNP $rs1564348$ in $SLC22A1$, and SNP $rs1013442$ in $BDNF\text{-}AS$. This alignment underscores the model's ability to successfully detect true associations, thereby effectively controlling false negatives. Conversely, SNPs not identified as DPLs either by our approach or by prior studies can be reasonably regarded as unrelated to the disease, indicating that the model also maintains strong control over false positives.

3. Extension to gene-environment interactions

Our Bayesian hierarchical mixture framework integrates the mechanisms by which gene–environment interactions, as well as the isolated and joint effects of genes, contribute to disease susceptibility, while accommodating potential population stratification. A distinctive feature of the model is its ability to infer the number of latent genetic subpopulations.

To capture the influence of environmental variables, the proposed semiparametric specification employs Dirichlet process-based finite mixtures at the individual level, jointly modeling genetic profiles and case–control status. These mixtures are linked through a structured dependence encoded via hierarchical matrix-normal distributions, enabling the model to account for correlations induced by environmental exposures. The framework extends the gene–gene interaction model and Bayesian hypothesis testing methodology developed in Section 2 to detect the effects of genes, environmental factors, and their interactions.

Computation is performed via a parallel MCMC scheme that leverages the model's conditional independence structure, combining Gibbs sampling with Transformation-based MCMC (TMCMC) for efficient high-dimensional updates. Environmental covariates influence individual-level Dirichlet process mixtures, allowing for subject-specific modulation

of genotype distributions. The prior hierarchy accommodates locus-specific, gene-specific, and environment-dependent parameters. Parallel updates are applied to gene–environment-specific components, while interaction parameters are updated centrally using TMCMC. Figure 2 presents a schematic representation of the proposed framework.

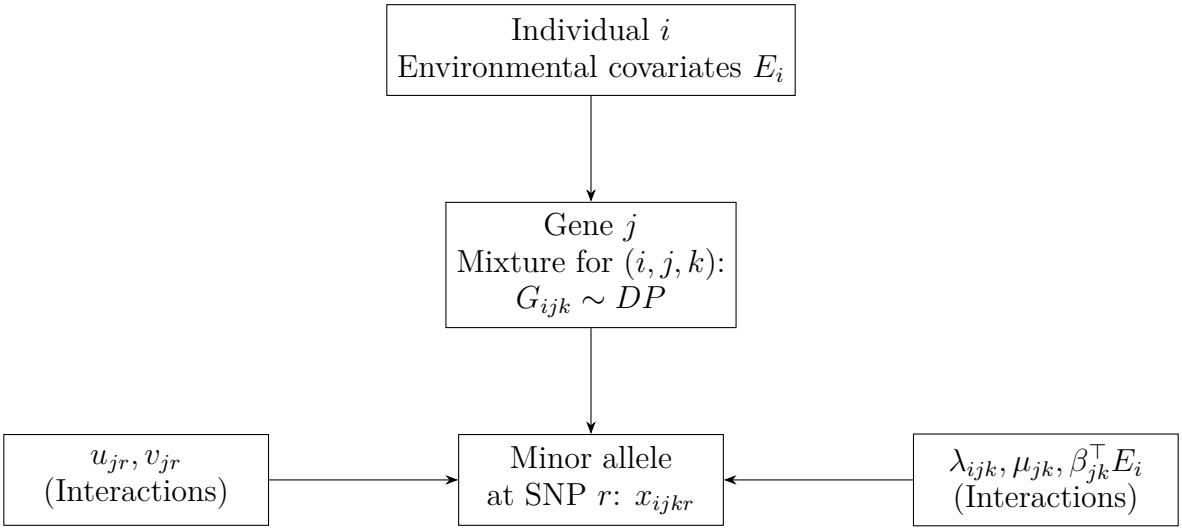


Figure 2: Diagram of the extended Bayesian framework incorporating gene-environment interactions

3.1. Modeling genotypic sub-populations with mixture models driven by Dirichlet processes

Let E_i denote the set of environmental variables associated with the i-th individual. We model the case-control genotype data, together with environmental information, using our Bayesian semiparametric model.

Let $x_{ijkr} = (x^1_{ijkr}, x^2_{ijkr})$ denote the genotype at the r-th locus of the j-th gene for the i-th individual in the k-th group (case/control), and let $\mathbf{X}_{ijk} = (x_{ijk1}, x_{ijk2}, \ldots, x_{ijkL_j})$ denote the genotype information across all L_j loci of the j-th gene. Let p_{mijkr} denote the minor allele frequency at the r-th locus of the j-th gene for the i-th individual in the k-th group. The minor allele frequency represents the frequency at which the second most common allele occurs in a given population.

We assume the mixture distribution:

$$[\mathbf{X}_{ijk}] = \sum_{m=1}^{M} \pi_{mijk} \prod_{r=1}^{L_j} f(x_{ijkr} \mid p_{mijkr}), \tag{12}$$

where $f(\cdot \mid p_{mijkr})$ denotes the Bernoulli mass function:

$$f(x_{ijkr} \mid p_{mijkr}) = p_{mijkr}^{x^1_{ijkr}+x^2_{ijkr}} (1 - p_{mijkr})^{2-(x^1_{ijkr}+x^2_{ijkr})}, \tag{13}$$

and M is the maximum number of mixture components. The allocation variables z_{ijk} are such that:

$$[z_{ijk} = m] = \pi_{mijk}, \quad m = 1, \ldots, M. \tag{14}$$

We set $\pi_{mijk} = 1/M$ for all (i, j, k) and m, as this fixed weight approach has been shown to yield better performance than Dirichlet priors in learning about the true number of components.

This representation captures the possibility that different individuals, even within the same group and gene, may belong to different sub-populations, influenced by their environmental exposures E_i. This is a substantial extension from the model in Section 3, which did not account for environmental effects.

3.2. Summary of the results of MI data analysis with this new model

We applied the proposed model to the myocardial infarction (MI) dataset previously analyzed in Section 2.4, incorporating sex as an environmental covariate. The resulting inferences were consistent with established findings in the literature. Although gene–gene interactions were not statistically significant, SNP–SNP correlations, quantified via Euclidean distances between case and control groups, provided plausible explanations for discrepancies between our identified disease-predisposing loci (DPLs) and those reported in earlier studies.

Importantly, the Bayesian framework produced interpretable results despite the limited sample size of 200 individuals, underscoring the utility of hierarchical modeling with informative priors and the efficiency of the employed MCMC algorithms.

4. A general model for gene-gene and gene-environment interactions based on hierarchies of Dirichlet processes

As discussed in Section 3, gene–gene interactions alone are insufficient to explain the etiology of most complex diseases. Similarly, examining environmental factors in isolation from genetic variation is inadequate; biomedical evidence underscores the pivotal role of gene–environment interactions in elucidating complex disease mechanisms. Given the absence of a simple, additive relationship between genetic and environmental influences, linear or additive models commonly used to date are inadequate for modeling these interactions.

In Section 3, we introduced a Bayesian semiparametric model for case–control genotype data, employing Dirichlet process-based finite mixtures at the subject level. A hierarchical matrix-normal dependence structure linked these mixtures to capture correlations among genes under environmental influence. However, a potential limitation of this framework arises from its induced covariance structure: for individual i, the relevant gene–gene covariance matrix is $\tilde{\sigma}_{ii}A$, where A is a common gene–gene interaction matrix (in the absence of environmental variables) and $\tilde{\sigma}_{ii} = \sigma_{ii} + \phi$, with σ_{ii} denoting the i-th diagonal element of a symmetric positive-definite matrix unrelated to environmental variables, and $\phi \geq 0$ representing the effect of the environmental covariate E. This formulation assumes that environmental exposures modify gene–gene interactions in an identical manner across all individuals, which may be unrealistic when the magnitude and nature of exposure vary.

To address this limitation, we propose a Bayesian nonparametric framework for modeling joint gene–gene and gene–environment interactions, as developed in Bhattacharya (2019) (see also Bhattacharya and Bhattacharya (2024)). Like the earlier model, individual genotype distributions are represented via Dirichlet process-based finite mixtures; however, in place of the matrix-normal dependence structure, we introduce a hierarchy of Dirich-

let processes that flexibly captures nonparametric dependencies among genes induced by environmental covariates, case–control status, and inter-individual heterogeneity. This hierarchical construction overcomes the restrictive assumptions of the matrix-normal approach described in Section 3. Although conceptually related to the hierarchical Dirichlet process (HDP) of Teh *et al.* (2006), our model introduces an additional level of hierarchy, enhancing flexibility.

For computation, we develop a highly parallelizable MCMC algorithm that integrates modern parallel computing resources with Gibbs sampling, retrospective sampling, and Transformation-based MCMC (TMCMC). The Bayesian hypothesis testing procedures from our earlier framework are extended to this enriched setting.

Letting $s = 1, 2$ denote the two chromosomes, we define $y^s_{ijkr} = 1$ and 0 to indicate the presence and absence, respectively, of the minor allele for the i-th individual in group $k \in \{0, 1\}$ (with $k = 1$ denoting the case group), at the r-th locus of the j-th gene, for $i = 1, \ldots, N_k$; $r = 1, \ldots, L_j$; and $j = 1, \ldots, J$. Let $N = N_0 + N_1$, and let E_i denote a vector of environmental variables associated with individual i.

Again, before describing the components of the model in detail, we first present the schematic diagram in Figure 3.

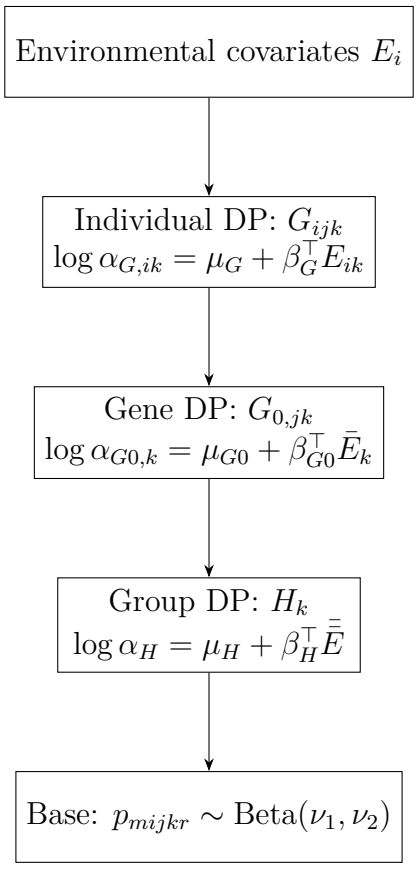


Figure 3: Schematic representation of the hierarchical Dirichlet process (HDP) model for gene-gene and gene-environment interactions

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.

4.1. Summary of the MI data analysis with the HDP model

Our analysis of the MI dataset revealed a strong effect of the sex variable, consistent with the findings in Section 3. Our hypothesis tests indicated no significant marginal effects of individual genes, in agreement with Section 3 where only weak marginal effects were observed.

Most notably, even though gene-gene correlations were generally weak, again consistent with Section 3 and Lucas *et al.* (2012), our tests detected that two genes, *AP006216.10* and *C6orf106*, exhibited broad, beneficial interactions with other genes that may help combat the disease. Furthermore, in the only subgroup for which all gene-gene interactions were found to be insignificant was the male cases. Hence, our results lend statistical support to the widely held belief that males may be more susceptible to heart attacks than females.

4.2. Summary and future directions

This work presents a unified Bayesian nonparametric framework for analyzing gene–gene and gene–environment interactions in case–control studies. The proposed approach is designed to accommodate multiple layers of uncertainty, a feature that distinguishes it from many existing methods that prioritize computational feasibility for large-scale datasets. Such differences in objectives necessarily lead to different performance criteria, making direct comparisons with standard approaches inappropriate. Both the simulated and real datasets analyzed here exhibit multiple subpopulations. While methods such as principal component analysis can infer subpopulation structure, most approaches require the number of subpopulations to be fixed a priori, which can lead to misestimation and inflated false positives Bhattacharjee *et al.* (2010). Since genetic interactions may differ across subpopulations, such errors can bias inference. Our method explicitly models this uncertainty, in contrast to De Iorio *et al.* (2015b) and De Iorio *et al.* (2015a), which do not address gene–gene or gene–environment interactions.

Existing approaches generally test only for the presence of interactions without quantifying their strength, whereas our framework enables classification of genes by the magnitude of their interactions. Many standard methods rely on heuristic definitions of main and interaction effects, for example, kernel-based methods Larson and Schaid (2013), Kullback–Leibler divergence Wan *et al.* (2010), entropy-based information gain Li *et al.* (2015), or genotype categories Yi *et al.* (2011), which can yield results sensitive to the chosen definition. In contrast, our framework models interactions using established statistical principles. Furthermore, most current models analyze pairwise SNP–SNP interactions via logistic regression, neglecting genes as functional units and lacking scalability to higher-order interactions. Two-stage approaches such as `BOOST` and Bayesian methods like BEAM or EpiBN operate only at the SNP level and overlook gene-level modeling Niel *et al.* (2015). Our

unified Bayesian approach simultaneously models uncertainties in both gene- and SNP-level interactions within a coherent probabilistic structure. Finally, the simulation datasets used to validate our method were generated under logistic models, which form the basis for most competing approaches. Given that our framework is nonparametric and fundamentally distinct from logistic regression, such simulation settings do not allow for direct performance comparisons.

The proposed methodology addresses several key challenges in genetic association studies, including population stratification, uncertainty in subgroup structure, and the joint modeling of genetic effects at both the SNP and gene levels. The model incorporates complex dependency structures through hierarchical Dirichlet process mixtures, and Bayesian hypothesis testing procedures are introduced to assess interaction significance and identify disease-predisposing loci. Computationally, the framework is highly scalable, leveraging parallelization, Gibbs sampling, and Transformation-based MCMC to efficiently analyze high-dimensional genomic data. Simulation studies and application to a myocardial infarction dataset demonstrated the accuracy, robustness, and interpretability of the approach, yielding results consistent with established findings while also uncovering novel patterns, including sex-specific susceptibility.

The flexibility of the proposed model allows for natural extensions to incorporate additional biological complexities, such as dynamic environmental effects or longitudinal data. Future work may extend the framework to handle time-to-event outcomes and integrate multi-omics data. Overall, this study demonstrates how Dirichlet process-based Bayesian nonparametric methods can advance the analysis of complex diseases by providing a principled, flexible, and computationally efficient alternative to traditional GWAS analyses, thereby contributing to a deeper understanding of the genetic architecture underlying complex traits.

Conflict of interest

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

References

Ahn, J., Mukherjee, B., Ghosh, M., and Gruber, S. B. (2013). Bayesian semiparametric analysis of two-phase studies of gene-environment interaction. *The Annals of Applied Statistics*, **7**, 543–569.

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. (2019). *Bayesian Nonparametric Approaches to Investigating Gene-Gene and Gene-Environment Interactions in Case-Control Studies*. Doctoral thesis, Indian Statistical Institute. Available at `https://www.researchgate.net/publication/361505764_Bayesian_Nonparametric_Approaches_to_Investigating_Gene-Gene_and_Gene-Environment_Interactions_in_Case-Control_Studies`.

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. (2024). Gene-gene and gene- environment interactions in case-control studies based on hierarchies of dirichlet processes. *Statistics and Applications*, **22**, 327–360.

Bhattacharya, S. (2008). Gibbs sampling based Bayesian analysis of mixtures with unknown number of components. *Sankhya. Series B*, **70**, 133–155.

Bonetta, L. (2010). Protein-protein interactions: Interactome under construction. *Nature*, **468**, 851–854.

Cordell, H. J. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, **11**, 2463–2468.

Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Reviews*, **10**, 392–404.

De Iorio, M., Elliott, L. T., Favaro, S., Adhikari, K., and Teh, Y. W. (2015a). Modeling population structure under hierarchical Dirichlet process. Available at "arXiv:1503.08278v1".

De Iorio, M., Favaro, S., and Teh, Y. W. (2015b). Bayesian inference on population structure: From parametric to nonparametric modeling. In *Nonparametric Bayesian Inference in Biostatistics*, pages 135–151. Springer International Publishing.

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.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinborough*, **52**, 399–433.

Hunter, D. J. (2005). Gene environment interactions in human diseases. *Nature Publishing Group*, **6**, 287–298.

Kempthorne, O. (1954). The correlation between relatives in a random mating population. *Proceedings of the Royal Society of London. Series B*, **143**, 103–113.

Khouri, M. J. (2005). Do we need genomic research for the prevention of common diseases with environmental causes? *American Journal of Epidemiology*, **161**, 799–805.

Ko, Y.-A., Saha Chaudhuri, P., Vokonas, P. S., Park, S. K., and Mukherjee, B. (2013). Likelihood ratio tests for detecting gene environment interaction in longitudinal studies. *Genetic Epidemiology*, **37**, 581–591.

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.

Li, J., Huang, D., Guo, M., Liu, X., Wang, C., Teng, Z., Zhang, R., Jiang, Y., Lv, H., and Wang, L. (2015). A gene-based information gain method for detecting gene-gene interactions in case-control studies. *European Journal of Human Genetics*, **23**, 1566–1572.

Lucas, G., Lluis-Ganella, C., Subirana, I., Masameh, M. D., and Gonzalez, J. R. (2012). Hypothesis-based analysis of gene-gene interaction and risk of myocardial infraction. *Plos One*, **7**, 1–8.

Mapp, C. (2003). The role of genetic factors in occupational asthma. *European Respiratory Journal*, **21**, 173–178.

Mukherjee, B., Ahn, J., Gruber, S. B., and Chatterjee, N. (2012). Testing gene environment interaction in large-scale association studies. *American Journal of Epidemiology*, **175**, 177–190.

Mukherjee, B., Ahn, J., Gruber, S. B., Ghosh, M., and Chatterjee, N. (2010). Bayesian sample size determination for case-control studies of gene-environment interaction. *Biometrics*, **66**, 934–948.

Mukherjee, B., Ahn, J., Gruber, S. B., Moreno, V., and Chatterjee, N. (2008). Testing gene-environment interaction from case-control data: A novel study of type-I error, power and designs. *Genetic Epidemiology*, **32**, 615–626.

Mukherjee, B. and Chatterjee, N. (2008). Exploiting gene-environment independence for analysis of case-control studies: An empirical-Bayes type shrinkage estimator to trade off Between bias and efficiency. *Biometrics*, **64**, 685–694.

Mukhopadhyay, S. and Bhattacharya, S. (2021). Bayesian MISE convergence rates of Polya urn-based density estimators: Asymptotic comparisons and choice of prior parameters. *Statistics: a Journal of Theoretical and Applied Statistics*, **55**, 120–151.

Musameh, M., Wang, W., Nelson, C., C.L-Ganella, Debiec, R., Subirana, I., Elosua, R., Balmforth, A., Ball, S., Hall, A., Kathiresan, S., Thompson, J., Lucas, G., Samani, N., and Tomaszewski, M. (2015). Analysis of gene-gene interactions among common variants in candidate cardiovascular genes in coronary artery disease. *Plos One*, **10**, 1–12.

Niel, C., Sinoquet, C., Dina, C., and Rocheleau, G. (2015). A survey about methods dedicated to epistasis detection. *Frontiers in Genetics*, **6**, 285.

Ottman, R. (2010). Gene environment interactions: Definitions and study designs. *Pubmed*, **6**, 764–770.

Phillips, P. C. (2008). Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Review Genetics*, **9**, 855–867.

Purcell, S. (2002). Variance components models for gene–environment interaction in twin analysis. *Twin Research*, **5**, 554–571.

Sanchez, B., Kang, S., and Mukherjee, B. (2012). A latent variable approach to study of gene-environment interactions in the presence of multiple correlated exposures. *Biometrics*, **68**, 466–476.

Scott, S. A. (2011). Personalizing medicine with clinical pharmacogenetics. *Genetics Medicine.*, **13**, 987–995.

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.

VanderWeele, T. J. (2009). Sufficient cause interactions and statistical interactions. *Epidemiology*, **20**, 6–13.

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L. S., and Yu, W. (2010). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *American Journal of Human Genetics*, **87**, 325–340.

Wang, X., Elston, R. C., and Zhu, X. (2010). The meaning of interaction. *Human Heredity*, **70**, 269–277.

Yi, N. (2010). Statistical analysis of genetic interactions. *Genetics Research*, **92**, 443–459.

Yi, N., Kaklamani, V. G., and Pasche, B. (2011). Bayesian analysis of genetic interactions in case-control studies, with application to adiponectin genes and colorectal cancer risk. *Annals of Human Genetics*, **75**, 90–104.

Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February, 2025; pp 167-176

Predictive Modeling of Maize Yield in Jammu Subtropical Zone using Weather Data and Penalized Regression

Manish Sharma[1], Amandeep Verma[1], Nishant Jasrotia[1], Sushil Sharma[2] and Divyam Sharma[1]

[1]*Division of Statistics and Computer Science, FBSc, SKUAST-Jammu*
[2]*Faculty of Agricultural Engineering, SKUAST-Jammu*

Received: 16 June 2025; Revised: 19 August 2025; Accepted: 21 August 2025

Abstract

The present study employed regression techniques such as Ordinary Least Squares (OLS), Ridge, and Lasso regression to analyse maize production data in relation to weather parameters from the subtropical region of Jammu. The models were evaluated across various training and testing splits based on performance metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). Among the models, the Lasso regression demonstrated the best overall performance with the lowest AIC (47.37) and BIC (50.76) values, indicating superior model fit and simplicity. The optimal regularisation parameter ($\alpha = 0.126$) and having minimum MSE, thus ensuring a well-balanced trade-off between model complexity and predictive accuracy. The Lasso regression model successfully identified key predictors influencing maize production, with maximum temperature and area being the most influential variables, followed by sunshine hours and relative humidity in the evening. Rainfall and minimum temperature were found to have minimal or no impact. Therefore, the proposed Lasso regression model, with its optimal alpha value and refined feature selection, serves as a robust and interpretable tool for predicting maize production in the subtropical region of Jammu.

Key words: Maize; Regularizations; Penalized regression; Weather; MSE.

1. Introduction

Statistical modelling tools can sometimes produce suitable models quite fast. The researchers used statistical models as baseline predictive models to assess the performance of advanced methods, even for situations where more adaptable machine-learning techniques (such regularization techniques and neural networks) can ultimately produce better results. The term regression was first introduced by a British anthropologist and meteorologist Sir Francis Galton (1886) in a paper entitled "Regression towards Mediocrity in Hereditary Stature". Since the regression analysis has emerged as a powerful statistical tool. With

Corresponding Author: Manish Sharma
Email: manshstat@gmail.com

the enhanced application of statistics in economics, industry, agriculture, social sciences, biology, medical sciences, psychology and education. According to Tibshirani (1996), the ordinary least square (OLS) estimates of the regression parameter in general have low bias but large variance and often have poor performance in both prediction and interpretation when the assumptions violated such as residual normality, homoscedasticity, independence, and linearity and the estimates may become skewed or undependable, especially when there are outliers and many correlations between the predictor variables. The prediction accuracy can sometimes be improved either by shrinking of some coefficient towards zero or by allowing a little bias to reduce the variance of the parameter estimates and predicted value. The outcomes of a linear regression can be severely distorted by outliers, which can result in incorrect interpretations and poor prediction accuracy. High correlation across predictor variables can lead to multicollinearity, which alters the variance of coefficient estimates and makes it challenging to isolate the impact of each predictor on the response variable. These problems may make yield projections less reliable and consider for exploring different modeling approaches. Penalized regression models like lasso and ridge regression have become effective alternatives for ordinary linear regression. By include a penalty term in the loss function, these models introduce regularization approaches that assist reduce the impact of outliers and multicollinearity. Through the efficient selection of important characteristics and the reduction of overfitting, this method not only increases model stability but also improves predictive performance.

One of the most adaptable developing crops, maize (*Zea mays L.*), has a wide range of modification under various agro-climatic situations. Maize has the largest genetic yield potential of all the cereals, it is referred to as the "Queen of cereals" worldwide. It is grown on roughly 150 million ha in about 160 nations, where there is a greater variety of soil, climate, biodiversity, and management techniques. This increases global grain production by 36 percent (782 million tonnes) Kiran *et al.* (2018).The maize crop is susceptible to the fluctuations of rainfall distribution as a whole. In Jammu and Kashmir, maize is widely planted in the kandi, karewa, and plain regions. It does well in loamy to sandy loam soils. Additionally, maize varieties that thrive in colder hilly and mountainous regions have been produced. In all such areas where the summer is long enough to support its cultivation and where frost does not arrive too early, it can be grown. When it is growing and developing, it needs a temperature of around 30°C, and when it is ripening, it needs a temperature of at least 20°C. A fertile, deeply tilled soil is necessary for maize. The soil is prepared in advance of the sowing season, which is April to May on the Jammu plain and May to June in the Kashmir valley, the kandi, and the state's mountainous regions. There are ten districts in the Jammu area of the UT of Jammu and Kashmir, maize is grown in almost all the districts of the Jammu region. In terms of production, the districts of Rajouri (110.20 thousand MT), Udhampur (70.11 thousand MT), and Poonch (69.59 thousand MT) have the largest concentration of maize. (Digest of Statistics, 2023-24). Jorvekar *et al.* (2024) conducted a study to evaluate and compare the performance of different regression models for agriculture crop yield prediction on the basis historical crop yield data, weather parameters and pesticides data features from various agricultural regions. Various regression models, including Linear Regression (LR), K Nearest neighbor Regression (KNR), Support Vector Regression (SVR), Decision Tree Regression (DTR), Random Forest Regression (RFR), Gradient Boosting Regression (GBR),Linear Model Lasso Regression, Elastic net Regression, Ridge Regression to predict crop yields for various crops. This study involved evaluating the performance

of these regression models based on several performance.Krishnadoss and Ramasmy (2024) used various machine learning models for crop yield prediction to make dynamic pre monsoon decisions. The input variables precipitation, temperature, evaporation, wind speed, and chemical use influence crop yield estimations. Jasrotia *et al.* forecast the production of walnut for Jammu and Kashmir using Time series models. Holt linear exponential Smoothing and Autoregressive Integrated Moving Average (ARIMA) model have been applied and it shows that ARIMA(1,2,1) is appropriate model for forecasting on the basis of minimum value of information criterion as compared to other models. Based on the forecast provided by the proposed model, there is a projected 56.69 percent increase in walnut production for the year 2035 with respect to 2022. Gupta *et al.* compared classification techniques through statistical as well as artificial neural network models for the primary data related to 140 rice genotypes from the trial laid in SKUAST, Jammu on the basis of maturity. And, the characters like yield per plant, number of days for 50 percent flowering, number of days for full flowering, plant height, number of effective tillers per plant, panicle length, grain length, grain width and ratio of grain length and grain width acts as supporting variables for classification.

2. Material and methods

The study was undertaken based on secondary data related to area and production of maize from three decades with effect from 1992 to 2023, 31 years of subtropical region of Jammu.To assess the performance of the model, different proportions of training and testing data were utilized related to 31 years of data of different parameters. The data pertains to weather-based components such as; Maximum temperature, minimum temperature, Rainfall, Relative humidity morning, Relative humidity evening and sunshine (hrs.) were also collected from digest of statistics published by Directorate of Economics and Statistics, UT Administration of Jammu and Kashmir and Agro-metrological unit of SKUAST-Jammu. In order to estimate the relation and prediction of maize production of Jammu division through weather parameters the following statistical models are applied for handling the problem of overfitting and challenges of influential observations in the data.

2.1. The least square method

The least square method used in regression is relatively straightforward. Imagine a scatterplot of data points that form a linear trend. An OLS linear regression procedure builds a line of best fit that would serve as the most accurate way of depicting the spread of the data points with a single line. The least squares property states that the line fit in the OLS method will have the smallest value of the summed squared deviations of each data point from the line.

2.2. Penalized regression models

The two most common techniques are ridge regression given by Hoerl and Kennard (1970). and lasso regression Tibshirani (1996). The predictor variables are all kept in the model through penalized regression techniques, but the regression coefficients are regularized by reducing them to zero or a value equal to zero. Shrinkage or regularization methods are other names for penalized regression.

(i) Ridge regression model

Ridge regression shrinks the regression coefficients, so that variables, with a minor contribution to the outcome, have their coefficients close to zero. The shrinkage of the coefficients is achieved by penalizing the regression model with a penalty term called L2-norm, which is the sum of the squared coefficients. Mathematical form of ridge regression model is:

$$\min\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{i=1}^{p}\beta_j x_{ij}\right)^2+\alpha 2\|\beta_j\|\right\}$$

where y is the dependent variable, x is the covariate, β is the corresponding coefficient, and α_2 represents the L2 norm penalty. The amount of the penalty can be fine-tuned using a constant called Alpha (α). Selecting a good value for α is critical. When $\alpha = 0$, the penalty term has no effect, and ridge regression will produce the classical least square coefficients. However, as α increases to infinity, the impact of the shrinkage penalty grows, and the ridge regression coefficients will get close to zero.

(ii) Least absolute shrinkage selection operator (LASSO)

It shrinks the regression coefficients toward zero by penalizing the regression model with a penalty term called L1-norm, which is the sum of the absolute coefficients. In the case of lasso regression, the penalty has the effect of forcing some of the coefficient estimates, with a minor contribution to the model, to be exactly equal to zero. This means that lasso can be also seen as an alternative to the subset selection methods for performing variable selection in order to reduce the complexity of the model. As in ridge regression, selecting a good value of α for the lasso is critical. One obvious advantage of lasso regression over ridge regression is that it produces simpler and more interpretable models that incorporate only a reduced set of predictors. The mathematical model of the of the LASSO Regression is

$$\min\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{i=1}^{p}\beta_j x_{ij}\right)^2+\alpha_1\|\beta\|\right\}$$

where y is the dependent variable, x is the covariate, β is the corresponding coefficient, and α_1 represents the L1 norm penalty. However, neither ridge regression nor the lasso will universally dominate the other. Generally, lasso might perform better in a situation where some of the predictors have large coefficients, and the remaining predictors have very small coefficients. The penalized model is trained using a different subset of the data observations, called the training set and rest is used as testing set. The performance of the proposed model *i.e.* OLS, RR and Lasso checked through the MSE, RMSE, AIC and BIC.

(a) Mean squared error (MSE)

The Mean Squared Error (MSE) measures how close a regression line is to a set of data points. It is a risk function corresponding to the expected value of the squared error loss. The Mean Squared Error is calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

where, n is sample size, actual is the actual data value and forecast the predicted data value

(b) Root mean squared error (RMSE)

RMSE is employed while assessing models. If we have a sample of n observations $y(Yi, i = 1, 2..., n)$, and n matching model predictions y, then RMSE is defined as the square root of the mean squared error (MSE), which is provided by

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

(c) Akaike's information criterion (AIC)

An important statistic for identifying and assessing statistical models is AIC. The likelihood function, L, and the number of hyperparameters estimated from the model, n, are used to determine this criterion, which was proposed by Akaike in 1979. Its is calculated as $AIC = -2logL + 2n$ where, L is the value of the likelihood and n is the number of estimated parameters.

(d) Bayesian information criterion (BIC)

The criterion was proposed by Schwarz (1978) using Bayesian likelihood maximization. Schwarz further demonstrated the BIC's validity by showing that it is independent of prior distribution. $SBIC = -2logL + nlogT$, where T is the total number of observations, is the formula for the BIC. The fitted model performs better when these data have a lower value.

3. Results and discussion

In this study, regression techniques like Ordinary Least Squares (OLS), Ridge, and Lasso regression were applied to the data of maize production with whether parameters from the subtropical region of Jammu. The subtropical region of Jammu shows moderate variability in maize cultivation as shown in table 1 with an average area of 21.15 thousand hectares and production of 38.74 thousand metric tonnes. Climatic conditions remain fairly stable, with average minimum and maximum temperatures of 16.32°C and 28.96°C, respectively. Rainfall displays notable variation, averaging 1240.30 mm with a high coefficient of variation (16.16%).

The normality tests conducted for the Jammu (subtropical) region revealed mixed results across variables as shown in table 2. Area and minimum temperature show slight deviations from normality as per Shapiro-Wilk and Anderson-Darling tests. Sunshine hours significantly deviate from normality across all three tests, indicating strong non-normal distribution. Production, rainfall, and relative humidity (morning) largely follow a normal distribution with high p-value. These findings suggest that while most climatic variables are normally distributed, sunshine data requires transformation or non-parametric handling in statistical modeling.

To evaluate the model performance, the dataset was divided into different training

Table 1: Descriptive statistics for subtropical region of Jammu

Variable	Minimum	Maximum	Average	*SD*	*CV* (%)
Area (000 hectares)	15.71	25.03	21.15	2.63	12.43
Production (000 MT)	23.70	48.24	38.74	5.32	13.73
Minimum Temperature (°C)	15.15	17.22	16.32	0.57	3.51
Maximum Temperature (°C)	28.00	29.96	28.96	0.58	2.00
Rainfall (mm)	769.14	1730.12	1240.30	200.48	16.16
Sunshine (hrs.)	4.36	6.79	6.14	0.40	6.54
Relative Humidity Morning (%)	74.98	87.74	80.76	2.48	3.08
Relative Humidity Evening (%)	43.05	55.04	49.98	2.81	5.54

Table 2: Test of normality for area, production, and different weather parameters for subtropical region of Jammu using Shapiro-Wilk (*S-W*), Kolmogorov-Smirnov (*K-S*), and Anderson-Darling (*A-D*) tests

Variable	Shapiro-Wilk (*S-W*)		Kolmogorov-Smirnov (*K-S*)		Anderson-Darling (*A-D*)	
	Statistic	p-value	Statistic	p-value	Statistic	p-value
Area (000 hectare)	0.93	0.05	0.76	0.60	0.78*	0.04
Production (000 MT)	0.96	0.41	0.54	0.93	0.31	0.52
Minimum Temperature	0.93*	0.04	0.89	0.41	0.86*	0.02
Maximum Temperature	0.96	0.25	0.77	0.59	0.35	0.46
Total Rainfall (mm)	0.99	0.99	0.36	1.00	0.16	0.94
Relative Humidity (Morning)	0.97	0.54	0.65	0.79	0.37	0.41
Relative Humidity (Evening)	0.93	0.08	1.13	0.16	0.85*	0.02
Sunshine (hrs.)	0.69**	0.00	1.56**	0.01	2.85**	0.00
Sunshine (hrs.)	0.70**	0.00	1.56*	0.01	2.86**	0.00

and testing ratios: 80:20, 70:30, and 60:40. Here, the first value shows the percentage of data used for training the model, and the second value indicates the percentage used for testing. The basis for choosing the optimal ratio is obtaining the lowest MSE and RMSE for testing datasets. These evaluation criteria are essential because they provide insight on the model's accuracy and precision and capacity for generalization

The examination of various training and testing data splits for the OLS regression model determined that the 60:40 ratio is the optimal option for the subtropical region of Jammu, as shown in table 3. This ratio showed the best performance of the model on unknown data, yielding the lowest testing MSE (15.72) and RMSE (3.96) as compared to training MSE (16.36) and RMSE (4.04). The 80:20 ratio had the greatest testing MSE (21.43) and RMSE (4.63) than training MSE (13.87) and RMSE (3.72), indicating overfitting even though it displayed the lowest training error. When it came to testing MSE and RMSE, the 60:40 ratio performed better than the 70:30 ratio having testing MSE(17.45) and RMSE (4.18). Therefore, the 60:40 ratio is selected for the model and optimal option for the RR model in the subtropical area of Jammu, according to the criterion of choosing the ratio with the smallest testing MSE and RMSE. Its training MSE (16.38) and RMSE (4.05) and testing MSE (15.61) and RMSE (3.95) were the lowest, demonstrating the best model performance and generalization. The 80:20 ratio exhibited the highest testing MSE (19.29) and RMSE (4.39), indicating overfitting, despite having the lowest training MSE (14.54) and RMSE (3.81). The 70:30 ratios show overfitting having lower training MSE (14.75), RMSE (3.84) and higher testing MSE (16.84) and RMSE (4.10). The selected ratio is the best option for reducing errors on the testing dataset and promising improved generalization because

Table 3: The MSE, RMSE, AIC and BIC of regression models for subtropical region of Jammu w.r.t different ratios for training and testing datasets

Model	Statistics	60:40		70:30		80:20	
		Training	Testing	Training	Testing	Training	Testing
OLS	MSE	16.36	15.72	14.58	17.45	13.87	21.43
	RMSE	4.04	3.96	3.82	4.18	3.72	4.63
	AIC	64.31	49.81	70.27	42.60	77.11	35.45
	BIC	70.54	53.77	77.58	44.71	85.36	35.08
Ridge regression	MSE	16.38	15.61	14.75	16.84	14.54	19.29
	RMSE	4.05	3.95	3.84	4.10	3.81	4.39
	AIC	66.34	51.72	72.52	44.23	80.25	36.72
	BIC	73.46	56.24	80.77	46.66	89.67	36.28
Lasso regression	MSE	16.67	15.19	15.09	16.54	14.57	18.20
	RMSE	4.08	3.90	3.89	4.07	3.82	4.27
	AIC	62.65	47.37	70.99	42.06	76.30	32.31
	BIC	67.99	50.76	78.30	44.18	83.37	31.98

it shows a good trade-off between training and testing errors. The selcted ratio is the best option for the lasso regression model based on the criterion of choosing the ratio with the smallest testing MSE and RMSE for subtropical region of Jammu. Its testing MSE (15.19) and RMSE (3.90) were the lowest, in comparison to training MSE (16.67) and RMSE (4.08) which demonstrating the best model performance and generalization. The testing MSE (18.20) and RMSE (4.27) of the 80:20 ratio was reasonably high than training MSE (14.57) and RMSE (3.82), and showing overfitting. The 70:30 ratio performed poorer than the 60:40 ratio as it also shows overfitting having lower training MSE (15.09) and RMSE (3.89) and higher testing MSE (16.54) and RMSE (4.07). As a result, the common ratio selected was the best option for minimizing testing MSE and RMSE, and indicating strong generalisation with reduced overfitting for all the models.

AIC, BIC, and coefficient values based on the OLS, RR, and LR models analysis utilizing the selected ratio for the subtropical region of Jammu are shown in table 4, With an AIC (49.81) and a BIC (53.77) for OLS regression, 51.72 and a BIC (56.24) for RR, and the lowest AIC (47.37) and a BIC (50.76) for lasso regression, these results were obtained. Lasso showed the smallest testing AIC and BIC, suggesting superior model performance and generalization, the LR model is chosen as the best-performing model for predicting maize production in the subtropical region.

The variation in the coefficients of several meteorological variables with different log(alpha) values in a Lasso regression model is illustrated in fig. 1. The coefficients are displayed on the y-axis, and log(alpha) values are represented on the x-axis. The significance of each meteorological variable in the model is indicated by the line that corresponds to it. The optimal alpha value (0.126), is selected to minimize the mean squared error, which can be observed by the dashed line. The coefficients for every variable are visible at this ideal alpha, emphasizing their relative significance. The most important variables are found using this optimum alpha value in the Lasso regression, producing a more precise and understandable model.

The relationship between the log of the regularization parameter (alpha) and the mean

Table 4: Regression coefficient estimates of different parameters for subtropical region of Jammu with respect to OLS, ridge and lasso regression models

Parameters	OLS	Ridge regression	Lasso regression
Intercept	-143.53	-90.48	-51.66
Area (X1)	1.02*	1.01*	1.03*
Max temp. (X2)	3.92*	3.46*	2.58*
Min temp. (X3)	0.18	0.03	0
Rainfall (X4)	0.00	0.00	0.0001*
RH (morning) (X5)	-0.24	-0.30	-0.35
RH (evening) (X6)	0.36*	0.35*	0.29*
Sunshine (hrs.) (X7)	1.70*	1.35*	0.30*
R^2	0.70		
AIC	49.81	51.72	47.37
BIC	53.77	56.24	50.76

squared error (MSE) in a lasso regression model is depicted in fig. 2. The x-axis represents log(alpha) values, and the y-axis shows the MSE. The dotted line indicates how MSE varies with different log(alpha) values, and the red dot marks the best alpha value (0.126), which minimizes the MSE. At this optimal alpha, the model achieves the best balance between bias and variance, resulting in the lowest prediction error. Thus, model is proposed at best alpha (0.126) and Lasso model describes the effect of various features on the target variable is shown in the fig. 3. The coefficients for each attribute are shown by the bars, and larger values denote greater significance. Maximum temperature (2.58*) is the most significant predictor, with the largest positive coefficient; area (1.03*) comes in second. While relative humidity morning (0.35) has a slight negative influence, features like sunshine (0.30*) and relative humidity evening (0.29*) have moderately good effects. The coefficients for rainfall (0.0001*) are almost zero, suggesting that they have little effect. whereas, coefficient for minimum temperature is reduced to zero indicating lesser effect on production than other variables. The optimum alpha value, which minimizes prediction error while setting some coefficients to zero, strikes a compromise between model complexity and accuracy when determining these coefficients.

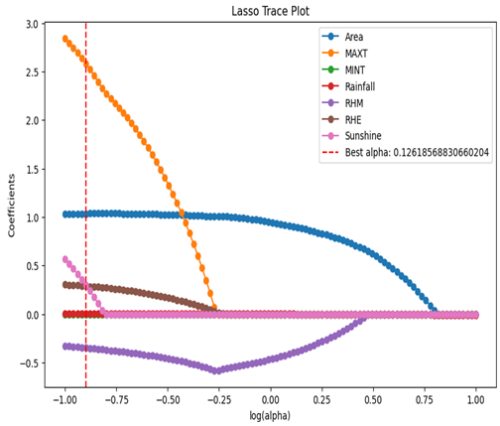


Figure 1: The curve for lasso trace of coefficients w.r.t log alpha

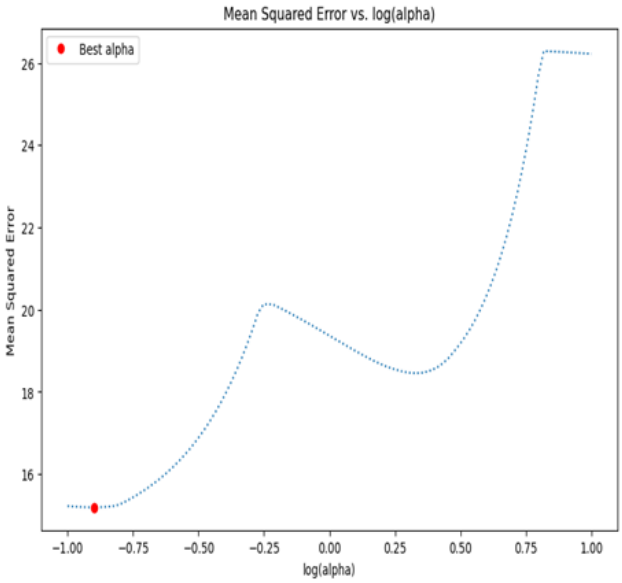


Figure 2: Plot between mean squared error and log alpha

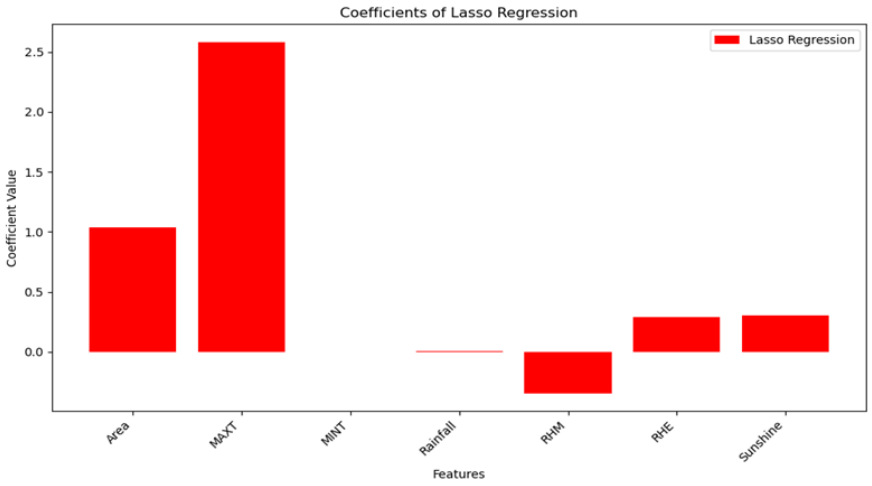


Figure 3: Plot for coefficients of lasso regression

4. Conclusion

Lasso regression outperformed OLS and Ridge models, offering the best predictive accuracy and simplicity, with the lowest AIC (47.37) and BIC (50.76) values. Key predictors identified by the Lasso model were maximum temperature and area, followed by sunshine hours and evening relative humidity; rainfall and minimum temperature had negligible influence. The model offers a reliable and interpretable framework for forecasting maize production, supporting data-driven agricultural planning in the subtropical region of Jammu.

Thus, proposed penalized (Lasso) regression model for maize production prediction is:

$$y = -51.66 + 1.03 \cdot \text{Area} + 2.58 \cdot \text{MAXT} + 0.29 \cdot \text{RH(Evening)} + 0.30 \cdot \text{Sunshine (hrs)}. \quad (1)$$

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

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika*, **66**, 237-242.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, **15**, 246-263.

Gupta, S., Sharma, M., Jasrotia, N. and Mahajan, S. (2025). Comparative analysis for classification of rice genotype using statistical and artificial neural network models on the basis of maturity. *Model Assisted Statistics and Applications*, 19, 342-349.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, **12**, 55-67.

Jasrotia, N., Sharma, M., Bhat, A., Mahajan, S., & Gupta, S. (2024). Forecasting models for the production of walnut in Jammu and Kashmir - A comparative study. *Special Proceedings of the 26th Annual Conference, Society of statistics and Computer Application*, 51–60.

Jorvekar., P. P., Wagh., S. K., and Prasad., J. R. (2024). Predictive modeling of crop yields: a comparative analysis of regression techniques for agricultural yield prediction. *Agricultural Engineering International: CIGR Journal*, **26**, 125-140.

Krishnadoss. N. and Ramasamy. L. K (2024). Crop yield prediction with environmental and chemical variables using optimized ensemble predictive model in machine learning. *Environmental Research Communications*, **6**, 1-11.

Kiran, A. S. Shashi, Umesh, K. B., and Shankara, M. H. (2018). Growth and instability in agriculture - A case of maize production in India. In *Conference, July 28-August 2, 2018, Vancouver, British Columbia 277535, International Association of Agricultural Economists*.

Kumar, S., Attri, S. D., and Singh, K. K. (2019). Comparison of Lasso and stepwise regression technique for wheat yield prediction. *Journal of Agrometeorology*, **21**, 188-192.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461-464.

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

Publisher
Society of Statistics, Computer and Applications
Registered Office: I-1703, Chittaranjan Park, New Delhi- 110019, INDIA
Mailing Address: B-133, Ground Floor, C.R. Park, New Delhi-110019, INDIA
Tele: 011-40517662
https://ssca.org.in/
statapp1999@gmail.com
2025

Printed by : Galaxy Studio & Graphics
Mob: +91 9818 35 2203, +91 9582 94 1203
Email: galaxystudio08@gmail.com