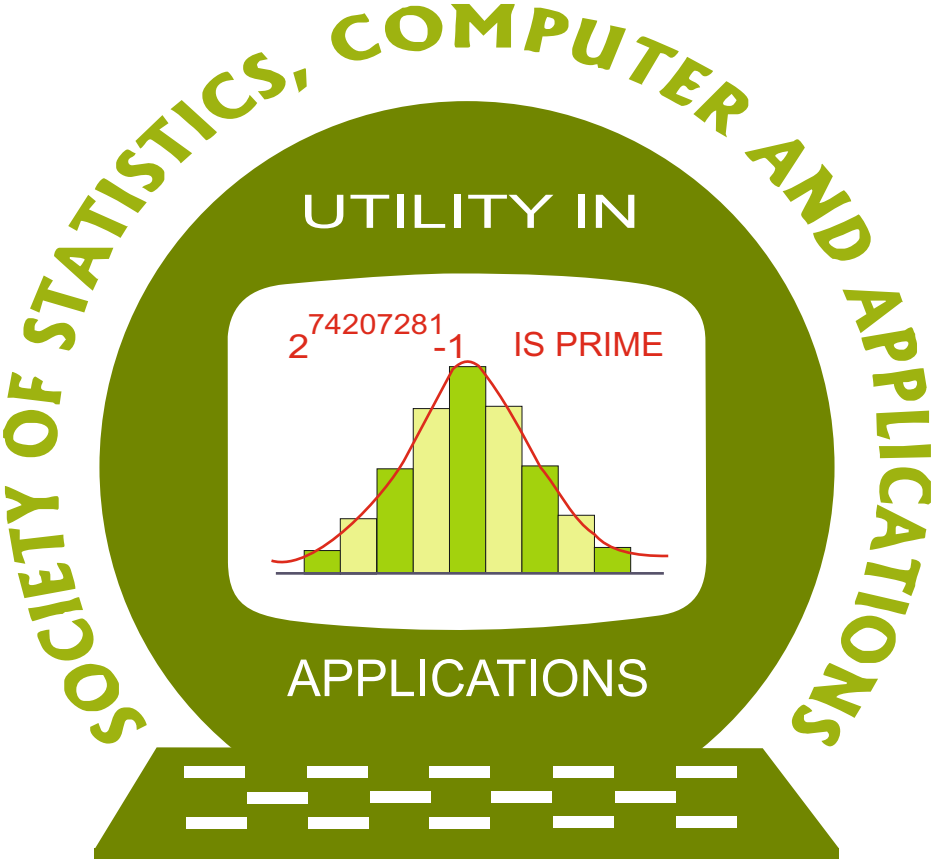
ISSN 2454-7395(online)

STATISTICS AND APPLICATIONS



FOUNDED 1998

Journal of the Society of Statistics, Computer and Applications

https://ssca.org.in/journal.html
Volume 20, No. 2, 2022 (New Series)

Society of Statistics, Computer and Applications

Council and Office Bearers

Founder President
Late M.N. Das

President
V.K. Gupta

Executive President
Rajender Parsad

Patrons

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

Vice Presidents

A. Dhandapani
P. Venkatesan
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

Council Members

Anil Kumar Yadav
B. Re. Victor Babu
Manish Sharma
Manisha Pal
Piyush Kant Rai
Praggya Das
Rajni Jain
Rakhi Singh
Ranjit Kumar Paul
Raosaheb V. Latpate
S.A. Mir
Sapam Sobita Devi
V. Srinivasa Rao
V.M. Chacko
Vishal Deo

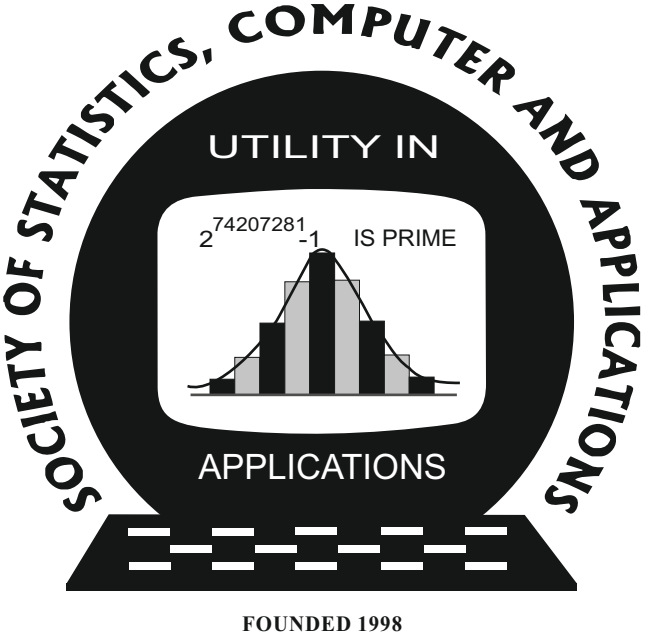
Ex-Officio Members (By Designation)

Director General, Central Statistics Office, Government of India, New Delhi
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

Statistics and Applications

ISSN 2454-7395(online)



Journal of the Society of Statistics, Computer and Applications

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

Volume 20, No. 2, 2022 (New Series)

Statistics and Applications

Volume 20, No. 2, 2022 (New Series)

Nripes Kumar Mandal, Flat No. 5, 141/2B, South Sinthee Road, Kolkata-700050; mandalnk2001@yahoo.co.in
P. Venkatesan, Professor Computational Biology SRIHER, Chennai, Adviser, CMRF, Chennai; venkaticmr@gmail.com
Pritam Ranjan, Indian Institute of Management, Indore - 453556; MP, India; pritam.ranjan@gmail.com
Ramana V. Davuluri, Department of Biomedical Informatics, Stony Brook University School of Medicine, Health Science Center Level 3, Room 043 Stony Brook, NY 11794-8322, USA; ramana.davuluri@stonybrookmedicine.edu; ramana.davuluri@gmail.com
S. Ejaz Ahmed, Faculty of Mathematics and Science, Mathematics and Statistics, Brock University, ON L2S 3A1, Canada; sahmed5@brocku.ca
Sanjay Chaudhuri, Department of Statistics and Applied Probability, National University of Singapore, Singapore -117546; stasc@nus.edu.sg
Sat N. Gupta, Department of Mathematics and Statistics, 126 Petty Building, The University of North Carolina at Greensboro, Greensboro, NC -27412, USA; sngupta@uncg.edu
Saumyadipta Pyne, Health Analytics Network, and Department of Statistics and Applied Probability, University of California Santa Barbara, USA; spyne@ucsb.edu, SPYNE@pitt.edu
Snigdhansu Chatterjee, School of Statistics, University of Minnesota, Minneapolis, MN -55455, USA; chatt019@umn.edu
T.V. Ramanathan; Department of Statistics; Savitribai Phule Pune University, Pune; madhavramanathan@gmail.com
Tapio Nummi, Faculty of Natural Sciences, Tampere University, Tampere Area, Finland; tapio.nummi@tuni.fi
Tathagata Bandyopadhyay, Indian Institute of Management Ahmedabad, Gujarat; tathagata.bandyopadhyay@gmail.com, tathagata@iima.ac.in
Tirupati Rao Padi, Department of Statistics, Ramanujan School of Mathematical Sciences, Pondicherry University, Puducherry; drtrpadi@gmail.com
V. Ramasubramanian, ICAR-IASRI, Library Avenue, PUSA, New Delhi – 110012; ram.vaidhyanathan@gmail.com

Corrigendum

With apology, we wish to share that in the following volumes and issues of Statistics and Applications, the ISSN Number mentioned on the title page of each individual paper is 2452-7395 (online). The correct ISSN Number, though, is 2454-7395 (online). The cover page of these issues does mention the correct ISSN Number as 2454-7295 (online). It is requested to please read the ISSN Number on each individual paper in these issues as 2454-7395 (online) in place of 2452-73975 (online)

1. Vol. 18, No. 2 (Special Issue), 2020 (New Series)
2. Vol. 18, No. 1, 2020 (New Series)
3. Vol. 17, No. 2, 2019 (New Series)
4. Vol. 17, No. 1, 2019 (New Series)
5. Vol. 16, No. 1, 2018 (New Series)
6. Vol. 15, Nos. 1 & 2, 2017 (New Series)
7. Vol. 14, Nos. 1 & 2, 2016 (New Series)

CONTENTS

Statistics and Applications ISSN 2454-7395 (online)
Volume 20, No. 2, 2022 (New Series)

1. Design a Soft Computing DECEL Model to Optimize Water Usage in Irrigation Management 1-13
Pradeep H.K., Jasma Balasangameshwara, K. Rajan, Archana B.K.
2. Symbolic Data Analysis vs Classical Data Analysis: A Comparative Study 15-25
Dipanka Bora and Hemanta Saikia
3. Identifying the Time of a Permanent Shift in the Normal Process Mean with Memory Type Control Chart 27-40
R. A. Kapase and V. B. Ghute
4. A-optimal Designs for Cubic Polynomial Models with Mixture Experiments in Three Components 41-55
Mahesh Kumar Panda and Rushi Prasad Sahoo
5. A Stochastic Modeling of a Monthly Rainfall of Hillsborough County Using Frechet Distribution 57-71
H. J. Patel and M. N. Patel
6. Transmuted Sine - G Family of Distributions: Theory and Applications 73-92
K.M. Sakthivel and J. Rajkumar
7. Constant Block-Sum Designs Through Confounded Factorials 93-101
Sudhir Gupta
8. Use of Change Point Analysis in Seasonal ARIMA Models for Forecasting Tourist Arrivals in Sri Lanka 103-121
B.R.P.M. Basnayake and N.V. Chandrasekara
9. A Survey on Cyclic Solution of Block Designs 123-133
Shyam Saurabh and Kishore Sinha
10. Robustness of Bayes Estimation of Coefficient of Variation for Normal Distribution for a Class of Moderately Non-Gamma Prior Distributions 135-146
Priyanka Aggarwal and Samridhi Mehta
11. Almost Unbiased Dual Exponential Type Estimators of Population Mean Using Auxiliary Information 147-156
Sajad Hussain, Manish Sharma, Banti Kumar and Vilayat Ali Bhat
12. Theory and Applicability of the Weighted Modified Lindley Distribution 157-175
Christophe Chesneau, Lishamol Tomy and Jiju Gillariose

13. Bayesian Credible Intervals for Generalized Inverse Weibull Distribution 177-188
Kamaljit Kaur, Sangeeta Arora and Kalpana K. Mahajan
14. Robust Parameter Design Using 20 Run Plackett-Burman Design 189-201
Renu Kaul and Sanjoy Roy Chowdhury
15. Inference on P (X < Y) for Morgenstern Type Bivariate Exponential Distribution Based on Record Values 203-218
Manoj Chacko and Shiny Mathew
16. Clustering using Skewed Data via Finite Mixtures of Multivariate Lognormal Distributions 219-237
Deepana, R. and Kiruthika, C.
17. Simultaneous Testing Procedure for the Ordered Pair-Wise Comparisons of Location Parameters of Exponential Populations under Heteroscedasticity 239-249
Jatesh Kumar, Amar Nath Gill and Anju Goyal
18. Some Results of Auto-Relevation Transform in Reliability Analysis 251-263
Dileep Kumar M. and P. G. Sankaran
19. Modeling and Analysis of Competing Risks Cure Rate Regression Model with Weibull Distribution 265-277
P.G. Sankaran and Rejani P.P.
20. Improvisation of Dataset Efficiency in Visual Question Answering Domain 279-289
Sheerin Sitara Noor Mohamed and Kavitha Srinivasan

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 1– 13

Design a Soft Computing DECEL Model to Optimize Water Usage in Irrigation Management

Pradeep H.K.[1], Jasma Balasangameshwara[2], K.Rajan[3] and Archana B.K.[4]
[1]Department of Computer Science & Engineering, JSS Academy of Technical Education, Bengaluru, Visvesvaraya Technological University, Belagavi, Karnataka, India.
[2]Department of Computer Science & Technology, Dayananda Sagar University, Bengaluru, Karnataka, India.
[3]ICAR - Indian Institute of Soil and Water Conservation, Ooty, India.
[4]Department of Electronics & Communication Engineering, JSS Academy of Technical Education, Bengaluru, Karnataka, India.

Received: 24 December 2020; Revised: 20 May 2021; Accepted: 26 May 2021

Abstract

Finite Automata (FA) and soft computing techniques have potential to improve agricultural water management practices. The existing irrigation systems suffer from low water productivity. This issue can be ameliorated through Dirt texture, Evapotranspiration and Crop Evolution based Land specific (DECEL) model. The soft computing models such as K-Nearest Neighbor (KNN) and linear regression prediction methods are used in the DECEL irrigation framework. The results exhibited that, the KNN algorithm obtained accuracy of 95.88% over dirt texture classification and 99.98% accuracy on crop coefficient prediction. The reference evapotranspiration is predicted using linear regression method.

Key words: Dirt texture; Evapotranspiration; Crop coefficient; Machine learning; Finite automata.

1. Introduction

The global food requirement increases about 60% by the year 2050 due to growing population (Alexandratos and Bruinsma, 2012). Currently irrigated land can only satisfy 40% of the expected global food requirement by the year 2050. Agriculture sector uses 70% of the available water (Provenzano and Sinobas, 2014). Currently only 16% of the cultivable area is irrigated due to adoption of conventional irrigation approaches (Alexandratos and Bruinsma, 2012; Playan *et al.*, 2014). The arid and semi-arid regions are currently expanded to 36% and global warming trend further expands the aridity area (Safriel *et al.*, 2005; Alcamo *et al.*, 2007; Arnell *et al.*, 2011). The efficiency and economic outcome is the vital concern of irrigation system (Burt *et al.*, 2005; Chartzoulakis *et al.*, 2015). The performance of irrigation system depends on timely supply of exactly required volume of water. The water transformation through soil and crop are expressed using the metrics such as evaporation, transpiration, infiltration, runoff and deep percolation. Evaporation is the process of water transformation from liquid to vapour. The transpiration is the process of water passed from crop stomata to atmosphere in the form of vapour. Evapotranspiration (ET) is the combined process of surface evaporation and crop transpiration. The infiltration is the process of water entry in the surface of soil. The deep percolation is the infiltrated water which moves beyond the root zone. The water moves out of the land is called runoff (Burt *et al.*, 1997). The dirt

Corresponding Author: Pradeep H.K.
Email: hkpradeep.research@gmail.com

properties, weather conditions and crop coefficient play crucial role in irrigation system (Dabach *et al.,* 2011; Soulis and Elmaloglou, 2018).

The rest of the paper is structured as follows. The Section 2 describes the evolution of various irrigation methods. The irrigation automation framework is outlined in Section 3. The soft computing approaches and their results are discussed in Section 4. Finally, the conclusions and future research directions are summarized in Section 5.

2. Related Work

The surface irrigation method is most extensively used technique and this approach is popular due to low initial cost and energy demand despite the low irrigation efficiency. Basin, border and furrow are generally practiced surface irrigation techniques (Raghuwanshi *et al.,* 2011). The sprinkler irrigation framework comprises of pipe network in which water flows with force through nozzles and it simulates precipitation with the help of overhead spraying. The solid set, linear and hand move, centre pivot, wheel line, gun type and hose-pull are various sprinkler irrigation techniques. In drip irrigation, water is supplied via pipe network in a fixed model and water is slowly emitted to each plant to the root zone (Tindula *et al.,* 2013). The evolution of first-generation irrigation technology was started with multi-client electronic hydrants for utilization at dispensation network. The second-generation irrigation technology was variable frequency pump. The micro irrigation method was the third generation in irrigation technology wherein WP was increased but marginally installed due to high initial investment (Pradeep *et al.,* 2021a). The sub surface drip irrigation (SDI) was the fourth generation in irrigation technology invented to solve the issues of surface drip irrigation specifically to eliminate emitter clogging issue. The fifth generation in irrigation technology was deficit irrigation invented to supply reduced amount of water without affecting the yield based on crop growth stage (Levidow *et al.,* 2014; Kang *et al.,* 2017). Intelligent irrigation is the emerging area which addresses the low water productivity issue (Pradeep *et al.,* 2019; Pradeep *et al.,* 2020; Krishnashetty *et al.,* 2021; Pradeep *et al.,* 2021b). The evolution of irrigation methods are presented in Table1.

Table 1: Progress of irrigation techniques

Approach	Benefits	Implementation
Multi-client hydrants	Dispensation unit	Mostly used
Frequency pumps	Pumping plant	Mostly used
Drip & Sprinkler	Water control and irrigation scheduling	Marginally deployed
Sub surface drip	Water control and irrigation scheduling	Minimal
Deficit irrigation	Water control and irrigation scheduling	Minimal
Intelligent irrigation	High water productivity and economy	New era

3. Irrigation Automation Model

The Finite Automata (FA) is a core concept of intelligent computing. In this paper, the deterministic variant of FA (DFA) is used to design the automated irrigation framework

provided with some rules that permit the automaton to handle the symbols, according to the rules to generate the output. There are only two possible outcomes over the input passed to the FA, “accept” or “reject”. In FA model the states are represented by circles. Arcs between the states are labeled by inputs. The States may have a self loop for some of the input symbols. In FA model one of the states is designated as start state, indicated by an arrow leading to that state without origin state and its necessary to have one or more states as final or accepting states, indicated by double circle. For all valid input string the FA should halt at one of the designated final state. In the present study an irrigation automation framework is proposed which is represented in Figure 1.

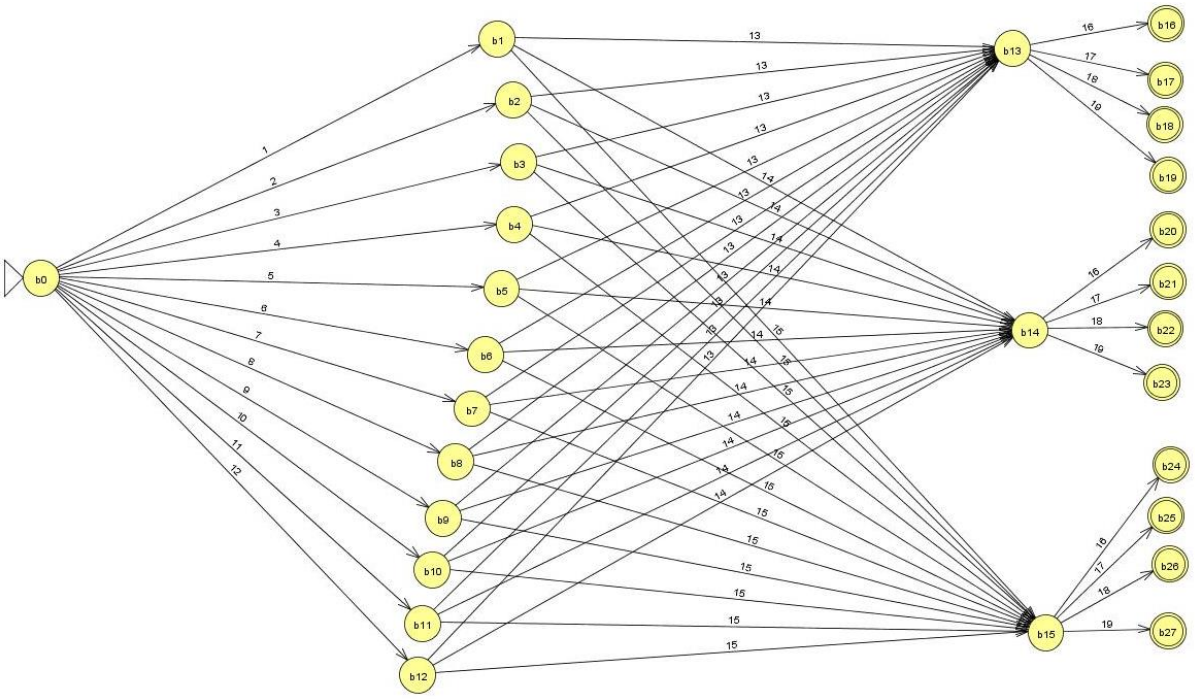


Figure 1: DFA model for irrigation automation

The input variables are dirt texture, evapotranspiration, and crop evolution coefficient data for specific land. The automated irrigation framework variables are reported in Table 2.The United States Department of Agriculture (USDA) has defined twelve major soil texture classes considering the combination of sand, silt and clay fractions, which are highlighted in Table 3. The set of soil texture input parameters are represented in the model as {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}. The weather data is classified as warm, temperate, and polar, reported in the Table 4 and also represented as input variables {13, 14, 15} in Figure 1. The crop evolution coefficient depends on the crop growth stage. They are represented in the model as {16, 17, 18, 19, 20} and reported in Table 5. The accepting states determine volume of water required for the given input pattern.

Table 2: Variables of automated irrigation framework

DFA Attributes	Description
States	$Q = \{b_0, b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8, b_9, b_{10}, b_{11}, b_{12}, b_{13}, b_{14}, b_{15}, b_{16}, b_{17}, b_{18}, b_{19}, b_{20}, b_{21}, b_{22}, b_{23}, b_{24}, b_{25}, b_{26}, b_{27}\}$

Input symbols	Soil texture variables = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12} Evapotranspiration data variables = { 13, 14, 15} Crop evolution input parameters = {16, 17, 18, 19, 20}
Start state	b_0
Final states	F = {b_{16}, b_{17}, b_{18}, b_{19}, b_{20}, b_{21}, b_{22}, b_{23}, b_{24}, b_{25}, b_{26}, b_{27}}

Table 3: Dirt classification transitions

Current State	Input	Next State	Soil texture
b_0	1	b_1	Sand
b_0	2	b_2	Loamy sand
b_0	3	b_3	Sandy loam
b_0	4	b_4	Loam
b_0	5	b_5	Silty loam
b_0	6	b_6	Silt
b_0	7	b_7	Clay loam
b_0	8	b_8	Sandy clay loam
b_0	9	b_9	Silty clay loam
b_0	10	b_{10}	Sandy clay
b_0	11	b_{11}	Silty clay
b_0	12	b_{12}	Clay

3.1. Reference evapotranspiration

The reference evapotranspiration is an important metric to understand the crop water requirements to obtain satisfactory yield. The reference evapotranspirationplays vital role to compute irrigation water requirements. To estimate reference evapotranspirationthe weather data such as temperature (T), wind speed (WS), solar radiation (SR), sunshine hours (SS), relative humidity (RH), rainfall (RF) and vapour pressure (VP) are key input variables (Allen and Pruitt, 1991). The most widely used model for estimation of reference evapotranspirationis FAO-56 Penman-Monteith method (Allen *et al.*, 1998).

$$ET_0 = \frac{0.408\Delta(R_n - G) + \gamma \frac{900}{T+273} U_2 (e_s - e_a)}{\Delta + \gamma (1+0.34U_2)} (1)$$

where,

ET_0 = Reference ET (mm day $^{-1}$),

e_s = Saturation vapor stress (kPa),

e_a = Actual vapor stress (kPa),

Δ = Incline of the saturation vapor stress function (kPa °C^{-1}),

G = Dirt heat flux density (MJm^{-2} day^{-1}),

γ = Psychometric constant (kPa °C^{-1})

R_n = Net radiation (MJm^{-2} day^{-1}),

T = Average air temperature (°C)

U_2 = Mean wind speed at 2 m height for 24-h (m s^{-1}) and

$e_s - e_a$ = Vapor stress loss (kPa).

3.2. Dataset

The observed weather dataset of metrological station, University of Agriculture Sciences, GKVK, Bengaluru is used for prediction of reference evapotranspiration. The Colorado Maize crop evolution water requirement data set is used for prediction of crop coefficient. The soil texture classification dataset is created using USDA triangle soil texture classification reference model.

Table 4: Weather data classification transitions

Current state	Input	Next state	Weather classification
b_1	13	b_{13}	Warm
b_1	14	b_{14}	Temperate
b_1	15	b_{15}	Polar
b_2	13	b_{13}	Warm
b_2	14	b_{14}	Temperate
b_2	15	b_{15}	Polar
b_3	13	b_{13}	Warm

b_3	14	b_{14}	Temperate
b_3	15	b_{15}	Polar
b_4	13	b_{13}	Warm
b_4	14	b_{14}	Temperate
b_4	15	b_{15}	Polar
b_5	13	b_{13}	Warm
b_5	14	b_{14}	Temperate
b_5	15	b_{15}	Polar
b_6	13	b_{13}	Warm
b_6	14	b_{14}	Temperate
b_6	15	b_{15}	Polar
b_7	13	b_{13}	Warm
b_7	14	b_{14}	Temperate
b_7	15	b_{15}	Polar
b_8	13	b_{13}	Warm
b_8	14	b_{14}	Temperate
b_8	15	b_{15}	Polar
b_9	13	b_{13}	Warm
b_9	14	b_{14}	Temperate
b_9	15	b_{15}	Polar
b_{10}	13	b_{13}	Warm
b_{10}	14	b_{14}	Temperate
b_{10}	15	b_{15}	Polar
b_{11}	13	b_{13}	Warm
b_{11}	14	b_{14}	Temperate
b_{11}	15	b_{15}	Polar

b_{12}	13	b_{13}	Warm
b_{12}	14	b_{14}	Temperate
b_{12}	15	b_{15}	Polar

Table 5: Transition table represents crop growth evolution classification

Current state	Input	Next state	Crop evolution
q13	16	q16	initial stage
q13	17	q17	development stage
q13	18	q18	mid-season
q13	19	q19	late season
q14	16	q16	initial stage
q14	17	q17	development stage
q14	18	q18	mid-season
q14	19	q19	late season
q15	16	q16	initial stage
q15	17	q17	development stage
q15	18	q18	mid-season
q15	19	q19	late season

4. Results and Discussions

The DFA irrigation framework is reviewed using Java Formal Languages and Automata Package (JFLAP) tool. (Rodger *et al.,* 2006). The model is validated for the pattern "11316". The variable '1' indicates sandy soil texture, '13' indicates warm weather and '16' indicates initial stage of crop. The tracing of sample pattern amp is represented in Figure 2. The pattern "11316", demands high water supply because of sandy soil texture, warm weather and initial stage of crop. Hence for pattern 1, the model halts at state q16, which indicates high crop-water requirement for the given input condition.

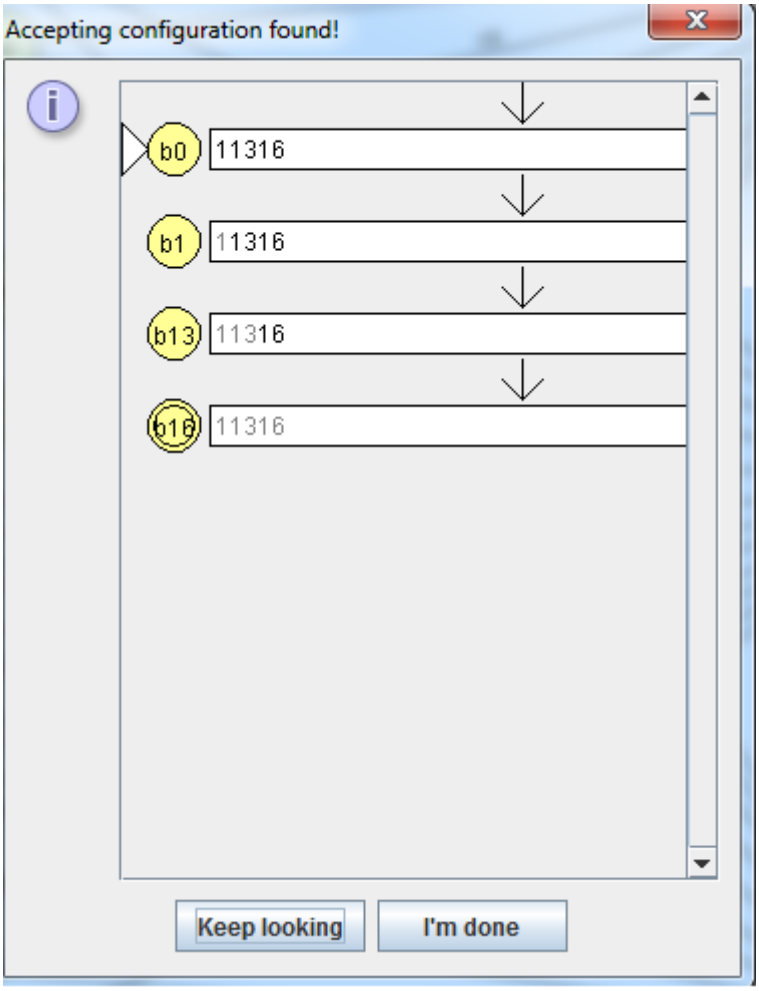


Figure 2: Tracing over input pattern "1 13 16"

4.1. Evapotranspiration prediction

The evapotranspiration is an important metric to understand the crop water requirements to obtain satisfactory yield. The linear regression method is used for prediction of evapotranspiration, which determines water requirement considering weather data. The relevant data instances are reported in Table 6. The models are analyzed using statistical performance measures such as Mean Absolute Error (MAE) and coefficient of correlation (R). The different weather input variable combinations are reported in Table 7. The prediction accuracy is highlighted in Figure 3.

Table 6: Evapotranspiration estimation sample instances

Maximum temperature	Minimum Temperature	Vapor Pressure	Relative Humidity	Wind Speed	Bright Sun Shine Hours	Evapotranspiration
32.4	21.8	17.8	80	4.4	2.6	3.5

32.6	22.0	16.5	84	4.0	7.6	3.9
33.6	22.4	19.1	85	3.8	8.8	4.2
34.2	20.8	19.4	87	4.8	8.0	4.3
31.8	21.2	17.7	82	9.0	9.7	4.5
34.2	21.2	18.8	85	7.7	8.7	4.7

Table 7: Statistical analysis of linear regression model over different input combination

Input combination No.	Input Variables	Statistical Analysis	
		MAE	R^2
1	Max.Temp, Min.Temp, Vapor pressure, Relative humidity, Wind speed, Bright sunshine hours	0.16	0.90
2	Max.temp, Min.Temp, Vapor pressure, Relative humidity, Wind speed	0.27	0.82
3	Max. Temp, Min.Temp, Vapor pressure, Relative humidity	0.27	0.82
4	Max. Temp, Min.Temp, Vapor pressure	0.32	0.74
5	Max. Temp, Min.Temp	0.32	0.73

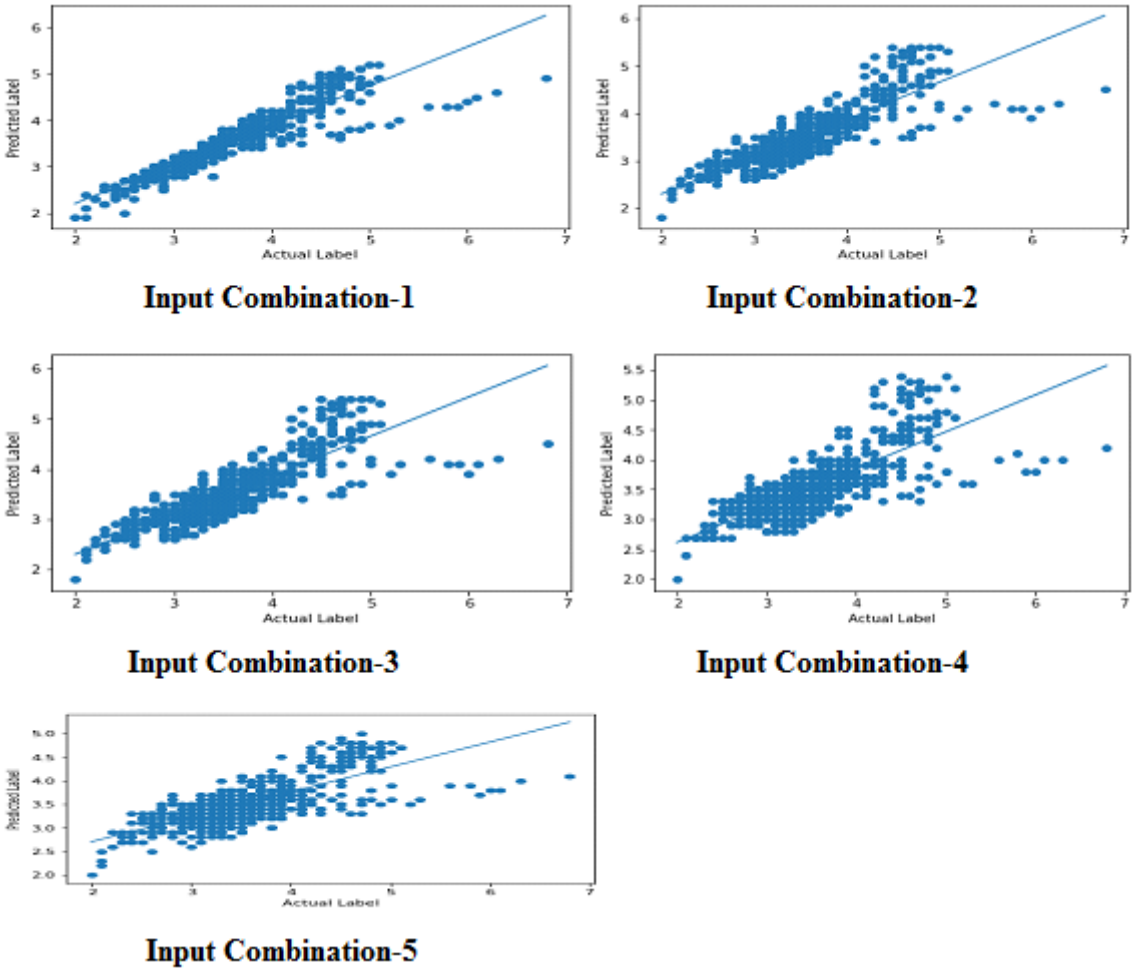


Figure 3: Linear regression model prediction analysis over different input combinations

4.2. Dirt texture and crop-evolution coefficient prediction using K-NN algorithm

The dirt texture determines the water holding capacity of soil and which helps to increase the water productivity of irrigation system. The crop evolution-based coefficient indicates the crop water requirement based on the plant growth stage and it supports for computing water budget in irrigation automation. The K-NN algorithm is applied to predict dirt texture and crop coefficient. The sand, silt and clay fraction are input variables for soil texture classification, which are reported in Table 8. The Maize crop growth stage water requirement is input for crop coefficient prediction. The experimental results exhibited the accuracy of 95.88% over soil texture prediction and 99.98% accuracy over crop coefficient prediction and reported in Table 9.

Table 8: Dirt texture classification sample instances

Sand	Silt	Clay	Type
91	6	3	Sand
50	20	30	Sandy clay loam
15	55	30	Silty clay loam
40	10	50	Clay

Table 9: Dirt texture and crop coefficient estimation accuracy

Algorithm	Prediction	Input	Accuracy
K-NN	Dirt texture	Sand, silt and clay fraction	95.88%
K-NN	Crop-coefficient	Crop growth stage and crop species	99.98%

5. Conclusion

In the proposed research work the finite automata and soft computing concepts are integrated to design a DECEL model to optimize water usage in irrigation management. The automated irrigation framework is proposed using deterministic finite state machine, linear regression and K-NN algorithm. The proposed irrigation automation framework predicts the water requirement considering soil texture, evapotranspiration and weather data. The linear regression model experimental results proved that the best input features combination for prediction of reference evapotranspiration are Max.Temp, Min.Temp, vapor pressure, relative humidity, wind speed and bright sunshine hours. The soil texture class and crop coefficient values are predicted using K-NN algorithm and results exhibited the 95.88% and 99.98% accuracy respectively. As far as we know, the proposed DECEL model is a novel idea, which is designed to increase water productivity in irrigation system. The proposed research work opens the future research on development of efficient intelligent irrigation system and also deployment in the field.

Acknowledgments

We would like to thank Dr. M.S.Sheshshayee, Professor, Department of Crop Physiology, University of Agricultural Sciences, Bengaluru, Karnataka, India for providing the source of weather data set information.

References

Allen, R.G., Pereira, L.S., Raes, D., *et al.*(1998). *Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements.* FAO Irrigation and Drainage Paper No. 56.

Albaji, M., Boroomand-Nasab, S., Naseri, A. and Jafari, S.(2010). Comparison of different irrigation methods based on the parametric evaluation approach in Abbas plain: Iran. *Journal of Irrigation and Drainage Engineering*, **136(2),** 131-136.

Albaji, M., Golabi, M., Nasab, S. B. and Jahanshahi, M.(2014). Land suitability evaluation for surface, sprinkler and drip irrigation systems. *Transactions of the Royal Society of South Africa*, **69(2),** 63-73.

Alcamo, J., Flörke, M. and Märker, M.(2007). Future long-term changes in global water resources driven by socio-economic and climatic changes. *Hydrological Sciences Journal*, **52(2)**, 247-275.

Alexandratos, N. and Bruinsma, J.(2012). *World Agriculture Towards* 2030/2050:*The* 2012*Revision,* **12(3).** FAO, Rome: ESA Working paper.

Arnell, N. W., van Vuuren, D. P. and Isaac, M.(2011). The implications of climate policy for the impacts of climate change on global water resources. *Global Environmental Change*, **21(2)**, 592-603.

Boroomand-Nasab, S., Landi, A., Behzad, M., Tondrow, M., Albaji, M. andJazaieri, A. (2008). Land suitability evaluation for surface, sprinkle and drip irrigation methods in Fakkehplain, Iran. *Journal of Applied Sciences*, **8(20)**, 3646-3653.

Burt, C. M., Clemmens, A. J., Strelkoff, T. S., Solomon, K. H., Bliesner, R. D., Hardy, L. A., Howell, T. A. and Eisenhauer, D. E.(1997). Irrigation performance measures: efficiency and uniformity. *Journal of Irrigation and Drainage Engineering*, **123(6),** 423-442.

Burt, C. M., Mutziger, A. J., Allen, R. G. and Howell, T. A. (2005). Evaporation research: Review and interpretation. *Journal of Irrigation and Drainage Engineering*, **131(1),** 37-58.

Çetin, O. and Kara, A. (2019).Assessment of water productivity using different drip irrigation systems for cotton.*Agricultural Water Management*, **223,** 105693–105693.

Chartzoulakis, K. and Bertaki, M.(2015). Sustainable water management in agriculture under climate change. *Agriculture and Agricultural Science Procedia*, **4,** 88-98.

Dabach, S., Lazarovitch, N., Šimůnek, J. and Shani, U.(2013). Numerical investigation of irrigation scheduling based on soil water status. *Irrigation Science*, **31(1)**, 27-36.

Ghamarnia, H., Arji, I., Sepehri, S., Norozpour, S. and Khodaei, E.(2011). Evaluation and comparison of drip and conventional irrigation methods on sugar beets in a semiarid region. *Journal of Irrigation and Drainage Engineering*, **138(1)**, 90-97.

Jiang, Q., Fu, Q. and Wang, Z. (2011). Study on delineation of irrigation management zones based on management zone analyst software. *Computer and Computing Technologies in Agriculture*, 419-427.

Kang, S., Hao, X., Du, T., Tong, L., Su, X., Lu, H., Li, X., Huo, Z., Li, S. and Ding, R.(2017). Improving agricultural water productivity to ensure food security in China under changing environment: From research to practice. *Agricultural Water Management*, **179**, 5-17.

Krishnashetty, P. H., Balasangameshwara, J., Sreeman, S., Desai, S., & Kantharaju, A. B. (2021). Cognitive computing models for estimation of reference evapotranspiration: A review. *Cognitive Systems Research*, **70**, 109-116.

Levidow, L., Zaccaria, D., Maia, R., Vivas, E., Todorovic, M. and Scardigno, A.(2014). Improving water-efficient irrigation: Prospects and difficulties of innovative practices. *Agricultural Water Management*, **146**, 84-94.

Niu, C. J., Deng, W., Gu, S. X., Chen, G. and Liu, S. S.(2017). Real-time irrigation forecasting for ecological water in artificial wetlands in the Dianchi Basin. *Journal of Information and Optimization Sciences*, **38(7)**, 1181-1196.

Playan, E., Salvador, R., López, C., Lecina, S., Dechmi, F. and Zapata, N.(2013). Solid-set sprinkler irrigation controllers driven by simulation models: Opportunities and bottlenecks. *Journal of Irrigation and Drainage Engineering*, **140(1)**, 04013001.

Pradeep, H. K., Balasangameshwara, J., & Jagdadeesh, P. (2020). Soil moisture based automatic irrigation system to improve water productivity: Automatic irrigation system. *Journal of AgriSearch*, **7(4)**, 220-222.

Pradeep, H. K., Balasangameshwara, J., Sheshshayee, M. S., Rajan, K., & Jagadeesh, P. (2021a). Irrigation Practices and Soft Computing Applications: A Review. *Statistics and Applications*, **19(2)**, 181–198.

Pradeep, H. K., Balasangameshwara, J., Sheshshayee, M. S., Rajan, K., & Jagadeesh, P. (2021b). Automation of USDA Triangle Soil Texture Classification Using Finite State Machine: A Novel Conceptual Modeling Approach. *Statistics and Applications*, **19(2)**, 255 - 266.

Pradeep, H. K., Jagadeesh, P., Sheshshayee, M. S., & Sujeet, D. (2019). Irrigation System Automation Using Finite State Machine Model and Machine Learning Techniques. *International Conference on Intelligent Computing and Communication* (pp. 495-501). Springer, Singapore.

Provenzano, G. and Sinobas, L. R.(2014). Special Issue on trends and challenges of sustainable irrigated agriculture.*Journal of Irrigation and Drainage Engineering*, **140(9)**. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000773

Raghuwanshi, N. S., Saha, R., Mailapalli, D. R. and Upadhyaya, S. K.(2010). Infiltration evaluation strategy for border irrigation management. *Journal of Irrigation and Drainage Engineering*, **137(9)**, 602-609.

Rodger, S. H. andGramond, E. (1998). JFLAP: An aid to studying theorems in automata theory. *ACM SIGCSE Bulletin Inroads*, **30**, 302–302.

Safriel, U., Adeel, Z., Niemeijer, D., Puigdefabregas, J., White, R., Lal, R., Winsolow, M., Ziedler, J., Prince, S., Archer, E. and King, C.(2006). Dryland systems. In *Ecosystems and Human Well-being. Current State and Trends*, **1,** 625-656. Island Press.

Shiri, J., Keshavarzi, A., Kisi, O. and Karimi, S. (2017). Using soil easily measured parameters for estimating soil water capacity: Soft computing approaches. *Computers and Electronics in Agriculture*, **141,** 327-339.

Singh, A., Ganapathysubramanian, B., Singh, A. K. and Sarkar, S. (2016). Machine learning for high-throughput stress phenotyping in plants. *Trends in Plant Science*, **21(2),** 110-124.

Soulis, K. X. and Elmaloglou, S.(2018). Optimum soil water content sensors placement for surface drip irrigation scheduling in layered soils. *Computers and Electronics in Agriculture*, **152**, 1-8.

Tindula, G. N., Orang, M. N. and Snyder, R. L.(2013). Survey of irrigation methods in California in 2010. *Journal of Irrigation and Drainage Engineering*, **139(3)**, 233-238.

Torres-Rua, A. F., Ticlavilca, A. M., Walker, W. R. and McKee, M.(2012). Machine learning approaches for error correction of hydraulic simulation models for canal flow schemes. *Journal of Irrigation and Drainage Engineering*, **138(11)**, 999-1010.

Wu, W., Li, A. D., He, X. H., Ma, R., Liu, H. B. andLv, J. K. (2018). A comparison of support vector machines, artificial neural network and classification tree for identifying soil texture classes in southwest China. *Computers and Electronics in Agriculture*, **144**, 86–93.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 15-25

Symbolic Data Analysis *vs* Classical Data Analysis: A Comparative Study

Dipanka Bora[1] and Hemanta Saikia[2]
[1]Teaching Associate of Agricultural Statistics, FGI College of Agricultural Sciences, Hengbung, Manipur
[2]Hemanta Saikia, Assistant Professor of Statistics, College of Sericulture, Assam Agricultural University, Jorhat-13, Assam

Received: 28 April 2021; Revised: 07 June 2021; Accepted: 09 June 2021

Abstract

A symbolic data set is a combination of symbolic values. The analysis of these symbolic values is known as symbolic data analysis. It is an extension of the standard classical data analysis where symbolic data tables are used as input and symbolic objects are made output as a result. Symbolic data may arise in all branches of science and social science after aggregating a base data set over individual entries that together constitute a category of interest. This study attempts to bring into notice the use of symbolic data analysis and compare its outcome with standard classical data analysis. Different statistical tools have been used for comparative analysis of the symbolic and classical data *viz.* descriptive statistics, covariance, and correlation. To apply these statistical tools in both symbolic and classical data analysis set up, a well-known *Iris* flower data set is being used. The outcome of the study shows that there is a little difference in the results of descriptive statistics for the univariate case between classical data analysis and symbolic data analysis. However, in bivariate statistics computation though the directions of the covariance and correlation values (*i.e.* positive or negative) are the same, yet symbolic data analysis gives comparatively lesser magnitude values than the classical data analysis.

Key words: Data analysis; Descriptive statistics; Interval-valued variables; Symbolic data.

1. Introduction

When we deal with classical data set that data may be either univariate or bivariate or multivariate. In the case of univariate classical data, a single random variable is considered (*e.g.* production of rice). For bivariate data, two random variables (*e.g.* amount of fertilizer and production of wheat) are studied simultaneously in respect of their distribution. Similarly, more than two random variables concerning their distributions for a multivariate classical data are considered (*e.g.* monthly information of temperature, rainfall, humidity, *etc.*). Usually, classical data analysis seeks to describe the descriptive statistics and determine the reliability of inferential statistics. It is based on the repeatedly measured properties of the same objects or only one value per object.

Statistically, classical data on $\boldsymbol{P}$ random variables are represented by a single point (say) in $\boldsymbol{P}$-dimensional space. For instance, the observed values for the random variable $Y = (Y_1, Y_2, \ldots, Y_p)$ for a single individual. This type of data can be analyzed using classical techniques for $n = 150$ observations (say) with $\boldsymbol{P} = 30$ variables. When the size of n becomes

Corresponding Author: Dipanka Bora
Email: dipankabora666@gmail.com

very large (*e.g.* $\boldsymbol{n}$ = 50000 and $\boldsymbol{P}$ = 90), standard classical analysis can be knotty. Again, consider a random variable "Type of Disease ($\boldsymbol{Y}$)" with $\boldsymbol{Y}$ = {Diabetes, High Blood Pressure, Cancer} then a classical response from the respondent could be $\boldsymbol{Y}$ = Diabetes, or $\boldsymbol{Y}$ = High Blood Pressure, or $\boldsymbol{Y}$ = Cancer. It can be noticed that each observation consists of only one value or data point. Now if a respondent has two diseases say Diabetes and High Blood Pressure [*i.e.*$\boldsymbol{Y}$= {Diabetes, High Blood Pressure}] then the typical classical data set format can't accommodate this information. In such situations, symbolic data analysis can be ready to lend a hand. However, there are two possible issues. The first one is how the data set can be prepared to a size that allows analysis to proceed appropriately. The second issue is, to attain the first one, it is essential to consider what we want to learn or extract from the given data set. Symbolic data may arise in all branches of science and social science (*e.g.* from medical, industry, government experiments, and other data collection pursuits) in a variety of different ways. It may arise after aggregating a base data set over individual entries that together constitute a category of interest to the researcher (Diday and Fraiture, 2008). Furthermore, they may arise as an outcome of aggregating very large data set into a smaller manageable sized data set or aggregating into a data set that provides information about categories of interest (Diday and Fraiture, 2008). More specifically, we could say that a symbolic value typically represents the set of individuals who satisfy the description of the associated symbolic concept or category. A symbolic value may include lists, intervals, categories and so on. A more elaborate discussion of symbolic values is provided in the following section.

Symbolic data analysis is an extension of standard data analysis where symbolic data tables are used as input and symbolic objects are made output as a result. The data units are called symbolic since they are more complex than standard ones, as they donot only contain values or categories but also include internal variation and structure (Billard and Diday, 2006). Suppose we have a data set that can be structured like a classical data set. This data can be aggregated to a manageable size and categorized with specializing decision to construct symbolic data sets. Like classical data, symbolic data set can also have three types of variables *viz.* interval-valued, multi-valued and modal variables. The multi-valued variables are the different attributes of the symbolic data set which can have a relation with other variables. Suppose for a field experiment on rice, presence and absence of fungal disease, and the number of spraying for treatment can be jointly considered as multi-valued variables. The interval-valued variable has the maximum and minimum value of the observation, where the values of the observation are varied. The modal variables are multistate variables with a frequency, probability or weight attached to a specific value in the data. Usually, these weights are capacities, creditabilities, necessities or possibilities.

Now let us explain how a symbolic dataset is created from the classical data set. The Table 1 consists of a set of classical data which represents the different varieties of rice along with the information of season, production, tillers per hill, duration and grain size.

As mentioned, a symbolic value may be lists, intervals, categories, *etc.*; from Table 1, the season variable may be considered as a concept to construct a symbolic data set. It could be described by considering rice season (*i.e.* Sali, Ahu, and Bodo) as the concept. The set of seasons is the extent and the different characteristics of rice are the intent. Thus, using the different seasons of rice as a concept a symbolic data table is constructed (*cf.* Table 2). In Table 2, the variable production, number of tillers, and durations are interval-valued variables. For these interval-valued variables, other variables vary within the respective symbolic values. Here except grain size, all the variables are quantitative interval-valued variables. The variable grain size is qualitative *viz.* Big or Small. To transform this qualitative

variable into symbolic values, first, we calculated the ratio of big and small grains corresponding to the rice season. Thereafter, these ratio values are assigned as symbolic values to the variable grain size to make the variable as an interval-valued variable.

Table 1: Classical data set of rice varieties

Variety	Season	Production (kg/ha)	Tillers/hill (number)	Duration (days)	Grain Size
Ranjit	Sali	75	25	155	Big
Kushal	Sali	65	20	150	Big
Satyaranjan	Sali	54	18	135	Small
Lachit	Ahu	42	13	115	Small
Luit	Ahu	36	11	105	Small
Silarai	Ahu	45	15	125	Big
Bishnuprasad	Bodo	66	21	160	Big
Jyotiprasad	Bodo	60	19	170	Big
Joymoti	Bodo	78	26	175	Small

(*Source: Leaflet of Regional Agricultural Research Station, Titabar, Assam Agricultural University, Jorhat*)

The following symbolic data table obtained using the season as a concept from Table 1.

Table 2: Symbolic data set of rice varieties

Season	Production (kg/ha)	Tillers/hill (number)	Duration (days)	Grain Size
Sali	[54,75]	[18,25]	[135,155]	(0.67B, 0.33S)
Ahu	[36,45]	[11,15]	[105,125]	(0.33B, 0.67S)
Bodo	[60,78]	[19,26]	[160,175]	(0.67B, 0.33S)

From the above discussion, we understand that classical values can be qualitative or quantitative. In contrast, symbolic values can be single-valued, interval-valued, and multi-valued with or without logical dependency rules. However, we have especially focused on interval-valued variables in this study. These days the researchers are more acquainted with the classical data and its modeling, so the importance of symbolic data analysis is always quarantined. Therefore, this study attempts to bring into notice the symbolic data analysis and compare its outcome with classical data analysis. Different statistical tools have been used for comparative analysis of symbolic as well as classical data *viz.* descriptive statistics, covariance, and correlation.

2. Review of Literature

Statistical data analysis always plays an important role in determining useful and effective information on real-life situations. In the words of Tukey (1962), data analysis is the "*Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyze data*." In an era of big data, the prominence of symbolic data analysis is

indispensable through which one could summarize big data into smaller data set of manageable size. Several authors have already added some value to the literature through their contribution.

Symbolic objects are the basic elements for knowledge representation in symbolic data analysis (Prediger, 1997). A method of clustering for a set of symbolic data where individuals are described by symbolic variables of various types: interval, categorical multi-valued or modal variables are presented by Brito (2003). Billard and Diday (2003) summarized large datasets into a more manageable size and tried to get maximum knowledge inherent in the entire dataset as much as possible. A similar study was performed by Diday and Esposito (2003) introducing symbolic objects and constitutes an explanatory output for data analysis. Mballo and Diday (2005) studied the reliability of the Kolmogorov-Smirnov criterion to build the decision tree on interval-valued variables to extract symbolic objects from the decision tree and to induced the data table of symbolic objects for higher study of symbolic data analysis. Brito *et al*. (2006) introduced partitioning clustering methods for objects described by interval data. Appice *et al.* (2006) generalized symbolic data analysis aimed at some standard statistical data mining methods, which has developed for classification tasks in the case of symbolic objects. Brito (2007*a*) discussed some issues that arise when trying to apply classical data analysis techniques to symbolic data and addressed the vital question of the measurement of dispersion and also the result of different possible choices in the design of multivariate methods. Diday (2008) observed that databases are now ubiquitous in industrial companies and public administrations and they often grow to an enormous size. In symbolic data analysis, these categorical and numerical are considered to be the new statistical units. The next step is to get these higher-level units and then to describe them by taking care of their internal variation. Domingues *et al.* (2010) introduced a new linear regression method for interval-valued data. Fraiture *et al.* (2011) worked on symbolic data analysis and explained how the classical data models to take into account more complete and complex information. Primental *et al.* (2012) used common tools of symbolic data analysis to reduce the data without losing much information. They used information about researchers of institutions from Brazil through the tools of symbolic data. The main goal was to analyze the scientific production of Brazilian institutions. Brito (2007*b*) worked on modeling and analyzing interval data and discussed some issues that arose when applying classical data analysis techniques to interval data. She put a special focus on the notions of dispersion, association and linear combinations of interval variables and presented some methods that have been proposed for analyzing clustering, discriminant analysis, linear regression and interval time series analysis. Some Indian statisticians also worked on the field of symbolic data analysis. Dinesh *et al.* (2005) studied symbolic data analysis literature revealed that symbolic distance measures are playing a major role in solving pattern recognition and analysis problems. Guru *et al.* (2011) proposed a new model to grade cured tobacco leaves using symbolic data. Doreswamy and Narasegouda (2014) proposed an object-oriented data model using symbolic data analysis which provides a sensor data repository for storing and managing sensor data.

3. Data and Methodology

3.1. Data

The relevant data for comparative analysis between classical and symbolic data analysis is collected from the source *https://en.m.wikipedia.org/wiki/Iris_flower_data_set*. This data set is popularly known as *Iris* flower data set or Fisher's *Iris* data set. It is a multivariate data set introduced by British statistician and biologist Fisher (1936) in his paper entitled "*The*

Use of Multiple Measurements in Taxonomic Problems as an Example of Linear Discriminant Analysis". The data set consists of 50 samples from each of the three species of *Iris viz. Setosa*, *Virginica*, and *Versicolor*. The four different features were measured from each sample and they are the length and width of the sepals and petals of *Iris* flower in centimeters.

3.2. Methodology

Keeping the objective of the study in mind, the formulae of descriptive statistics, covariance and correlation are presented in the subsequent sections for symbolic data. As we believe that the formulae of the above statistical measures for classical data analysis set up are well known to the readers. Also, noticed that we have considered the methodology of symbolic data analysis for interval-valued variables only. The single-valued and multi-valued variables of symbolic data are not considered in this study for comparison.

3.2.1. Descriptive statistics for interval–valued variables

Let us define the interval-valued variable $Y_j = Z$ and the 'Z' contains the interval of 'u' number of observation. For 'u' number of observation of 'Z' interval, the values of $Z(u) = [a_u, b_u]$ for $u \in E = \{1,\ldots.., m\}$. Here a_u is the minimum value, b_u is the maximum value of the specified observation and 'm' is the total number of observations. Now for a interval-valued variable 'Z' the symbolic mean and symbolic variances are calculated by

$$\text{Symbolic Mean}\left(S_m\right) = \frac{1}{2m}\sum_{u\in E}\left(b_u + a_u\right) \tag{1}$$

$$\text{Symbolic Variance}\left(S_v\right) = \frac{1}{3m}\sum_{u\in E}\left(b_u^2 + b_u a_u + a_u^2\right) - \frac{1}{4m^2}\left[\sum_{u\in E}\left(b_u + a_u\right)\right]^2 \tag{2}$$

$$\text{Symbolic Standard Deviation } (SD_j) = \sqrt{S_v} \tag{3}$$

3.2.2. Bivariate statistics for interval–valued variables

Let $Z_1(u)$ and $Z_2(u)$ are two symbolic observations on the space $Z(u) = Z_1(u) \times Z_2(u)$. The $Z_1(u)$ contains interval symbolic variables $[a_{1u}, b_{1u}]$ at 'u' observation and $Z_2(u)$ contains interval symbolic variables $[a_{2u}, b_{2u}]$ at 'u' observation for each $u \in E$. Here 'a_{1u}' is the minimum value and 'b_{1u}' is the maximum value of Z_{1u} interval symbolic variables. Similarly, 'a_{2u}' is the minimum value and 'b_{2u}' is the maximum value of Z_{2u} interval symbolic variables. Now the symbolic covariance function between $Z_1(u)$ and $Z_2(u)$ interval-valued symbolic variables is defined as

$$Cov\left(Z_1, Z_2\right) = \frac{1}{4m}\left\{\sum_{u\in E}\left(b_{1u} + a_{1u}\right)\left(b_{2u} + a_{2u}\right)\right\} - \frac{1}{4m^2}\left\{\sum_{u\in E}\left(b_{1u} + a_{1u}\right)\right\}\left\{\sum_{u\in E}\left(b_{2u} + a_{2u}\right)\right\} \tag{4}$$

Once we have a covariance function, we can easily calculate the symbolic correlation between the interval–valued variables Z_1 and Z_2. It is defined as

$$r\left(Z_1, Z_2\right) = \frac{Cov\left(Z_1, Z_2\right)}{\sqrt{Var\left(S_{v1}\right)} \times \sqrt{Var\left(S_{v2}\right)}} \tag{5}$$

where $r(Z_1, Z_2)$ is the symbolic correlation between interval-valued variables Z_1 and Z_2, variance (S_{v1}) is the symbolic variance of $Z_1(u)$ and variance (S_{v2}) is the symbolic variance of $Z_2(u)$.

4. Results and Discussion

A classical data set is a group of contents of a single database table where every column of the table represents a particular variable and each row corresponds to a given member of the data set. On the other hand, the symbolic data set is a combination of symbolic values *viz.* intervals, lists, categories and so on. We have constructed the symbolic data set from the classical data set in case of the univariate and bivariate case using the concept of interval-valued symbolic values. To do that statistical language *R* is being used with package *RSDA* (*cf.* Appendix-A) and the corresponding symbolic data is presented in Table 3.

Table 3: Symbolic data set of *Iris* flower data

Species	Sepal Length $[Z_1(u)]$	Sepal Width $[Z_2(u)]$	Petal Length $[Z_3(u)]$	Petal Width $[Z_4(u)]$
	$[a_{1u}, b_{1u}]$	$[a_{2u}, b_{2u}]$	$[a_{3u}, b_{3u}]$	$[a_{4u}, b_{4u}]$
Setosa	[4.3, 5.8]	[2.3, 4.4]	[1.0, 1.9]	[0.1, 0.6]
Versicolor	[4.9, 7.0]	[2.0, 3.4]	[3.0, 5.1]	[1.0, 1.8]
Virginica	[4.9, 7.9]	[2.2, 3.8]	[4.5, 6.9]	[1.4, 2.5]

Table 3 represented the symbolic data set of the classical *Iris* flower data set, which is having interval-valued variables. The variables *viz.* Sepal length, Sepal width, Petal Length and Petal width in the table contain [minimum value, maximum value] corresponding to the number of observations or species. In this symbolic data set, the factor species is considered as a concept and accordingly we have four symbolic variables are $Z_1(u)$, $Z_2(u)$, $Z_3(u)$ and $Z_4(u)$.

4.1. Univariate statistics of classical and symbolic data for *Iris* flower data set

The descriptive statistics for classical data analysis of *Iris* flower data set are presented to compare with the symbolic data analysis. We have used the usual statistical tools to calculate the descriptive statistics presented in Table 4.

Table 4: Descriptive statistics of *Iris* flower data set for classical data analysis

Descriptive Statistics	Sepal Length (Y_1)	Sepal Width (Y_2)	Petal Length (Y_3)	Petal Width (Y_4)
Mean	5.8433	3.0573	3.7580	1.1933
Variance	0.6856	0.1899	3.1162	0.5810
Standard deviation	0.8280	0.4358	1.7653	0.7622

Likewise, to calculate the descriptive statistics for symbolic data analysis of *Iris* flower data set, some of the basic and essential computations we have to perform at first. Thereafter, the symbolic mean (S_m), symbolic variance (S_v) and symbolic standard deviation (S_{sd}) are calculated using the formula given in equation (1), (2) and (3) respectively. In Table 5, the total values for each of the four variables are obtained by $\sum_{i=1}^{4}(b_{iu} + a_{iu})$ and

$\sum_{i=1}^{4}\left[b_{iu}^2+\left(b_{iu}\times a_{iu}\right)+a_{iu}^2\right]$. Now if we look at the values of descriptive statistics from both the classical and symbolic data analysis table, the values are not differing too much. It means that symbolic data analysis gives almost the same mean, variance and standard deviation values which we have computed from the classical data analysis for *Iris* flower data.

Table 5: Descriptive statistics of *Iris* flower data set for symbolic data analysis

Sepal Length [$Z_1(u)$]	a_{1u}	b_{1u}	$b_{1u}+a_{1u}$	$[b_{1u}^2+(a_{1u}\times b_{1u})+a_{1u}^2]$	S_m	S_v	S_{sd}
Setosa	4.3	5.8	10.1	77.07	5.8	0.75	0.86
Versicolor	4.9	7.0	11.9	107.31			
Virginica	4.9	7.9	12.8	125.13			
Total			34.8	309.51			
Sepal Width [$Z_2(u)$]	a_{2u}	b_{2u}	$b_{2u}+a_{2u}$	$[b_{2u}^2+(a_{2u}\times b_{2u})+a_{2u}^2]$	S_m	S_v	S_{sd}
Setosa	2.3	4.4	6.7	34.77	3.02	0.32	0.56
Versicolor	2.0	3.4	5.4	22.36			
Virginica	2.2	3.8	6.0	27.64			
Total			18.1	84.77			
Petal Length [$Z_3(u)$]	a_{3u}	b_{3u}	$b_{3u}+a_{3u}$	$[b_{3u}^2+(a_{3u}\times b_{3u})+a_{3u}^2]$	S_m	S_v	S_{sd}
Setosa	1.0	1.9	2.9	6.51	3.73	3.37	1.83
Versicolor	3.0	5.1	8.1	50.31			
Virginica	4.5	6.9	11.4	98.91			
Total			22.4	155.73			
Petal Width [$Z_4(u)$]	a_{4u}	b_{4u}	$b_{4u}+a_{4u}$	$[b_{4u}^2+(a_{4u}\times b_{4u})+a_{4u}^2]$	S_m	S_v	S_{sd}
Setosa	0.1	0.6	0.7	0.43	1.23	0.49	0.76
Versicolor	1.0	1.8	2.8	6.04			
Virginica	1.4	2.5	3.9	11.71			
Total			7.4	18.18			

4.2. Bivariate statistics of classical and symbolic for *Iris* flower data set

For comparing the bivariate statistics between classical and symbolic data set of *Iris* flower data, we computed the covariance and correlation between the variables. The results are presented in the following tables.

Table 6: Bivariate statistics of *Iris* flower data set for classical data analysis

Bivariate Statistics	Y_1Y_2	Y_1Y_3	Y_1Y_4	Y_2Y_3	Y_2Y_4	Y_3Y_4
Covariance	-0.0422	1.2658	0.5123	-0.3275	-0.1205	1.2854
Correlation	-0.1176	0.8718	0.8179	-0.4284	-0.3654	0.9627

The symbolic covariance between Sepal Length $Z_1(u)$ and Sepal Width $Z_2(u)$, Sepal Length $Z_1(u)$ and Petal Length $Z_3(u)$, *etc.* for interval-valued symbolic variables are obtained

by using the equation (4). Likewise, we have calculated all the possible symbolic covariance and correlation between the variables and the results can be seen in Table 7.

In Table 7, column-wise the $Z_1(u) \times Z_2(u)$ represents the symbolic covariance (-0.1025) and symbolic correlation (-0.2097) of Sepal Length and Sepal Width, the $Z_1(u) \times Z_3(u)$ represents the symbolic covariance (0.98) and symbolic correlation (0.6168) of Sepal Length and Petal Length, *etc.* Now let us compare the results of classical data analysis (*cf.* Table 6) and symbolic data analysis (*cf.* Table 7) from the computation of bivariate statistics. It is observed that though the directions of the values (*i.e.* positive or negative) are the same in covariance and correlation, the symbolic data analysis gives comparatively lesser magnitude values than classical data analysis. This is a result of the loss of information from the data in every step of processing. In symbolic data analysis for interval-valued variables, we deal with the maximum and minimum values of the data set instead of considering all the values like in classical real-valued data. Thus, the covariance and correlation results from symbolic data analysis have revealed lesser magnitude values in comparison to the classical data analysis results. If this is factual then the higher level of statistical analysis like multiple regression, clustering, factor analysis, *etc.* based on symbolic data might mislead the researchers to draw an appropriate inference from the data.

Table 7: Bivariate statistics of *Iris* flower data set for symbolic data analysis

Bivariate Statistics	$Z_1(u) \times Z_2(u)$	$Z_1(u) \times Z_3(u)$	$Z_1(u) \times Z_4(u)$	$Z_2(u) \times Z_3(u)$	$Z_2(u) \times Z_4(u)$	$Z_3(u) \times Z_4(u)$
Covariance	–0.1025	0.9800	0.3725	–0.2981	–0.1197	1.1597
Correlation	–0.2097	0.6168	0.6090	–0.2878	–0.3003	0.8950

For both univariate and bivariate statistics, we have computed mean, variance, and standard deviation, covariance and correlation. Though the simple linear regression analysis is not performed, the information available from univariate and bivariate statistics can easily be attained considering Y as the dependent variable and X as an independent variable in terms of classical set up like

$$\left(Y - \overline{Y}\right) = r \frac{S_y}{S_x}\left(X - \overline{X}\right) \tag{6}$$

Based on the equation (6), the equivalent symbolic linear regression equation between Sepal Length ($Z_1(u)$) and Petal Length ($Z_3(u)$) can easily be fitted. Let us consider that Sepal Length is dependent on Petal Length and then it is defined as

$$[Z_1(u) - \overline{Z}_1(u)] = r\left(Z_1, Z_3\right)\frac{S_{v1}}{S_{v3}}[Z_3(u) - \overline{Z}_3(u)] \tag{7}$$

where $r(Z_1, Z_3)$ represents the correlation and S_{v1} and S_{v3} represents the standard deviation of Sepal Length ($Z_1(u)$) and Petal Length ($Z_3(u)$) respectively.

5. Conclusion

The extension of classical exploratory data analysis to the analysis of interval-valued symbolic data raises a few pertinent questions. How to compute dispersion precisely for different types of symbolic data (*i.e.* single-valued and multi-vaued)? How to define linear combinations between the symbolic variables? Whether the properties which we have usually

considered for classical real-valued data in terms of linear models will be valid for symbolic data? It is because the way to assess the central tendency and dispersion of symbolic data is not comparable with classical real-valued data. There may be some alternatives to attain these questions which we need to explore for greater attention of the researchers across the globe. However, the choice of an alternative way shall depends on the type of symbolic data to be used subsequently. Regarding interval-valued symbolic data, the one important issue is application of statistical models. Without statistical modeling, parameter estimation and testing of hypothesis are not possible. So the challenge is in front of the researchers who wants to explore symbolic data analysis beyond the classical framework of real-valued data. In today's era of big data, where data storage and analytics is a big challenge, the exploration of symbolic data analysis in solving the problem of big data may open a new window in front of the researchers. Furthermore, the development of appropriate user-friendly statistical software to analyze the symbolic data will go a long way in tackling the challenges posed by big data.

References

Appice, A., Amato, C. D., Esposito, F. and Malebra, D. (2006). Classification of symbolic objects: A lazy learning approach. *Intelligent Data Analysis*, **10**(**4**), 301-324.

Arroyo, J. and Mate, C. (2009). Forecasting histogram time series with K-nearest neighbours methods. *International Journal of Forecasting*, **25**(**1**), 192-207.

Billard, L. and Diday, E. (2003). Symbolic Data Analysis: Definitions and examples. *Technical Report*, 62 pages, at (http://www.stat.uga.edu/faculty/LYNNE/Lynne.html).

Billard, L. and Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic data Analysis. *Journal of the American Statistical Association*, **98**(**462**), 470-487.

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis, Conceptual Statistics and Data Mining.* West Sussex, England: John Wiley & Sons Ltd.

Brito, P., De Carvalho, F. A. T. and Bock, H. H. (2006). Dynamic clustering for interval data based on L2 distance. *Computational Statistics*, **21**(**2**), 231-250.

Brito, P. (2003). Hierarchical and pyramidal clustering for symbolic data. *Journal of the Japanese Society of Computational Statistics*, **15**(**2**), 231-244.

Brito, P. (2007*a*). *On the Analysis of Symbolic Data*. Brito, P., Cucumel, G., Bertrand, P. and Carvalho, F. de. (Eds.) *Selected Contributions in Data Analysis and Classification* (pp. 13-22). Springer Nature, Switzerland.

Brito, P. (2007*b*). *Modelling and Analysing Interval Data*. Springer, Decker, R. and Lenz, H. J. (Eds.) (2006). *Advances in Data* Analysis. Berlin, Heidelberg.

Cury, A., Diday, E. and Cremona, C. (2010). Application of symbolic data analysis for structural modification assessment. *Engineering Structures*, **32**(**3**), 762-775.

Diday, E. and Esposito, F. (2003). An introduction to symbolic data analysis and the SODAS software. *Intelligent Data Analysis*, **7**(**6**), 583-601.

Diday, E. and Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*. John Wiley and Sons Ltd.: West Sussex, England.

Diday, E. and Vrac M. (2005). Mixture decomposition of distributions by copulas in the symbolic data analysis framework. *Discrete Applied Mathematics*, **147**(**1**), 27-41.

Dinesh, M. S., Gowda, K. C. and Nagabhusan, P. (2005). *Fuzzy-Symbolic Analysis for Classification of Symblic Data*. S. K. Pal (Ed.) *Pattern Recognition and Machine Intelligence: First International Conference* (pp. 338-343). Kolkata, India: Springer.

Diday, E. (2008). *Symbolic Data Analysis and the SODAS Software.* Belgium: John Wiley & Sons, Ltd.

Domingues, M. A. O, De Souza, R. M. C. R. and Cysneiors, F. J. A. (2010). A robust method for linear regression of symbolic interval data. *Pattern Recognition Letters*, **31**(**13**), 1991-1996.

Doreswamy and Narasegouda S. (2014). Symbolic data analysis for development. In *Proceeding of the International Conference on Frontiers of the Intelligent Computing: Theory and Applications* (FICTA-2013), Switzerland: Springer, 435-442.

Esposito, F., Malebra, D. and Lisi, F. (1998). Flexible matching of boolean symbolic objects. *Proc. NTTS*, **98**.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis. *Annals of Eugenics*, **7**(**2**), 179-188.

Fraiture, M. N. and Brito, P. (2011). Far beyond the classical data models: Symbolic data analysis. *Statistical Analysis and Data Mining*, **4**(**2**), 157-170.

Guru, D. S., Mallikarjuna, P. B. and Shenoi, M. M. (2011). Min-max representation of features for grading cured tobacco leaves. *Statistics and Applications*, **9**(**1&2**),15-29.

Kaytoue, M., Kuznetsov, S. O., Napoli, A. and Polaillon, G. (2011, September). Symbolic data analysis and formal concept analysis. In XVIIIeme rencontres de la Société Francophone de Classification-SFC 2011.

Lauro, C. N. and Palumbo, F. (2000). Principal component analysis of interval data: a symbolic data analysis approach. *Computational Statistics*, **15**(**1**), 73-87.

Mballo, C. and Diday, E. (2005). Decision trees on interval-valued variables. *The Electronic Journal of Symbolic Data Analysis*, **3(1)**, 1723-5081.

Nagabhusan, P. and Kumar, R. P. (2007). *Histogram PCA*. Liu Derong, H. Z. (Ed.) *Advances in Neural Networks - ISNN 2007* (pp. 1012-1022). Nanjing, China, Springer.

Prediger, S. (1997). *Symbolic Objects in Formal Concept.* Darmstadt: Frachbereich Mathematik, Technische Hochschulle.

Primental, B. A., Nobrega, J. P. and De Souza R. M. C. R. (2012). *Using Weighted Clustering and Symbolic Data to Evaluate Institutes's Scientific Production*. In Proceedings of International Conference on Artificial Neutral Networks (pp. 435-442). Springer, Berlin, Heidelberg.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, **33**(**1**), 1-67.

Appendix-A

Classical Data to Symbolic Data in *R* using the Package '*RSDA*'

```
library(RSDA)

S=classic.to.sym(data=iris, concept="Species", variables = c(Sepal.Length, Sepal.Width,
Petal.Length, Petal.Width))

S       # to get the symbolic output
```

Appendix-B

Some Basic Computation of Symbolic Data Analysis for Bivariate Statistics

Table B1: Procedure to calculate the symbolic covariance ($Z_i(u)$)

Species	$(b_{1U} + a_{1U}) \times (b_{2U} + a_{2U})$	$(b_{1U} + a_{1U}) \times (b_{3U} + a_{3U})$	$(b_{1U} + a_{1U}) \times (b_{4U} + a_{4U})$	$(b_{2U} + a_{2U}) \times (b_{3U} + a_{3U})$	$(b_{2U} + a_{2U}) \times (b_{4U} + a_{4U})$	$(b_{3U} + a_{3U}) \times (b_{4U} + a_{4U})$
Setosa	67.67	29.29	7.07	19.43	4.69	2.03
Versicolor	64.26	96.39	33.32	43.74	15.12	22.68
Verginica	76.80	145.92	49.92	68.40	23.40	44.46
Total	208.73	271.6	90.31	131.57	43.21	69.17

The symbolic covariance between Sepal Length ($Z_1(u)$) and Sepal Width ($Z_2(u)$) is calculated using the equation (4).

$$Cov(Y_1, Y_2) = \frac{1}{4\times 3}\{208.73\} - \frac{1}{4\times 3^2}[(34.8)\times(18.1)] = -0.1025 \quad \text{(B1)}$$

where, $\sum(b_{1u} + a_{2u})\times(b_{3u} + a_{4u}) = 208.73$, $\sum(b_{1u} + a_{1u}) = 34.8$ & $\sum(b_{2u} + a_{2u}) = 18.1$

Now the symbolic correlation (*cf.* equation (5)) between Sepal Length ($Z_1(u)$) and Sepal Width ($Z_2(u)$) is calculated by

$$r(Y_1, Y_2) = \frac{Cov(Y_1, Y_2)}{\sqrt{Variance(S_{v1})\times Variance(S_{v2})}} \quad \text{(B2)}$$

$$r(Y_1, Y_2) = \frac{-0.0125}{\sqrt{0.7500\times 0.3186}} = -0.2097$$

Similarly, we have calculated all the possible covariance and correlations among the variables for *Iris* flower data set using symbolic data analysis and the results are presented in Table 7.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 27-40

Identifying the Time of a Permanent Shift in the Normal Process Mean with Memory Type Control Chart

R. A. Kapase and V. B. Ghute
Department of Statistics, Punyashlok Ahilyadevi Holkar Solapur University, Solapur, India.

Received: 20 March 2021; Revised: 03 June 2021; Accepted: 09 June 2021

Abstract

Control chart is a valuable statistical process control (SPC) tool used for monitoring the process performance. When control chart gives the out-of-control signal, the search initiates to identify the sources responsible for the special cause of variation. But control chart does not give the exact time when the process change begun. The time when the process change appears first in the process called change point. Knowing the change point in the process helps to identify the special cause of variation. This article discusses the approach based on the maximum likelihood estimator of process change to identify the time of a permanent shift in the normal mean with EWMA and MA control charts.

Key words: Statistical quality control; Change point; maximum Likelihood; EWMA control chart; Moving average control chart; Average run length.

1. Introduction

Control charts distinguish between the special cause of variation and the common cause of the variation in the process. To improve and control the process control charts are widely used in the manufacturing industries. Once control chart issues a signal that the special cause is present in the process. Process professionals should initiate a search for the special cause of the variation which could be quite difficult. The search depends on the professional's knowledge and experience. To quality improvement it is necessary to bring the process back into the statistical control. One essential step would help to quality improvement is that knowing the starting time the special cause of variation appears first in the process. Once it is possible to identify the exact time when the process happens due to special cause of variation appears first in the process, there may not be delay finding the occurrence of the special cause of the variation in the process. As a result, the special cause of variation can be identified more quickly, and the corrective action can be taken to eliminate the sources of the special cause of variation which leads to process improvement.

In recent years change point estimation in control charts has received a great deal of attention, as the change point estimation procedure simplify the effort to search for and identify special causes in statistical process monitoring. Hinkley (1970) considered inference about the point in a sequence of random variables at which the probability distribution changes. They compared asymptotic distribution of the MLE and likelihood ratio statistic with some finite sample distributions. Samuel *et al.* (1998a) proposed a method of maximum likelihood estimator to identify the time of step change in the normal mean with $\overline{X}$ control chart. Samuel *et al*. (1998b) considered the step change in the normal process variance.

Corresponding Author: R. A. Kapase
Email: rakapase@sus.ac.in

Samuel and Pignatiello (1998) estimated the change point in the rate parameter of the Poisson process. Nedumaran *et al.* (2000) considered the time of the step change in the multivariate process with chi-square control chart. Pignatiello and Samuel (2001a) considered the change point in the normal process mean in SPC applications. Pignatiello and Samuel (2001b) estimated the step change point in the process fraction nonconforming. They have estimated MLE of a change point when a step change occurred in the fraction nonconforming. Park and Park (2004) considered the time of step change in the normal process mean and variance when $\overline{X}$ and S control charts used simultaneously. Khoo (2004) determined the permanent shift in the process mean with CUSUM control chart. Perry *et al.* (2005) estimated the time of step change in the rate parameter of the Poisson distribution with linear trend and monotonic change, respectively. Fahmy and Elsayed (2006) estimated the maximum likelihood estimator of the change point when Shewhart control chart is used under linear trend disturbance. Perry and Pignatiello (2006) estimated the time of a linear trend change in the normal process mean. Perry *et al.* (2007) considered the monotonic change in the non-conformity level p, when the process is modeled by binomial distribution. Gazanfari *et al.* (2008) used clustering approach to identify the time of a step change in Shewhart control charts. Noorossana et al. (2009) estimated the step change point in the process non-conformity proportion when process is modelled by geometric distribution. Dogu and Kocakoc (2011) proposed change point model for generalized variance control chart. Zandi *et al.* (2011) estimated MLE of a change point for a linear trend disturbance in the process non-conformity.

There are many situations in which the sample size used for process monitoring is one (Montgomery (2012)). An individual control chart is usually used to monitor shifts in the process mean when it is not possible to form subgroups. Shewhart individual X chart have been extensively used in monitoring the process mean. The main drawback of Shewhart X chart is that it uses only information of the last sample observation and ignores the information of the process which makes it insensitive to small shifts in process mean. An alternative to detect small shifts is to use the memory type chart as like Cumulative sum (CUSUM) chart, exponentially weighted moving average (EWMA) chart or moving average (MA) chart. These charts consider the past as well as current information about the process, which makes charts very sensitive to small shifts in process parameters. Relative to CUSUM chart, the EWMA and MA charts are quite basic. The EWMA chart uses a weighted average as the chart statistic while the time weighted MA chart is based on simple moving averages. Kapase and Ghute (2018) estimated the time of a step change in the normal process mean with Tukey's control chart and individual X control chart and compared both the control charts in detecting the occurrence of the special cause in the process.

In this paper, we describe the application of change point estimators to memory type control charts namely EWMA and MA control charts based on individual observations using an approach developed by Samuel *et al.* (1998a). The remainder of this paper is organized as follows: In Section 2, change point estimation procedure is given. Section 3 provides the details of EWMA and MA charts. In Sections 4 and 5, we analyze the performance of the change point estimator for EWMA and MA control charts respectively. Section 6 provides numerical examples to demonstrate the use of estimator each with EWMA control chart and MA control chart. It is shown that change point estimator works well with EWMA and MA control charts. Some conclusions are given in Section 7.

2. Change Point Estimator

It is assumed that the process initially is in-control with a known value of mean μ_0 and variance σ_0^2. Following an unknown point in time a change in the process mean occurs from μ_0 to an out-of-control state mean $\mu_1 = \mu_0 + \delta\sigma_0/\sqrt{n}$ where n is the subgroup size and δ is the unknown magnitude of the change. Here we consider the case of individual observations (Subgroup size $n=1$). It is assumed that σ_0 does not change while shift occurs in μ_0. It is also assumed that once this step change in the process mean occurs, the process remains at the new level of μ_1 until special cause has been identified and eliminated.

We will consider the process move to the out-of-control state at observation *T*. This out of signal can be obtained when a point is plotted beyond the control limits. Assuming this is not a false alarm, this is the point at which process professionals should initiate a search for special cause of variation. Let $X_1, X_2, ..., X_\tau$ be the observations from the in-control process, while $X_{\tau+1}, X_{\tau+2}, ..., X_T$ are the observations when the process changed, so that τ is the point where process change happened. This point τ is point in time when the shift in the process mean appears for first time and then process gets changed. Identifying this point in time when process change appears for first time the change point estimator works uniquely.

The data with subgroup size $n=1$, the change point estimator is derived based on the method of the maximum likelihood estimator (Samuel *et al*. (1998) and Khoo (2004)).We denote the MLE of the change point τ as $\hat{\tau}$. For given single observations the MLE of τ is the value of τ which maximizes the logarithm of the likelihood function. The probability density function of the observation X, which follows normal distribution with mean μ and variance σ^2.

$$f(x,\mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma}(x-\mu)^2}, \quad -\infty < x < \infty$$

The likelihood function (apart from constant) is

$$L(\tau, \mu_1 / x) = \prod_{i=1}^{\tau} e^{-(x_i-\mu_0)/2\sigma_0^2} + \prod_{i=\tau+1}^{T} e^{-(x_i-\mu_0)/2\sigma_0^2}$$

$$\log L(\tau, \mu_1 \mid x) = -\frac{1}{2\sigma_0^2}\left[\sum_{i=1}^{\tau}(x_i-\mu_0)^2 + \sum_{i=\tau+1}^{T}(x_i-\mu_1)^2\right]$$

$$= -\frac{1}{2\sigma_0^2}\left[\sum_{i=1}^{T} x_i^2 - 2\mu_0\sum_{i=1}^{\tau} x_i + \tau\mu_0^2 - 2\mu_1\sum_{i=\tau+1}^{T} x_i + (T-\tau)\mu_1^2\right] \tag{1}$$

There are two unknown parameters τ and μ_1 in the equation (1). If the change point τ is known the MLE of μ_1 is $\overline{X}_{T,t} = \sum_{i=\tau+1}^{T} \frac{X_i}{T-\tau}$.

Substituting this in the equation (1) follows:

$$\log L(\tau, \mu_1 \mid x) = -\frac{1}{2\sigma_0^2}\left[\sum_{i=1}^{T} x_i^2 - 2\mu_0 \sum_{i=1}^{\tau} x_i + \tau \mu_0^2 - \frac{2(\sum_{i=\tau+1}^{T} x_i)^2}{T-t} + \frac{(\sum_{i=\tau+1}^{T} x_i)^2}{T-t}\right]$$

$$= -\frac{1}{2\sigma_0^2}\left[\sum_{i=1}^{T} x_i^2 - 2\mu_0 \sum_{i=1}^{\tau} x_i + \tau \mu_0^2 - (T-t)\,\bar{X}_{T,t}^2\right]$$

$$= -\frac{1}{2\sigma_0^2}\left[\sum_{i=1}^{T} x_i^2 - 2\mu_0 \sum_{i=1}^{T} x_i + T\mu_0^2 - (T-\tau)(\bar{X}_{T,t} - \mu_0)^2\right]$$

It follows that the value of τ which maximizes the log-likelihood function is

$$\hat{\tau} = \underset{0 \le t < \mathrm{T}}{\arg\max}\left\{(T-t)(\bar{X}_{T,t} - \mu_0)^2\right\}$$

3. EWMA and MA Control Charts

Assume that $X_1, X_2, X_3, \ldots$ denote independent and identically distributed observations with an in-control mean μ_0 and standard deviation σ_0. We assume that both the parameters are known. In practice, μ_0 and σ_0 are estimated from the observed historical data. The EWMA control statistic for individual observations is defined as

$$Z_i = \lambda X_i + (1-\lambda) Z_{i-1}, \qquad \text{for } i = 1, 2, \ldots \text{ and } Z_0 = \mu_0,$$

where X_i is the current observation and $0 < \lambda \le 1$ is the smoothing parameter.
The exact control limits for the EWMA chart are

$$LCL = \mu_0 - L\sigma_0 \sqrt{\frac{\lambda}{(2-\lambda)}[1-(1-\lambda)^{2i}]}$$

$$UCL = \mu_0 + L\sigma_0 \sqrt{\frac{\lambda}{(2-\lambda)}[1-(1-\lambda)^{2i}]}, \quad \text{for } i = 1,2,\ldots$$

where $L > 0$ determines the width of the control limits.

The moving average statistic of span w at time i for a sequence of observations iscomputed as

$$M_i = \frac{X_i + X_{i-1} + \ldots + X_{i-w+1}}{w}, \text{ for } i \ge w$$

For periods $i < w$, we compute the average of available observations. In other words, average of all observations up to period i defines moving average.
The control limits for the moving average control chart are as follows:

$$UCL / LCL = \mu_0 \pm \frac{3\sigma_0}{\sqrt{w}} \text{ for } i \ge w \text{ and}$$

$$UCL / LCL = \mu_0 \pm \frac{3\sigma_0}{\sqrt{i}} \text{ for } i < w$$

4. Change Point Estimator Used with EWMA Control Chart

We consider EWMA control chart to study the performance of the estimator. We used Monte Carlo simulation to study the performance of the change point estimator. The change point of the process is simulated at observation $\tau = 100$ .The first 100 individual observations are randomly generated from standard normal distribution. Then starting from observation 101, the individual observations are randomly generated from changed process with normal distribution with mean δ and standard deviation 1 until the EWMA chart gives an out-of-control signal. At this point $\hat{\tau}$ is computed. This procedure is repeated a total number of $N = 10000$ times for each of the values of $\delta = 0.5, 1.0, 1.5, 2.0, 2.5, 3.0$ with different values of parameters $L = 2.86, \lambda = 0.2$ which has in-control $ARL_0 = 370.37$, same as the Shewhart control chart and $(L = 2.615, \lambda = 0.05), (L = 2.814, \lambda = 0.1), (L = 2.998, \lambda = 0.25)$ have in-control ARL_0 = 500. The average of the estimates obtained using the estimator from the 10,000 simulation runs is computed with their standard error.

Tables 1-4 show $E(T)$, the expected number of observations at which the control chart signals a change in the process mean that occurred at time 100. Thus, $E(T) = ARL + 100$. We show that $\bar{\hat{\tau}}$ the average change point estimate obtained using MLE change point estimator with their standard error.

Table 1: Average change point estimates for δ and standard errors when used with a EWMA control chart, $ARL_0 = 500, L = 2.615, \lambda = 0.05, \tau = 100$ and $N = 10000$ independent simulation trials

δ	0.5	1.0	1.5	2.0	2.5	3.0
$E(T)$	123.17	107.16	103.73	102.40	101.75	101.40
$\bar{\hat{\tau}}$	104.10	99.84	99.87	99.86	99.85	99.88
$s.e.(\bar{\hat{\tau}})$	0.2086	0.0532	0.0275	0.0161	0.0111	0.0084

In Table 1, we can see that the expected number of observations required to detect the change in the process mean for the magnitude of the shift $\delta = 1.0$ is 107.16. The average change point estimate is 99.84. Since the change point is simulated at point 100, the average change point estimate should be close to 100. For the magnitude of the shift $\delta = 2.0$ the control chart issues signal at 102.40, that of average change point estimate is 99.86 which is close to 100.

In Table 2, we can see that the control chart gives out of control signal at 108.16 on an average of 10000 simulation trial for the magnitude of the shift $\delta = 1.0$. The average change point estimate is 99.54 which is close to 100. For the magnitude of the shift $\delta = 2.5$ the control chart issues signal at 101.92 and that of average change point estimate is 99.90. For $\delta = 3.0$, $E(T) = 101.50$ and that of the average change point estimate is 99.92 which is again close to 100.

In Table 3, we can see that for the magnitude of the shift $\delta = 1.0$ the expected number of observations at the point when control chart gives out of control signal at 110.36 on an average of 10000 simulation trial. The average change point estimate is 100.09 which are close to 100. For the magnitude of the shift $\delta = 2.5$ the control chart issues signal at 102.08 and that of average change point estimate is 99.94. For $\delta = 3.0$, $E(T) = 101.62$ and that of the average change point estimate is 99.92 which is again close to 100.

In Table 4, we can see that for the magnitude of the shift $\delta = 1.0$ the expected number of observations at the point when control chart gives out of control signal at 108.80 on an average of 10000 simulation trial. The average change point estimate is 99.85 which are close to 100. For the magnitude of the shift $\delta = 1.5$ the control chart issues signal at 104.32 on an average of the 10000 simulation trials and that of average change point estimate is 99.94. For $\delta = 3.0$, $E(T) = 101.53$ and that of the estimated change point is 99.93 on an average of total 10000 simulation trial which is again close to 100.

Table 2: Average change point estimates for δ and standard errors when used with a EWMA control chart, $ARL_0 = 500, L = 2.814, \lambda = 0.1, \tau = 100$ and $N = 10000$ independent simulation trials

δ	0.5	1.0	1.5	2.0	2.5	3.0
$E(T)$	128.61	108.16	104.13	102.64	101.92	101.50
$\bar{\hat{\tau}}$	105.45	99.54	99.91	99.90	99.90	99.92
$s.e.(\bar{\hat{\tau}})$	0.2471	0.0527	0.0258	0.0158	0.0106	0.0076

Table 3: Average change point estimates for δ and standard errors when used with a EWMA control chart, $ARL_0 = 500,\ L = 2.998,\ \lambda = 0.25,\ \tau = 100$ and $N = 10000$ independent simulation trials

δ	0.5	1.0	1.5	2.0	2.5	3.0
$E(T)$	147.41	110.36	104.8	102.93	102.08	101.62
$\bar{\hat{\tau}}$	103.03	100.09	99.98	99.91	99.94	99.92
$s.e.(\bar{\hat{\tau}})$	0.2235	0.0506	0.0247	0.0142	0.0092	0.0069

Table 4: Average change point estimates for δ and standard errors when used with a EWMA control chart, $ARL_0 = 370.37, L = 2.86, \lambda = 0.2, \tau = 100$ and $N = 10000$ independent simulation trials

δ	0.5	1.0	1.5	2.0	2.5	3.0
$E(T)$	135.00	108.80	104.32	102.71	101.95	101.53
$\bar{\hat{\tau}}$	103.99	99.85	99.94	99.88	99.91	99.93
$s.e.(\bar{\hat{\tau}})$	0.2382	0.0521	0.0251	0.0152	0.01	0.0072

The precision of the change point estimator for process mean can be examining the probability that $\hat{\tau}$ within m observations of the exact change point. Table 5 contains the results for the case where $L = 2.86, \lambda = 0.2$. For the magnitude of the shift $\delta = 0.5$, the probability that the change point estimator correctly identified the actual time of change in the process mean 9% of the simulation trials. The change point estimator correctly identified the actual time of the change time of the change in 45% and 55% of the simulation trials within 6 and 9 observations respectively.

It can also be seen that for the value of the shift in the parameter 1.5, the estimator is within 2 (6) observations of the actual process change point in 83% (96%) of the simulation trials. For the magnitude of the shift $\delta = 2.0$, the probability that the change point estimator correctly identified the exact time of change in process mean is 62% of simulation trials. The probability that the estimator correctly identified the time of the change within 3 (7) observations is 95% (99%) of the simulation trials.

Table 5: Precision of estimator for δ when used with EWMA control chart $\tau = 100$ and $L = 2.86,\ \lambda = 0.2$ and $N = 10000$ independent simulation trials

δ	0.5	1.0	1.5	2.0	2.5	3.0
$P[\hat{\tau} - \tau = 0]$	0.09	0.27	0.44	0.62	0.75	0.84
$P[\lvert \hat{\tau} - \tau \rvert \le 1]$	0.19	0.48	0.71	0.80	0.86	0.96
$P[\lvert \hat{\tau} - \tau \rvert \le 2]$	0.26	0.58	0.83	0.85	0.95	0.98
$P[\lvert \hat{\tau} - \tau \rvert \le 3]$	0.32	0.67	0.89	0.95	0.97	0.995
$P[\lvert \hat{\tau} - \tau \rvert \le 4]$	0.37	0.73	0.93	0.97	0.98	0.997
$P[\lvert \hat{\tau} - \tau \rvert \le 5]$	0.41	0.80	0.95	0.98	0.989	0.9987
$P[\lvert \hat{\tau} - \tau \rvert \le 6]$	0.45	0.83	0.96	0.987	0.993	0.9989
$P[\lvert \hat{\tau} - \tau \rvert \le 7]$	0.49	0.86	0.97	0.991	0.994	0.999
$P[\lvert \hat{\tau} - \tau \rvert \le 8]$	0.52	0.88	0.98	0.994	0.995	0.999
$P[\lvert \hat{\tau} - \tau \rvert \le 9]$	0.55	0.91	0.987	0.995	0.997	0.999
$P[\lvert \hat{\tau} - \tau \rvert \le 10]$	0.59	0.92	0.99	0.997	0.997	1
$P[\lvert \hat{\tau} - \tau \rvert \le 11]$	0.61	0.94	0.996	0.999	0.999	1

5. Change Point Estimator Used with MA Control Chart

In this section, we consider MA control chart to study how well the estimator performs. As with the EWMA control chart, we used Monte Carlo simulation to study the performance of the change point estimator. The change point of the process is simulated at observation $\tau = 100$. The first 100 in-control individual observations are randomly generated from standard normal distribution. Then starting from observation 101, the individual observations are randomly generated from normal distribution with mean δ and standard deviation 1 until the moving average chart gives an out-of-control signal. At this point $\hat{\tau}$ is computed. This procedure is repeated a total number of $N = 10000$ times for each of the values of $\delta = 0.5, 1.0, 1.5, 2.0, 2.5, 3.0$ with different values of moving average span $w = 1, 2, 3$. The average of the estimates obtained using the estimator from the 10000 simulation runs is computed with their standard error.

Table 6-8 shows $E(T)$, the expected number of observations at which the control chart signals a change in the process mean that occurred at time 100. Thus, $E(T) = ARL + 100$. We show that $\bar{\hat{\tau}}$ the average change point estimate obtained using MLE change point estimator with their standard error.

Table 6: Average change point estimates for δ and standard errors when used with MA control chart, $w = 1,\ \tau = 100$ and $N = 10000$ independent simulation trials

δ	0.5	1.0	1.5	2.0	2.5	3.0
$E(T)$	256.11	143.70	115.10	106.34	103.23	102.02
$\bar{\hat{\tau}}$	108.5	100.85	100.18	99.95	99.85	99.82
$s.e.(\bar{\hat{\tau}})$	0.251	0.057	0.027	0.018	0.015	0.013

In Table 6, we can see that for the magnitude of the shift $\delta = 1.0$ the control chart issues signal at 143.70 on an average of 10000 simulation trial. The average change point estimate is 100.85 which are close to 100. For the magnitude of the shift $\delta = 1.5$ the control chart issues signal at 115.10 and that of average change point estimate is 100.18. For $\delta = 3.0$, the expected number to issue a signal from control chart is 102.02. The average change point estimate is 99.82.

Table 7: Average change point estimates for δ and standard errors when used with MA control chart, $w = 2,\ \tau = 100$ and $N = 10000$ independent simulation trials

δ	0.5	1.0	1.5	2.0	2.5	3.0
$E(T)$	203.58	122.63	107.58	103.62	102.23	101.65
$\bar{\hat{\tau}}$	109.59	100.93	99.99	99.79	99.73	99.75
$s.e.(\bar{\hat{\tau}})$	0.252	0.057	0.031	0.021	0.018	0.016

In Table 7, it is seen that for the magnitude of the shift $\delta = 1.5$ the control chart issues signal at 107.58. The average change point estimate is 99.99 which are close to 100. For the magnitude of the shift $\delta = 2.0$ the control chart issues signal at 103.62. The average change point estimate is 99.79. For $\delta = 3.0$, the expected number to issue a signal from control chart is 101.65. The average change point estimate is 99.75 which are again close to 100.

Table 8: Average change point estimates for δ and standard errors when used with MA control chart, $w = 3,\ \tau = 100$ and $N = 10000$ independent simulation trials

δ	0.5	1.0	1.5	2.0	2.5	3.0
$E(T)$	183.63	116.52	105.95	103.15	102.11	101.62
$\bar{\hat{\tau}}$	109.66	100.79	99.89	99.69	99.71	99.74
$s.e.(\bar{\hat{\tau}})$	0.2413	0.0575	0.0319	0.024	0.019	0.017

In Table 8, we can see that for the magnitude of the shift $\delta = 1.5$ the expected number of observations at the point when control chart gives out of control signal at 105.95 on an average of 10000 simulation trial. The average change point estimate is 99.89 which are close to 100. For the magnitude of the shift $\delta = 2.5$ the control chart issues signal at 102.11 on an average of the 10000 simulation trials and that of average change point estimate is 99.71. For $\delta = 3.0$, $E(T) = 101.62$ and that of the estimated change point is 99.74 on an average of total 10000 simulation trial.

We next consider the observed frequency with which the estimator of the time of step change is within *m* observations of the exact change point, for $m = 0,1,...,11$. This indicates the precision of the proposed estimator. This table contains the results for the case where $w = 2$. For the magnitude of the shift $\delta = 0.5$, the precision that the change point estimator correctly identified the actual time of change in the process mean 9% of the simulation trials, same as EWMA control chart. The change point estimator correctly identified the actual time of the change within the 2(6) is 26% and 47% of the simulation trials within the respectively.

It can also be seen that for the value of the shift in the parameter 1.0, the estimator is within 3(9) observations of the actual process change point in 68% (91%) of the simulation trials. For the value of the shift 2.0 the estimator correctly identified the actual change point is 61%. The precision of the estimator within the 4(8) observations of the actual change point is 97% (99%) of the total simulation trials. The precision of estimated time of the change within *m* observations of the actual change point should increase as *m* increases.

Table 9: Precision of estimator for δ when used with MA control chart $\tau = 100$ and $w = 2$ and $N = 10000$ independent simulation trials

δ	0.5	1.0	1.5	2.0	2.5	3.0
$P[\lvert\hat{\tau}-\tau\rvert=0]$	0.09	0.27	0.45	0.61	0.75	0.84
$P[\lvert\hat{\tau}-\tau\rvert\le 1]$	0.19	0.47	0.70	0.83	0.90	0.93
$P[\lvert\hat{\tau}-\tau\rvert\le 2]$	0.26	0.60	0.82	0.92	0.95	0.97
$P[\lvert\hat{\tau}-\tau\rvert\le 3]$	0.33	0.68	0.88	0.95	0.97	0.98
$P[\lvert\hat{\tau}-\tau\rvert\le 4]$	0.38	0.75	0.92	0.97	0.98	0.984
$P[\lvert\hat{\tau}-\tau\rvert\le 5]$	0.42	0.80	0.94	0.98	0.988	0.988
$P[\lvert\hat{\tau}-\tau\rvert\le 6]$	0.47	0.84	0.96	0.983	0.990	0.990
$P[\lvert\hat{\tau}-\tau\rvert\le 7]$	0.50	0.87	0.97	0.989	0.992	0.992
$P[\lvert\hat{\tau}-\tau\rvert\le 8]$	0.53	0.89	0.98	0.991	0.995	0.994
$P[\lvert\hat{\tau}-\tau\rvert\le 9]$	0.56	0.91	0.983	0.993	0.996	0.997
$P[\lvert\hat{\tau}-\tau\rvert\le 10]$	0.59	0.92	0.986	0.997	0.999	0.999
$P[\lvert\hat{\tau}-\tau\rvert\le 11]$	0.61	0.94	0.99	0.999	1	1

6. Examples of Application

This section provides numerical examples to demonstrate the use of estimator each with EWMA control chart and MA control chart. The change point estimator works well with EWMA and MA control charts.

Example-1: EWMA Control Chart

In this example, we consider the data of a production process for forged piston rings used in the illustrative example of (Samuel *et al*., 1998a). The in-control process follows a normal distribution with mean 100 and standard deviation 5. Each subgroup has n = 4 observations. The EWMA control chart with $\lambda = 0.1$ and $L = 2.703$ is considered. From the original data of 27 subgroups given in (Samuel et al., 1998a), only the first 20 subgroups are required before the EWMA chart signals an out-of-control, since $Z_i > UCL$. Table 10 summarizes the 20 subgroup averages and the corresponding EWMA statistics.

Table 10: Subgroup averages and the corresponding EWMA statistics

Subgroup (i)	$\overline{X}_i$	EWMA Z_i	UCL	LCL
1	100.45	100.045	101.351	98.648
2	97.45	100.15	101.818	98.181
3	102.45	97.95	102.122	97.877
4	100.675	102.272	102.339	97.660
5	98.550	100.462	102.502	97.497
6	102.95	98.990	102.626	97.373
7	98.825	102.537	102.722	97.277
8	101.325	99.075	102.798	97.201
9	103.075	101.5	102.858	97.141
10	99.600	102.727	102.900	97.094
11	98.825	99.522	102.943	97.056
12	97.950	98.737	102.974	97.025
13	100.425	98.197	102.998	97.001
14	96.075	99.99	103.018	96.981
15	101.225	96.59	103.034	96.965
16	103.075	101.41	103.046	96.953
17	101.925	102.96	103.057	96.942
18	101.350	101.867	103.065	96.934
19	103.575	101.572	103.072	96.927
20	102.925	**103.509**	**103.077**	96.922

Table 11 summarizes the reverse cumulative subgroup averages and C_t values for $t = 1, 2, \ldots, 19$ and T = 20. The value of t which gives the maximum C_t value is the estimator of the last subgroup from the in-control process. From the results in Table 11, we observed that the maximum value of $C_t$is 32.998 and it happens at t = 15. Thus, it is estimated that subgroup 16 is the first subgroup obtained from the shifted process and that subgroup 15 is the last subgroup from the in-control process.

However, the EWMA chart enables an out-of-control signal to be detected earlier, *i.e*., at subgroup 20 compared to the time when the $\overline{X}$ chart first detected an out-of-control signal, *i.e*., at subgroup 27 (see Samuel *et al*., 1998a).

Table 11: The computed $\overline{\overline{X}}_{20,t}$ and the corresponding C_t values

Subgroup (i)	$\overline{X}_i$	T	$\overline{\overline{X}}_{20,t}$	Ct
1	100.45	0	100.634	8.058
2	97.45	1	100.644	7.891
3	102.45	2	100.821	12.160
4	100.675	3	100.726	8.964
5	98.55	4	100.729	8.511
6	102.95	5	100.874	11.475
7	98.825	6	100.726	7.387
8	101.325	7	100.872	9.900
9	103.075	8	100.835	8.366
10	99.6	9	100.631	4.384
11	98.825	10	100.734	5.394
12	97.95	11	100.946	8.065
13	100.425	12	101.321	13.965
14	96.075	13	101.449	14.703
15	101.225	14	102.345	32.994
16	**103.075**	**15**	**102.569**	**32.998**
17	101.925	16	102.442	23.863
18	101.35	17	102.615	20.514
19	103.575	18	103.2475	21.09251
20	102.92	19	102.92	8.5264

Example 2: MA Control Chart

In this example, we consider the data of 20 observations coming from normal distribution with mean 10 and last 10 observations with mean 11 with common standard deviation 1. This data is used in the example of (Montgomery, 2012). The moving average control chart with subgroup size $n = 1$is considered. For purpose of the use of the change point estimator with given data we take large value of $w = 6$. From the original data of 30observations given in (Montgomery, 2012), only the first 28 observations are required before the moving average chart signals an out-of-control, since M_i>UCL. Table A.1 (Appendix) summarizes the 28 observations and the corresponding moving average statistics.

Table A.2 shows the reverse cumulative averages and corresponding values of C_t for $t = 1, 2, \ldots, 27$ and $T = 28$. The value of t which maximizes the value of C_t is the estimator of the last observation from the in-control process. From Table A.2, we can observe that at observation $t = 22$ the maximum value of C_t is **6.025.** Thus, it is estimated that observation 23 is the first observation obtained from the shifted process and that observation 22 is the last observation from the in-control process.

From this example we can observe the moving average control chart signals out-of-control at $t = 28$ and that the change point estimator identified the change in the process at

$t = 22$. This shows that the change point estimator fairly works with moving average control chart to identify the out-of-control signal earlier.

7. Conclusion

Control charts are used to detect whether or not a process has changed. When a control chart detects the shift in a process, process professionals initiate a search to find the special causes of variation in the process. When the process gets changed, the process change is not usually known to the process professionals. However, the process professionals knew when the change started in the process; it will help to provide the scope of the search window at what time the process gets changed. Subsequently, it helps to eliminate the sources of the special causes.

In this paper, an estimator based on the maximum likelihood estimator method is used with EWMA control chart and MA chart to find the step change that occurred in the normal process mean. The results show that the change point estimator is helpful to detect the change in the process. The EWMA and MA control charts are effective to detect the small shifts in the process mean. The change point estimator also performs well to detect the small changes with EWMA and moving average control chart.

Acknowledgment

The authors would like to thank the editor, the managing editor, and the referee for their comments which greatly improved the paper. The first author would like to thank the Department of Science and Technology for supporting this research by Inspire Fellowship No. DST/INSPIRE Fellowship/2016/IF160410.

References

Dogu, E. and Kocakoc, I. D. (2011). Estimation of change point in generalized variance control chart. *Communication in Statistics- Simulation and Computation*, **40**(**3**), 345-363.

Fahmy, H. M. and Elsayed, E. A. (2006). Drift time detection and adjustment procedures for processes subject to linear trend. *International Journal of Production Research*, **44**(**16**), 3257–3278.

Gazanfari, M., Alaeddini, A., Niaki, S. T. A., Aryanezhad, M. B. (2008). A clustering approach to identify the time of a step change in Shewhart control charts. *Quality and Reliability Engineering International*, **24**(**7**), 765–778.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, **57(1)**, 1–17.

Kapase, R. A. and Ghute, V.B. (2018). Estimating the period of a step-change in single observation data. *International Journal of Agricultural and Statistical Sciences*, **14**(**2**), 433-438.

Khoo, M. B. C. (2004). Determining the time of a permanent shift in the process mean of CUSUM control charts. *Quality Engineering,* **17**(**1**), 87-93.

Montgomery, D. C. (2012). *Introduction to Statistical Quality Control*. John Wiley & Sons.

Nedumaran, G., Pignatiello Jr., J. J. and Calvin, J. A. (2000). Identifying the time of a step-change with χ^2 control charts. *Quality Engineering,* **13**(**2**), 153-159.

Noorossana, R., Saghaei, A., Paynabar, K. and Abdi, S. (2009). Identifying the period of a step change in high-yield processes. *Quality and Reliability Engineering International*, **25**(**7**), 875-883.

Park, J. and Park, S. (2004). Estimation of the change point in the X-bar and S control charts. *Communications in Statistics-Simulation and Computation,* **33**(**4**), 1115-1132.

Perry, M. B. and Pignatiello Jr., J. J. (2006). Estimation of the change point of a normal process mean with a linear trend disturbance in SPC. *Quality Technology & Quantitative Management,* **3**(**3**), 325–334.

Perry, M. B., Pignatiello Jr., J. J. and Simpson, J. R. (2005). Estimation of the change point of a Poisson rate parameter with a linear trend disturbance. *Quality and Reliability Engineering International*, **22**(**4**), 371–384.

Perry, M. B., Pignatiello Jr., J. J. and Simpson, J. R. (2007). Change point estimation for monotonically changing Poisson rates in SPC. *International Journal of Production Research,* **45**(**8**), 1791–1813.

Pignatiello Jr., J. J. and Samuel, T. R. (2001a). Estimation of the change point of a normal process mean in SPC applications. *Journal of Quality Technology*, **33**(**1**), 82-95.

Pignatiello Jr., J. J. and Samuel, T. R. (2001b). Identifying the time of a step-change in the process fraction nonconforming. *Journal of Quality Engineering*, **13**(**3**), 357-365.

Samuel, T. R., Pignatiello Jr., J. J. and Calvin, J. A. (1998a). Identifying the time of a step-change with X-bar control charts. *Quality Engineering,* **10**(**3**), 521-527.

Samuel, T. R., Pignatiello Jr., J. J. and Calvin, J. A. (1998b). Identifying the time of a step-change in a normal process variance. *Quality Engineering*, **10**(**3**), 529-538.

Samuel, T. R. and Pignatiello Jr., J. J. (1998). Identifying the time of a step-change in a Poisson rate parameter. *Journal of Quality Engineering*, **10**(**4**), 673-681.

Zandi, F., Niaki, S. T. A., Nayeri, M. A. and Fathi, M. (2011). Change point estimation of the process fraction nonconforming with a linear trend in statistical process control. *International Journal of Computer Integrated Manufacturing*, **24**(**10**), 939–947.

APPENDIX

Table A.1: Averages and the corresponding moving average statistics

Subgroup (i)	$\overline{X}_i$	Moving Average M_i	UCL	LCL
1	9.45	9.45	13	7
2	7.99	8.72	12.12	7.88
3	9.29	8.91	11.73	8.27
4	11.66	9.722	11.50	8.5
5	12.16	9.814	11.34	8.66
6	10.18	9.518	11.22	8.78
7	8.04	9.853	11.22	8.78
8	11.46	10.055	11.22	8.78
9	9.20	10.23	11.22	8.78
10	10.34	9.708	11.22	8.78
11	9.03	9.923	11.22	8.78
12	11.47	10.335	11.22	8.78
13	10.51	9.991	11.22	8.78
14	9.40	10.138	11.22	8.78
15	10.08	9.976	11.22	8.78

16	9.37	10.241	11.22	8.78
17	10.62	10.04	11.22	8.78
18	10.31	9.716	11.22	8.78
19	8.52	9.956	11.22	8.78
20	10.84	10.083	11.22	8.78
21	10.90	10.338	11.22	8.78
22	9.33	10.123	11.22	8.78
23	12.29	10.453	11.22	8.78
24	11.50	10.95	11.22	8.78
25	10.60	10.91	11.22	8.78
26	11.08	10.95	11.22	8.78
27	10.38	10.863	11.22	8.78
28	11.62	**11.245**	**11.22**	8.78

Source: Montgomery D. C. (2012). *Introduction to Statistical Quality Control*

Table A.2: The computed $\overline{\overline{X}}_{28,t}$ and the corresponding C_t values

Subgroup (*i*)	$\overline{X}_i$	*t*	$\overline{\overline{X}}_{28,t}$	C_t
1	9.45	0	10.242	0
2	7.99	1	10.272	0.023
3	9.29	2	10.36	0.356
4	11.66	3	10.402	0.639
5	12.16	4	10.329	0.180
6	10.18	5	10.336	0.199
7	8.04	6	10.440	0.858
8	11.46	7	10.391	0.466
9	9.20	8	10.451	0.870
10	10.34	9	10.457	0.874
11	9.03	10	10.536	1.553
12	11.47	11	10.481	0.970
13	10.51	12	10.480	0.899
14	9.40	13	10.552	1.433
15	10.08	14	10.585	1.645
16	9.37	15	10.679	2.475
17	10.62	16	10.684	2.337
18	10.31	17	10.718	2.485
19	8.52	18	10.938	4.832
20	10.84	19	10.948	4.486
21	10.90	20	10.962	4.143
22	9.33	21	10.971	3.715
23	**12.29**	**22**	**11.245**	**6.025**
24	11.50	23	11.036	3.145
25	10.60	24	10.920	1.834
26	11.08	25	11.026	1.843
27	10.38	26	11.000	1.146
28	11.62	27	11.620	1.896

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No.2, 2022 (New Series), pp 41-55

A-optimal Designs for Cubic Polynomial Models with Mixture Experiments in Three Components

Mahesh Kumar Panda and Rushi Prasad Sahoo
Department of Statistics, Central University of Odisha, Koraput 763004

Received: 13 April 2021; Revised: 17 July 2021; Accepted: 20 July 2021

Abstract

This article discusses A-optimal minimum support designs for the three different forms of cubic polynomial mixture models *i.e.* full cubic, cubic without 3-way effect, and special cubic mixture models in three ingredients. The necessary and sufficient conditions for the proposed designs have been confirmed by the equivalence theorem.

Key words: A-optimal design; Mixture models; Equivalence theorem.

AMS Subject Classification: 62K05

1. Introduction

The importance of mixture experiments is increasing gradually, because it is utilized in many disciplines such as pharmaceutical science, food science, chemical science, and textile science, *etc*. Let us consider a mixture experiment having q ingredients with mixture proportions denoted by $x_1, x_2, ..., x_q$ then the factor space consisting of these ingredient proportions can be represented by a $(q-1)-$dimensional set χ given by

$$\chi = \left\{ \boldsymbol{x} = (x_1, x_2, ..., x_q)' \in R^q \mid \sum_{i=1}^{q} x_i = 1,\ 0 \le x_i \le 1,\ i = 1, 2, ..., q \right\}. \quad (1)$$

Let the observed response may be represented as $y = \eta(\boldsymbol{x}) + \varepsilon(\boldsymbol{x})$, where $\eta(\boldsymbol{x})$ is the expected response and $\varepsilon(\boldsymbol{x})$ is the random error observed at $\boldsymbol{x}$. We also assume that $\varepsilon(\boldsymbol{x})$ are *i.i.d.* random variables with mean 0 and variance σ^2. To describe the relationship between the response of interest and the ingredient proportions, in any mixture experiment, various mixture models have already been introduced in the literature *e.g.* Scheffè's canonical polynomial models, Becker's models, log contrast models, *etc*. Among these models, the canonical polynomial models are frequently used for the analysis of mixture data related to real-life problems.

In general, the optimal designs are constructed based on a certain optimality criterion to make the predicted response closer to the mean response over a certain region of interest. For the pioneering work on optimal designs for mixture experiments, one can refer to the work of Kiefer and Wolfowitz (1959), and Kiefer (1961). Afterward, many researchers have put their attention towards the discipline of optimal designs for mixture experiments [see Aggrawal *et*

Corresponding Author: Mahesh Kumar Panda
E-mail: mahesh2123ster@gmail.com

al. (2011), Singh and Panda (2011), Goos and Syafitri (2014), Mandal and Pal (2017), and Pal and Mandal (2021), *etc*.].

Kiefer (1961) obtained D-optimal designs for Scheffè's models of degrees one, two, and three. For Scheffè's linear model in q mixture ingredients, a saturated design that assigns a weight $1/q$ to each vertex of the simplex region is a D-optimal design. Again a minimum-point design supported by points of $\{q, 2\}$ simplex-lattice with equal mass assigned to each support point is D-optimum for Scheffè's quadratic mixture model. Kiefer (1961) obtained the saturated D-optimal designs for the full cubic model, the cubic model without 3-way effect, and the special cubic model when $q = 3$. Later on, Mikaeili (1989) obtained the D-optimal designs for the cubic model without 3-way effect. Farrell *et al*. (1967) and Lim (1990) derived the D-optimal designs for the general cubic polynomial model with two and three mixture components respectively. Mikaeli (1993) investigated the D-optimal designs for the full cubic model on the set χ.

For Scheffè's cubic canonical polynomial model in this effect, we see that most of the existing works focus solely on D-optimality. However, to date, no research work has been done concerning the A-optimal designs for the cubic polynomial models and it was still an open problem. The advantage of D-optimal design is that all the support points involved are associated with equal weight whereas in the case of A-optimality, the weights associated with different support points, in general, are different. Again, the weights vary when the number of mixture components varies. Thus, obtaining an A-optimal design for all the different forms of cubic mixture canonical polynomial models is comparatively much more complicated in comparison to the D-optimal design. In this article, we study the problem of finding A-optimal minimum support designs for the three different forms of cubic polynomial mixture models in three ingredients.

The article is structured as follows. In Section 2, a brief discussion on the A-optimal design and equivalence theorem is presented. Section 3 obtains A-optimal designs for the three different forms of the cubic model of mixture experiments *i.e.* full cubic model, cubic model without 3-way effect, and special cubic model when q = 3. The article ends with some discussions and conclusions in Section 4.

2. A-Optimal Design and Equivalence Theorem

Let us consider a regression model of the form

$$\eta(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\beta}\,,\ \boldsymbol{x} \in \chi\,, \tag{2}$$

where $\eta(\boldsymbol{x})$ denotes the expected response, $\boldsymbol{x}$ is the input variable, and $\boldsymbol{f}(\boldsymbol{x})$ is the regression function.

Again, let us consider an approximate design (Kiefer, 1974) of the following form

$$\xi = \begin{Bmatrix} \boldsymbol{x}_{(1)} & \ldots & \boldsymbol{x}_{(m)} \\ r_1 & \ldots & r_m \end{Bmatrix}, \quad \boldsymbol{x}_{(i)} \in \chi\,, \qquad 0 < r_i < 1,\ \sum_{i=1}^{m} r_i = 1,$$

where $\boldsymbol{x}_{(1)},...,\boldsymbol{x}_{(m)}$ are different design points over χ and r_i is the weight assigned to the point $\boldsymbol{x}_{(i)}$, $i=1,2,...,m$. Denote Δ as the set of all approximate designs with non-singular information matrix

$$\boldsymbol{M}(\xi)=\sum_{i=1}^{m} r_i \boldsymbol{f}(\boldsymbol{x}_{(i)})\boldsymbol{f}'(\boldsymbol{x}_{(i)})$$

on χ.

Definition 1: A design $\xi^* \in \Delta$ with an information matrix $\boldsymbol{M}(\xi)$ for model (2) is called A-optimal design if it minimizes Trace $(\boldsymbol{M}^{-1}(\xi))$ over Δ.

Definition 2:A minimum support design for any regression model having p parameters is supported on exactly p distinct support points [see Goos and Vandebroek (2001)].

The following equivalence theorem established by Fedorov (1971) provides the necessary and sufficient conditions for the determination of A-optimal design over the simplex region χ.

Theorem 1: A design $\xi^* \in \Delta$ is A-optimal for model (2) if and only if

$$\underset{\boldsymbol{x}\in\chi}{Max}\ d(\boldsymbol{x},\xi^*)=\text{Trace}(\boldsymbol{M}^{-1}(\xi^*)) \tag{3}$$

where $d(\boldsymbol{x},\xi)=\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{M}^{-2}(\xi)\boldsymbol{f}(\boldsymbol{x})$. Moreover, the supremum exists at the support point of ξ^*.

Selection of support points: Kiefer (1961) considered the design ξ_a (for $0<a<1/2$) which puts equal mass $\frac{1}{10}$ on each of the vertices $x_i=1,\ x_j=x_k=0$; each of the six points $x_i=1-x_j=a,\ x_k=0$, and $x_1=x_2=x_3=1/3$. He proved that the design ξ_a for $a=(1-5^{-\frac{1}{2}})/2$ is D-optimum for the full cubic model when $q=3$. Similarly, he showed that the design ξ_a (excluding the point $x_1=x_2=x_3=1/3$) in which each point is supported by a mass $\frac{1}{9}$ is D-optimum for the cubic model without 3-way effect in three ingredients for $a=(1-5^{-\frac{1}{2}})/2$. Further, he showed that the simplex centroid design which assigns mass $\frac{1}{7}$ to each of the support points is D-optimum for the special cubic model when $q=3$. We, therefore, propose the following subclasses (D_1, D_2, D_3) of designs ξ_a to find the minimum support A-optimal design.

Model (Subclass)	x_1	x_2	x_3	Weight
Full Cubic Model (D_1)	1	0	0	r_1
	0	1	0	
	0	0	1	
	a	$1-a$	0	r_2
	a	0	$1-a$	
($0 < a < 1/2$)	0	a	$1-a$	
	$1-a$	a	0	
	$1-a$	0	a	
	0	$1-a$	a	
	1/3	1/3	1/3	r_3
Cubic Model without 3-way effect (D_2)	1	0	0	r_1
	0	1	0	
	0	0	1	
	a	$1-a$	0	r_2
	a	0	$1-a$	
($0 < a < 1/2$)	0	a	$1-a$	
	$1-a$	a	0	
	$1-a$	0	a	
	0	$1-a$	a	
Special Cubic Model (D_3)	1	0	0	r_1
	0	1	0	
	0	0	1	
	a	$1-a$	0	r_2
	a	0	$1-a$	
	0	a	$1-a$	
	1/3	1/3	1/3	r_3

Here we assume that a weight of r_1 is associated with each of the vertices, a weight of r_2 is associated with each of the design points $x_i = 1 - x_j = a$, $x_k = 0$ (for full cubic and cubic model without 3-way effect) and $x_i = a$, $x_k = 0$ (for special cubic), and finally a weight of r_3 is associated with each of the midpoints of 2-dimensional faces such that the total weights add to unity. We can concentrate on the above class of designs because the A-optimality criterion is invariant for all three components. Consequently, the optimum design will also be invariant *w.r.t* x_1, x_2, and x_3.

In the next section, we obtain the A-optimal designs for the three different forms of Scheffè's cubic polynomial model when q =3.

3. A-Optimal Designs for Cubic Models for Mixture Experiments

3.1. Full cubic model

The expected response for a full cubic model (see Cornell (2002)) can be represented as

$$\eta_1 = \boldsymbol{f}_1'(\boldsymbol{x})\boldsymbol{\beta}_1 = \sum_{i=1}^{q}\beta_i x_i + \sum_{i<j=1}^{q}\beta_{ij}x_i x_j + \sum\sum_{i<j}^{q}\delta_{ij}x_i x_j (x_i - x_j) + \sum\sum\sum_{i<j<k}\beta_{ijk}x_i x_j x_k \tag{4}$$

where $\boldsymbol{f}_1(\boldsymbol{x})$ and $\boldsymbol{\beta}_1$ are column vectors of length $\frac{q(q+1)(q+2)}{6}$ and are defined by

$$\boldsymbol{f}_1(\boldsymbol{x}) = \left(x_1, x_2, ..., x_q, x_1x_2, x_1x_3, ..., x_{q-1}x_q, x_1x_2(x_1 - x_2), .x_1x_3(x_1 - x_3),\right.$$

$$\left. ..., x_{q-1}x_q(x_{q-1} - x_q), x_1x_2x_3, x_1x_2x_4, ..., x_{q-2}x_{q-1}x_q\right)';$$

$$\boldsymbol{\beta}_1 = \left(\beta_1, \beta_2, ..., \beta_q, \beta_{12}, \beta_{13}, ..., \beta_{q-1q}, \delta_{12}, \delta_{13}, ..., \delta_{q-1q}, \beta_{123}, \beta_{124}, ..., \beta_{q-2q-1q}\right)'.$$

The non-singular information matrix for the model (4) is given by

$$\boldsymbol{M}(\xi) = \sum_{i=1}^{m} r_i \boldsymbol{f}_1(\boldsymbol{x}_{(i)})\boldsymbol{f}_1'(\boldsymbol{x}_{(i)}) \tag{5}$$

The next theorem obtains the A-optimal minimum support design for model (4) when q = 3.

Theorem 2: For q = 3, the design ξ_1 with support points from {3, 3} simplex-lattice that assigns a weight of 0.0612 to the 3 vertices (1, 0, 0), (0, 1, 0), (0, 0, 1); a weight of 0.0933 to the 6 points (1/3, 2/3, 0), (1/3, 0, 2/3), (0, 1/3, 2/3), (2/3, 1/3, 0), (2/3, 0, 1/3), (0, 2/3, 1/3); and a weight of 0.2567 to the centroid point (1/3, 1/3, 1/3) is the A-optimal minimum support design for the full cubic polynomial model with mixture experiments on χ.

Proof: According to the equivalence theorem in equation (3), if ξ^* is the A-optimal design then the infimum of Trace$(\boldsymbol{M}^{-1}(\xi))$ and the supremum of $d(\boldsymbol{x}, \xi)$ both exists at the support points of ξ^*. Based on this result, we search for support points of A-optimal design *i.e.* Min Trace$(\boldsymbol{M}^{-1}(\xi))$ for the full cubic model over the subclass D_1 by considering different values of 'a' subject to the linear constraint that the sum of the weights is equal to 1.

Let us consider the proposed design ξ_a for the full cubic model given in Section 2. The inverse of the information matrix of the form (5) for the design ξ_a is

$$\boldsymbol{M}^{-1}(\xi_a)=\begin{pmatrix} \frac{1}{r_1}\boldsymbol{I}_3 & \boldsymbol{A}_1' & \boldsymbol{A}_2' & g_6\boldsymbol{1}_3 \\ \boldsymbol{A}_1 & g_2\boldsymbol{I}_3+g_3\boldsymbol{J}_3 & \boldsymbol{A}_3' & g_4\boldsymbol{1}_3 \\ \boldsymbol{A}_2 & \boldsymbol{A}_3 & \boldsymbol{A}_4 & 0.\boldsymbol{1}_3 \\ g_6\boldsymbol{1}_3' & g_4\boldsymbol{1}_3' & 0.\boldsymbol{1}_3' & g_7 \end{pmatrix} \tag{6}$$

where

$$g_1=\frac{1}{2ar_1(1-a)}, g_2=\frac{2r_1+r_2}{4a^2(1-a)^2r_1r_2},\ g_3=\frac{1}{4a^2(1-a)^2r_1},$$

$$g_4=-\frac{3(r_1+2r_2(3a(a-1)+1))}{2a^2(1-a)^2r_1r_2},\ g_5=\frac{r_1+(1-2a)^2r_2}{2a^2(2a^2-3a+1)^2r_1r_2},$$

$$g_6=\frac{3(3a^2-3a+1)}{ar_1(1-a)}, g_7=\frac{27\left(54a^2(a-1)^2r_1r_2+(2r_2(3a(a-1)+1)^2+r_1)r_3\right)}{2a^2(1-a)^2r_1r_2r_3},$$

and $\boldsymbol{I}_3$ is the identity matrix of order 3, $\boldsymbol{J}_3$ is a matrix of order 3×3 in which each entry is 1, $\boldsymbol{1}_3$ is a column vector of order 3×1,

$$\boldsymbol{A}_1=\begin{pmatrix} -g_1 & -g_1 & 0 \\ -g_1 & 0 & -g_1 \\ 0 & -g_1 & -g_1 \end{pmatrix}, \qquad \boldsymbol{A}_2=\begin{pmatrix} -g_1 & g_1 & 0 \\ -g_1 & 0 & g_1 \\ 0 & -g_1 & g_1 \end{pmatrix},$$

$$\boldsymbol{A}_3=\begin{pmatrix} 0 & g_3 & -g_3 \\ g_3 & 0 & -g_3 \\ g_3 & -g_3 & 0 \end{pmatrix}, \qquad \boldsymbol{A}_4=\begin{pmatrix} g_5 & g_3 & -g_3 \\ g_3 & g_5 & g_3 \\ -g_3 & g_3 & g_5 \end{pmatrix}.$$

Next, the trace of $\boldsymbol{M}^{-1}(\xi_a)$ is obtained as

$$\text{Trace}\,(\boldsymbol{M}^{-1}(\xi_a))=\frac{3\left(2+\frac{r_1+r_2}{(a-1)^2a^2r_2}+\frac{r_1+(1-2a)^2r_2}{a^2(1-3a+2a^2)^2r_2}\right)}{2r_1}$$

$$+27\left(\frac{54(a-1)^2a^2r_1r_2+(r_1+2(3a(a-1)+1)^2r_2)r_3}{2a^2(a-1)^2r_1r_2r_3}\right) \tag{7}$$

Now, the problem becomes minimizing equation (7) subject to the restriction of weights $3r_1+6r_2+r_3=1$. To solve this problem, we use the Lagrangian multiplier method and set the Lagrangian function as

$$\psi=\text{Trace}\ (\boldsymbol{M}^{-1}(\xi_a))+\lambda(3r_1+6r_2+r_3-1).$$

By taking the partial derivatives of ψ *w.r.t* r_1, r_2, r_3, and λ, and set them equal to 0, we get

$$\frac{3a(a-1)\left(a(a-1)(\lambda r_1^2-82)-54\right)-30}{a^2(a-1)^2 r_1^2}=0, \tag{8}$$

$$-\frac{3(11+40a(a-1))}{2a^2(2a^2-3a+1)^2 r_2^2}+6\lambda=0, \tag{9}$$

$$-\frac{729}{r_3^2}+\lambda=0, \tag{10}$$

$$3r_1+6r_2+r_3=1. \tag{11}$$

The algebraic derivations for solving equations (8) – (11) are lengthy and tedious, thus we numerically compute possible optimal values of r_1, r_2, r_3 (rounded off to the fourth place of the decimal) and the corresponding value of Trace$(\boldsymbol{M}^{-1}(\xi_a))$ for different values of a, which are tabulated in Table 1.

Table 1: Trace$(\boldsymbol{M}^{-1}(\xi_a))$ and corresponding weights of full cubic model for different values of a

a	r_1	r_2	r_3	Trace$(\boldsymbol{M}^{-1}(\xi_a))$
0.01	0.1581	0.0853	0.0137	3.8758×10^6
0.05	0.1407	0.0854	0.0654	170599.0
0.10	0.1213	0.0856	0.1224	48687.3
0.20	0.0902	0.0867	0.2095	16614.4
*0.28	0.0724	0.0891	0.2484	11819.3
**0.33	0.0612	0.0933	0.2567	11061.0
0.40	0.0471	0.1048	0.2297	13817.8
0.45	0.0306	0.1248	0.1591	28810.3
0.49	0.0080	0.1557	0.0419	414412.0

*Corresponding D-optimal design, $\left(a=\frac{1-5^{-1/2}}{2}=0.276...\right)$

* *Simplex lattice design (a = 1/3)

From Table 1, we observe that the support points of simplex lattice design *i.e.* ξ_1 are the possible support points of the A-optimal design for the full cubic model.

The next step is to prove the necessary and sufficient condition *i.e.* $\underset{\boldsymbol{x}\in\chi}{\text{Max}}\, d(\boldsymbol{x},\xi_1) = \text{Trace}(\boldsymbol{M}^{-1}(\xi_1))$ has been established as (A1) in Appendix A. In this case, we obtain the value of $\boldsymbol{M}^{-1}(\xi_1)$ by substituting a = 1/3 in equation (6).

We now obtain the A-optimal design for the cubic model without 3-way effect.

3.2. Cubic model without 3-way effect

The expected response for a cubic model without 3-way effect (see Cornell (2002)) is as follows:

$$\eta_2 = \boldsymbol{f}_2'(\boldsymbol{x})\boldsymbol{\beta}_2 = \sum_{i=1}^{q}\beta_i x_i + \sum_{i<j=1}^{q}\beta_{ij}x_i x_j + \sum\sum_{i<j}^{q}\delta_{ij}x_i x_j(x_i - x_j) \tag{12}$$

where $\boldsymbol{f}_2(\boldsymbol{x})$ and $\boldsymbol{\beta}_2$ are column vectors of length q^2 and are defined by

$$\boldsymbol{f}_2(\boldsymbol{x}) = \big(x_1, x_2,..., x_q, x_1x_2, x_1x_3,..., x_{q-1}x_q, x_1x_2(x_1 - x_2), x_1x_3(x_1 - x_3),$$
$$..., x_{q-1}x_q(x_{q-1} - x_q)\big)'$$
$$\boldsymbol{\beta}_2 = \big(\beta_1, \beta_2,..., \beta_q, \beta_{12}, \beta_{13},..., \beta_{q-1q}, \delta_{12}, \delta_{13},..., \delta_{q-1q}\big)'.$$

The non-singular information matrix for the model (12) is given by

$$\boldsymbol{M}(\xi) = \sum_{\text{i}=1}^{m} r_{\text{i}}\, \boldsymbol{f}_2(\boldsymbol{x}_{(\text{i})})\, \boldsymbol{f}_2'(\boldsymbol{x}_{(\text{i})}) \tag{13}$$

In the next theorem, we obtain an A-optimal minimum support design for the model (12) when q = 3.

Theorem 3: For q =3, the design ξ_2 with support points from the corresponding D-optimal design that assigns a weight of 0.0980 to the vertices (1, 0, 0), (0, 1, 0), (0, 0, 1); a weight of 0.1177 to the 6 points (a, 1- a, 0), (a, 0, 1-a),(0, a, 1- a), (1- a, a, 0), (1- a, 0, a),(0, 1 - a, a), with $a = (1-5^{-\frac{1}{2}})/2$ is the A-optimal minimum support design for the cubic polynomial model without 3-way effect on χ.

Proof: Following the similar arguments in Theorem 2, we search for support points of A-optimal design *i.e.* Min $\text{Trace}(\boldsymbol{M}^{-1}(\xi))$ for the cubic model without 3-way effect over the subclass D_2. Here we consider the proposed design ξ_a for the cubic model without 3-way effect given in Section 2. The inverse of the information matrix of the form (13) for the design ξ_a is

$$\boldsymbol{M}^{-1}(\xi_a)=\begin{pmatrix} \frac{1}{r_1}\boldsymbol{I}_3 & \boldsymbol{A}_1' & \boldsymbol{A}_2' \\ \boldsymbol{A}_1 & g_2\boldsymbol{I}_3+g_3\boldsymbol{J}_3 & \boldsymbol{A}_3' \\ \boldsymbol{A}_2 & \boldsymbol{A}_3 & \boldsymbol{A}_4 \end{pmatrix} \tag{14}$$

which is again a submatrix of the information matrix in equation (6). Next, the trace of $\boldsymbol{M}^{-1}(\xi_a)$ is obtained as

$$\text{Trace}\,(\boldsymbol{M}^{-1}(\xi_a))=\frac{(6a(a-1)+3)r_1+3(1-2a)^2(a^2(a-1)^2+1)r_2}{a^2(2a^2-3a+1)^2 r_1 r_2}. \tag{15}$$

Now, the problem becomes minimizing equation (15) subject to the restriction of weights $3r_1+6r_2=1$. To solve this problem, we use the Lagrangian multiplier method and set the Lagrangian function as

$$\psi=\text{Trace}\ (\boldsymbol{M}^{-1}(\xi_a))+\lambda(3r_1+6r_2-1)$$

By taking the partial derivatives of ψ *w.r.t* r_1, r_2, and λ, and set them equal to 0, we get

$$-\frac{3(a^2(a-1)^2+1)}{a^2(a-1)^2 r_1^2}+3\lambda=0, \tag{16}$$

$$\frac{-3-6a(a-1)}{a^2(2a^2-3a+1)^2 r_2^2}+6\lambda=0, \tag{17}$$

$$3r_1+6r_2=1. \tag{18}$$

Next, by solving equations (16) – (18), we numerically compute possible optimal values of r_1, r_2 (rounded off to the fourth place of the decimal) and the corresponding value of Trace $(\boldsymbol{M}^{-1}(\xi_a))$ for different values of a, which are tabulated in Table 2.

Table 2: Trace $(\boldsymbol{M}^{-1}(\xi_a))$ and corresponding weights of cubic model without 3-way effect for different values of a

a	r_1	r_2	Trace $(\boldsymbol{M}^{-1}(\xi_a))$
0.01	0.1372	0.0980	541681.0
0.05	0.1337	0.0998	24850.5
0.10	0.1285	0.1024	7539.0
0.20	0.1142	0.1096	3072.7
*0.28	0.0980	0.1177	2708.1

**0.33	0.0815	0.1259	3194.5
0.40	0.0560	0.1387	5866.4
0.45	0.0310	0.1512	18037.5
0.49	0.0067	0.1633	375443.0

*Corresponding D-optimal design, $\left(a = \frac{1-5^{-1/2}}{2} = 0.276...\right)$

** Simplex lattice design (a = 1/3) excluding the centroid

From Table 2, we observe that the support points of the corresponding D-optimal design *i.e.* ξ_2 are the possible support points of the A-optimal design for the cubic model without 3-way effect.

The next step is to prove the necessary and sufficient condition *i.e.* $\underset{x \in \chi}{\text{Max}}\, d(\boldsymbol{x}, \xi_2) = \text{Trace}\,(\boldsymbol{M}^{-1}(\xi_2))$ has been established as (A2) in Appendix A. In this case, we obtain the value of $\boldsymbol{M}^{-1}(\xi_2)$ by substituting a =0.276393 in equation (14).

In the next part, we obtain the A-optimal minimum support design for a special cubic model.

3.3. Special cubic model

The expected response for the special cubic model (see Cornell (2002)) is as follows:

$$\eta_3 = \boldsymbol{f}_3'(\boldsymbol{x})\boldsymbol{\beta}_3 = \sum_{i=1}^{q} \beta_i x_i + \sum_{i<j=1}^{q} \beta_{ij} x_i x_j + \sum \sum_{i<j<k} \sum \beta_{ijk} x_i x_j x_k \tag{19}$$

where $\boldsymbol{f}_3(\boldsymbol{x})$ and $\boldsymbol{\beta}_3$ are column vectors of length $\frac{q(q^2+5)}{6}$ and are defined as

$$\boldsymbol{f}_3(\boldsymbol{x}) = \left(x_1, x_2, ..., x_q, x_1x_2, x_1x_3, ..., x_{q-1}x_q, x_1x_2x_3, x_1x_2x_4, ..., x_{q-2}x_{q-1}x_q\right)';$$

$$\boldsymbol{\beta}_3' = \left(\beta_1, \beta_2, ..., \beta_q, \beta_{12}, \beta_{13}, ..., \beta_{q-1q}, \beta_{123}, \beta_{124}, ..., \beta_{q-2q-1q}\right)'.$$

The non-singular information matrix for the model (19) is as follows:

$$\boldsymbol{M}(\xi) = \sum_{i=1}^{m} r_i \boldsymbol{f}_3(\boldsymbol{x}_{(i)}) \boldsymbol{f}_3'(\boldsymbol{x}_{(i)}) \tag{20}$$

The next theorem obtains the A-optimal minimum support design for the model (19) when q = 3.

Theorem 4: For q = 3, the weighted simplex-centroid design ξ_3 that assigns a weight of 0.0546 to the vertices (1, 0, 0), (0, 1, 0), (0, 0, 1); a weight of 0.1629 to the barycentre of depth 1 *i.e.*(1/2, 1/2, 0), (1/2, 0, 1/2), (0, 1/2, 1/2); and a weight of 0.3476 to the centroid point (1/3, 1/3, 1/3) is the A-optimal minimum support design for the special cubic polynomial model with mixture experiments on χ.

Proof: Following the similar arguments in Theorem 2, we search for support points of A-optimal design *i.e.* Min Trace $(\boldsymbol{M}^{-1}(\xi))$ for the special cubic model over the subclass D_3. Here we consider the proposed design ξ_a for the special cubic model given in Section 2. The inverse of the information matrix of the form (20) for the design ξ_a is

$$\boldsymbol{M}^{-1}(\xi_a)=\begin{pmatrix} \frac{1}{r_1}\boldsymbol{I}_3 & \boldsymbol{A}_1' & \boldsymbol{\alpha} \\ \boldsymbol{A}_1 & \boldsymbol{A}_2 & \boldsymbol{\rho} \\ \boldsymbol{\alpha}' & \boldsymbol{\rho}' & h_3 \end{pmatrix} \tag{21}$$

where

$$h_1=\frac{r_1+r_2(2a(a-1)+1)}{(a-1)^2a^2r_1r_2},\ h_2=\frac{1}{(a-1)^2r_1},h_3=-\frac{3[r_1+r_2(a(5a-4)+1)]}{a^2(a-1)^2r_1r_2},$$

$$h_3=\frac{27r_1+9[5+a(a-1)(27a(a-1)+26)]r_2}{a^2(a-1)^2r_1r_2}+\frac{729}{r_3},\ h_4=\frac{1}{ar_1(1-a)},h_5=\frac{1}{a^2r_1},$$

$$\boldsymbol{A}_1=\begin{pmatrix} \frac{1}{(a-1)r_1} & -\frac{1}{ar_1} & 0 \\ \frac{1}{(a-1)r_1} & 0 & -\frac{1}{ar_1} \\ 0 & \frac{1}{(a-1)r_1} & -\frac{1}{ar_1} \end{pmatrix},\boldsymbol{A}_2=\begin{pmatrix} h_1 & h_2 & h_4 \\ h_2 & h_1 & h_5 \\ h_4 & h_5 & h_1 \end{pmatrix},$$

$$\boldsymbol{\alpha}=\begin{pmatrix} \frac{3(1-3a)}{(a-1)r_1} \\ \frac{3(3a^2-3a+1)}{ar_1(1-a)} \\ \frac{3(2-3a)}{ar_1} \end{pmatrix},\quad \text{and} \qquad \boldsymbol{\rho}=\begin{pmatrix} -\frac{3[r_1+r_2(a(5a-4)+1)]}{a^2(a-1)^2r_1r_2} \\ -\frac{3[r_1+r_2(7a(a-1)+2)]}{a^2(a-1)^2r_1r_2} \\ -\frac{3[r_1+r_2(a(5a-6)+2)]}{a^2(a-1)^2r_1r_2} \end{pmatrix}.$$

Next, the trace of $\boldsymbol{M}^{-1}(\xi_a)$ is obtained as

$$\text{Trace}\,(\boldsymbol{M}^{-1}(\xi_a))=\frac{729}{r_3}+\frac{30r_1+6[8+a(a-1)(41a(a-1)+40)]r_2}{a^2(a-1)^2r_1r_2} \tag{22}$$

To minimize equation (22) subject to the restriction of weights $3r_1+3r_2+r_3=1$, we set the Lagrangian function as

$$\psi=\text{Trace}\ (\boldsymbol{M}^{-1}(\xi_a))\ +\lambda(3r_1+3r_2+r_3-1)$$

Now, taking the partial derivatives of ψ with respect to r_1, r_2, r_3 and λ, and set them equal to 0, we get

$$\frac{-48+3a(a-1)[-80+a(a-1)(-82+r_1^2\lambda)]}{a^2(a-1)^2 r_1^2}=0 \tag{23}$$

$$\frac{6[8+a(a-1)(41a(a-1)+40)]}{a^2(a-1)^2 r_1 r_2}$$

$$-\frac{30r_1+6[8+a(a-1)(41a(a-1)+40)]r_2}{a^2(a-1)^2 r_1 r_2^2}+3\lambda=0 \tag{24}$$

$$-\frac{729}{r_3^2}+\lambda=0 \tag{25}$$

$$3r_1+3r_2+r_3=1 \tag{26}$$

Next, by solving equations (23) – (26), we numerically compute possible optimal values of r_1, r_2, r_3 (rounded off to the fourth place of decimal) and corresponding value of Trace $(\boldsymbol{M}^{-1}(\xi_a))$ for different values of a, which are tabulated in Table 3.

Table 3: Trace $(\boldsymbol{M}^{-1}(\xi_a))$ and corresponding weights of special cubic model for different values of a

a	r_1	r_2	r_3	Trace $(\boldsymbol{M}^{-1}(\xi_a))$
0.01	0.1818	0.1474	0.0124	4.69727×10^6
0.05	0.1650	0.1483	0.0601	201592.0
0.10	0.1455	0.1495	0.1149	55204.2
0.20	0.1111	0.1527	0.2086	16758.6
0.28	0.0891	0.1556	0.2657	10322.6
0.33	0.0753	0.1580	0.2998	8106.8
0.40	0.0627	0.1608	0.3294	6716.7
0.45	0.0567	0.1623	0.3429	6199.2
*0.50	0.0546	0.1629	0.3476	6033.5

* Simplex centroid design and corresponding D-optimal design (a = 1/2)

From Table 3, we observe that the support points of simplex centroid design *i.e.* ξ_3 are the possible support points of the A-optimal design for the full cubic model.

The next step is to prove the necessary and sufficient condition *i.e.* $\underset{\boldsymbol{x}\in\chi}{\text{Max}}\, d(\boldsymbol{x}, \xi_3) =$ Trace $(\boldsymbol{M}^{-1}(\xi_3))$ has been established as (A3) in Appendix A. In this case, we obtain the value of $\boldsymbol{M}^{-1}(\xi_3)$ by substituting a = 1/2 in equation (21).

4. Discussions and Conclusions

In comparison to D-optimal designs for models for mixture experiments, obtaining A-optimal designs for models with mixture experiments involves more challenges, as the support points in general, are associated with different weights. The present article obtains A-optimal minimum support designs for the three different forms of the cubic model of mixture experiments when the mixture involves three ingredients. We find that the design points of {3, 3} simplex- lattice and simplex-centroid designs are the support points of the obtained A-optimal designs for the full cubic and special cubic models respectively. In the case of the cubic model without 3-way effect, the support points of the corresponding D-optimal designs are the support points of A-optimal designs. One may apply this result to the case of mixture experiments having $q \geq 4$ ingredients. Of course, the task may be complicated for computing the inverse of the information matrix.

Acknowledgments

The authors would like to thank the anonymous referee and the editor for the very constructive suggestions made on an earlier version of this article.

References

Aggrawal, M. L., Singh, P. and Panda, M. K. (2011). A-optimal designs for an additive cubic model, *Statistics and Probability Letters*,**81**(**2**), 259-266.

Cornell, J. A. (2002). *Experiments with Mixtures – Designs, Models, and the Analysis of Mixture Data*. 3rd Edition. Wiley, New York.

Farrell, R. H., Kiefer, J. and Walbran, J. (1967). Optimum multivariate designs. *Proceedings of Fifth Berkeley Symposium on Mathematics, Statistics, and Probability*,**1**, 113-138, University of California Press, Berkley.

Fedorov, V. V. (1971). Design of experiments for linear optimality criteria. *Theory of Probability and its Applications*, **16**(**1**), 189 – 195.

Goos, P. and Syafitri, U. (2014). V-optimal mixture designs for the q^{th} degree model. *Chemometrics and Intelligent Laboratory Systems*,**136**, 173-178.

Goos, P.and Vandebroek, M. (2001). D-optimal response surface designs in the presence of random block effects. *Computational Statistics and Data Analysis*,**37**(**4**), 433-453.

Kiefer, J. (1961). Optimal designs in regression problems II. *The Annals of Mathematical Statistics*, **32**, 298-325.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *The Annals of Statistics*,**2**, 849-879.

Kiefer, J. and Wolfowitz, J. (1959). Optimal designs in regression problems. *The Annals of Mathematical Statistics*, **30**, 271-294.

Lim, Y. B. (1990). D-optimal design for cubic polynomial regression on the *q*-simplex. *Journal of Statistical Planning and Inference*, **25**, 141-152.

Mandal, N. K. and Pal, M. (2017). Optimum mixture designs in some constrained experimental regions. *Communication in Statistics – Theory and Methods*, **46**(**9**), 4240-4249.

Mikaeili, F. (1989). D-optimum design for cubic without 3-way effect on the simplex. *Journal of Statistical Planning and Inference*, **21**, 107-115.

Mikaeili, F. (1993). D-optimum design for full cubic on q-simplex. *Journal of Statistical Planning and Inference*,**35**, 121-130.

Pal, M. and Mandal, N. K. (2021). Optimum designs for parameter estimation in mixture experiments with group synergism. *Communication in Statistics – Theory and Methods*, **50**(**9**), 2001-2014.

Singh, P. and Panda, M.K. (2011). Optimal design for second degree *K*-model for mixture experiments based on weighted simplex centroid design. *Metron-International Journal of Statistics*, **69**(**3**), 251-263.

APPENDIX A

Proof of Theorem 2:

$$d(\boldsymbol{x},\xi_1) = b_1\sum_{i=1}^{3} x_i^2 + b_2\sum_{i<j}^{3} x_i x_j - b_3\left(\sum_{i<j}^{3} x_i^2 x_j + \sum_{i<j}^{3} x_i x_j^2\right) + b_4\sum_{i<j}^{3} x_i^2 x_j^2$$

$$- b_5\left(\sum_{i<j}^{3} x_i^3 x_j + \sum_{i<j}^{3} x_i x_j^3\right) - b_6\sum_{i<j}^{3} x_i^3 x_j^3 + b_7\left(\sum_{i<j}^{3} x_i^4 x_j^2 + \sum_{i<j}^{3} x_i^2 x_j^4\right) - b_8 x_1 x_2 x_3$$

$$+ b_9(x_1^2 x_2 x_3 + x_1 x_2^2 x_3 + x_1 x_2 x_3^2) + b_{10}(x_1^3 x_2 x_3 + x_1 x_2^3 x_3 + x_1 x_2 x_3^3)$$

$$+ b_{11}(x_1^4 x_2 x_3 + x_1 x_2^4 x_3 + x_1 x_2 x_3^4)$$

$$- b_{11}(x_1^3 x_2^2 x_3 + x_1^3 x_2 x_3^2 + x_1 x_2^3 x_3^2 + x_1^2 x_2^3 x_3 + x_1^2 x_2 x_3^3 + x_1 x_2^2 x_3^3)$$

$$- b_{12}(x_1^2 x_2^2 x_3 + x_1^2 x_2 x_3^2 + x_1 x_2^2 x_3^2) + b_{13} x_1^2 x_2^2 x_3^2$$

where

$b_1 = 11061,$ $b_2 = 10794.5,$ $b_3 = 129898,$ $b_4 = 730343,$
$b_5 = 97276.5,$ $b_6 = 2.67057\times 10^6,$ $b_7 = 1.33528\times 10^6,$ $b_8 = 289304,$
$b_9 = 2.04729\times 10^6,$ $b_{10} = 473639,$ $b_{11} = 380398,$ $b_{12} = 9.93829\times 10^6,$
$b_{13} = 4.81873\times 10^7.$

By using Matlab, the value of $d(\boldsymbol{x}, \xi_1)$, at all the support points can be seen to be is equal to Trace $(\boldsymbol{M}^{-1}(\xi_1)) = 11061$. Again using the standard maximize function in Matlab, we find that

$$\underset{\boldsymbol{x}\in\chi}{Max}\ d(\boldsymbol{x},\xi_1) = 11061 \qquad \text{(A1)}$$

over the simplex region χ. Thus equivalence theorem is verified and this proves Theorem 2.

Proof of Theorem 3:

$$d(\boldsymbol{x},\xi_2) = c_1\sum_{i=1}^{3} x_i^2 - c_2\left(\sum_{i<j}^{3} x_i^2 x_j + \sum_{i<j}^{3} x_i x_j^2\right) + c_3\left(\sum_{i<j}^{3} x_i^2 x_j^2\right) - c_4\left(\sum_{i<j}^{3} x_i^3 x_j + \sum_{i<j}^{3} x_i x_j^3\right)$$

$$+c_5\left(\sum_{i<j}^{3}x_i^4x_j^2+\sum_{i<j}^{3}x_i^2x_j^4\right)-c_6\sum_{i<j}^{3}x_i^3x_j^3+c_7\left(x_1^2x_2x_3+x_1x_2^2x_3+x_1x_2x_3^2\right)$$
$$+c_8\left(x_1^3x_2x_3+x_1x_2^3x_3+x_1x_2x_3^3-x_1^2x_2^2x_3-x_1^2x_2x_3^2-x_1x_2^2x_3^2\right)$$
$$+c_9\left(x_1^4x_2x_3+x_1x_2^4x_3+x_1x_2x_3^4-x_1^2x_2^3x_3-x_1^2x_2x_3^3-x_1x_2^2x_3^3-x_1^3x_2^2x_3\right.$$
$$\left.-x_1^3x_2x_3^2-x_1x_2^3x_3^2\right)+c_{10}x_1^2x_2^2x_3^2$$

where

$c_1=2708.1,\quad c_2=18961,\quad c_3=153524,\quad c_4=40643.2,$
$c_5=451458,\quad c_6=902917,\quad c_7=60953.9,\quad c_8=230319,$
$c_9=169365,\quad c_{10}=508094.$

By using Matlab, the value of $d(\boldsymbol{x},\xi_2)$, at all the support points can be seen to be is equal to Trace$(\boldsymbol{M}^{-1}(\xi_2))$ = 2708.1. Again using standard maximize function in Matlab, we find that

$$\underset{\boldsymbol{x}\in\chi}{Max}\ d(\boldsymbol{x},\xi_2)=2708.1 \tag{A2}$$

over the simplex region χ. Thus equivalence theorem is verified and this completes the proof of Theorem 3.

Proof of Theorem 4:

$$d(\boldsymbol{x},\xi_3)=a_1\sum_{i=1}^{3}x_i^2+a_2\sum_{i<j}^{3}x_ix_j-a_3\left(\sum_{i<j}^{3}x_i^2x_j+\sum_{i<j}^{3}x_ix_j^2\right)+a_4\sum_{i<j}^{3}x_i^2x_j^2-a_5x_1x_2x_3$$
$$+a_6\left(x_1^2x_2x_3+x_1x_2^2x_3+x_1x_2x_3^2\right)-a_7\left(x_1^2x_2^2x_3+x_1^2x_2x_3^2+x_1x_2^2x_3^2\right)+a_8x_1^2x_2^2x_3^2$$

where

$a_1=6033.45,\quad a_2=8714.99,\quad a_3=81138,\quad a_4=337960,$
$a_5=201717,\quad a_6=1.26788\times10^6,\quad a_7=5.80653\times10^6,\quad a_8=2.83064\times10^7.$

By using Matlab, the value of $d(\boldsymbol{x},\xi_3)$ at all the support points can be seen to be is equal to Trace$(\boldsymbol{M}^{-1}(\xi_3))$ = 6033.5. Again using the standard maximize function in Matlab, we find that

$$\underset{\boldsymbol{x}\in\chi}{Max}\ d(\boldsymbol{x},\xi_3)=6033.5 \tag{A3}$$

over the simplex region χ. Thus equivalence theorem is verified and this proves Theorem 4.

Statistics and Applications {ISSN 2454-7395(online)}
Volume 20, No. 2, 2022 (New Series), pp 57-71

A Stochastic Modeling of a Monthly Rainfall of Hillsborough County Using Frechet Distribution

H. J. Patel and M. N. Patel
Department of Statistics, School of Sciences, Gujarat University, Ahmedabad-380009, Gujarat, India

Received: 16 May 2021; Revised: 06 August 2021; Accepted: 08 August 2021

Abstract

In this paper we investigate the prediction problem for the monthly rainfall of Hillsborough County at United States of Florida by Markov chain model. We have used the monthly rainfall data from January 1915 to June 2016. Then the data is divided in 11 states and hence, 11 x 11 transition probability matrix (TPM) is prepared. The truncated Frechet distribution is used for the data in each state. To estimate the parameter of the distribution, method of moment and Bayes estimation are used. Using the estimate of the parameter in 11 states prediction method is developed based on Markov chain approach. To validate the proposed method, we have simulated the monthly rainfall for the same period of the original data and the results are compared. The simulated results come out almost similar to the original data. To predict monthly rainfall for future 5000 and 10,000 months a simulation study is also carried out and the results are shown.

Key words: Markov chain; Truncated Frechet model; Rainfall; Bayes estimation; Simulation.

1. Introduction

There will be a high impact of advancement in human necessity on natural events such as rainfall, temperature, precipitation, wind flow et cetera. Since decades there was a very complex pattern observed in climate change which was difficult to predict the parameters by the meteorologists or the hydrologists. There is still an intense scope of research is available in hydrology and meteorology. The hydrological data mainly consists of water and its application such as precipitation, rainfall, humidity level and water storage level of the dam. A rainfall is the one of the natural sources for getting water for drinking, agriculture and industrial use purpose.

The analyses of hydrological and meteorological data have a great importance amongst the scientists and researchers. Researchers must ensure the collection of hydrological data should be efficient and effective which meet the requirements (Stewart, 2015). A data from hydrological networks is used by public and private sectors for variety of applications like designing, operating and maintaining the multipurpose water management systems (USGS, 2006). Three essential elements of life are fresh water, food and house. The data related to rainfall, precipitation, temperature, humidity, wind speed is essential for the planning of any hydrological event. Analysis of rainfall data found useful in cropping pattern, providing drinking water and construction of roads, dams, bridges and culverts. Such analysis will provide useful information to farmers, water resources planner and engineers to assess the availability and requirement of storage of water.

Corresponding author: H. J. Patel
Email: harshil.dare421@gmail.com

There are multiple research studies have been done on rainfall data and its analysis. The analysis of dry and wet spells received a special attention of many scientists, which is another aspect of the rainfall analysis. Singh and Ranade (2009) analyzed wet and dry spells and their extremes across India. Harsha (2017) describes the analysis of rainfall data in Mangalore. The classical procedure is being used for the analysis of rainfall data. To test the random fluctuations and the presence of climate changes in the yearly rainfall data run test and Kolmogorov-Smirnov two sample tests are used. G. Di Baldassarre *et al.* (2006) have used the generalized extreme value distribution to analyze rainfall extremes of northern central Italy based on L-moments and investigate its statistical properties. Nyatuame *et al.* (2014) have performed the statistical analysis for the monthly and yearly rainfall data of Volta region, Ghana using Latin squared design and analysis of variance. For trend analysis of rainfall data the linear regression model is used. Arvind *et al.* (2017) has performed statistical analysis for a rain gauge station in Trichy district. They studied various statistical distributions to analyze the rainfall data.

Not significant work has been done for the statistical analysis using stochastic process Markov chain modeling under various types of distribution, which motivate us to consider this kind of research.

We have used monthly rainfall data of Hillsborough County (latitude 27°54'36.00" N and longitude -82°20'60.00" W) at United States of Florida is considered for the period of January 1915 to June 2016. The data is taken from the pertinent website: https://www.swfwmd.state.fl.us/resources/data-maps/rainfall-summary-data-region. A separate spread sheet is available for the monthly rainfall data. Then the monthly rainfall data of Hillsborough County was concatenated for the period of January 1915 to June 2016 from that web page.

In hydrological research studies multiple statistical approaches have been applied for the estimation. The objective of this study is to develop a statistical model based on Markov chain to estimate and predict the month wise rainfall of the mentioned time period. The span of the rainfall data used is 0.00 to 19.06 mm. To consider the analysis based on the Markov chain we have bifurcated the data into some small numbers of intervals which we called the states of the Markov chain. 11 states are prepared from the data and which are shown in Table A.1.

In Section 3, a transition probability matrix for a Markov chain model is prepared. The truncated Frechet distribution is considered for the rainfall of each states and the estimate of the parameter of the distribution is obtained using the method of moments in Section 4. In Section 5 we have used a Bayesian approach to estimate the parameter of the distribution. A simulation study is considered in Section 6. A detailed algorithm is prepared for estimation and prediction of present and future rainfall data. Discussion about the estimated results is provided in Section 7. The conclusion is presented in the Section 8.

2. Model Creation

The rainfall of the Hillsborough lies between 0.00 mm to 19.06 mm from the period of January 1915 to June 2016 is taken. For the Markov chain model the determination of states is the first aspect. The states should non-overlapping subsets of entire data. Based on the range of our data we have constructed 11 subsets such that each subset possesses sufficient numbers of observations. Looking at the data we have considered the subsets having different length. These subsets we considered as states of our Markov chain model, which are displayed in Table A.1.

3. Markov Chain Model

A discrete parameter Markov process is known as a Markov chain. Here time space is considered as discrete. The Markov chain models are much valuable mechanism in stochastic process, which also indicates that when present value is known then the historical and future values are independent. Sericola (2013) mentioned that the present state of the procedure is known then the best future prediction can be made using very less parameters of Markov chain model.

Mahanta *et al.* (2019) applied Markov chain model for the daily temperature data of Dhaka and Chittagong stations of Bangladesh. The Markov chain model have been used as a process to search its reliability and obtain failure free operational process for long term period can be established specifically for sugar mills by Sharma and Vishwakarma (2014). Zakaria *et al.* (2019) have used the Markov chain model based on the initial state as well as transition from one state to another state for the forecasting pattern of the air pollution index of Miri, Sarawak.

Jain (1986) have also implemented the Markov model for the seasonal variation in patients who are suffering from asthma. Zhou *et al.* (2018) proposed a Markov chain model which provides prediction of daily bike production and attraction of stations with better predictive accuracy based on the daily data collected from Zhongshan city. Al-Anzi and AbuZeina (2016) have provided the hidden Markov Models (HMM) can be used for the natural language processing (NLP) applications. Patel and Patel (2020a, 2020b, 2021) have considered a first order Markov chain model for the prediction of daily high temperature and daily low temperature.

In this study, the 101 years of monthly rainfall of the Hillsborough County is being considered in millimeter (mm). A data of 1218 (=N) observations is taken for the creation of Markov chain model.

Let Z_t, $t = 1, 2, ..., N$ be the rainfall for the month t, and the states are U_1, U_2, U_3 ... U_{11}.

If $P[Z_{t+1} = U_j \mid Z_1 = U_1, ..., Z_{t-1} = U_{t-1}, Z_t = U_i] = P[Z_{t+1} = U_j \mid Z_t = U_i]$, then such model is called first order Markov chain model with 11 states. Here, $P[Z_{t+1} = U_j \mid Z_t = U_i]$ is independent of time t. This transition probability is denoted by pij, $i,j = 1, 2, ... ,11$, which denotes the probability that the monthly rainfall is on any month will belong to state U_j, given that it was in the state U_i a month before. Thus, 11×11 TPM, $M = [m_{ij}]$ is prepared.

The transition frequency from state U_i to U_j denotes the total number of months having rainfall in state U_j from the rainfall of earlier month in state U_i. Such transition frequencies are calculated for each state and hence, transition frequency matrix is prepared which is shown in Table A.2.

Using the transition frequency matrix a transitional probability matrix (TPM) is obtained, dividing by row total of each row to its cell values. Then the value of $(i, j)^{th}$ cell is called transition probability of j^{th} state from i^{th} state. The TPM is given in Table A.3. The cumulative TPM is provided in Table A.4.

4. Truncated Frechet Distribution for Monthly Rainfall

Very limited research work has been done about the analysis of the hydrological data using Markov chain approach along with statistical distribution. Various types of statistical distributions like exponential distribution, Weibull distribution, Gamma distribution, extreme value distribution, Frechet distribution are used to analyze the data related to meteorological data like temperature, as well as hydrological data like rainfall, wind flow, water storage capacity and precipitation. Patel and Patel (2020 a, 2020 b, 2021) have considered the truncated exponential distribution and generalized exponential distribution for the analysis of the data related to daily low and high temperature of the Ahmedabad, Gujarat, India.

In this paper we have considered Truncated Frechet distribution to analyze the monthly rainfall data the Hillsborough County. Frechet distribution is named after a French mathematician Maurice Rene Frechet, who developed it in 1920 as a maximum value distribution. Frechet distribution is a special case of generalized extreme value distribution which is also named as extreme value type II distribution. This distribution is also referred as inverse Weibull distribution. Kotz and Nadarajah (2000) describe this distribution and discussed its various application in different fields such as rainfall, wind speeds, track race records, natural calamities and so on. Ramos *et al.* (2017) have presented the parameter estimation for the Frechet distribution in the presence of cure fraction.

Recently Ramos *et al.* (2020) have considered various methods of classical and Bayesian estimation of the parameters of the Frechet distribution. They have described the application of this distribution for five real data sets related to the minimum flow of water on Piracicaba river in Brazil.

The probability density function (pdf) of Frechet distribution:

$$g\,(x,\,\propto) = \frac{\propto}{\sigma}\left(\frac{x}{\sigma}\right)^{-1-\propto} e^{-\left(\frac{x}{\sigma}\right)^{-\propto}};\ \propto > 0;\ \sigma > 0;\ x > 0. \tag{1}$$

We have used truncated Frechet distribution to analyze the monthly rainfall data considering σ=1 in equation (1).

$$g\,(x,\,\propto) = \propto x^{-1-\propto} e^{-x^{-\propto}};\ \propto > 0;\ \sigma > 0;\ x > 0. \tag{2}$$

From equation (1) the pdf of truncated Frechet distribution whose range lies between a and b is obtained by:

$$g\,(x,\propto) = \frac{\mathrm{f}\,(\mathrm{x},\ \propto)}{F(b,\ \propto) - F\,(a,\ \propto)},\ 0 < a < x < b,\ x > 0;\ \propto > 0. \tag{3}$$

where $F\ (x, \propto) = e^{-x^{-\propto}}$ can be represented as and the equation (3) can be re-written as

$$g\,(x\,/a < x < b) = \frac{\propto x^{-1-\propto} e^{-x^{-\propto}}}{e^{-b^{-\propto}} - e^{-a^{-\propto}}},\ a < x < b,\ x > 0. \tag{4}$$

The cumulative distribution function for Frechet distribution is represented as follows:

$$G\,(x\,/a < x < b) = \frac{e^{-x^{-\propto}} - G\,(a)}{G\,(b) - G\,(a)},\ a < x < b \tag{5}$$

Mean of the truncated Frechet distribution is given by

$$E(X) = \frac{\Gamma(1-\frac{1}{\propto})}{e^{-b^{-\propto}} - e^{-a^{-\propto}}} \quad (6)$$

The value of $\propto_j$ is estimated by using the method of moment by equating the observed mean with the mean of the truncated Frechet distribution of the j^{th} state, for j = 1, 2, ... , 11. The moment estimates of the parameters of 11 states are shown in Table A.5. For fitting of the truncated Frechet distribution in each state, the chi-square test of goodness of fit is performed and found that the p-values for each state appeared as > 0.05. The graph of state wise observed and expected frequencies is given below. Based on the Figure B.1 we also confirm that the Frechet distribution works well for the monthly rainfall data of each state.

5. Bayes Estimation

The Bayesian method has been applied to assess the parameters of a hydrological model. The Bayesian method also provides an estimate of uncertainty of model parameters by using prior probability distribution of the parameters. Rainfall data contains significant uncertainty, the Bayesian method has been used by several researchers to consolidate rainfall uncertainty in model calibration (Sun *et al.* (2017)). Engeland and Gottschalk (2002) have used Bayesian approach for estimation of parameters in a regional hydrological model for NOPEX area in southern Sweden. Badjana *et al.* (2017) have used Bayesian approach to investigate the long term trend in annual rainfall, annual rainfall duration and annual maximum rainfall for seven stations at Kara river basin, West Africa. The trend analysis was performed by fitting the Log normal, Normal and Generalised extreme value distribution to the annual rainfall data.

The similar type of research work around Bayesian analysis and statistical modeling can be found in, for example, Fortin *et al.* (1997), P.H.A.J.M Van Gelder (1996) and Noortwijk *et al.* (1998). Morita (1993) has applied the Bayesian estimates as the symptomatic tool for the clinical practice. Various priors of the Bayes estimators based on the power law distribution, of the double Gamma-Exponential distribution has the minimum posterior standard error as well as minimum Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC) by Sultan *et al.* (2014).

Verma *et al.* (2019) has proved that Bayesian technique is quite helpful if any prior data information is available, which reduces the variability for making the effective clinically meaningful decisions.

In this section Bayes estimates of the parameters of truncated Frechet distribution under squared error loss function are derived for 11 states. The prior distribution for the j^{th} state is considered as exponential distribution with mean θ_j having pdf

$$\pi_j(\propto_j) = \frac{1}{\theta_j} e^{\frac{-\propto_j}{\theta_j}};\ \propto_j > 0,\ \theta_j > 0,\ j = 1,2, \dots ,12. \quad (7)$$

That is $\propto_j \sim Exp\ (mean\ \theta_j),\ j = 1,2, \dots ,12.$ (8)

The likelihood function based on the observations $x_{1j}, x_{2j}, \ldots x_{n_j j}$ of the j^{th} state is given by

$$L(\underline{x} / \alpha_j) = \prod_{i=1}^{n_j} \frac{\alpha_j x_{ij}^{1-\alpha_j} e^{-x_{ij}^{-\alpha_j}}}{e^{-b^{-\alpha_j}} - e^{-a^{-\alpha_j}}} \tag{9}$$

Using likelihood function and prior distribution, the posterior distribution of θ_j for j^{th} state is obtained as:

$$h(\alpha_j / \underline{x}) \sim L(\underline{x} / \alpha_j)\, \pi(\alpha_j)$$

$$= \frac{\alpha_j^{n_j} \prod_{i=1}^{n_j} x_{ij}^{-\alpha_j} e^{-\sum_{i=1}^{n_j} x_{ij}^{-\alpha_j}}}{(e^{-b^{-\alpha_j}} - e^{-a^{-\alpha_j}})^{n_j} \prod_{i=1}^{n_j} x_{ij}} \frac{1}{\theta_j} e^{\frac{-\alpha_j}{\theta_j}} \tag{10}$$

$$\sim \frac{\frac{1}{\theta_j} \alpha_j^{n_j} e^{-(\sum_{i=1}^{n_j} log x_{ij} + \frac{1}{\theta}) \alpha_j} e^{-\sum_{i=1}^{n_j} x_i^{-\alpha_j}}}{(e^{-b^{-\alpha_j}} - e^{-a^{-\alpha_j}})^{n_j}}, \theta_j > 0 \tag{11}$$

Under squared error loss function the Bayes estimator of θ_j is nothing but mean of its posterior distribution.

That is, $\widehat{\alpha_j}_{Bayes} = E_h(\alpha_j / \underline{x}), j = 1,2, \ldots, 11.$ (12)

$$\widehat{\alpha_j}_{Bayes} = \int_0^{\infty} \frac{\frac{1}{\theta_j} \alpha_j^{n_j+1} e^{-(\sum_{i=1}^{j} log x_{ij} + \frac{1}{\theta}) \alpha} e^{-\sum_{i=1}^{n_j} x_i^{-\alpha_j}}}{k(e^{-b^{-\alpha_j}} - e^{-a^{-\alpha_j}})^{n_j}} d\alpha_j \tag{13}$$

where

$$k = \int_0^{\infty} \frac{\frac{1}{\theta_j} \alpha_j^{n_j} e^{-(\sum_{i=1}^{n_j} log x_{ij} + \frac{1}{\theta}) \alpha_j} e^{-\sum_{i=1}^{n_j} x_i^{-\alpha_j}}}{(e^{-b^{-\alpha_j}} - e^{-a^{-\alpha_j}})^{n_j}} d\alpha_j$$

is a function of $\underline{x}$, independent of α_j.

Here Bayes estimate cannot be simplified and obtained in a closed form. So, we use the important sampling method, proposed by Kundu *et al.* (2009).
We rewrite the posterior distribution of α_j as

$$h(\widehat{\alpha_j}_{Bayes} / \underline{x}) = Gamma(n_j + 2, \sum_{i=1}^{n_j} log\, x_{ji} + \frac{1}{\theta_j}) \omega(\alpha_j) \tag{14}$$

where $\omega(\alpha_j) = \frac{e^{-\sum_{i=1}^{j} x_i^{-\alpha}}}{\theta\, (e^{-b^{-\alpha}} - e^{-a^{-\alpha}})^{n_j}}$ (15)

Using important sampling the Bayes estimates of the α_j can be obtained by following algorithm:

Step 1: Generate $\propto_j$ from Gamma ($n_j + 2$, $\sum_{i=1}^{n_j} log\, x_{ji} + \frac{1}{\beta_j}$) distribution.
Step 2: Repeat the above steps S=1000 times to generate ($\propto_{j1}, \propto_{j2}, \ldots \propto_{jS}$).
Step 3: Compute the S values of $\omega(\alpha_j)$ using the values of $\propto_j$ in Step 2.
Step 4: The Bayes estimate of parameter $\propto_j$ is given by

$$\widehat{\propto_j}_{Bayes} = \frac{\sum_{i=1}^{S} \propto_{ji} * \omega(\alpha_{ji})}{\sum_{i=1}^{S} \omega(\alpha_{ji})}$$

The values of the Bayes estimates of the parameters obtained for all the states are given in Table A.6.

6. Simulation and Prediction

In this section we check the performance of the proposed methods of prediction. We consider a simulation to check whether the simulated results are approximately accurate to the original data or not. To estimate the monthly rainfall, the moment estimates and Bayes estimates of the parameters of the 11 states are used.

6.1. Simulation algorithm

The simulations algorithm steps are mentioned below:

1. Let us consider the initial state as the state observed for the first value of the rainfall data. say j (j = 1, 2, 3, … 11). Generate the uniform random number from uniform distribution U (0, 1), say rnx.
2. To decide the next state, say l, the random value (rnx) is compared with the cumulative transition probabilities of the state j, till the random value (rnx) outstrip the cumulative transition probability of the state.
3. Let us consider the relevant values of a parameter for l-th state from Table A.6.
4. Insert the value of a parameter in the cumulative distribution function of truncated Frechet distribution.

$$F(x / a_j < x < b_j) = \frac{\propto_j x^{-1-\propto_j} e^{-x^{-\propto_j}}}{e^{-b^{-\propto_j}} - e^{-a^{-\propto_j}}}, \text{j} = 1,2, \ldots, 11. \quad (16)$$

Here (a_j, b_j) are the lower and upper limits of the j^{th} state respectively. Replace $F(x / a_j < x < b_j)$ by the random number between 0 to 1 in Equation (16).

5. Solving the Equation (16) we get the estimate of rainfall for next month.
6. Continue the step 1 to step 6 by considering initial state j=l till we have 1218 estimated rainfall values.

In similar manner, prediction for future monthly rainfall is being done using the above steps considering the initial state j as the state of the last rainfall value of the data. The simulation is continuing for next 5000 and 10,000 months. The estimation is carried out for the monthly rainfall of the Hillsborough County of the same period from January 1915 to June 2016 under the proposed methods. The average rainfall obtained from both the methods reflect almost close to each other.

The simulated results obtained from moment estimates and Bayes estimates, the descriptive statistics (minimum, maximum, average and standard deviation) are presented in Table A.7. A comparison of state wise frequencies obtained through the proposed methods is made with frequencies obtained based on the actual data. The results are shown in the Table A.8 and Table A.9. The prediction is being carried out for the future months. Using 101 years of monthly rainfall data of Hillsborough County the next 5000 and 10,000 months of rainfall can be predicted through method of moments and Bayes estimation.

A prediction is being made for number of months and percentage for the rainfall, higher than 0.50mm, 2.00mm, 4.00mm, 7.00mm, 9.00mm and 11.00mm as well as the numbers of months having rainfall below 0.50mm, 2.00mm, 4.00mm, 7.00mm, 9.00mm and 11.00mm of Hillsborough County (refer Table A.10 and Table A.11).

7. Results and Discussion

The state wise frequency obtain from the method of moments (MOM) and Bayes estimation are almost near to the actual data. Prediction made by method of moments is almost similar to the results obtained under the Bayes estimation. The prediction done under the proposed methods are near to actual value which reveals that the prediction based on truncated Frechet distribution under the Markov model is appropriate.

Based on Table A.8, The state wise frequency and percentage results achieved thorough both these methods are completely identical to each other. The outcome obtained through method of moments and Bayes estimation for most of the states are very much similar in frequency and percentage values of the observed data.

The Table A.10 and Table A.11 exhibits that there are approximately 95% chances of having < 11.00mm rainfall during the next 5,000 and 10,000 months. In a similar way there are only around 5% probability of having > 11.00mm rainfall during the next 5,000 and 10,000 days.

8. Conclusion

In this paper we have analysed monthly rainfall data of Hillsborough County at United states of Florida. The overall data is divided into 11 states. We have applied the truncated Frechet distribution for the monthly rainfall data for each state. Two types of approaches have been jointly used *viz*: 1. Markov chain model and 2. Distribution theory approach. To estimate the parameters of the distribution we have used method of moments and method of Bayes estimation. In case of the Bayes estimation important sampling is used to estimate the parameters. Simulation study is considered to judge the performance of the proposed methods and for prediction of future monthly rainfall. The models work good for estimation and prediction of the monthly rainfall. This work may be useful to water resource management to take precautionary steps in advance on the basis of future predictions.

Acknowledgments

The authors sincerely appreciate the valuable comments and suggestions provided by the editor(s) during their thorough review process, which makes this research article more profound.

References

Al-Anzi, F. S. and AbuZeina, D. (2016). A survey of Markov chain models in Linguistics applications. *Computer Science and Information Technology (CS & IT)*, 53-62.

Arvind, G., Kumar, P. A., Karthi, S. G. and Suribabu, C. R. (2017). Statistical analysis of 30 years rainfall data: A case study. *IOP Conf. Series: Earth and Environmental Science*, **80(1)**.

Badjana, H. M., Renard, B., Helmschrot, J., Edjame K. S., Afouda, A. and Wala, K. (2017). Bayesian trend analysis in annual rainfall total, duration and maximum in the Kara River, basin (West Africa). *Journal of Hydrology: Regional Studies*, **13**, 255-273.

Baldassarre, G. Di., Casterllarin, A. and Brath, A. (2006). Relationships between statistics of rainfall extremes and mean annual precipitation: an application for design-storm estimation in northern central Italy. *Hydrology and Earth System Sciences Discussions, European Geosciences Union*, **10(4)**, 589-601.

Engeland, K. and Gottschalk, L. (2002). Bayesian estimation of parameters in a regional hydrological model. *Hydrological and Earth System Sciences*, **6(5)**, 883-898.

Fortin, V., Bernier, J. and Bobée, B. (1997). Simulation, Bayes, and bootstrap in statistical hydrology. *Water Resources Research*, **33(3)**, 439-448.

Harsha, S. (2017). Role of statistics in monitoring rainfall. *Journal of Environmental Science, Toxicology and Food Technology*, **11(3)**, 2319-2402.

https: //www.swfwmd.state.fl.us/resources/data-maps/rainfall-summary-data-region.

Jain, S. (1986). Markov chain model and its application. *Computers and Biomedical Research*, **19(4)**, 374-378.

Kotz, S. and Nadarajah S. (2000). *Extreme Value Distributions: Theory and Application*. World Scientific.

Mahanta, J., Nath, S. K. and Rashid, Md. H. (2019). Using Markov analysis to study the Impact of temperature in Bangladesh. *Asia Pacific Journal of Energy and Environment*, **6(2)**, 69-76.

Morita, S. (1993). Use of Bayesian estimation as a diagnostic tool in clinical practice. *The Japanese Journal of Anesthesiology*, **42(5)**, 733-737.

Nyatuame, M., Owusu-Gyimah, V. and Ampiaw, F. (2014). Statistical analysis of rainfall trend for volta region in Ghana. *International Journal of Atmospheric Sciences*, http://dx.doi.org/10.1155/2014/203245.

P. H. A. J. M. Van Gelder (1996). How to deal with wave statistical and model uncertainties in the design of vertical breakwaters? In H. G. Voortman *Probabilistic Design Tools for Vertical Breakwaters, Proceedings Task* 4 *Meeting, Hannover, Germany,* (*MAST III/PROVERBS: MAS3-CT95-0041*), 1-13.

Patel, H. J. and Patel M. N. (2020a). A stochastic model of daily temperature for Ahmedabad, India. *International Journal of Environmental Studies*, **77(6)**, 916-927.

Patel, H. J. and Patel M. N. (2020b). Statistical analysis of daily low temperature of Ahmedabad city using stochastic process. *International Journal of Agricultural and Statistical Sciences*, **16(2)**, 573-582.

Patel, H. J. and Patel M. N. (2021). A generalized exponential model based analysis of daily low temperature data of Ahmedabad city using Markov Chain approach. *International Journal on Emerging Technologies*, **12(1)**, 1-6.

Ramos, P. L., Louzada, F., Ramos., E. and Dey, S. (2020). The Frechet distribution: Estimation and application – An overview. *Journal of Statistics and Management Systems*, **23(3)**, 549-578.

Ramos, P., Nascimento, D. and Louzada, F. (2017). The long term Frechet distribution: Estimation, properties and its application. *Biom Biostat International Journal*, **6**, 00170.

Sericola, B. (2013). *Markov Chains: Theory, Algorithms and Applications*. ISTE Ltd and John Wiley & Sons Inc.

Sharma, S. P. and Vishwakarma, Y. (2014). Application of Markov process in performance analysis of feeding system of sugar industry. *Journal of Industrial Mathematics*.

Singh, N. and Ranade, A. (2009). *Climatic and Hydroclimatic Features of Wet and Dry Spells and Their Extremes Across India*. Research Report, Indian Institute of Tropical Meteorology, Pune.

Stewart, B. (2015). Measuring what we manage – the importance of hydrological data to water resources management. *Hydrological Sciences and Water Security: Past, Present and Future*, **366**, 80-85.

Sun, S., Leonhardt, G., Sandoval, S., Bertrand-Krajewski, J. L. and Rauch, W. (2017). A Bayesian method for missing rainfall estimation using a conceptual rainfall-runoff model. *Hydrological Sciences Journal*, **62(15)**, 2456-2468

United States Geological Survey (USGS) (2006). Benefit of the USGS stream gaging program – Users and uses of US stream flow data. http://water.usgs.gov/osw/pubs/nhwc_report.pdf

Van, Noortwijk, J. M. and Van, Gelder, P. H. A. J. M. (1998). Bayesian estimation of quantiles for the purpose of flood prevention. *Coastal Engineering Proceedings*, **27**, 3529-3541. doi: 10.9753/icce.v26.%p

Verma, V., Mishra, A. K. and Narang R. (2019). Application of Bayesian analysis in medical diagnosis. *Journal of the Practice of Cardiovascular Sciences*, **5(3)**, 136-141.

Zakaria, N. N., Othman, M., Sokkalingam, R., Daud, H., Abdullah, L. and Kadir, E. A. (2019). Markov Chain model development for forecasting. *Air Pollution Index of Miri, Sarawak. Sustainability*, **11**.

Zhou, Y., Wang, L., Zhong, R. and Tan, Y. (2018). A Markov Chain based demand prediction model for stations in bike sharing systems. *Mathematical Problems in Engineering Systems,* **2018**, 1-8.

Appendix A

Table A.1: States for rainfall of Hillsborough

States	Rainfall
1	0.00--0.50
2	0.51--1.00
3	1.01--1.50
4	1.51--2.00
5	2.01--3.00
6	3.01--4.00
7	4.01--5.50
8	5.51--7.00
9	7.01--9.00
10	9.01--11.00
11	11.01--19.06

Table A.2: Transition frequency matrix for rainfall of Hillsborough

i \ j	Transition frequency											Row Total
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	
1	10	9	13	5	14	13	6	5	4	0	0	79
2	11	9	14	6	12	14	7	9	7	2	2	93
3	8	16	12	19	14	13	10	5	3	0	1	101
4	8	13	7	15	19	12	14	7	6	2	0	103
5	15	17	14	13	30	23	19	11	9	6	5	162
6	8	12	11	13	18	16	18	19	6	7	5	133
7	7	9	10	17	16	13	17	17	21	8	8	143
8	7	4	6	4	16	13	23	21	21	13	10	138
9	2	2	9	7	12	9	12	22	32	21	7	135
10	0	0	1	1	8	8	10	10	18	18	9	83
11	3	9	4	2	3	2	6	5	8	5	6	53

Table A.3: Transition probability matrix (TPM) for rainfall of Hillsborough

	Transition probability										
$i \times j$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	0.1266	0.1139	0.1646	0.0633	0.1772	0.1646	0.0759	0.0633	0.0506	0.0000	0.0000
2	0.1183	0.0968	0.1505	0.0645	0.1290	0.1505	0.0753	0.0968	0.0753	0.0215	0.0215
3	0.0792	0.1584	0.1188	0.1881	0.1386	0.1287	0.0990	0.0495	0.0297	0.0000	0.0099
4	0.0777	0.1262	0.0680	0.1456	0.1845	0.1165	0.1359	0.0680	0.0583	0.0194	0.0000
5	0.0926	0.1049	0.0864	0.0802	0.1852	0.1420	0.1173	0.0679	0.0556	0.0370	0.0309
6	0.0602	0.0902	0.0827	0.0977	0.1353	0.1203	0.1353	0.1429	0.0451	0.0526	0.0376
7	0.0490	0.0629	0.0699	0.1189	0.1119	0.0909	0.1189	0.1189	0.1469	0.0559	0.0559
8	0.0507	0.0290	0.0435	0.0290	0.1159	0.0942	0.1667	0.1522	0.1522	0.0942	0.0725
9	0.0148	0.0148	0.0667	0.0519	0.0889	0.0667	0.0889	0.1630	0.2370	0.1556	0.0519
10	0.0000	0.0000	0.0120	0.0120	0.0964	0.0964	0.1205	0.1205	0.2169	0.2169	0.1084
11	0.0566	0.1698	0.0755	0.0377	0.0566	0.0377	0.1132	0.0943	0.1509	0.0943	0.1132

Table A.4: Transition cumulative probability matrix for rainfall of Hillsborough

	Transition cumulative probability										
$i \times j$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	0.1266	0.2405	0.4051	0.4684	0.6456	0.8101	0.8861	0.9494	1.0000	1.0000	1.0000
2	0.1183	0.2151	0.3656	0.4301	0.5591	0.7097	0.7849	0.8817	0.9570	0.9785	1.0000
3	0.0792	0.2376	0.3564	0.5446	0.6832	0.8119	0.9109	0.9604	0.9901	0.9901	1.0000
4	0.0777	0.2039	0.2718	0.4175	0.6019	0.7184	0.8544	0.9223	0.9806	1.0000	1.0000
5	0.0926	0.1975	0.2840	0.3642	0.5494	0.6914	0.8086	0.8765	0.9321	0.9691	1.0000
6	0.0602	0.1504	0.2331	0.3308	0.4662	0.5865	0.7218	0.8647	0.9098	0.9624	1.0000
7	0.0490	0.1119	0.1818	0.3007	0.4126	0.5035	0.6224	0.7413	0.8881	0.9441	1.0000
8	0.0507	0.0797	0.1232	0.1522	0.2681	0.3623	0.5290	0.6812	0.8333	0.9275	1.0000
9	0.0148	0.0296	0.0963	0.1481	0.2370	0.3037	0.3926	0.5556	0.7926	0.9481	1.0000
10	0.0000	0.0000	0.0120	0.0241	0.1205	0.2169	0.3373	0.4578	0.6747	0.8916	1.0000
11	0.0566	0.2264	0.3019	0.3396	0.3962	0.4340	0.5472	0.6415	0.7925	0.8868	1.0000

Table A.5: Moment estimates of $\propto_j$ for each state

State	Moment Estimates ($\widehat{\propto}$)
1	1.00024008
2	1.00486095
3	1.00532035
4	1.00425636
5	1.01327133
6	1.00793604
7	1.01226263
8	1.00726517
9	1.00860501
10	1.00545760
11	1.04569409

Table A.6: Bayes estimates of $\propto_j$ for each state

State	Bayes Estimates ($\widehat{\propto}$)
1	1.001338
2	1.005927
3	1.006125
4	1.002706
5	1.053751
6	1.009213
7	1.049640
8	1.008237
9	1.010347
10	1.006579
11	1.073141

Table A.7: Descriptive statistics for observed and simulated results

Statistics	Observed	Simulations using Moment estimates			Simulations using Bayes estimates		
	N=1218	N=1218	N=5000	N=10,000	N=1218	N=5000	N=10,000
Minimum rainfall (mm)	0.00	0.11	0.09	0.09	0.11	0.09	0.09
Maximum rainfall (mm)	19.06	19.05	19.06	19.06	19.05	19.06	19.06
Average rainfall (mm)	4.38	4.60	4.70	4.63	4.60	4.70	4.63
Standard deviation of rainfall (mm)	3.39	4.27	4.40	4.27	4.27	4.40	4.27

Table A.8: Observed and estimated frequencies based on method of moments and Bayes estimates

Sr. No.	State	Actual data	Method of Moments		
		N=1218	N=1218	N=5000	N=10,000
1	0.0-0.5	79 (6.49%)	81 (6.65%)	355 (7.10%)	676 (6.76%)
2	0.51-1.0	93 (7.64%)	102 (8.37%)	405 (8.10%)	824 (8.24%)
3	1.01-1.50	101 (8.29%)	94 (7.72%)	414 (8.28%)	815 (8.15%)
4	1.51-2.0	103 (8.46%)	105 (8.62%)	418 (8.36%)	840 (8.40%)
5	2.01-3.00	162 (13.30%)	168 (13.79%)	662 (13.24%)	1350 (13.50%)
6	3.01-4.00	133 (10.92%)	155 (12.73%)	563 (11.26%)	1172 (11.72%)
7	4.01-5.50	143 (11.74%)	141 (11.58%)	593 (11.86%)	1164 (11.64%)
8	5.51-7.00	138 (11.33%)	117 (9.61%)	509 (10.18%)	1041 (10.41%)
9	7.01-9.00	135 (11.08%)	123 (10.10%)	526 (10.52%)	1052 (10.52%)
10	9.01-11.00	78 (6.40%)	78 (6.40%)	302 (6.04%)	619 (6.19%)
11	11.01-19.06	53 (4.35%)	54 (4.43%)	253 (5.06%)	447 (4.47%)

Table A.9: Observed and estimated frequencies based on Bayes estimates

Sr. No.	State	Actual data	Bayes		
		N=1218	N=1218	N=5000	N=10,000
1	0.00-0.50	79 (6.49%)	81 (6.65%)	355 (7.10%)	676 (6.76%)
2	0.51-1.00	93 (7.64%)	102 (8.37%)	405 (8.10%)	824 (8.24%)
3	1.01-1.50	101 (8.29%)	94 (7.72%)	414 (8.28%)	815 (8.15%)
4	1.51-2.00	103 (8.46%)	105 (8.62%)	418 (8.36%)	840 (8.40%)
5	2.01-3.00	162 (13.30%)	168 (13.79%)	662 (13.24%)	1350 (13.50%)
6	3.01-4.00	133 (10.92%)	155 (12.73%)	563 (11.26%)	1172 (11.72%)
7	4.01-5.50	143 (11.74%)	141 (11.58%)	593 (11.86%)	1164 (11.64%)
8	5.51-7.00	138 (11.33%)	117 (9.61%)	509 (10.18%)	1041 (10.41%)
9	7.01-9.00	135 (11.08%)	123 (10.10%)	526 (10.52%)	1052 (10.52%)
10	9.01-11.00	78 (6.40%)	78 (6.40%)	302 (6.04%)	619 (6.19%)
11	11.01-19.06	53 (4.35%)	54 (4.43%)	253 (5.06%)	447 (4.47%)

Table A.10: Frequency of simulated observations with different ranges based on method of moments

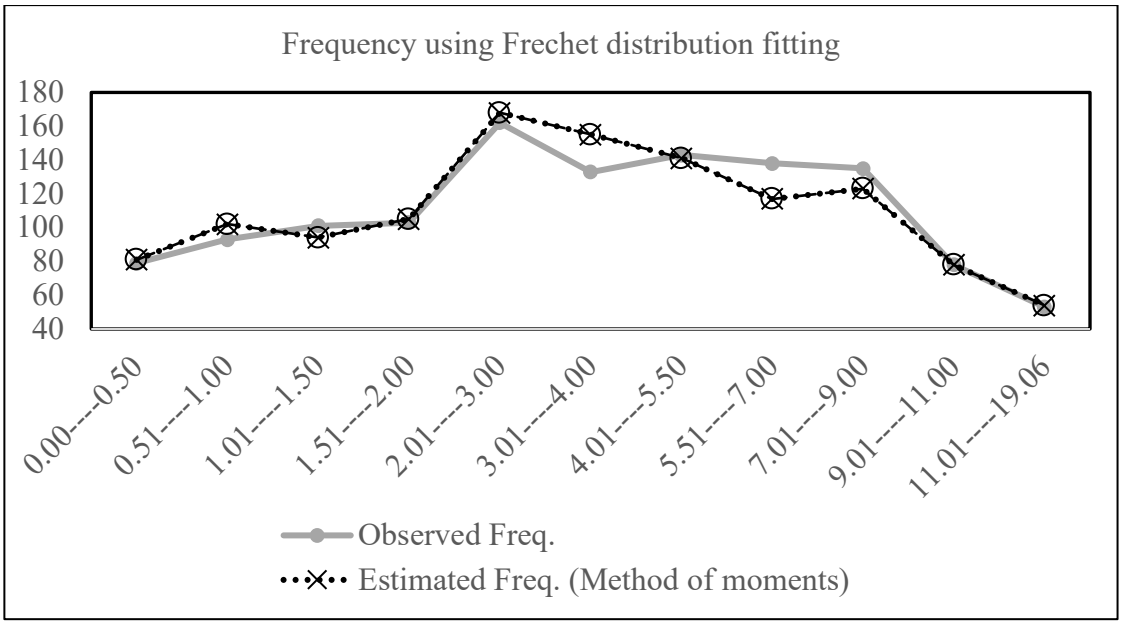
Sr. No.	Rainfall (mm)	Method of Moments		
		N=1218	N=5000	N=10,000
1	<0.50	81 (6.65%)	355 (7.10%)	676 (6.76%)
2	<2.00	382 (31.36%)	1592 (31.84%)	3155 (31.55%)
3	<4.00	705 (57.88%)	2817 (56.34%)	5677 (56.77%)
4	<7.00	963 (79.06%)	3919 (78.38%)	7882 (78.82%)
5	<9.00	1086 (89.16%)	4445 (88.90%)	8934 (89.34%)
6	<11.00	1164 (95.57%)	4747 (94.94%)	9553 (95.53%)
7	>0.50	1137 (93.35%)	4645 (92.90%)	9324 (93.24%)
8	>2.00	836 (68.64%)	3408 (68.16%)	6845 (68.45%)
9	>4.00	513 (42.12%)	2183 (43.66%)	4323 (43.23%)
10	>7.00	255 (20.94%)	1081 (21.62%)	2118 (21.18%)
11	>9.00	132 (10.84%)	555 (11.10%)	1066 (10.66%)
12	>11.00	54 (4.43%)	253 (5.06%)	447 (4.47%)

Table A.11: Frequency of simulated observations with different ranges based on Bayes estimates

Sr. No.	Rainfall (mm)	Bayes estimates		
		N=1218	N=5000	N=10,000
1	<0.50	81 (6.65%)	355 (7.10%)	676 (6.76%)
2	<2.00	382 (31.36%)	1592 (31.84%)	3155 (31.55%)
3	<4.00	705 (57.88%)	2817 (56.34%)	5677 (56.77%)
4	<7.00	963 (79.06%)	3919 (78.38%)	7882 (78.82%)
5	<9.00	1086 (89.16%)	4445 (88.90%)	8934 (89.34%)
6	<11.00	1164 (95.57%)	4747 (94.94%)	9553 (95.53%)
7	>0.50	1137 (93.35%)	4645 (92.90%)	9324 (93.24%)
8	>2.00	836 (68.64%)	3408 (68.16%)	6845 (68.45%)
9	>4.00	513 (42.12%)	2183 (43.66%)	4323 (43.23%)
10	>7.00	255 (20.94%)	1081 (21.62%)	2118 (21.18%)
11	>9.00	132 (10.84%)	555 (11.10%)	1066 (10.66%)
12	>11.00	54 (4.43%)	253 (5.06%)	447 (4.47%)

Appendix B

Table B.1: State wise observed and expected frequency of monthly rainfall data of Hillsborough County



Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 73–92

Transmuted Sine - G Family of Distributions: Theory and Applications

K.M. Sakthivel and J. Rajkumar
Department of Statistics
Bharathiar University, Coimbatore, Tamil Nadu, India.

Received: 28 April 2021; Revised: 17 July 2021; Accepted: 10 August 2021

Abstract

In this paper, we introduced transmuted sine - G family and it's mathematical properties. Recently, the statistical relevance and applicability of trigonometric distributions got much attention among researchers for modeling various real-time phenomena. This paper contributes to a core area of statistics by investigating a new trigonometric family of probability distributions defined from the alliance of the families known as transmuted and sine - G family with the inspiring name of transmuted sine generating (TS - G) family. The characteristics of this new family are studied through analytical, graphical and numerical approaches. In addition, we observe the fact that the TS - G family can generate original, simple and pliant trigonometric models for statistical purposes. This fact is revealed with the special TS - G model based on the Weibull model and discussed maximum likelihood estimation with real time application.

Key words: Sine - G family; Transmuted family; Weibull distribution; Maximum likelihood estimation.

AMS Subject Classifications: 60E05, 62E10, 62F10

1. Introduction

Statistical distribution is very useful in describing and predicting real-world phenomena. Life time distributions are playing a vital role in many area of research such as economics, engineering, finance, medicine, biological science, amongst others. Further analyzing life time data are imperative. Recent developments focus on designing and generating new families of probability distributions that extend well-known probability distributions and at the same time provide great flexibility in modeling real time data. In recent years, the generalization of probability distributions has attracted many statisticians. For example, exponentiated family of distributions proposed by Gupta *et al.* (1999). The exponentiated family of probability distributions provides flexibility by adding one more parameter to the base distribution. Several classes of probability distributions have been introduced by adding one or more parameters to generate new family of probability distributions in the statistical literature. Examples of such families are the exp - G family by Gupta *et al.* (2001),

Corresponding Author: K.M.Sakthivel
Email: sakthithebest@gmail.com

Weibull - G family by Bourguignon *et al.* (2014), Topp - Leone generated (TL - G) family by Al-Shomrani *et al.* (2016), a new extended alpha power transformed - G by Ahmad *et al.* (2020), a new alpha power transformed - G by Elbatal *et al.* (2019), truncated inverted Kumaraswamy - G by Bantan *et al.* (2019) , type II general inverse exponential - G by Jamal *et al.* (2020), exponentiated truncated inverse Weibull - G by Almarashi *et al.* (2020) and type II power TL - G by Bantan *et al.* (2020). The quadratic rank transmutation map was introduced by Shaw *et al.* (2007) to generate new family of distributions. Recent studies have highlighted the statistical relevance and applicability of trigonometric distributions for modeling many phenomena. Kharazmi and Saadatinik (2016) introduced Hyperbolic Cosine - F (HCF) family and Sakthivel *et al.* (2020) proposed Hyperbolic Cosine Rayleigh distribution and studied some of it's mathematical properties with application to breaking stress of carbon fibers. Kumar *et al.* (2015) and Souza (2015) introduced the sine - G family with use of the sine function. This paper introduced a new family of distribution namely transmuted sine - G (TS - G) family. The characteristics of this family are studied through graphical and numerical approaches.

In this paper, we introduce a new probability distribution namely transmuted sine - G family and studied some of it's properties. In section 2, we present the transmuted probability models. Section 3 discuss sine - G family. In section 4, we propose some transmuted sine - G family of distributions. In section 5, present statistical properties of transmuted sine Weibull distribution. The reliability analysis of proposed model is discussed in section 6. The maximum likelihood estimation for the parameters of transmuted sine Weibull distribution is presented given in section 7. The application of transmuted sine Weibull (TS Weibull) distribution is studied using real data set in section 8 and conclusions of this work is presented in section 9.

2. Transmuted Distribution

The quadratic rank transmutation map introduced by Shaw *et al.* (2007). The cumulative distribution function (cdf) $F(x)$ and probability density function (pdf) $f(x)$ are defined as follows:
The cdf of transmuted family of distribution is defined as

$$F(x) = (1+\lambda)G(x) - \lambda G^2(x)\,; \qquad |\lambda| \leq 1 \tag{1}$$

The pdf of transmuted family of distribution is defined as

$$f(x) = g(x)[(1+\lambda) - 2\lambda G(x)]\,; \qquad |\lambda| \leq 1 \tag{2}$$

where λ is a parameter of transmutation; $G(x)$ is the cdf and $g(x)$ is the pdf of the baseline distribution respectively.

3. Sine - G Family

The method of generating new family of probability distributions using sine transformation was introduced by Kumar *et al.* (2015) and Souza (2015). The cdf and pdf of sine -

G family distribution can be obtained as follows:
The cdf of sine - G distribution is defined as

$$G(x) = \sin\left(\frac{\pi}{2}H(x)\right) \tag{3}$$

The pdf of sine - G distribution is defined as

$$g(x) = \frac{\pi}{2}h(x)\cos\left(\frac{\pi}{2}H(x)\right) \tag{4}$$

where $h(x)$ and $H(x)$ are pdf and cdf of baseline distribution respectively.

4. Proposed Model

4.1. Transmuted sine family

This paper contributes to the subject by investigating a new trigonometric family of probability distributions defined from the alliance of the families known as transmuted distribution and sine - G family and it is named as transmuted sine - G family (TS - G).
The cdf of transmuted sine - G family is defined as

$$F(x) = (1+\lambda)\sin\left(\frac{\pi}{2}H(x)\right) - \lambda\left[\sin\left(\frac{\pi}{2}H(x)\right)\right]^2 ; \qquad |\lambda| \leq 1 \tag{5}$$

The pdf of transmuted sine - G family is

$$f(x) = \frac{\pi}{2}h(x)\cos\left(\frac{\pi}{2}H(x)\right)\left[(1+\lambda) - 2\lambda\sin\left(\frac{\pi}{2}H(x)\right)\right] ; \qquad |\lambda| \leq 1 \tag{6}$$

where λ is a parameter of transmutation. The $h(x)$ and $H(x)$ are pdf and cdf of baseline distribution respectively.

4.2. Transmuted sine exponential distribution

The cdf of exponential distribution with parameter θ is

$$H(x) = 1 - e^{-\theta x} ; \quad x > 0 ; \theta > 0 \tag{7}$$

The pdf of exponential distribution with parameter θ is

$$h(x) = \theta e^{-\theta x} ; \quad x > 0 ; \theta > 0 \tag{8}$$

where θ is a rate parameter.
The cdf of transmuted sine exponential family is given by

$$F(x) = (1+\lambda)\sin\left(\frac{\pi}{2}\left(1-e^{-\theta x}\right)\right) - \lambda\left[\sin\left(\frac{\pi}{2}\left(1-e^{-\theta x}\right)\right)\right]^2 ; \quad x > 0 ; |\lambda| \leq 1, \theta > 0 \tag{9}$$

The pdf of transmuted sine exponential family is given by

$$f(x) = \frac{\pi}{2}\left(\theta e^{-\theta x}\right)\cos\left(\frac{\pi}{2}\left(1-e^{-\theta x}\right)\right)\left[(1+\lambda) - 2\lambda\sin\left(\frac{\pi}{2}\left(1-e^{-\theta x}\right)\right)\right]; \quad x>0\,;|\lambda|\leq 1, \theta>0 \tag{10}$$

where θ is a rate parameter and λ is parameter of transmutation.
The r^{th} moment is defined as

$$E\left(X^r\right) = \theta\sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^i}{(2i)!}\left[(1+\lambda)\gamma_{k,r} - 2\lambda\delta_{j,l,r}\right] \tag{11}$$

where (12)

$$\gamma_{k,r} = \sum_{k=0}^{2i}\left(\begin{array}{c} 2i \\ k \end{array}\right)\frac{(-1)^k r!}{(\theta(k+1))^{r+1}}$$

and (13)

$$\delta_{j,l,r} = \sum_{j=0}^{\infty}\sum_{l=0}^{2i+2j+1}\left(\frac{\pi}{2}\right)^{2j+1}\left(\begin{array}{c} 2i+2j+1 \\ l \end{array}\right)\frac{(-1)^i(-1)^l}{(2j+1)!}\frac{r!}{(\theta(l+1))^{r+1}}$$

The moment generating function is defined as

$$M_X(t) = \theta\sum_{i=0}^{\infty}\sum_{r=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^i}{(2i)!}\frac{t^r}{r!}\left[(1+\lambda)\gamma_{k,r} - 2\lambda\delta_{j,l,r}\right] \tag{14}$$

The first four moments are given below

$$E\left(X\right) = \theta\sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^i}{(2i)!}\left[(1+\lambda)\gamma_{k,1} - 2\lambda\delta_{j,l,1}\right] \tag{15}$$

$$E\left(X^2\right) = \theta\sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^i}{(2i)!}\left[(1+\lambda)\gamma_{k,2} - 2\lambda\delta_{j,l,2}\right] \tag{16}$$

$$E\left(X^3\right) = \theta\sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^i}{(2i)!}\left[(1+\lambda)\gamma_{k,3} - 2\lambda\delta_{j,l,3}\right] \tag{17}$$

$$E\left(X^4\right) = \theta\sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^i}{(2i)!}\left[(1+\lambda)\gamma_{k,4} - 2\lambda\delta_{j,l,4}\right] \tag{18}$$

The r^{th}moment of TS exponential distribution is expressed as series. One can easily verify the convergence of this series by using Cauchy ratio test.

4.3. Transmuted sine modified Weibull distribution

The modified Weibull distribution was introduced by Zaindin *et al.* (2009) and the cdf of modified Weibull distribution is given by

$$H(x) = \left(1-e^{-\alpha x-\beta x^{\theta}}\right)\;; \quad x>0\,;\alpha,\beta\geq 0, \theta>0 \tag{19}$$

such that $\alpha+\beta>0$. Here β and θ are shape parameters and α is a scale parameter. The pdf of modified Weibull distribution is given by

$$h(x)=\left(\alpha+\theta\beta x^{\theta-1}\right)e^{-\alpha x-\beta x^{\theta}}\;;\quad x>0\,;\alpha,\beta\geq 0,\theta>0 \tag{20}$$

The cdf of transmuted sine modified Weibull distribution is

$$\begin{aligned}F(x) &= (1+\lambda)\sin\left(\frac{\pi}{2}\left(1-e^{-\alpha x-\beta x^{\theta}}\right)\right)-\lambda\left[\sin\left(\frac{\pi}{2}\left(1-e^{-\alpha x-\beta x^{\theta}}\right)\right)\right]^{2};\\ &\qquad x>0\,;|\lambda|\leq 1,\alpha,\beta\geq 0,\theta>0\end{aligned} \tag{21}$$

The pdf of transmuted sine modified Weibull distribution is defined as

$$\begin{aligned}f(x)=\frac{\pi}{2}\left(\alpha+\theta\beta x^{\theta-1}\right)e^{-\alpha x-\beta x^{\theta}}\cos\left(\frac{\pi}{2}\left(1-e^{-\alpha x-\beta x^{\theta}}\right)\right)\\ \left[(1+\lambda)-2\lambda\sin\left(\frac{\pi}{2}\left(1-e^{-\alpha x-\beta x^{\theta}}\right)\right)\right];\\ x>0\,;|\lambda|\leq 1,\alpha,\beta\geq 0,\theta>0\end{aligned} \tag{22}$$

The r^{th} moment of TS modified Weibull distribution is defined as

$$E\left(X^{r}\right) = \sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^{i}}{(2i)!}\left[(1+\lambda)\gamma_{k,l,r}-2\lambda\delta_{j,m,n,r}\right] \tag{23}$$

where

$$\gamma_{k,l,r} = \sum_{k=0}^{2l}\sum_{l=0}^{\infty}\binom{2i}{k}\frac{(-1)^{k+l}(\beta k)^{l}}{l!}\left[\frac{(r+l\theta)!}{k\,(\alpha k)^{r+l\theta}}+\frac{\beta\theta\,(r+l\,(\theta+1)-1)}{(\alpha k)^{r+l(\theta+1)}}\right]$$

and

$$\begin{aligned}\delta_{j,m,n,r} = \sum_{j=0}^{\infty}\sum_{m=0}^{2i+2j+1}\sum_{n=0}^{\infty}\left(\frac{\pi}{2}\right)^{2j+1}\binom{2i+2j+1}{m}\frac{(-1)^{j+m+n}\,(\beta m)^{n}}{(2j+1)!}\\ \left[\frac{(r+n\theta)!}{m\,(\alpha m)^{r+n\theta}}+\frac{\beta\theta\,(r+n\theta)!}{(\alpha m)^{r+n(\theta+1)}}\right]\end{aligned}$$

The moment generation function of TS modified Weibull distribution is defined as

$$M_{X}(t) = \sum_{i=0}^{\infty}\sum_{r=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^{i}}{(2i)!}\frac{t^{r}}{r!}\left[(1+\lambda)\gamma_{k,l,r}-2\lambda\delta_{j,m,n,r}\right] \tag{24}$$

The first four moments are given below

$$E\left(X\right) = \sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^{i}}{(2i)!}\left[(1+\lambda)\gamma_{k,l,1}-2\lambda\delta_{j,m,n,1}\right] \tag{25}$$

$$E\left(X^{2}\right) = \sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^{i}}{(2i)!}\left[(1+\lambda)\gamma_{k,l,2}-2\lambda\delta_{j,m,n,2}\right] \tag{26}$$

$$E\left(X^{3}\right) = \sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^{i}}{(2i)!}\left[(1+\lambda)\gamma_{k,l,3}-2\lambda\delta_{j,m,n,3}\right] \tag{27}$$

$$E\left(X^{4}\right) = \sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^{i}}{(2i)!}\left[(1+\lambda)\gamma_{k,l,4}-2\lambda\delta_{j,m,n,4}\right] \tag{28}$$

4.4. Weibull distribution

The Weibull distribution was introduced by Swedish physicist Waloddi Weibull (1951). He applied it on modelling yield strength of materials. The Weibull distribution is popular and widely used in many fields such as engineering, reliability, failure analysis, lifetime analysis, material science, quality control, physics, medicine, meteorology, hydrology and others. However, there are cases when standard Weibull distribution fails to model data adequately enough for certain types of data. Hence, it is necessary to apply generalized Weibull distribution because of it's flexibility and suitability for such type of data. Later, the importance of this type of generalization has been proved in recent years on various problems.

The cdf of Weibull distribution is given by

$$H(x) = 1 - e^{-\eta x^{\theta}}; \quad x > 0; \eta, \theta > 0 \tag{29}$$

The pdf of Weibull distribution is given by

$$h(x) = \eta\theta x^{\theta-1} e^{-\eta x^{\theta}}; \quad x > 0; \eta, \theta > 0 \tag{30}$$

where θ is a shape parameter and η is a scale parameter.

4.5. Transmuted sine Weibull family

The cdf of transmuted sine Weibull (TS Weibull) distribution is given as

$$F(x) = (1+\lambda)\sin\left(\frac{\pi}{2}\left(1 - e^{-\eta x^{\theta}}\right)\right) - \lambda\left[\sin\left(\frac{\pi}{2}\left(1 - e^{-\eta x^{\theta}}\right)\right)\right]^2; \quad x > 0; \eta, \theta > 0, |\lambda| \leq 1 \tag{31}$$

The pdf of transmuted sine Weibull (TS Weibull) distribution is given as

$$f(x) = \frac{\pi}{2}\eta\theta x^{\theta-1} e^{-\eta x^{\theta}} \cos\left(\frac{\pi}{2}\left(1 - e^{-\eta x^{\theta}}\right)\right) \left[(1+\lambda) - 2\lambda\sin\left(\frac{\pi}{2}\left(1 - e^{-\eta x^{\theta}}\right)\right)\right]; x > 0; \eta, \theta > 0, |\lambda| \leq 1 \tag{32}$$

where θ is a shape parameter, η is a scale parameter and λ is transmuting parameter.

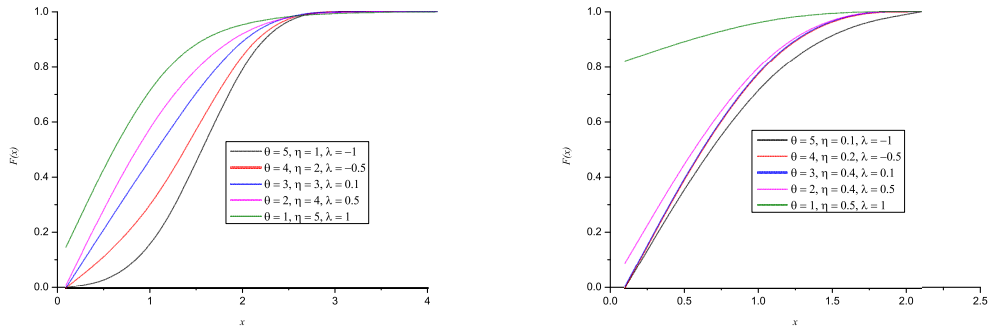


Figure 1: Plots for cdf of TS Weibull distribution for different values of the parameters.

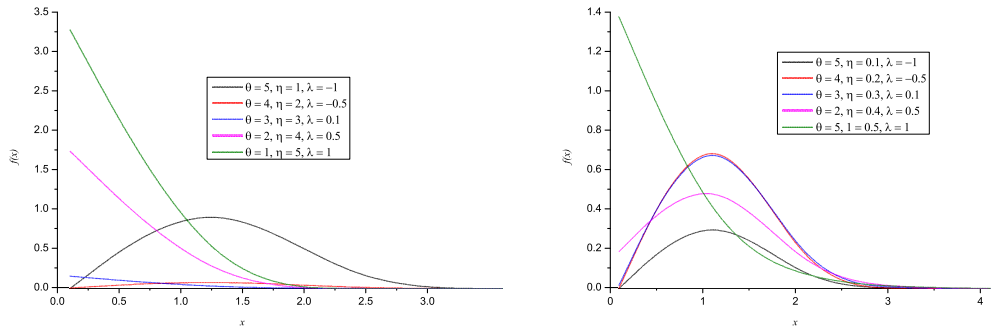


Figure 2: Plots for pdf of TS Weibull distribution for different values of the parameters.

5. Statistical Properties

5.1. Moments

We obtained an expression for the r^{th} moment of TS Weibull distribution as

$$E\left(X^r\right) = \eta\theta \sum_{i=0}^{\infty} \left(\frac{\pi}{2}\right)^{2i+1} \frac{(-1)^i}{(2i)!} \left[(1+\lambda)\gamma_{k,r} - 2\lambda\delta_{j,l,r,}\right] \tag{33}$$

where

$$\gamma_{k,r} = \sum_{k=0}^{2i} \binom{2i}{k} \frac{(-1)^k}{\eta\theta(k+1)} \frac{\left(\frac{r}{\theta}\right)!}{(\eta(k+1))^{\frac{r}{\theta}}}$$

and

$$\delta_{j,l,r} = \sum_{j=0}^{\infty} \sum_{l=0}^{2i+2j+1} \frac{(-1)^j(-1)^l}{(2j+1)!} \binom{2i+2j+1}{l} \left(\frac{\pi}{2}\right)^{2j+1}$$

The first four moments are given below

$$E(X) = \eta\theta\sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^i}{(2i)!}\left[(1+\lambda)\gamma_{k,1}-2\lambda\delta_{j,l,1}\right] \quad (34)$$

$$E\left(X^2\right) = \eta\theta\sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^i}{(2i)!}\left[(1+\lambda)\gamma_{k,2}-2\lambda\delta_{j,l,2}\right] \quad (35)$$

$$E\left(X^3\right) = \eta\theta\sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^i}{(2i)!}\left[(1+\lambda)\gamma_{k,3}-2\lambda\delta_{j,l,3}\right] \quad (36)$$

$$E\left(X^4\right) = \eta\theta\sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^i}{(2i)!}\left[(1+\lambda)\gamma_{k,4}-2\lambda\delta_{j,l,4}\right] \quad (37)$$

The r^{th}moment of TS Weibull distribution is expressed as series. One can easily verify the convergence of this series by using Cauchy ratio test.

Variance

$$V(X) = \eta\theta\sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^i}{(2i)!}\left[(1+\lambda)\gamma_{k,2}-2\lambda\delta_{j,l,2}\right] - \left[\eta\theta\sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^i}{(2i)!}\left[(1+\lambda)\gamma_{k,1}-2\lambda\delta_{j,l,1}\right]\right]^2 \quad (38)$$

The moment generating function of TS Weibull distribution is given below

$$M_X(t) = \sum_{r=0}^{\infty}\frac{t^r}{r!}\eta\theta\sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^i}{(2i)!}\left[(1+\lambda)\gamma_{k,1}-2\lambda\delta_{j,l,r}\right] \quad (39)$$

The cumulant generating function of TS Weibull distribution is given below

$$K_X(t) = \log\left[\sum_{t=0}^{\infty}\frac{t^t}{r!}\eta\theta\sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^i}{(2i)!}\left[(1+\lambda)\gamma_{k,1}-2\lambda\delta_{\mathrm{j},l,r}\right]\right] \quad (40)$$

The characteristic function of TS Weibull distribution is given below

$$\phi_X(t) = \sum_{r=0}^{\infty}\frac{(it)^r}{r!}\eta\theta\sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^i}{(2i)!}\left[(1+\lambda)\gamma_{k,1}-2\lambda\delta_{j,l,r}\right] \quad (41)$$

5.2. Quantile function

The quantile function of TS Weibull distribution is given by

$$Q(p) = \left[\frac{\lambda\sum_{i=0}^{\infty}\sum_{k=0}^{4i+2}\frac{(-1)^{i+k}}{(2i+1)!}\left(\frac{\pi}{2}\right)^{4i+2}\binom{4i+2}{k}(\eta k) - (1+\lambda)\sum_{i=0}^{\infty}\sum_{k=0}^{2i+1}\frac{(-1)^{i+k}}{(2i+1)!}\left(\frac{\pi}{2}\right)^{2i+1}\binom{2i+1}{k}(\eta k)}{(\log p)}\right]^{1/\theta} \quad (42)$$

5.3. Generalized entropy

The generalized entropy of TS Weibull distribution is given below

$$GE(w,\psi) = \frac{\eta\theta\sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^i}{(2i)!}\left[(1+\lambda)\gamma_{k,\psi}-2\lambda\delta_{j,l,\psi}\right]}{\psi(\psi-1)\left[\eta\theta\sum_{i=0}^{\infty}\left(\frac{\pi}{2}\right)^{2i+1}\frac{(-1)^i}{(2i)!}\left[(1+\lambda)\gamma_{k,1}-2\lambda\delta_{j,l,1]}\right]\right]^{\psi}}-1 \quad (43)$$

5.4. Asymptotic behaviours

The asymptotic behaviours of transmuted sine Weibull distribution is given below

$$\begin{aligned}
\lim_{x\to 0} f(x:\eta,\theta,\lambda) &= 0 \text{ and } \lim_{x\to\infty} f(x:\eta,\theta,\lambda)=0\\
\lim_{x\to 0} f(x:\eta,\theta,\lambda) &= \frac{\pi}{2}\eta\theta\lim_{x\to 0}x^{\theta-1}e^{-\eta x^{\theta}}\cos\left(\frac{\pi}{2}\left(1-e^{-\eta x^{\theta}}\right)\right)\\
&\quad\times\left[(1+\lambda)-2\lambda\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x^{\theta}}\right)\right)\right]\\
&= \frac{\pi}{2}\eta\theta\times(0)=0\\
\Rightarrow \lim_{x\to 0} f(x:\eta,\theta,\lambda) &= 0 \qquad (44)
\end{aligned}$$

$$\begin{aligned}
\lim_{x\to\infty} f(x:\eta,\theta,\lambda) &= \frac{\pi}{2}\eta\theta\lim_{x\to\infty}x^{\theta-1}\lim_{x\to\infty}e^{-\eta x^{\theta}}\\
&\quad\times\lim_{x\to\infty}\cos\left(\frac{\pi}{2}\left(1-e^{-\eta x^{\theta}}\right)\right)\\
&\quad\times\lim_{x\to\infty}\left[(1+\lambda)-2\lambda\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x^{\theta}}\right)\right)\right]\\
&= \frac{\pi}{2}\eta\theta\times(0)=0\\
\Rightarrow \lim_{x\to\infty} f(x:\eta,\theta,\lambda) &= 0 \qquad (45)
\end{aligned}$$

5.5. Order statistics

The pdf of j^{th} order statistic of TS Weibull distribution is given by

$$\begin{aligned}
f_{X_{(j)}}(x) &= \frac{n!}{(j-1)!(n-j)!}\frac{\pi}{2}\eta\theta x^{\theta-1}e^{-\eta x^{\theta}}\\
&\quad\cos\left(\frac{\pi}{2}\left(1-e^{-\eta x^{\theta}}\right)\right)\\
&\quad\left[(1+\lambda)-2\lambda\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x^{\theta}}\right)\right)\right]\\
&\quad\left[(1+\lambda)\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x^{\theta}}\right)\right)-\lambda\left[\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x^{\theta}}\right)\right)\right]^2\right]^{j-1}\\
&\quad\left[1-\left[(1+\lambda)\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x^{\theta}}\right)\right)-\lambda\left[\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x^{\theta}}\right)\right)\right]^2\right]^{n-j}\right]
\end{aligned}$$

The pdf of largest order statistic $X_{(n)}$ is given by

$$\begin{aligned} f_{X_{(n)}}(x) &= n\frac{\pi}{2}\eta\theta x^{\theta-1}e^{-\eta x^{\theta}}\cos\left(\frac{\pi}{2}\left(1-e^{-\eta x^{\theta}}\right)\right) \\ &\quad \left[(1+\lambda)-2\lambda\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x^{\theta}}\right)\right)\right] \\ &\quad \left[(1+\lambda)\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x^{\theta}}\right)\right)-\lambda\left[\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x^{\theta}}\right)\right)\right]^2\right]^{n-1} \end{aligned}$$

The pdf of smallest order statistic $X_{(1)}$ is given by

$$\begin{aligned} f_{X_{(1)}}(x) &= n\frac{\pi}{2}\eta\theta x^{\theta-1}e^{-\eta x^{\theta}}\cos\left(\frac{\pi}{2}\left(1-e^{-\eta x^{\theta}}\right)\right)\left[(1+\lambda)-2\lambda\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x^{\theta}}\right)\right)\right] \\ &\quad \left[1-\left[(1+\lambda)\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x^{\theta}}\right)\right)-\lambda\left[\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x^{\theta}}\right)\right)\right]^2\right]\right]^{n-1} \end{aligned}$$

5.6. Stochastic ordering

Stochastic ordering of positive continuous random variables are an important tool for judging their comparative behaviour. A random variable X is said to be smaller than a random variable Y then

(1) Stochastic order ($X \leq_{st} Y$) if $F_X(x) \geq F_Y(x)$ for all x
(2) Hazard rate order ($X \leq_{hr} Y$) if $h_X(x) \geq h_Y(x)$ for all x
(3) Mean residual life order ($X \leq_{mrl} Y$) if $m_X(x) \leq m_Y(x)$ for all x
(4) Likelihood ratio order ($X \leq_{lr} Y$) if $f_X(x)/f_Y(x)$ decreases in x.

The following implications based on these properties are illustrated by Yadav *et al.* (2019) and Shaked *et al.* (1995).

$$(X \leq_{lr} Y) \Rightarrow (X \leq_{hr} Y) \Rightarrow (X \leq_{mrl} Y)$$

and hence

$$(X \leq_{hr} Y) \Rightarrow (X \leq_{st} Y)$$

The following theorem shows that the TS Weibull random variable is ordered with respect to the strongest likelihood ratio ordering.

Theorem 1:

Let $X \sim TSW(\eta_1, \theta_1, \lambda_1)$ and $Y \sim TSW(\eta_2, \theta_2, \lambda_2)$ the following under conditions

(i) $\eta_1 = \eta_2, \lambda_1 = \lambda_2$ and $\theta_1 > \theta_2$
(ii) $\eta_1 = \eta_2, \lambda_1 > \lambda_2$ and $\theta_1 = \theta_2$
(iii) $\eta_1 > \eta_2, \lambda_1 = \lambda_2$ and $\theta_1 = \theta_2$

then $(X \leq_{lr} Y)$ and hence it implies other ordering.

Proof:

$$\frac{f_X(x)}{f_Y(y)} = \frac{\frac{\pi}{2}\eta_1\theta_1 x^{\theta_1-1}e^{-\eta x^{\theta_1}}\cos\left(\frac{\pi}{2}\left(1-e^{-\eta_1 x^{\theta_1}}\right)\right)\left[(1+\lambda_1)-2\lambda_1\sin\left(\frac{\pi}{2}\left(1-e^{-\eta_1 x^{\theta_1}}\right)\right)\right]}{\frac{\pi}{2}\eta_2\theta_2 x^{\theta_2-1}e^{-\eta_2 x^{\theta_2}}\cos\left(\frac{\pi}{2}\left(1-e^{-\eta_2 x^{\theta_2}}\right)\right)\left[(1+\lambda_2)-2\lambda_2\sin\left(\frac{\pi}{2}\left(1-e^{-\eta_2 x^{\theta_2}}\right)\right)\right]}$$

Taking logarithm of both sides, we can write

$$\log\frac{f_X(x)}{f_Y(y)} = \log\left[\frac{\frac{\pi}{2}\eta_1\theta_1 x^{\theta_1-1}e^{-\eta x^{\theta_1}}\cos\left(\frac{\pi}{2}\left(1-e^{-\eta_1 x^{\theta_1}}\right)\right)\left[(1+\lambda_1)-2\lambda_1\sin\left(\frac{\pi}{2}\left(1-e^{-\eta_1 x^{\theta_1}}\right)\right)\right]}{\frac{\pi}{2}\eta_2\theta_2 x^{\theta_2-1}e^{-\eta_2 x^{\theta_2}}\cos\left(\frac{\pi}{2}\left(1-e^{-\eta_2 x^{\theta_2}}\right)\right)\left[(1+\lambda_2)-2\lambda_2\sin\left(\frac{\pi}{2}\left(1-e^{-\eta_2 x^{\theta_2}}\right)\right)\right]}\right].$$

Taking partial derivative on both sides, we write

$$\begin{aligned}\frac{d}{dx}\log\frac{f_X(x)}{f_Y(y)} &= \frac{(\theta_1-1)\,x^{\theta_1-2}}{x^{\theta_1-1}} - \frac{(\theta_2-1)\,x^{\theta_2-2}}{x^{\theta_2-1}} - \eta_1\theta_1 x^{\theta_1-1} + \eta_2\theta_2 x^{\theta_2-1} \\ &\quad - \frac{\frac{\pi}{2}\eta_1\theta_1 x^{\theta_1-1}e^{-\eta x^{\theta_1}}\sin\left(\frac{\pi}{2}\left(1-e^{-\eta_1 x^{\theta_1}}\right)\right)}{\cos\left(\frac{\pi}{2}\left(1-e^{-\eta_1 x^{\theta_2}}\right)\right)} \\ &\quad + \frac{2\lambda_1\cos\left(\frac{\pi}{2}\left(1-e^{-\eta_1 x^{\theta_1}}\right)\right)\frac{\pi}{2}\eta_1\theta_1 x^{\theta_1-1}e^{-\eta x^{\theta_1}}}{1+\lambda_1-2\lambda_1\sin\left(\frac{\pi}{2}\left(1-e^{-\eta_1 x^{\theta_1}}\right)\right)} \\ &\quad + \frac{\frac{\pi}{2}\eta_2\theta_2 x^{\theta_2-1}e^{-\eta_2 x^{\theta_2}}\sin\left(\frac{\pi}{2}\left(1-e^{-\eta_2 x^{2}}\right)\right)}{\cos\left(\frac{\pi}{2}\left(1-e^{-\eta_2 x^{\theta_2}}\right)\right)} \\ &\quad + \frac{2\lambda_2\cos\left(\frac{\pi}{2}\left(1-e^{-\eta_2 x^{\theta_2}}\right)\right)\frac{\pi}{2}\eta_2\theta_2 x^{\theta_2-1}e^{-\eta_2 x^{\theta_2}}}{1+\lambda_2-2\lambda_2\sin\left(\frac{\pi}{2}\left(1-e^{-\eta_2 x^{\theta_2}}\right)\right)}\end{aligned}$$

It can be easily verified that under conditions (i), (ii) and (iii) then $\frac{d}{dx}\log\frac{f_X(x)}{f_Y(y)} < 0$. This means that $(X \leq_{lr} Y)$ and hence $(X \leq_{hr} Y), (X \leq_{mrl} Y)$ and $(X \leq_{st} Y)$.

6. Reliability Analysis

The survival function of TS Weibull distribution is defined as

$$R(t) = 1 - \left[(1+\lambda)\sin\left(\frac{\pi}{2}\left(1-e^{-\eta t^{\theta}}\right)\right) - \lambda\left[\sin\left(\frac{\pi}{2}\left(1-e^{-\eta t^{\theta}}\right)\right)\right]^2\right] \tag{46}$$

The hazard rate function is

$$h(t) = \frac{\frac{\pi}{2}\eta\theta t^{\theta-1}e^{-\eta t^{\theta}}\cos\left(\frac{\pi}{2}\left(1-e^{-\eta t^{\theta}}\right)\right)\left[(1+\lambda)-2\lambda\sin\left(\frac{\pi}{2}\left(1-e^{-\eta t^{\theta}}\right)\right)\right]}{1-\left[(1+\lambda)\sin\left(\frac{\pi}{2}\left(1-e^{-\eta t^{\theta}}\right)\right) - \lambda\left[\sin\left(\frac{\pi}{2}\left(1-e^{-\eta t^{\theta}}\right)\right)\right]^2\right]} \tag{47}$$

(a) $\theta = 2, \eta = 1$ and $\lambda = 1$ hazard shape is linear.
(b) $\theta = 1, \eta = 1$ and $\lambda = 1$ hazard shape is increasing decreasing increasing function.
(c) $\theta = 0.5, \eta = 1$ and $\lambda = -1$ hazard shape is decreasing function.
(d) $\theta = 0.5, \eta = 1$ and $\lambda = 1$ hazard shape is increasing function.
(e) $\theta = 1.2, \eta = 1$ and $\lambda = -1$ hazard shape is inverse bathtab function.
(f) $\theta = 1.2, \eta = 1.5$ and $\lambda = -1$ hazard shape is unimodal function.

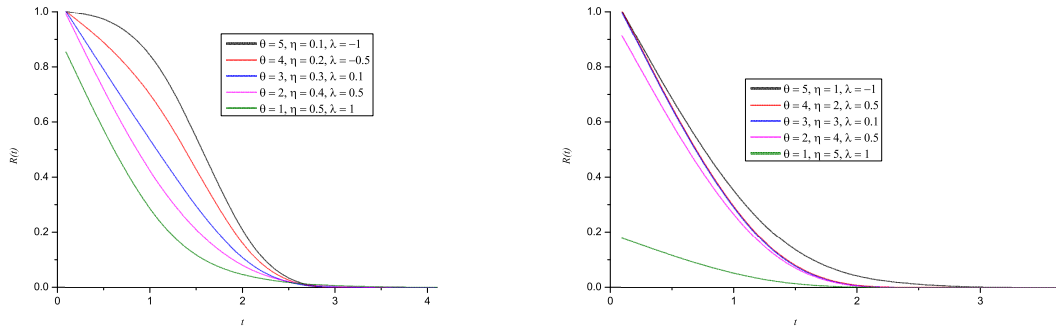


Figure 3: Plots for reliability function of TS Weibull distribution for different values of the parameters.

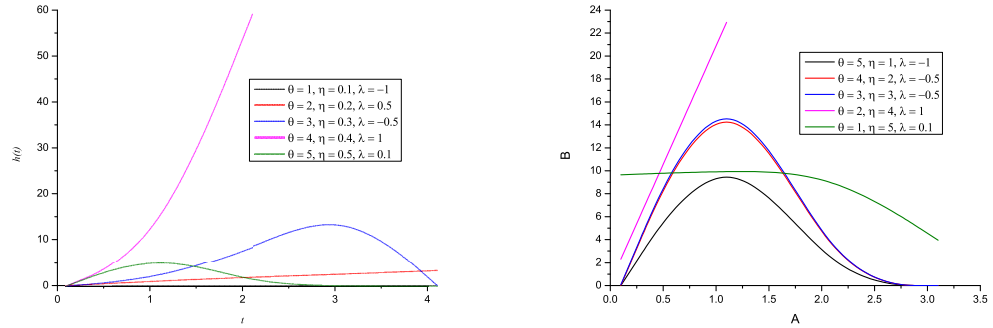


Figure 4: Plots for hazard function of TS Weibull distribution for different values of the parameters.

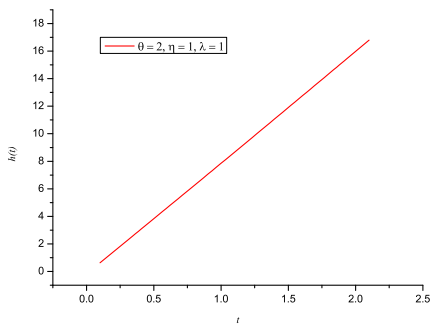


Figure 5: (a)

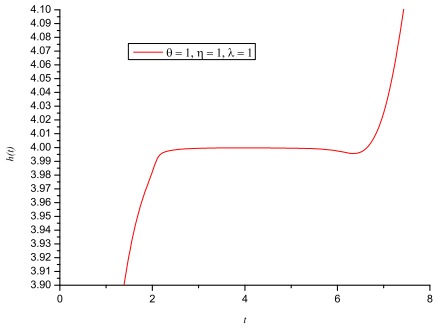


Figure 6: (b)

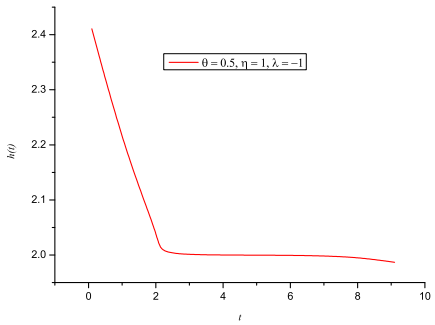


Figure 7: (c)

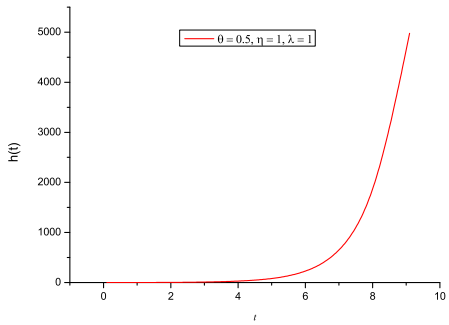


Figure 8: (d)

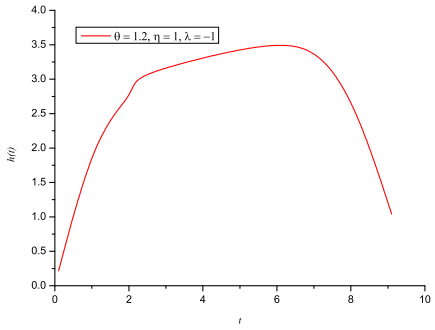


Figure 9: (e)

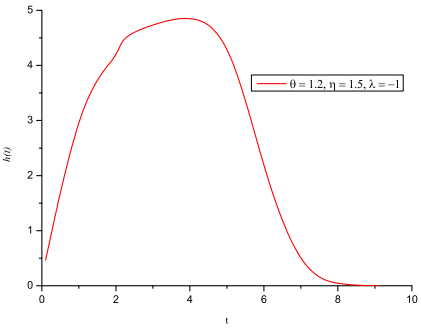


Figure 10: (f)

The above figures 5 - 10 of hazard function of TS Weibull distribution for different values of the parameters takes shapes as (a) linear, (b) increasing decreasing increasing (IDI), (c) decreasing, (d)increasing, (e) inverse bathtub and (f) unimodal shapes respectively.

7. Maximum Likelihood Estimation

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from transmuted sine Weibull distribution. The likelihood function is given by

$$\begin{aligned} L(x:\eta,\theta,\lambda) &= \prod_{i=1}^{n} \frac{\pi}{2}\eta\theta x_i^{\theta-1} e^{-\eta x_i^{\theta}} \cos\left(\frac{\pi}{2}\left(1-e^{-\eta x_i^{\theta}}\right)\right) \\ &\quad \left[(1+\lambda) - 2\lambda \sin\left(\frac{\pi}{2}\left(1-e^{-\eta x_i^{\theta}}\right)\right)\right] \end{aligned} \tag{48}$$

$$\begin{aligned} l(x:\eta,\theta,\lambda) &= \log\left(\frac{\pi}{2}\right) + \log(\eta) + \log(\theta) + (\theta-1)\sum_{i=1}^{n} x_i - \eta\sum_{i=1}^{n} x_i^{\theta} \\ &\quad + \sum_{i=1}^{n} \log\left[\cos\left(\frac{\pi}{2}\left(1-e^{-\eta x_i^{\theta}}\right)\right)\right] \\ &\quad + \sum_{i=1}^{n} \log\left[(1+\lambda) - 2\lambda \sin\left(\frac{\pi}{2}\left(1-e^{-\eta x_i^{\theta}}\right)\right)\right] \end{aligned} \tag{49}$$

$$\begin{aligned} \frac{\partial l(x:\eta,\theta,\lambda)}{\partial \eta} =& \frac{1}{\eta} - \sum_{i=1}^{n} x_i^{\theta} - \sum_{i=1}^{n} \frac{\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x_i^{\theta}}\right)\left(\frac{\pi}{2}\right)\right) x_i^{\theta} e^{-\eta x_i^{\theta}}}{cos\left(\frac{\pi}{2}(1-e^{-\eta x_i^{\theta}})\right)} \\ & - \sum_{i=1}^{n} \frac{2\lambda \cos\left(\frac{\pi}{2}\left(1-e^{-\eta x_i^{\theta}}\right)\right)\left(\frac{\pi}{2}\right) x_i^{\theta} e^{-\eta x_i^{\theta}}}{1+\lambda-2\lambda\ sin\left(\frac{\pi}{2}(1-e^{-\eta x_i^{\theta}})\right)} \end{aligned} \tag{50}$$

$$\begin{aligned} \frac{\partial l(x:\eta,\theta,\lambda)}{\partial \theta} =& \frac{1}{\theta} + \sum_{i=1}^{n} x_i - \eta\sum_{i=1}^{n} x_i^{\theta} logx_i - \sum_{i=1}^{n} \frac{\frac{\pi}{2}\eta x_i^{\theta} logx_i\ sin\left(\frac{\pi}{2}(1-e^{-\eta x_i^{\theta}})\right)}{cos\left(\frac{\pi}{2}(1-e^{-\eta x_i^{\theta}})\right)} \\ & \times - \sum_{i=1}^{n} \frac{2\lambda(\frac{\pi}{2}) e^{-\eta x_i^{\theta}} \eta x_i^{\theta} logx_i\ cos\left(\frac{\pi}{2}(1-e^{-\eta x_i^{\theta}})\right)}{1+\lambda-2\lambda\ sin\left(\frac{\pi}{2}(1-e^{-\eta x_i^{\theta}})\right)} \end{aligned} \tag{51}$$

$$\frac{\partial l(x:\eta,\theta,\lambda)}{\partial \lambda} = \sum_{i=1}^{n} \frac{1-2\ sin\left(\frac{\pi}{2}(1-e^{-\eta x_i^{\theta}})\right)}{1+\lambda-2\lambda\ sin\left(\frac{\pi}{2}(1-e^{-\eta x_i^{\theta}})\right)} \tag{52}$$

The maximum likelihood estimate $\hat{\Theta} = (\hat{\eta}, \hat{\theta}, \hat{\lambda})$ of $\Theta = (\eta, \theta, \lambda)$ Also as $n \to \infty$ the asymptotic distribution of the MLEs (η, θ, λ) are given by, see Rahman *et al.* (2018) and Zaindin *et al.* (2009).

$$\begin{pmatrix} \hat{\eta} \\ \hat{\theta} \\ \hat{\lambda} \end{pmatrix} \sim N\left[\begin{pmatrix} \eta \\ \theta \\ \lambda \end{pmatrix}, \begin{pmatrix} V_{11} & V_{12} & V_{13} \\ V_{21} & V_{22} & V_{23} \\ V_{31} & V_{32} & V_{33} \end{pmatrix}\right]$$

Where $\hat{V_{ij}} = V_{ij}$ the asymptotic variance-covariance matrix V of the estimates is obtained by inverting Hessian matrix. See Appendix. An approximate $100\,(1-\alpha)\,\%$ two sided confidence intervals for η, θ and λ are respectively is given by

$$\eta \in \left[\hat{\eta} - Z_{\frac{\alpha}{2}}\sqrt{V_{11}^{-1}}, \hat{\eta} + Z_{\frac{\alpha}{2}}\sqrt{V_{11}^{-1}}\right] \tag{53}$$

$$\theta \in \left[\hat{\theta} - Z_{\frac{\alpha}{2}}\sqrt{V_{22}^{-1}}, \hat{\theta} + Z_{\frac{\alpha}{2}}\sqrt{V_{22}^{-1}}\right] \tag{54}$$

$$\lambda \in \left[\hat{\lambda} - Z_{\frac{\alpha}{2}}\sqrt{V_{33}^{-1}}, \hat{\lambda} + Z_{\frac{\alpha}{2}}\sqrt{V_{33}^{-1}}\right] \tag{55}$$

Where Z_α is the α^{th} percentile of the standard normal distribution.

7.1. Simulation study

In this section, a simulation study is performed to test the performance of MLEs. We generate a random sample of size $n = 50, 100, 200$ and 500 from TS Weibull distribution for the values of $\theta = 2, \eta = 4$ and $\lambda = 0.5$. With replicated 1000 times. It is observed that the mean squared error (MSE) and average bias decreases when the sample size increases. Therefore, the maximum likelihood estimate converges to true value of the parameters of TS Weibull distribution.

Table 1: The MSE and average bias for the above given values of parameters

n	θ	η	λ
50	0.0138	0.0805	0.0223
	0.0185	0.0364	0.0194
100	0.0064	0.0390	0.0111
	0.0092	0.0193	0.0103
200	0.0032	0.0194	0.0055
	0.0054	0.0083	0.0044
500	0.0016	0.0097	0.0028
	0.0023	0.0047	0.0025

8. Application

The following data represents lifetimes of Kevlar 49/epoxy strands subjected to constant sustained pressure at 90 percent stress level until the strand failure studied by Barlow *et al.* (1984) and Pobocikova *et al.* (2018).
0.0251, 0.0886, 0.0891, 0.2501, 0.3113, 0.3451, 0.4763, 0.5650, 0.5671, 0.6566, 0.6748, 0.6751, 0.6753, 0.7696, 0.8375, 0.8391, 0.8425, 0.8645, 0.8851, 0.9113, 0.9120, 0.9836, 1.0483, 1.0596, 1.0773, 1.1733, 1.2570, 1.2766, 1.2985, 1.3211, 1.3503, 1.3551, 1.4595, 1.4880, 1.5728, 1.5733, 1.7083, 1.7263, 1.7460, 1.7630, 1.7746, 1.8275, 1.8375, 1.8503, 1.8808, 1.8878, 1.8881, 1.9316, 1.9558, 2.0048, 2.0408, 2.0903, 2.1093, 2.1330, 2.2100, 2.2460, 2.2878, 2.3203, 2.3470, 2.3513, 2.4951, 2.5260, 2.9911, 3.0256, 3.2678, 3.4045, 3.4846, 3.7433, 3.7455, 3.9143, 4.8073, 5.4005, 5.4435, 5.5295, 6.5541, 9.0960.

Table 2: Descriptive statistics

Min	Max	Mean	Vari	Lower Quantile	Median	Upper Quantile	Skewness	Kurtosis
0.0251	**9.0960**	**1.9592**	**2.4774**	**0.8982**	**1.7362**	**2.3041**	**2.0196**	**5.6004**

Table 3: ML estimates of the parameters with measures for model selection

Distribution	Parameter estimates	Log-lik	AIC	AICC	BIC
W(a,b)	$\hat{a}$=1.3256, $\hat{b}$=2.1328	-122.5247	249.0494	249.2094	253.7108
W(a,b,c)	$\hat{a}$=1.3169, $\hat{b}$=2.1228, $\hat{c}$=0.0058	-122.5141	251.0282	251.3525	258.0204
W(a,b,λ)	$\hat{a}$=1.0509, $\hat{b}$=1.4419, $\hat{\lambda}$=-0.7955	-121.4300	248.8600	249.1843	255.8522
SW(η,λ)	$\hat{\eta}$=1.2564257, $\hat{\lambda}$=0.2210405	-122.4717	248.9434	249.1078	253.6049
TSW(η,θ,λ)	$\hat{\eta}$=0.40025, $\hat{\theta}$=0.98819, $\hat{\lambda}$=-0.80737	-121.2775	242.5552	248.5552	255.5472

In Table 2, we observed that the empirical distribution is right skewed. Table 3 presents the ML estimates of the parameters along with the log-likelihood, AIC, AICC and BIC values. Further, we observed that the TS Weibull distribution provides better fits compared to 2-parameter Weibull, 3-parameter Weibull, transmuted Weibull distributions and sine-Weibull distribution. Hence, we can conclude that the TS Weibull distribution provides a better fit to the data than the other four suitable probability distributions.

9. Conclusion

In the present study, we have introduced transmuted sine Weibull distribution and we have derived some statistical properties such as moments, mean and variance for the proposed distribution. The behaviour of the pdf, cdf, reliability function and hazard function are explained through the graphical methods. We have discussed certain statistical properties like generalized entropy, asymptotic behaviour, order statistics and stochastic ordering for TS Weibull distribution. The parameter estimation is performed using the maximum likelihood method for the proposed distribution. Finally, we have shown TS Weibull distribution provides a better fit compared to 2-parameter Weibull, 3-parameter Weibull, transmuted Weibull and sine - Weibull distribution for real time data set.

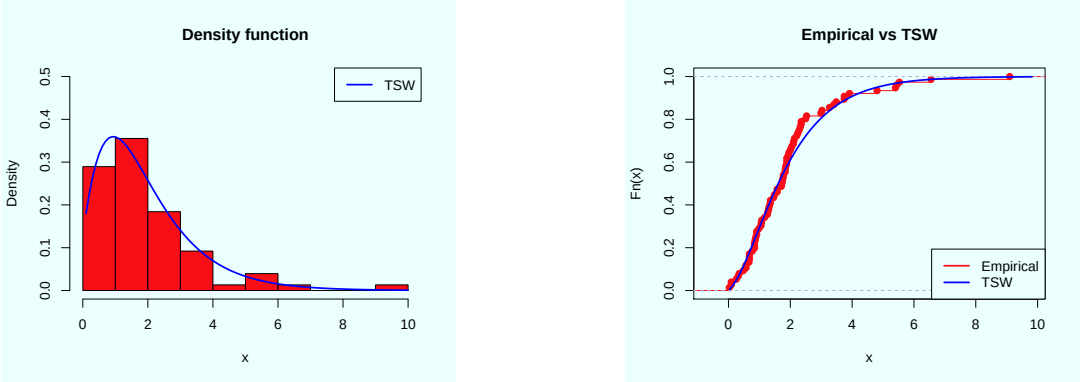


Figure 11: Plot of the empirical pdf and cdf of TS Weibull distribution.

References

Ahmad, Z., Elgarhy, M., Hamedani, G. G. and Butt, N. S. (2020). Odd Generalized NH Generated Family of Distributions with Application to Exponential Model. *Pakistan Journal of Statistics and Operation Research*, **16(1)**, 53-71.

Almarashi, A. M., Elgarhy, M., Jamal, F. and Chesneau, C. (2020). The exponentiated truncated inverse Weibull-generated family of distributions with applications. *Symmetry*, **12(4)**, 650.

Al-Shomrani, A., Arif, O., Shawky, A., Hanif, S. and Shahbaz, M. Q. (2016). Topp − Leone family of distributions: some properties and application. *Pakistan Journal of Statistics and Operation Research*, **12(3)**, 443-451.

Bantan, R. A., Jamal, F., Chesneau, C. and Elgarhy, M. (2019). Truncated inverted Kumaraswamy generated family of distributions with applications. *Entropy*, **21(11)**, 1089.

Bantan, R. A., Jamal, F., Chesneau, C. and Elgarhy, M. (2020). Type II power Topp-Leone generated family of distributions with statistical inference and applications. *Symmetry*, **12(1)**, 75.

Barlow, R. E., Toland, R. H. and Freeman, T. (1984). A Bayesian analysis of stress-rupture life of kevlar 49/epoxy spherical pressure vessels. In Proc. Conference on Applications of Statistics, Marcel Dekker, New York.

Bourguignon, M., Silva, R. B. and Cordeiro, G. M. (2014). The Weibull-G family of probability distributions. *Journal of Data Science*, **12(1)**, 53-68.

Elbatal, I., Ahmad, Z., Elgarhy, B. M. and Almarashi, A. M. (2019). A new alpha power transformed family of distributions: properties and applications to the Weibull model. *Journal of Nonlinear Sciences and Applications*, **12(1)**, 1-20.

Gupta, R. D. and Kundu, D. (1999). Theory and methods: Generalized exponential distributions. *Australian and New Zealand Journal of Statistics*, **41(2)**, 173-188.

Gupta, R. D. and Kundu, D. (2001). Exponentiated exponential family: an alternative to gamma and Weibull distributions. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*,**43(1)**, 117-130.

Jamal, F., Chesneau, C. and Elgarhy, M. (2020). Type II general inverse exponential family of distributions. *Journal of Statistics and Management Systems*, **23(3)**, 617-641.

Kharazmi, O. and Saadatinik, A. (2016). Hyperbolic cosine-F families of distributions with an application to exponential distribution. *Gazi University Journal of Science*, **29(4)**, 811-829.

Kumar, D., Singh, U. and Singh, S. K. (2015). A new distribution using sine function-its application to bladder cancer patients data. *Journal of Statistics Applications and Probability*, **4(3)**, 417-427.

Pal, M., Ali, M. M. and Woo, J. (2006). Exponentiated weibull distribution. *Statistica*, **66(2)**, 139-147.

Pobocikova, I., Sedliackova, Z. and Michalkova, M. (2018). Transmuted Weibull distribution and its applications. In MATEC Web of Conferences (Vol. 157, p. 08007). EDP Sciences.

Rahman, M. M., Al-Zahrani, B. and Shahbaz, M. Q. (2018). A general transmuted family of distributions. *Pakistan Journal of Statistics and Operation Research*, **14(2)**, 451-469.

Sakthivel, K. M. and Rajkumar, J. (2020). Hyperbolic Cosine Rayleigh Distribution and Its Application to Breaking Stress of Carbon Fibers. *Journal of Indian Society and Probability Statistics*, **21(2)**, 471-485.

Sarhan, A. M. and Zaindin, M. (2009). Modified Weibull distribution. *APPS. Applied Sciences*, **11**, 123-136.

Shaked, M., Shanthikumar, J. G. and Tong, Y. L. (1995). Stochastic orders and their applications. *SIAM Review*, **37(3)**, 477-478.

Shaw, W. T. and Buckley, I. R. (2007). The alchemy of probability distributions: Beyond gram-charlier and cornish-fisher expansions, and skew-normal or kurtotic-normal distributions. arXiv:0901.0434.

Souza, L. (2015). New trigonometric classes of probabilistic distributions (Doctoral dissertation, Thesis, Universidade Federal Rural de Pernambuco).

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, **18(3)**, 293-297.

Yadav, A. S., Maiti, S.S. and Saha, M. (2019). The inverse Xgamma distribution: statistical properties and different methods of estimation. *Annals of Data Science*, **8(2)**, 275-293.

Zaindin, M. and Sarhan, A. M. (2009). Parameters estimation of the modified Weibull distribution. *Applied Mathematical Sciences*, **3(11)**, 541-550.

APPENDIX

The Hessian matrix is given as

$$H = \begin{pmatrix} H_{11} & H_{12} & H_{13} \\ H_{21} & H_{22} & H_{23} \\ H_{31} & H_{32} & H_{33} \end{pmatrix}$$

Where the variance - covariance matrix V is obtained by

$$V = \begin{pmatrix} V_{11} & V_{12} & V_{13} \\ V_{21} & V_{22} & V_{23} \\ V_{31} & V_{32} & V_{33} \end{pmatrix} = \begin{pmatrix} H_{11} & H_{12} & H_{13} \\ H_{21} & H_{22} & H_{23} \\ H_{31} & H_{32} & H_{33} \end{pmatrix}^{-1}$$

$$H_{11} = E\left[-\frac{\partial logL}{\partial \eta^2}\right], H_{22} = E\left[-\frac{\partial logL}{\partial \theta^2}\right], H_{33} = E\left[-\frac{\partial logL}{\partial \lambda^2}\right],$$

$$H_{12} = H_{21} = E\left[-\frac{\partial^2 logL}{\partial \eta \partial \theta}\right], H_{13} = H_{31} = E\left[-\frac{\partial^2 logL}{\partial \eta \partial \lambda}\right], H_{23} = H_{32} = E\left[-\frac{\partial^2 logL}{\partial \theta \partial \lambda}\right].$$

$$H_{11} = \frac{1}{\eta^2} + \sum_{i=1}^{n} x_i^2 \left(\frac{\pi}{2}\right) e^{\eta x_i^\theta} \left(x_i^\theta\right) \left[\frac{A+B}{\left(e^{-\eta x_i^\theta}\right)}\right]$$

$$\begin{aligned} A &= \cos\left(\frac{\pi}{2}\left(1-e^{-\eta x_i^\theta}\right)\right) \\ &\quad \left(\cos\left(\frac{\pi}{2}\left(1-e^{-\eta x_i^\theta}\right)\right)\left(\frac{\pi}{2}\left(1-e^{-\eta x_i^\theta}\right)\right) + \sin\left(\frac{\pi}{2}\left(1-e^{-\eta x_i^\theta}\right)\right)\right) \end{aligned}$$

$$B = \left(\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x_i^\theta}\right)\right)\right)^2 \left(\frac{\pi}{2}\right)\left(1-e^{-\eta x_i^\theta}\right)$$

$$\begin{aligned} H_{22} = = \frac{1}{\theta^2}\eta \sum_{i=1}^{n} x_i^\theta (\log x_i)^2 + \sum_{i=1}^{n}\left(\frac{\pi}{2}\right)\eta (\log x_i)^2 x_i^\theta \left[\frac{C}{\left(\cos\left(\frac{\pi}{2}\left(1-e^{-nx_i^\theta}\right)\right)\right)^2}\right] \\ - \sum_{i=1}^{n} 2\lambda \log x_i e^{-\eta x_i^\theta} x_i^\theta \log x_i \left[\frac{D}{\left(1+\lambda-2\lambda\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x_i^\theta}\right)\right)\right)^2}\right] \end{aligned}$$

$$\begin{aligned} C = \cos\left(\frac{\pi}{2}\left(1-e^{-nx_i^\theta}\right)\right)\left[\eta x_i^\theta e^{-\eta x_i^\theta}\cos\left(\frac{\pi}{2}\left(1-e^{-nx_i^\theta}\right)\right) + \sin\left(\frac{\pi}{2}\left(1-e^{-nx_i^\theta}\right)\right)\right] \\ + x_i^\theta \sin\left(\frac{\pi}{2}\left(1-e^{-nx_i^\theta}\right)\right)^2 \frac{\pi}{2} e^{\eta x_i^\theta} \end{aligned}$$

$$D = \left(1+\lambda-2\lambda\sin\left(\frac{\pi}{2}\left(1-e^{-nx_i^\theta}\right)\right)\right)$$
$$\left[\left(\cos\left(\frac{\pi}{2}\left(1-e^{-nx_i^\theta}\right)\right)-x_i\eta\right)-\eta x_i^\theta\frac{\pi}{2}e^{-\eta x_i^\theta}\sin\left(\frac{\pi}{2}\left(1-e^{-nx_i^\theta}\right)\right)\right]$$
$$+2\lambda\left(\eta x_i^\theta\frac{\pi}{2}e^{-\eta x_i^\theta}\sin\left(\frac{\pi}{2}\left(1-e^{-nx_i^\theta}\right)\right)\right)^2 e^{\eta x_i^\theta}x_i^\theta$$

$$H_{33} = \sum_{i=1}^{n}\frac{\left(1-2sin\left(\frac{\pi}{2}(1-e^{-\eta x_i^\theta})\right)\right)^2}{\left(1+\lambda-2\lambda sin\left(\frac{\pi}{2}(1-e^{-\eta x_i^\theta})\right)\right)^2}$$

$$H_{21} = H_{12} = \sum_{i=1}^{n}x_i^\theta\log x_i + \sum_{i=1}^{n}\left[\frac{E}{\left(\cos\left(\frac{\pi}{2}\left(1-e^{-nx_i^\theta}\right)\right)\right)^2}\right]$$
$$-\sum_{i=1}^{n}\left[\frac{F}{\left(1+\lambda-2\lambda\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x_i^\theta}\right)\right)\right)^2}\right]$$

$$E = \left(\frac{\pi}{2}x_i^\theta\log x_i\right)\left[\left(\cos\left(\frac{\pi}{2}\left(1-e^{-\eta x_i^\theta}\right)\right)\right)^2\right]\left[\frac{\pi}{2}\eta e^{\eta x_i^\theta}+1\right]$$
$$+n\left(\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x_i^\theta}\right)\right)\right)^2 e^{-\eta x_i^\theta x_i^\theta}$$

$$F = \left(\left(\frac{\pi}{2}e^{-\eta x_i^\theta\log x_i}\right)\right)\left[-\left(1+\lambda-2\lambda\sin\left(\frac{\pi}{2}\left(1-e^{-\eta x_i^\theta}\right)\right)\right)\right]$$
$$\left[2\lambda\eta\left(\frac{\pi}{2}\right)\sin\left(\frac{\pi}{2}\left(1-e^{-nx_i^\theta}\right)\right)\left(\frac{\pi}{2}\right)e^{-\eta x_i^\theta}\right]$$
$$+\left[2\lambda\cos\left(\frac{\pi}{2}\left(1-e^{-nx_i^\theta}\right)\right)\left(1-\eta x_i^\theta\right)\right]$$
$$\left[2\lambda\cos\left(\frac{\pi}{2}\left(1-e^{-nx_i^\theta}\right)\right)^2\eta\left(\frac{\pi}{2}e^{-\eta x_i^\theta}x_i^\theta\right)\right]$$

$$H_{13} = \sum_{i=1}^{n}\left[\frac{G-H}{\left(\left(1+\lambda-2\lambda sin\left(\frac{\pi}{2}(1-e^{-\eta x_i^\theta})\right)\right)\right)^2}\right]$$

$$G = \left(1+\lambda-2\lambda sin\left(\frac{\pi}{2}(1-e^{-\eta x_i^\theta})\right)\right)\left(2x_i^\theta(\frac{\pi}{2}e^{-\eta x_i^\theta}cos\left(\frac{\pi}{2}(1-e^{-\eta x_i^\theta})\right)\right)$$

$$H = 2\lambda x_i^\theta(\frac{\pi}{2})e^{-\eta x_i^\theta}cos\left(\frac{\pi}{2}(1-e^{-\eta x_i^\theta})\right)\left(1-2sin\left(\frac{\pi}{2}(1-e^{-\eta x_i^\theta})\right)\right)$$

$$H_{23} = H_{32} = \left[\sum_{i=1}^{n}\frac{2cos\left(\frac{\pi}{2}(1-e^{-\eta x_i^\theta})\right)(\frac{\pi}{2})e^{-\eta x_i^\theta logx_i}}{\left(1+\lambda-2\lambda sin\left(\frac{\pi}{2}(1-e^{-\eta x_i^\theta})\right)\right)^2}\right]$$

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 93–101

Constant Block-Sum Designs Through Confounded Factorials

Sudhir Gupta
Department of Statistics
Northern Illinois University, DeKalb, Illinois, USA

Received: 19 April 2021; Revised: 21 August 2021; Accepted: 23 August 2021

Abstract

Confounded factorial designs are shown to provide a rich class of constant block-sum designs. The approach also provides a direct and straightforward proof of the necessary condition for existence of constant block-sum designs given recently by Khattree (2022).

Key words: Balanced incomplete block design; Group divisible design; Treatment contrast.

1. Introduction

Constant block-sum designs for quantitative treatment levels have been recently introduced by Khattree (2019a,b). In these designs, the sum of the treatment levels in each block is constant. Several methods of their construction have been presented by Khattree (2020). A general approach to determine whether or not a given design can be transformed into a constant block-sum design and its construction if it exists has been developed in Khattree (2022). He also discussed several individual examples, including two-associate class group divisible (GD) designs. Non-existence of constant block-sum balanced incomplete designs was established by Khattree (2019a, 2022). Bansal and Garg (2022) and Khattree (2022) derived some conditions for existence of partially balanced constant block-sum designs and gave further combinatorial methods of their construction. Gupta (2021) gave general results for GD designs with respect to the property of constant block-sum. He established non-existence of semi-regular and regular GD constant block-sum designs. He also discussed construction of singular GD constant block-sum designs and gave several illustrative examples.

Motivated by the results presented by Khattree (2022), the purpose of this paper is to study construction of constant block-sum designs using factorial designs. It is shown that the method of confounding provides a rich class of constant block-sum designs. The approach also provides a direct and straightforward proof of the necessary condition for existence of constant block-sum designs given by Khattree (2022).

2. Method of Construction

Consider an equireplicate confounded block design with parameters v, b, r, k, and let $\boldsymbol{\tau} = (\tau_1, \tau_2, \cdots, \tau_v)'$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_b)'$ respectively denote the $v \times 1$ and $b \times 1$ vectors

Corresponding Author: Sudhir Gupta
Email: sudhirgupta.stat@gmail.com

of treatment and block parameters. Let $\boldsymbol{h}'\boldsymbol{\tau}$ denote a treatment contrast that is partially or completely confounded in the design, $\boldsymbol{h}'\mathbf{1}_v = 0$, where $\mathbf{1}_a$ denotes a $a \times 1$ vector of 1's. Further, $\boldsymbol{s}'\boldsymbol{\tau}$ denotes a treatment contrast that is estimated with full efficiency in the design, i.e. it is not confounded in any of the replications of the design, $\boldsymbol{s}'\mathbf{1}_v = 0$. We will refer to factorial effects that are estimated with full efficiency as completely unconfounded effects.

To motivate the method of construction, we replace the ith treatment in the confounded design by the ith element of $\boldsymbol{h}$ and $\boldsymbol{s}$. In other words, the treatments in the design are replaced by the corresponding coefficients of the confounded and unconfounded contrasts. This is illustrated with the help of the following example.

Example 1: Consider the 2^3 partially confounded design of Table 1 having parameters $v = 8$, $b = 4,$, $r = 2$, $k = 4$. The designed is obtained by confounding the three-factor interaction $F_1F_2F_3$ in one replication and the two-factor interaction F_2F_3 in the other replication.

Table 1

$F_1F_2F_3$ confounded		F_2F_3 confounded	
Block 1	Block 2	Block 3	Block 4
000	001	000	001
101	010	011	010
110	100	100	101
011	111	111	110

Let $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, $\boldsymbol{u}_3$, $\boldsymbol{u}_{12}$, $\boldsymbol{u}_{13}$, $\boldsymbol{u}_{23}$, and $\boldsymbol{u}_{123}$ be the contrast coefficient vectors for the F_1, F_2, F_3 main effects and F_1F_2, F_1F_3, F_2F_3, $F_1F_2F_3$ interactions respectively,

$$\begin{bmatrix} \boldsymbol{u}_1' \\ \boldsymbol{u}_2' \\ \boldsymbol{u}_3' \\ \boldsymbol{u}_{12}' \\ \boldsymbol{u}_{13}' \\ \boldsymbol{u}_{23}' \\ \boldsymbol{u}_{123}' \end{bmatrix} = \begin{bmatrix} -1 & -1 & -1 & -1 & +1 & +1 & +1 & +1 \\ -1 & -1 & +1 & +1 & -1 & -1 & +1 & +1 \\ -1 & +1 & -1 & +1 & -1 & +1 & -1 & +1 \\ +1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 \\ +1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 \\ +1 & -1 & -1 & +1 & +1 & -1 & -1 & +1 \\ -1 & +1 & +1 & -1 & +1 & -1 & -1 & +1 \end{bmatrix} .$$

Also, the vector of treatment parameters can be written as,

$$\boldsymbol{\tau}' = (\tau_{000} \ \tau_{001} \ \tau_{010} \ \tau_{011} \ \tau_{100} \ \tau_{101} \ \tau_{110} \ \tau_{111}) ,$$

with τ_x denoting the effect of the treatment combination x. Now we replace the treatment combinations in each block by the corresponding $F_1F_2F_3$ contrast coefficients and obtain the design displayed in Table 2. The block sums are given in the last row of the table.

Table 2

	Replace treatment combinations by the corresponding $F_1F_2F_3$ contrast coefficients			
	Block 1	Block 2	Block 3	Block 4
	-1	$+1$	-1	$+1$
	-1	$+1$	-1	$+1$
	-1	$+1$	$+1$	-1
	-1	$+1$	$+1$	-1
Block sums	-4	$+4$	0	0

Similarly, Tables 3 and 4 give the designs obtained by replacing the treatment combinations in each block of the design by respectively the F_2F_3 and F_1F_2 contrast coefficients. Note that $F_1F_2F_3$ and F_2F_3 are partially confounded whereas F_1F_2 is not confounded and it is estimated without any loss of information.

Table 3

	Replace treatment combinations by the corresponding F_2F_3 contrast coefficients			
	Block 1	Block 2	Block 3	Block 4
	$+1$	-1	$+1$	-1
	-1	-1	$+1$	-1
	-1	$+1$	$+1$	-1
	$+1$	$+1$	$+1$	-1
Block sums	0	0	$+4$	-4

Table 4

	Replace treatment combinations by the corresponding F_1F_2 contrast coefficients			
	Block 1	Block 2	Block 3	Block 4
	$+1$	$+1$	$+1$	$+1$
	-1	-1	-1	-1
	$+1$	-1	-1	-1
	-1	$+1$	$+1$	$+1$
Block sums	0	0	0	0

Block sums are constant, being equal to zero, for the design of Table 4 corresponding to the F_1F_2 interaction estimated with full efficiency in the design. It can be verified that the block sums are also constant, being equal to zero, for the designs constructed similarly corresponding to the other four unconfounded effects F_1, F_2, F_3, and F_1F_3 respectively. However, block sums are not constant for the designs of Tables 2 and 3 corresponding to the partially confounded interactions $F_1F_2F_3$ and F_2F_3 respectively.

The pattern in block sums with respect to confounded and unconfounded contrasts observed in the above example holds true in general. A completely unconfounded contrast $\boldsymbol{s}'\boldsymbol{\tau}$ is estimated from within block comparisons, *i.e.* it is estimated orthogonal to blocks. Clearly, its contrast coefficients falling in any block must sum to zero in order for the corresponding block effect to be canceled out from within block comparisons. Thus, as observed in the above example, block sum for a completely unconfounded contrast must be zero for each and every block. Conversely, a partially or completely confounded contrast $\boldsymbol{h}'\boldsymbol{\tau}$ is mixed up with some block contrast implying non-constancy of block sums.

Lemma: Let block contents of a partially confounded design be replaced by corresponding coefficients of a treatment contrast. Then the property of constant block sum being equal to zero holds for all contrasts that are estimated with full efficiency. Furthermore, this property does not hold for the treatment contrasts that are partially or completely confounded in the design.

Although, neither the block contents of $+1$ and -1 nor the block sum of zero are helpful from a practical point of view, as will be seen later, useful constant block sum designs can be easily derived through this approach.

The above lemma is closely related to the main result of Khattree (2022). He proved that a necessary condition for existence of a constant block-sum design is that $\boldsymbol{w} \neq \mathbf{1}_v$ is an eigenvector of $\boldsymbol{A}$ corresponding to a zero eigenvalue, where

$$\boldsymbol{A} \;=\; \boldsymbol{N}\boldsymbol{N}' \;-\; \frac{rk}{v}\mathbf{1}_v\mathbf{1}_v' \,,$$

and $\boldsymbol{N}$ is the incidence matrix of the design. Gupta (2021) showed that the term $(rk/v)\mathbf{1}_v\mathbf{1}_v'$ in the expression of $\boldsymbol{A}$ is in fact redundant. Thus equivalently, a necessary condition for existence of constant block-sum design is that $\boldsymbol{w} \neq \mathbf{1}_v$ is an eigenvector of $\boldsymbol{N}\boldsymbol{N}'$ corresponding to a zero eigenvalue. Note that a treatment contrast is estimated with full efficiency if and only if its contrast coefficient vector is an eigenvector of $\boldsymbol{N}\boldsymbol{N}'$ with zero eigenvalue. Thus, estimation of a treatment contrast orthogonal to blocks provides a direct and straightforward proof of the necessary condition for existence of a constant block-sum design.

We now discuss constructions of constant block-sum designs. Let q denote the number of treatment contrasts that are estimated with full efficiency in a factorial design, and let these contrasts be denoted by

$$\boldsymbol{U}'\boldsymbol{\tau} \;=\; \begin{bmatrix} \boldsymbol{u}_1' \\ \boldsymbol{u}_2' \\ \vdots \\ \boldsymbol{u}_q' \end{bmatrix} \boldsymbol{\tau} \;\;,$$

where $\boldsymbol{u}_i' = (u_{i1} \;\; u_{i2} \;\; \cdots \, u_{iv})$, with $\boldsymbol{u}_i'\mathbf{1}_v = 0$, $i = 1, 2, \cdots, \; q$. Consider θ_u, a linear function of the q contrasts given by

$$\theta_u \;=\; \boldsymbol{C}'\boldsymbol{U}'\boldsymbol{\tau} \;=\; \sum_{i=1}^{q} (c_i\boldsymbol{u}_i')\,\boldsymbol{\tau} \;=\; \boldsymbol{t}_u'\boldsymbol{\tau} \;\;,$$

where

$$\begin{aligned}\boldsymbol{C}' &= (\, c_1 \quad c_2 \quad \cdots \quad c_q \,)\ , \\ \boldsymbol{t}'_u &= \left(\sum_{i=1}^{q} c_i u_{i1} \quad \sum_{i=1}^{q} c_i u_{i2} \quad \cdots \quad \sum_{i=1}^{q} c_i u_{iv} \right) \\ &= (t_{u1} \quad t_{u2} \quad \cdots \quad t_{uv}) \ ,\end{aligned}$$

and c_i's are some constants chosen such that all the elements of $\boldsymbol{t}_u$ are different from each other. Being a linear function of treatment contrasts that are estimated with full efficiency, the treatment contrast θ_u is also estimated with full efficiency in the design. Thus, using the Lemma, the property of constant block-sum holds when block contents of the design are replaced by corresponding coefficients of the treatment effects in the linear function θ_u, i.e. by the corresponding elements of $\boldsymbol{t}_u$. The $\boldsymbol{t}_u$ being a contrast coefficient vector, $\sum_{i=1}^{v} t_{ui} = 0$, which means that not all the t_{ui}'s are greater than zero. However, it is easily seen, *cf.* Khattree (2022), that the property of constant block-sum still holds if we add a constant value, say c_0, to all the elements of $\boldsymbol{t}_u$. Let $\boldsymbol{t}^*_u = (t_{u1} + c_0 \ \ t_{u2} + c_0 \ \ \cdots \ \ t_{uv} + c_0)$, where c_0 is chosen such that all the elements of $\boldsymbol{t}^*_u$ are greater than zero. Finally, the treatment combinations in the design are then replaced by the corresponding elements of $\boldsymbol{t}^*_u$ to arrive at a constant block-sum design. For illustration, we again consider the 2^3 partially confounded design of Example 1.

Example 1 contd.: Here we have five completely unconfouned contrasts, *i.e.* $q = 5$, given by

$$\boldsymbol{U} = (\, \boldsymbol{u}_1 \ \ \boldsymbol{u}_2 \ \ \boldsymbol{u}_3 \ \ \boldsymbol{u}_{12} \ \ \boldsymbol{u}_{13} \,) \ ,$$

and let $\boldsymbol{T}$ denote the vector of treatment combinations arranged in the lexicographic order, i.e. in increasing numerical order,

$$\boldsymbol{T}' = (000 \ \ 001 \ \ 010 \ \ 011 \ \ 100 \ \ 101 \ \ 110 \ \ 111).$$

Taking $\boldsymbol{C}' = (\, 0.44 \quad -0.10 \quad -0.08 \quad 0.18 \quad -0.20 \,)$ and $c_0 = 1.2$ gives,

$$\boldsymbol{t}^{*\prime}_u = (\, 0.92 \ \ 1.16 \ \ 0.36 \ \ 0.60 \ \ 1.84 \ \ 1.28 \ \ 2.00 \ \ 1.44 \,) \ .$$

Replacing the ith element of $\boldsymbol{T}$ in Table 1 by the ith element of $\boldsymbol{t}^{*\prime}_u$, $i = 1, 2, \cdots, v$, yields a design with a constant block-sum of $4c_0 = 4.8$. A very large number of distinct constant block-sum designs can be constructed in this fashion by choosing different values of $\boldsymbol{C}$ and the constant c_0. Tables 5 and 6 list five more solutions for the vector of treatment levels $\boldsymbol{t}^{*\prime}_u$ obtained by trial and error. Many more solutions can be easily constructed in this way.

Table 5: Further solutions for Example 1

$\boldsymbol{t}^{*\prime}_u$ No.	$\boldsymbol{t}^{*\prime}_u$
1	0.56 1.12 0.40 0.96 1.48 1.24 2.04 1.80
2	1.09 0.89 0.99 0.79 0.55 1.07 1.85 2.37
3	0.21 0.71 1.17 1.67 0.39 2.29 0.63 2.53
4	1.07 1.57 0.83 1.33 1.13 3.03 0.17 2.07
5	0.72 1.92 0.48 1.68 1.08 2.28 0.12 1.32

Table 6: The C' and c_0 corresponding to $t_u^{*\prime}$ listed in Table 5

$t_u^{*\prime}$ No.	c_1	c_2	c_3	c_4	c_5	c_0
1	0.44	0.10	0.08	0.18	-0.20	1.2
2	0.26	0.30	0.08	0.35	0.18	1.2
3	0.26	0.30	0.60	-0.18	0.35	1.2
4	0.20	-0.30	0.60	-0.18	0.35	1.4
5	0.00	-0.30	0.60	-0.18	0.00	1.2

The next two examples further illustrate the richness of confounded factorials as constant block-sum designs.

Example 2: We now consider a 2^4 partially confounded design presented in Table 7, having parameters $v = 16$, $b = 8$, $r = 2$, $k = 4$, obtained by confounding $F_1F_2F_3$ and $F_2F_3F_4$ in one replication and $F_1F_2F_4$ and $F_1F_3F_4$ in the other replication. Note that the generalized interactions F_1F_4 and F_2F_3 are also partially confounded in the design.

Table 7

$F_1F_2F_3$, $F_2F_3F_4$, F_1F_4 confounded				$F_1F_2F_4$, $F_1F_3F_4$ F_2F_3 confounded			
Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8
0000	0001	1001	1000	0000	0010	0011	0001
0110	0111	1111	1110	0111	0101	0100	0110
1011	1010	0010	0011	1001	1011	1010	1000
1101	1100	0100	0101	1110	1100	1101	1111

As before, let $\boldsymbol{T}$ be the vector of treatment combinations arranged in the lexicographic order. Further, let

$$\boldsymbol{J}_0 = \begin{pmatrix} -1 \\ +1 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{J}_2 = \begin{pmatrix} +1 \\ +1 \end{pmatrix}.$$

The contrast coefficient vectors $\boldsymbol{u}_i$, $\boldsymbol{u}_{i_1i_2}$, $\boldsymbol{u}_{i_1i_2i_3}$, and $\boldsymbol{u}_{1234}$ for the main effects and interactions, $i,\ i_1 < i_2 < i_3 = 1, 2, 3, 4$, are given by $\boldsymbol{f}_1 \otimes \boldsymbol{f}_2 \otimes \boldsymbol{f}_3 \otimes \boldsymbol{f}_4$ as below:

$$\boldsymbol{f}_1 \otimes \boldsymbol{f}_2 \otimes \boldsymbol{f}_3 \otimes \boldsymbol{f}_4 = \left[\begin{array}{l} \boldsymbol{u}_i \\ \boldsymbol{u}_{i_1i_2} \\ \boldsymbol{u}_{i_1i_2i_3} \\ \boldsymbol{u}_{1234} \end{array} \right. \text{where } \boldsymbol{f}_j = \boldsymbol{J}_0 \left\{ \begin{array}{l} \text{for } j = i \\ \text{for } j = i_1, i_2 \\ \text{for } j = i_1, i_2, i_3 \\ \text{for } j = 1, 2, 3, 4 \end{array} \right\}, \text{ and } f_j = \begin{array}{l} \boldsymbol{J}_2 \text{ otherwise} \\ j = 1, 2, 3, 4 \end{array} \left. \right]$$

The completely unconfounded $q = 9$ contrast coefficient vectors are given by,

$$\boldsymbol{U} = (\, \boldsymbol{u}_1 \ \ \boldsymbol{u}_2 \ \ \boldsymbol{u}_3 \ \ \boldsymbol{u}_4 \ \ \boldsymbol{u}_{12} \ \ \boldsymbol{u}_{13} \ \ \boldsymbol{u}_{24} \ \ \boldsymbol{u}_{34} \ \ \boldsymbol{u}_{1234} \,).$$

For instance, taking $\boldsymbol{C}' = (-0.22,\ 0.30\ -0.25\ 0\ 0\ 0\ 0\ -0.30\ -0.25)$ and $c_0 = 1.2$ gives,

$$\boldsymbol{t}_u^{*\prime} = (0.79\ 1.95\ 1.45\ 0.29\ 1.89\ 2.05\ 1.55\ 1.39\ 0.85\ 1.01\ 0.51\ 0.35\ 0.95\ 2.11\ 1.61\ 0.45),$$

which yields a design given in Table 8 with a constant block-sum of $4c_0 = 4.8$.

Table 8

Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8
0.79	1.95	1.01	0.85	0.79	1.45	0.29	1.95
1.55	1.39	0.45	1.61	1.39	2.05	1.89	1.55
0.35	0.51	1.45	0.29	1.01	0.35	0.51	0.85
2.11	0.95	1.89	2.05	1.61	0.95	2.11	0.45

Five more solutions are given in Tables 9 and 10.

Table 9: Further solutions for Example 2

$\boldsymbol{t}_u^{*\prime}$ No.	$\boldsymbol{t}_u^{*\prime}$															
1	0.64	2.30	1.80	0.14	2.24	1.90	1.40	1.74	1.20	0.86	0.36	0.70	0.80	2.46	1.96	0.30
2	1.99	1.91	3.41	1.49	3.51	1.59	3.09	3.01	2.41	0.49	0.99	0.91	2.09	2.01	2.51	0.59
3	1.02	1.88	2.28	1.42	2.92	3.38	3.78	3.32	1.92	2.38	0.78	0.32	4.02	4.88	3.28	2.42
4	1.02	1.88	2.28	1.42	0.92	1.38	1.78	1.32	1.92	2.38	0.78	0.32	2.02	2.88	1.28	0.42
5	1.27	0.19	0.43	1.03	0.47	2.07	1.83	0.71	1.17	2.29	2.53	0.93	2.57	1.97	1.73	2.81

Table 10: The $\boldsymbol{C}'$ and $\boldsymbol{c}_0$ corresponding to $\boldsymbol{t}_u^{*\prime}$ listed in Table 9

$\boldsymbol{t}_u^{*\prime}$ No.	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_0
1	-0.22	0.30	-0.25	0	0	0	0	-0.33	-0.50	1.3
2	-0.50	0.30	0	-0.50	0	-0.25	0	0	-0.46	2.0
3	0	1.00	-0.30	0	0.15	-0.50	0	-0.33	-0.10	2.5
4	0	0	-0.30	0	0.15	-0.50	0	-0.33	-0.10	1.5
5	0.50	0.27	0	0	0	0	0.12	-0.13	0.55	1.5

Example 3: 3^2 partially confounded factorial design with parameters $v = 9$, $b = 6$, $r = 2$, $k = 3$. Here the two main effects F_1 and F_2 have 2 $d.f.$ each, and the two-factor interaction F_1F_2 has 4 $d.f.$ The treatment combinations vector is given by,

$$\boldsymbol{T}' = (00\ \ 01\ \ 02\ \ 10\ \ 11\ \ 12\ \ 20\ \ 21\ \ 22)\ .$$

The 4 $d.f.$ F_1F_2 interaction has two components: the 2 $d.f.$ F_1F_2 component and the 2 $d.f.$ $F_1F_2^2$ component. The design of Table 11 below is obtained by confounding the 2 $d.f.$ F_1F_2 component in one replication and the 2 $d.f.$ $F_1F_2^2$ component in the other replication.

Table 11

2 *d.f.*	F_1F_2 confounded		2 *d.f.*	$F_1F_2^2$ confounded	
Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
00	10	02	00	21	01
12	01	20	11	10	12
21	22	11	22	02	20

The four contrasts corresponding to the two main effects that are completely unconfounded in the design are given by,

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{u}'_{1\ell} \\ \boldsymbol{u}'_{1q} \\ \boldsymbol{u}'_{2\ell} \\ \boldsymbol{u}'_{2q} \end{bmatrix} = \begin{bmatrix} -1 & -1 & -1 & 0 & 0 & 0 & +1 & +1 & +1 \\ +1 & +1 & +1 & -2 & -2 & -2 & +1 & +1 & +1 \\ -1 & 0 & +1 & -1 & 0 & +1 & -1 & 0 & +1 \\ +1 & -2 & +1 & +1 & -2 & +1 & +1 & -2 & +1 \end{bmatrix},$$

where ℓ and q respectively denote the linear and quadratic components. Taking $\boldsymbol{C}' = (0.50 \;\; 0 \;\; -0.20 \;\; -0.19)$ and $c_0 = 1.6$ yields a design with constant block-sum of $3c_0 = 4.8$. The treatment levels vector $\boldsymbol{t}_u^{*\prime}$, arranged in the order of treatment combinations in T is given by,

$$\boldsymbol{t}_u^{*\prime} = (1.09 \;\; 0.90 \;\; 0.71 \;\; 2.19 \;\; 2.00 \;\; 1.81 \;\; 2.09 \;\; 1.90 \;\; 1.71) \; .$$

Five more solutions for this example are listed in Table 12.

Table 12: Further solutions for Example 3

	$t_u^{*\prime}$									c_1	c_2	c_3	c_4	c_0
1	1.09	1.50	0.71	1.59	2.00	1.21	2.09	2.50	1.71	0.50	0.00	-0.19	-0.20	1.6
2	1.12	1.62	1.52	1.30	1.80	1.70	1.48	1.98	1.88	0.18	0.00	0.20	-0.10	1.6
3	0.30	0.80	0.70	1.30	1.80	1.70	2.30	2.80	2.70	1.00	0.00	0.20	-0.10	1.6
4	0.47	2.92	0.87	0.65	3.10	1.05	0.83	3.28	1.23	0.18	0.00	0.20	-0.75	1.6
5	2.17	0.12	2.57	2.35	0.30	2.75	2.53	0.48	2.93	0.18	0.00	0.20	0.75	1.8

The constant block-sum designs of this paper are derived by searching for a treatment levels vector $\boldsymbol{t}_u^*$ through trial and error. Also, in practice treatment levels would be determined by subject matter specialists based on their study objectives. Therefore, a systmatic method of finding $\boldsymbol{t}_u^*$ with treatment levels in line with the study objectives is highly important from a practical point of view and deserves further research.

Acknowledgements

Thanks are due to a referee for helpful comments that led to a much improved version of the paper.

References

Bansal, N. and Garg, D. K. (2022). Construction and existence of constant block sum PBIB designs. *Communications in Statistics - Theory and Methods*, **51(7)**, 2231–2241.

Gupta, S. (2021). Constant block-sum group divisible designs. *Statistics and Applications*, **19**, 141-148.

Khattree, R. (2019a). A note on the nonexistence of the constant block-sum balanced incomplete block designs. *Communications in Statistics - Theory and Methods*, **48(20)**, 5165–5168.

Khattree, R. (2019b). The Parshvanath yantram yields a constant-sum partially balanced incomplete block design. *Communications in Statistics - Theory and Methods*, **49(4)**, 841–849.

Khattree, R. (2020). On construction of constant block-sum partially balanced incomplete block designs. *Communications in Statistics - Theory and Methods*, **49(11)**, 2585–2606.

Khattree, R. (2022). On construction of equireplicated constant block-sum designs. *Communications in Statistics - Theory and Methods*, **51(13)**, 4434–4450.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 103-121

Use of Change Point Analysis in Seasonal ARIMA Models for Forecasting Tourist Arrivals in Sri Lanka

B. R. P. M. Basnayake and N. V. Chandrasekara
Department of Statistics and Computer Science, Faculty of Science, University of Kelaniya Sri Lanka

Received: 31 December 2020; Revised: 15 September 2021; Accepted: 21 September 2021

Abstract

Sri Lanka is a popular place that attracts foreign travelers, and the impact of the tourism industry has a major contribution to the Sri Lankan economy. The main objective of this study is to model the behavior and forecast tourist arrivals in Sri Lanka through a time-series approach with Change Point Analysis (CPA). Autoregressive Integrated Moving Average (ARIMA) was extended to Seasonal Autoregressive Integrated Moving Average (SARIMA) with the seasonality behavior of the tourist arrivals. The better performed models were identified using the minimum Akaike Information Criterion (AIC) while performance indicators of Mean Absolute Percentage Error (MAPE) and Normalized Root Mean Squared Error (NRMSE) were applied to evaluate the actual and fitted values. The model diagnostics were used to assess the goodness of fit of a selected model. Monthly data from January 2000 to December 2019 was used in the analysis and during this period a total of 20,217,026 tourists arrived in Sri Lanka. Moreover, there are certain decline periods of this volume mainly due to the impacts of civil war, Tsunami and many others. The findings indicate that the model ARIMA (2,1,2) $(3,1,4)_{[3]}$ captures the behavior well with a minimum MAPE of 0.1941 and NRMSE of 0.8800. Meanwhile, with the application of CPA (at most one change and pruned exact linear time), data was split into two separate windows, which are Window 1 (W1) from January 2000 to October 2011 and Window 2 (W2) from November 2011 to December 2019. In W1, the better model that was used in the prediction was ARIMA (1,1,1) $(4,1,1)_{[3]}$ with a MAPE and NRMSE of 0.1727 and 1.1190 respectively. According to the results, the better performed model (MAPE of 0.2740 and NRMSE of 0.8700) in W2, was ARIMA (0,1,1) $(3,1,3)_{[3]}$ and this model captured the behavior until April 2019. However, due to the Easter bomb attack in April, there was a sudden drop in the arrival of tourists in May and June 2019. Nevertheless, from this point onwards the predicted line captured the behavior of the actual values even though they did not coincide with each other. Again, in December 2019, the predicted and actual values were very close. Thus, this study will be a benefit for both the private and public sectors as it has a prominent impact on the economy of the country.

Key words: Tourism; Time-series; Change point analysis; Forecasting; Seasonal Autoregressive Integrated Moving Average.

1. Introduction

Tourism is a crucial scope that has a direct impact on the economy around the world. When a location becomes a major tourist destination this affects advantageously to a country. Some of them facilitate new job opportunities in different sectors like health, education, and

Corresponding Author: B.R.P.M. Basnayake
E-mail: pavithramalkibasnayake@gmail.com

agriculture, revealing the cultural and social values of the country to the world, earning profits, developing the infrastructure and many others.

According to the Sri Lanka Tourism Development Authority, modern commercial tourism was initiated in 1960 with a long history and in this period 18,969 tourists arrived in Sri Lanka. Nevertheless, the terrorist attacks from 1983 to 2009, had a negative impact on this industry for a long period. In addition, the Tsunami hazard that occurred in 2004, resulted in many deaths and property damages whereas it indicated a decline in the growth of the tourism industry. For more than 25 years, there were deprivations caused by the civil war and at the end of the war, there was a significant development in the industry of tourism in Sri Lanka. At the same time due to the Easter Sunday bomb attack in April 2019, the number of tourists who arrived in Sri Lanka decreased.

This study mainly focuses and attempts on forecasting tourist arrivals in Sri Lanka by identifying the patterns in arrivals using the time series models. Furthermore, sudden changes are identified by the change point detections.

In Sri Lanka, the planning and policy implementation activities related to tourism are implemented by the Tourism Development Authority. Therefore, this work will benefit the government as well as the private sector for their future investments and progress. Moreover, this study will support the sustainability of the tourism industry and the processes related to the conservation of resources such as wildlife, cultural heritages and other natural resources.

There are many studies conducted relevant to the prediction of tourist arrivals in many countries including Sri Lanka. However, there is no related work identified with the change point analysis (CPA) to predict the volume of tourists in Sri Lanka.

This paper is organized as follows. The subsequent section is a review of previous related works. Sections 3 and 4 consist of the methodology and data analysis respectively. Section 5 includes the discussion and section 6 consists of the conclusions of the study.

2. Literature Review

Different previous studies were conducted relevant to tourism in many countries with different techniques.

In 1984, Jozef suggested that Harrison's harmonic smoothing technique was more appropriate to predict the foreign tourists who arrived in Netherland compared to the decomposition technique and Box Jenkins generalized adaptive filtering. An exponentially weighted non-linear time series approach with a sine function in time was used by Chan (1993) to forecast the volume of tourist arrival in Singapore after de-seasonalizing data due to the seasonal behavior and model performance was evaluated from the Mean Absolute Percentage Error (MAPE). Using Seasonal Autoregressive Integrated Moving Average (SARIMA) and Multivariate Autoregressive Integrated Moving Average (MARIMA), Goh and Law (2002) forecasted the tourist demand for Hong Kong with ten arrival series and the non-stationary behavior was recognized from the Augmented Dickey-Fuller test. Lim *et al.* (2002) found that the number of tourist arrivals from Singapore to Australia followed an Autoregressive Integrated Moving Average (ARIMA) approach where arrivals to Malaysia and Hong Kong extended with the SARIMA method. Similarly, many studies applied the ARIMA and SARIMA techniques in forecasting the tourist arrivals such as Saayman and Saayman (2010); Singh (2013); Kumar and Sharma (2016); Chhorn and Chaiboonsri (2017), and many others.

Cho (2003) applied three techniques: Exponential smoothing, SARIMA and Artificial Neural Network (ANN) to identify the travel demand to Hong Kong from different countries and claimed that ANN exhibited better forecasting with minimum errors for the series with the fewer fluctuations and ARIMA approach was better for the arrival patterns with obvious patterns.

ANN and hybrid models were built as the alternatives to the ARIMA models in the study of Aslanargun *et al.* (2007) and they stated that the models with components of non-linear indicated better performance. Moreover, the studies of Law and Au (1999) and Pai *et al*. (2006) have used the data science concept in forecasting tourist arrivals.

Due to the impacts of the Civil war and political influence in Sri Lanka, there were ups and downs in the tourism industry from 2003 to 2009. After the end of the war in 2009, Sri Lanka became a significant tourist destination as per the study of Fernando *et al.* in 2017. They claimed that Sri Lanka needs to increase the accommodation and infrastructure facilities with the tourism workforce.

Arrivals from the Western European countries (UK, Germany, France, Italy and Netherland) to Sri Lanka were considered by Konarasinghe *et al.* (2016) as they were the main contributors to the market of tourism in Sri Lanka. The patterns in arrivals were detected using time series plots and Auto-Correlation Functions (ACF) with the decomposition techniques. They concluded the additive decomposition model was better and recommended the circular model to increase the accuracy in forecasting. Peiris (2016) conducted a study after identifying seasonality in the monthly data for the period from January 1995 to July 2016 with the Hegy test. In this study, the SARIMA (1,0,16) $(36,0,24)_{[12]}$ model was identified as a better performed model to forecast the arrivals of tourists in Sri Lanka. However, using the monthly time series data from June 2009 to December 2018, the study of Nyoni in 2019 identified the optimal model with the minimum MAPE of 8.6877% value in the SARIMA (0, 1, 1) $(0, 1, 1)_{[12]}$ to forecast tourist arrivals in Sri Lanka.

The change point detections are vital in practical situations such as in financial analysis, climatology, and many other areas (Eckley *et al*., 2011). The At Most One Change Point (AMOC) method was repeatedly applied to detect multiple change points in climate by Wang in 2006. In addition, a study conducted by Lund *et al*. (2007) claimed the shifts in time series can be pointed out by the AMOC method.

Bakka (2018) stated that the Pruned Exact Linear Method (PELT) performed well in the univariate Gaussian series compared to the Binary segmentation method. Chapman and Killick (2020) assessed the prediction with change points in software applications using the PELT method and suggested that CPA is very useful in the cases of a large amount of data.

However, in this study, SARIMA models were built for the seasonal difference of 3, 6 and 12 separately based on combined pre and post war eras. Further, this study used the CPA to detect the important changes in arrivals. Following the CPA, separate new time series models were built for the windows with different seasonal differences. Thereafter, an attempt was made to identify the appropriate models with the lowest Akaike Information Criterion (AIC) value for each seasonal difference and recognized the better model for the prediction of tourist arrivals using the performance measures MAPE and normalized root mean squared error (NRMSE).

3. Methodology

Month wise data from January 2000 to December 2019 was obtained from the website of Sri Lanka Tourism Development Authority. Initially, the behavior of the data was recognized with the time-series plots where basic features were identified using descriptive analysis. For further analysis, time-series data was split for training and testing in a non-random manner. The stationary or the non-stationary behavior was pointed out using the ACF and PACF plots with the number of cut-off lags. Furthermore, unit root tests were applied to check the stationarity. Applied unit root tests are:

Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test

H_0: The series is stationary

H_1: The series is not stationary

Augmented Dickey-Fuller test (ADF) test and Phillips-Perron test (PP) test

H_0: The series possesses a unit root (The series is not stationary)

H_1: The series do not possess a unit root (The series is stationary)

For these three tests, if the *p*-value is less than the considered significant level, H_0 (Null hypothesis) is rejected at the significance level.

The non-stationary data was converted to stationary through the application of different transformations. Seasonality features were identified with the patterns in ACF, PACF plots and using the Webel-Ollech (WO) test. The WO is an overall seasonality test that merged results from QS-test and the Kwman-test. This test identifies the seasonality in the series by the QS-test if the *p*-value is below 0.01 and by the Kwman-test if the *p*-value is lower than 0.002.

ARIMA models are wide-ranging applications in time series analysis to realize the behavior of data and for prediction. The general form of the ARIMA model illustrates below:

$$ARIMA\ (p,\ d,\ q) \tag{1}$$

where p is the number of parameters in the autoregressive (AR) model, d is the differencing degree, q is the number of parameters in the Moving Average (MA) model. However, with the seasonality behavior, the ARIMA was extended to SARIMA. The general form of the SARIMA model is in Equation 2:

$$ARIMA\ (p,d,\ q)\ (P,D,Q)_s \tag{2}$$

where p is the number of parameters in the autoregressive (AR) model, d is the differencing degree, q is the number of parameters in the MA model, P is the number of parameters in the seasonal AR model, D is the seasonal differencing degree, Q is the number of parameters in seasonal MA model and s is the period of seasonality.

The parameters of ARIMA and SARIMA models are identified by the Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF). For the built SARIMA models, Akaike Information Criterion (AIC) was used to identify the better model with the minimum AIC value. Model assumptions of heteroscedasticity, autocorrelation and normality of residuals (model diagnostics) were evaluated using the tests ARCH, Ljung-Box and Jarque-Bera respectively for the selected models and they are below:

Heteroscedasticity: ARCH test

H_0: There is no heteroscedasticity in the residuals

H_1: There is heteroscedasticity in the residuals

Autocorrelation: Ljung-Box Test on Residuals

H_0: There is no autocorrelation in the residuals

H_1: There is autocorrelation in the residuals

Normality: Jarque –Bera Test

H_0: Residuals are normally distributed

H_1: Residuals are not normally distributed

The ARCH test is used to identify the behavior of the error term variance and if the residuals are homoscedastic then the *p*-value of the test is greater than the considered significant level. Also, in time series modeling the error terms should be free of autocorrelation and if there is no autocorrelation then the *p*-value is greater than the significance level for the Ljung-Box tests. Jarque-Bera test is a goodness of fit test which is used to detect the normality behavior of the residuals and in the presence of the normality for this test, the *p*-value is greater than the given significance level.

Then the selected model was used to predict the values in the test set (final 10% of data). The model accuracy was identified using Mean Absolute Percentage Error (MAPE) and Normalized Root Mean Squared Error (NRMSE) and calculated using Equations 3 and 4.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|e_t|}{y_t} \times 100 \tag{3}$$

$$NRMSE = \frac{\sqrt{\sum_{t=1}^{n}\frac{(e_t)^2}{n}}}{\bar{y}} \tag{4}$$

where t is the time-period, Y_t is the actual value, $\bar{y}$ is the average of the observations and $e_t = y_t - \hat{y}_t$ is the error in the period t and n is the number of observations.

Detection of change points in modeling and prediction of time series is an important task. CPA is useful in identifying whether a change or more than one change has occurred in the data and at which time the changes have occurred. CPA is performed on a time ordered series to detect the changes that occurred (Hackl, 2013) where it identifies the multiple changes with smaller shifts. The CPA can consider both mean and variance changes and, in this study, AMOC was applied to detect the single change point (Eckley *et al.*, 2011). This method is based on the likelihood-ratio approach and the hypothesis of the change is as follows:

H_0: No change point

H_1: A singel change point

To find the test statistic from both hypothesis (null and alternative), the maximum log-likelihood is calculated and is compared it with a threshold value to accept or reject the null hypothesis. If the test statistic is greater than the considered threshold, the null hypothesis is rejected.

PELT method was used to detect multiple change points as this method is fast and accurate than the binary segmentation and other methods (Wambui *et al.*, 2015). This approach minimizes the general penalized likelihood from the Schwarz information criterion (SIC) (Yao, 1988) and points out the appropriate model. In the PELT method, the linear functions of the number of change points are the linear penalties as follows:

$$\text{pen}(T) = \beta\,|T| \tag{5}$$

where T is a set of indexes and β is a smoothing parameter that controls the goodness of fit and complexity.

The pruning rule:

If the $\left[\min V\left(T, y_{0,\ldots,t}\right) + \beta\,|T|\right] + c\left(y_{t,\ldots,s}\right) \geq \left[\min V\left(T, y_{0,\ldots,t}\right) + \beta\,|T|\right]$ then t cannot be the concluding change point before T (Truong *et al.*, 2020).

where y is a signal, t < s < T, t and s are indexes and V is a function of y and T.

In this study, both AMOC and PELT methods were employed to identify the change points considering both mean and variance change in data using the package "changepoint" in R software (Killick & Eckley (2014)). This package calculates the number of change points with their optimal positions for given penalty functions and assumed test statistics.

4. Data Analysis

This section consists of the descriptive analysis of the data and results obtained through the SARIMA and CPA approaches for each case. Case I consists of the data from January 2000 to December 2019 while case II describes the approach with the application of CPA for the data.

4.1. Case I

Figure 1 illustrates the number of tourists who arrived from January 2000 to December 2019. There was a significant decline in the period from the year 2000 to mid of 2011. This huge decline may be due to the destructive terrorism phenomenon which was experienced in Sri Lanka. However, there is a gradual increase from the end of the year 2011 to the end of the year 2019.

According to Figure 2, there are two outliers in the boxplot as December 2018 and February 2019. However, to implement the continuity of time series data points outliers are not removed. The right-skewed data imply that most numbers of tourist arrivals are relatively small, and only a few are long. There were no missing values in the dataset.

The minimum value of the tourist arrival was 11,758 in September 2001 while the maximum amount was 253,169 in the month of December 2018. On average 84,238 tourists arrived in Sri Lanka. From the year 2000 to 2019 total of 20,217,026 tourists arrived and the Standard deviation value was 61132.3. This means that 68% of the total tourist arrivals are between 23,106 and 145,370.

Data was split with an initial 90% for the training set from January 2000 to December 2017 and the remaining 10% for the testing set from January 2018 to December 2019 for testing. There is a clear upward trend in the training data. Therefore, the series was not stationary, and

it was identified through the unit root tests. PP test indicated that the series was stationary where the ADF test and KPSS tests suggested it was not stationary at a 5% level of significance. Therefore, the log transformation was applied to the original data to make the data smoother (stabilize the variance). Then, differencing was applied to stabilize the mean of a time series by removing changes in the level of a time series (to reduce the trend). The trend was eliminated after this transformation. Through the unit root tests, the differenced log transformed training set was stationary at a 5% level of significance.

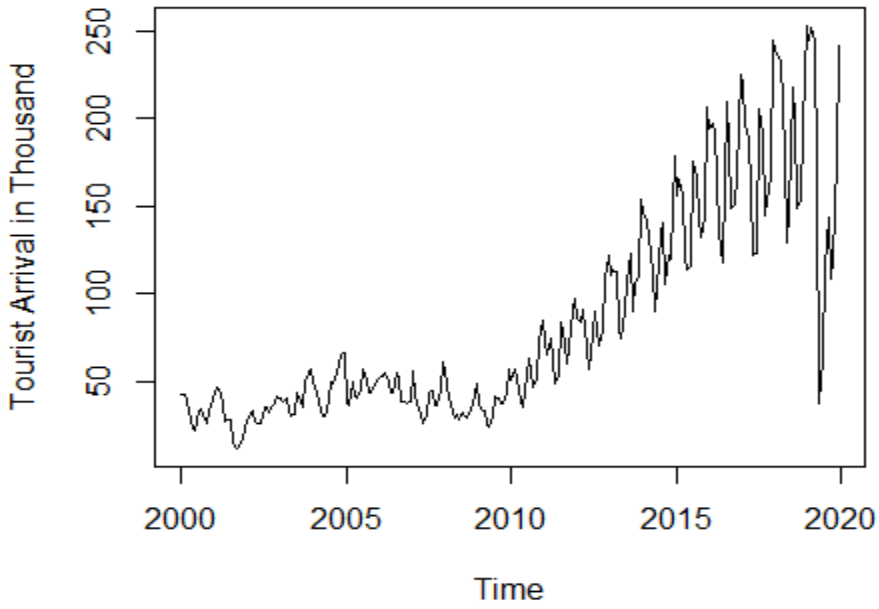


Figure 1: Time-series plot of tourist arrivals in thousand from 2000 to 2019

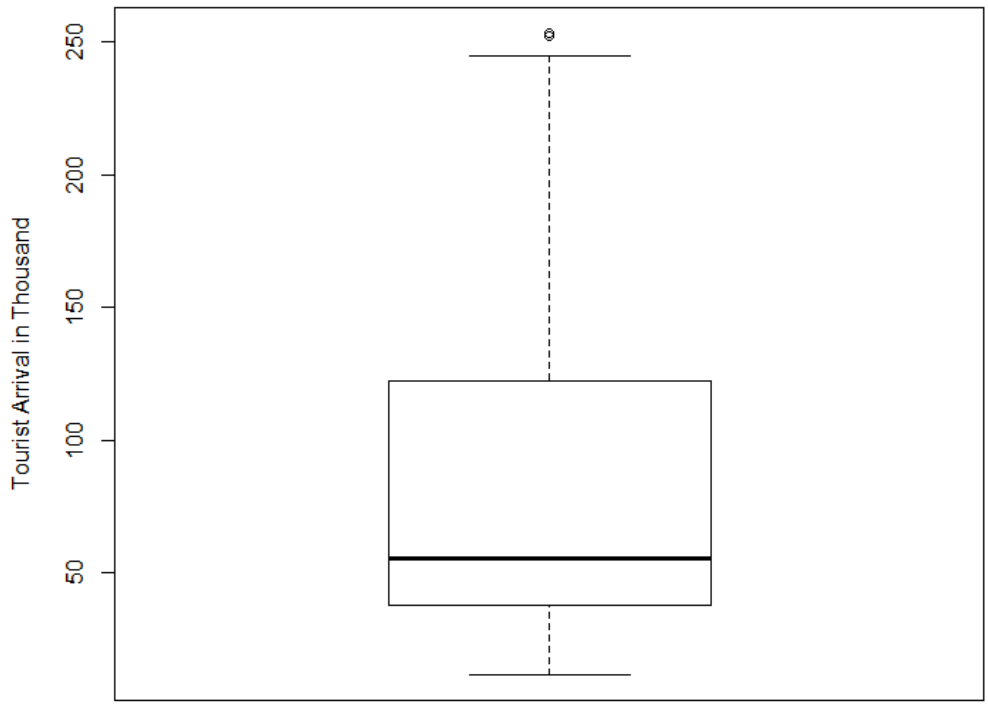
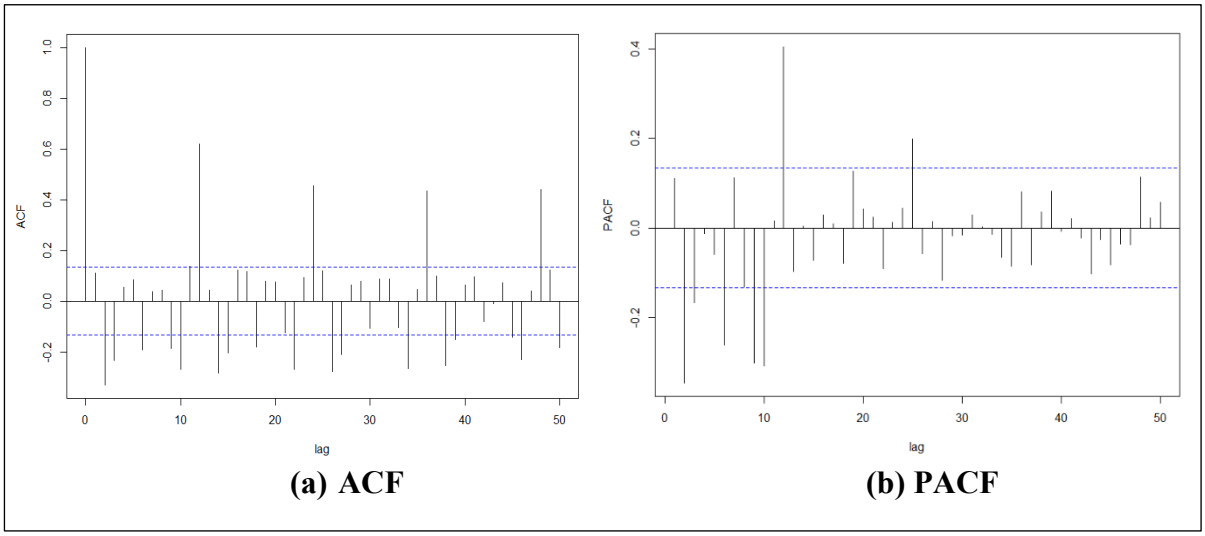


Figure 2: Boxplot of tourist arrivals

The ACF and PACF plots were used to identify the seasonal and non-seasonal lags as in Figures 3 (a) and 3 (b) respectively.



(a) ACF **(b) PACF**

Figure 3: ACF and PACF plots of the transformed series of tourist arrivals

From the ACF plot, it is visible that the significant lags are 2, 3, 6, 9, 10, 12, 14, 15 and 18 and from the PACF plot, lags 1, 2, 6, 9, 10 and 12 are significant (Figure 3). The seasonality of the transformed data was identified using the WO test. Through the ACF and PACF plots, different seasonal lags were recognized as 3, 6 and 12. Subsequently, candidate models for each case of seasonal differences were identified.

Initially, the transformed data was seasonally differed by 3, 6 and 12 separately. By considering the results of unit root tests, all the series were stationary at a 5% level of significance. From the ACF and PACF plots of the seasonality differed series, significant non-seasonal lags and seasonal lags were obtained as in Table 1.

Table 1: Seasonal lags and non-seasonal lags from ACF plots and PACF plots for Case I

	ACF		PACF	
Seasonal difference	Seasonal lags	Non-seasonal lags	Seasonal lags	Non-seasonal lags
3	3,9,12	1.2	3,6,9	1,2
6	6,12,18	1,2,3	6,18	1,2,3
12	12, 24	-	12,24,36	-

The models (relevant to seasonal difference by 3) were built from identified lags as in Table 1 in Annexure. In the table, the minimum AIC was –229.76 in the model ARIMA (2,1,2) $(3,1,4)_{[3]}$. For this model, the Jarque-Bera test violated the normality assumption and satisfied the assumptions of homoscedasticity and the absence of autocorrelation in residuals through the ARCH and Ljung-Box tests respectively. Aryani *et al.* (2018) stated that even the residual normality assumption of the ARIMA model is violated (it reflected the data with high volatility), the model can be used in forecasting. Hence, the model was used to forecast as all the other candidate models violated the assumption of normality in residuals.

Using the seasonal lags and non-seasonal lags, appropriate models were recognized similar to the procedure in Table 1 in Annexure for the seasonal difference of 6 and 12. The

minimum AIC was –236.53 in the model ARIMA (1,1,1) $(3,1,2)_{[6]}$ and through the model diagnostic tests, it only violated the normality assumption. Therefore, the aforementioned model was used in prediction among the models with the seasonal difference of 6.

The minimum AIC was –240.44 in the model ARIMA (0,1,0) $(2,1,1)_{[12]}$ (relevant to seasonal difference by 12) and in here, all the models violated the assumption of normality. Therefore, this model was used in forecasting tourist arrivals among all the models with the seasonal difference of 12.

4.2. Case II

In Case II, the CPA (both AMOC and PELT methods) were employed to identify the change points. There are many penalty functions that can be applied in the CPA which are AIC, Bayesian information criterion (BIC), SIC and Hannan-Quinn. In addition, the assumed test statistic can take Normal, Gamma, Exponential and Poisson distributions (Killick and Eckley (2014)). This study used AIC, BIC and SIC methods for the penalty functions in CPA (Wambui *et al.*, 2015). The test of fit for the probability distribution of Gamma was identified using the test of variance ratio for Gamma distributions (see Villasenor and Gonzalez-Estrada, 2015; Gonzalez-Estrada, 2020). For the test of fit, the test statistic value was 2.0834 and the *p*-value was 0.1407. They indicated that the null hypothesis of data follow a Gamma distribution was not rejected at a 5% level of significance. As data follow a Gamma distribution, the assumed test statistic was considered with the Gamma distribution in this study.

Multiple change points were not identified from the PELT method. From both AMOC and PELT methods only one change point was identified for all the information criteria. It is the 142nd observation as in Figure 4 and the change point was detected in the month of October 2011. From 2000 to mid of 2011, there was a decline period in tourist arrivals and that might be due to the impact of civil war. However, from the end of 2011, there is an increase in tourist arrivals. Therefore, theoretically and practically, it can be concluded that a better change point was identified from the AMOC and PELT methods.

Therefore, Window 1 (W1) was built based on the data from January 2000 to October 2011 and Window 2 (W2) was based on the time period from November 2011 to December 2019.

4.2.1. Window 1

In Window 1 (Figure 5), the minimum value of the tourist arrival was 11,758 in September 2001 while the maximum amount was 84,627 in the month of December 2010. On average 42,311 tourists had arrived in Sri Lanka from January 2000 to October 2011 which was a total of 6,008,152.

Observations from January 2000 to October 2010 were used as the training set while observations from November 2010 to October 2011 were used as the testing set.

ADF and PP tests exhibited that the training set was stationary at a 5% level of significance. However, the KPSS test indicated that the series was not stationary. The difference transformation for the log transformed variable was used to make data smooth and to remove the trend. The transformed series was stationary at a 5% level of significance.

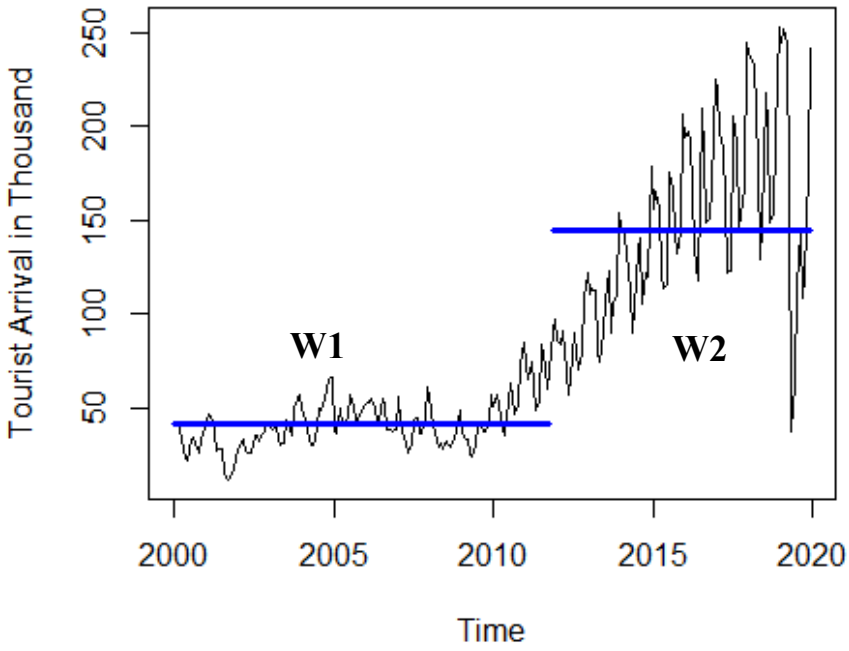


Figure 4: Detection of Change point using AMOC and PELT methods

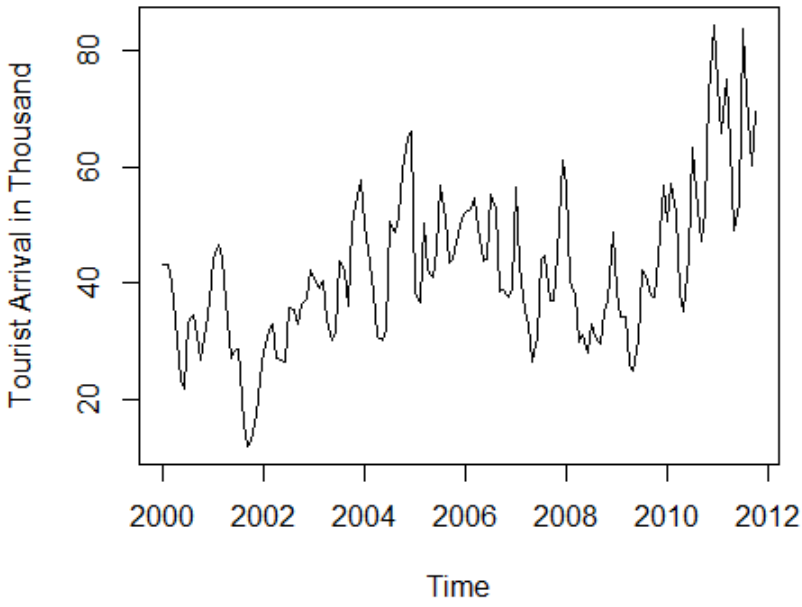


Figure 5: Time-series plot of tourist arrivals from January 2000 to October 2011

The WO test indicated the seasonality feature. Then the ACF and PACF plots were used to identify the seasonal and non-seasonal lags.

From Figure 6, the ACF plot indicates that the significant lags are 2, 3, 6, 11, 12, 24, 36 and 48 and the PACF plot indicates that the significant lags are 2, 6, 8, 9, 10 and 12. From the ACF and PACF plots, three seasonal differences were identified as 3, 6 and 12. The models were built separately for each seasonal difference of 3, 6 and 12.

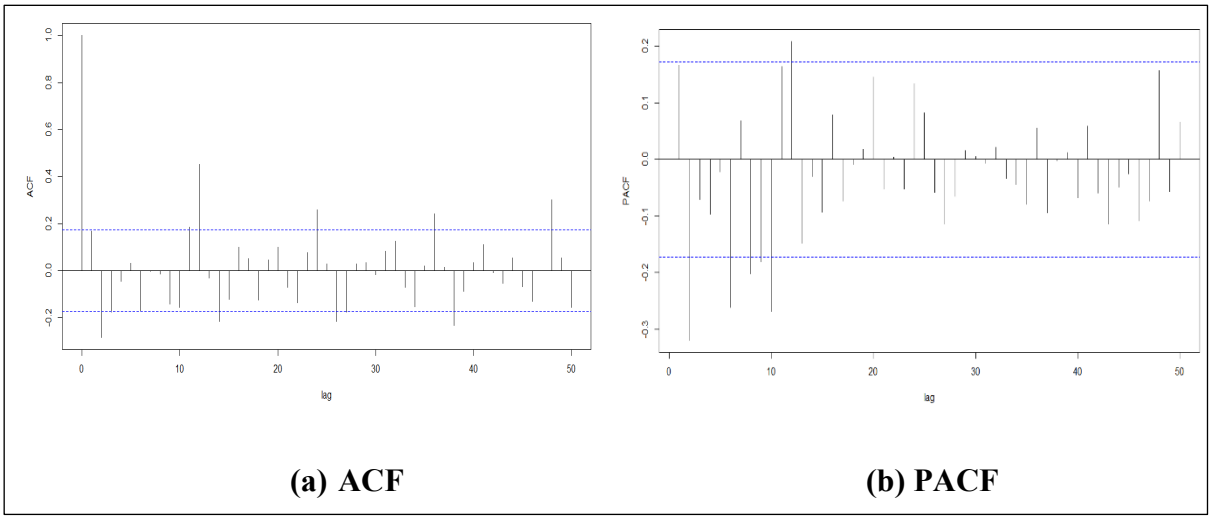


Figure 6: ACF and PACF plots of the transformed series of tourist arrivals

The seasonally differed series by 3, 6 and 12 were stationary at a 5% level of significance. As in Table 2, found the significant seasonal and non-seasonal lags from ACF and PACF plots to identify a better model in forecasting. Candidate models were built for each seasonal difference (same task as in Table 1 in Annexure) for Window 1.

Table 2: Seasonal lags and non-seasonal lags from ACF plots and PACF plots for W1

	ACF		PACF	
Seasonal difference	Seasonal lags	Non-seasonal lags	Seasonal lags	Non-seasonal lags
3	3,6,12,24	1,2	6,9,12	1,2
6	6,12,24,36	1,2,3	6,12	1,2
12	12,24,36,48	1,2,3	12	1,2

The minimum AIC (relevant to seasonal difference by 3) was in the ARIMA (1,1,1) $(4,1,1)_{[3]}$ which of –94.70. From the model diagnostic tests of ARCH, Ljung-Box and Jarque-Bera, this model satisfied all the assumptions in residuals.

The lowest AIC was in the model ARIMA (2,1,2) $(1,1,4)_{[6]}$ with a –103.39 value (relevant to seasonal difference by 6). However, the model violated the assumption of normality whereas all the other candidate models violated that assumption. Therefore, the aforementioned model was used to forecast the arrival of tourists among models with the seasonal difference of 6.

The seasonality differed series by 12 was stationary and the least AIC was –107.57 in the model ARIMA (2,1,2) $(1,1,4)_{[12]}$. This model violated the normality assumption. However, the model was used for forecasting as all other candidate models violated the assumption of normality in residuals.

4.2.2. Window 2

Window 2 was built based on the time period from November 2011 to December 2019. There is an upward trend in Figure 7. There is a sudden drop in May 2019 due to the Easter bombings on 28th April 2019 on Easter Sunday. In Window 2, the minimum value of the tourist arrival was 37,802 in May 2019 while the maximum amount was 253,169 in

December 2018. On average 144,989 tourists arrived in Sri Lanka from 2011 November to 2019 December while a total of 14,208,874.

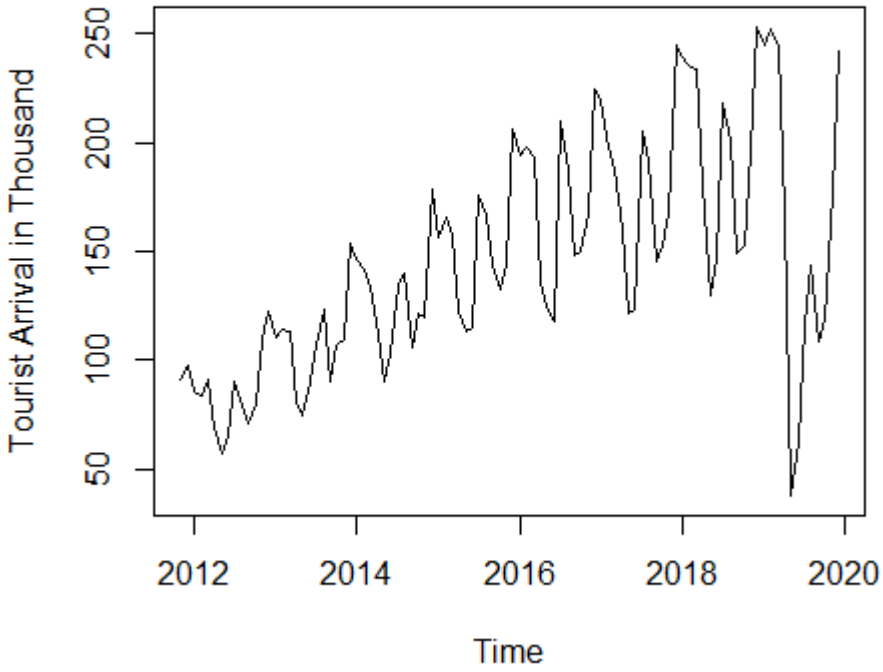


Figure 7: Time-series plot of tourist arrivals from 2011 of November to 2019 of December

The train set is from November 2011 to December 2018. The test set is from January 2019 to December 2019. The train set consists of a clear upward trend. In addition, it seems to have a seasonal pattern. The series was not stationary and it was examined through the unit root tests.

The difference transformation for the log transformed variable was used to make data smooth and to remove the trend. The transformed series was stationary at a 5% level of significance. The seasonality behavior was identified using the WO test. Significant lags from ACF and PACF plots (Figures 8) were used to build the models. From ACF plot, significant lags are 2, 3, 9, 10, 12, 14, 15, 22, 24, 26, 27, 34, 36, 39 and 48 while PACF plot illustrates the significant lags as 2, 3, 6, 7, 9, 10, 11, 12 and 18. The seasonality was identified in 3, 6 and 12 seasonal differences. Therefore, applied the seasonal differences in 3, 6 and 12 separately and identified seasonal and non-seasonal lags for each case as in Table 3.

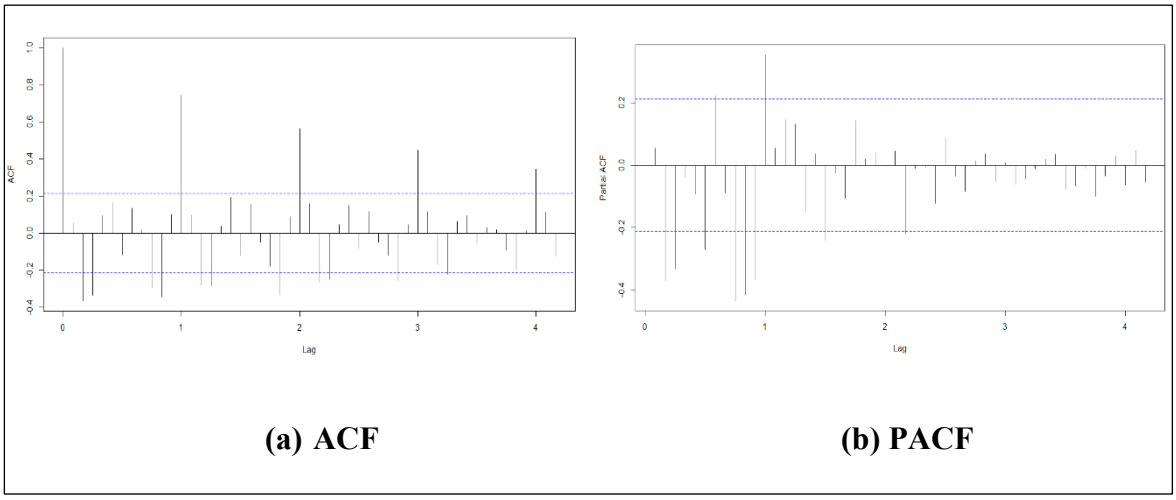


Figure 8: ACF plot of the transformed series of tourist arrivals

Table 3: Seasonal lags and non-seasonal lags from ACF plots and PACF plots for W2

	ACF		PACF	
Seasonal difference	Seasonal lags	Non-seasonal lags	Seasonal lags	Non-seasonal lags
3	3,9,12,15	1,2	3,6,9,12,18	1,2
6	12,24,36	1,2,3	6,12,18	1,2,3
12	12,24,36,48	1,2,3	12	1,2,3

Seasonally differed series by 3 was stationary and the least AIC was –146.09 in the model ARIMA (0,1,1) $(3,1,3)_{[3]}$ and it satisfied all the model diagnostics assumptions of the model.

Seasonally differed series by 6 was stationary and the minimum AIC was –166.12 in the model ARIMA (2,1,3) $(1,1,6)_{[6]}$. Moreover, this model satisfied all the model diagnostics assumptions.

Seasonally differed series by 12 was stationary and among all the candidate models, minimum AIC was –155.55 in the ARIMA (0,1,1) $(1,1,1)_{[12]}$ model. Even the normality assumption is violated, used this model in forecasting. The next least AIC value –154.23 was in model ARIMA (1,1,2) $(1,1,1)_{[12]}$ and it satisfied all the assumptions of model diagnostics.

5. Discussion

In this section, the better performed models in forecasting tourist arrivals were identified for Case I and Case II.

5.1. Case I

Table 4 indicates the performance measures, assumption satisfaction or violation of model diagnostics with their AIC values for better performed models in each seasonal difference. ARIMA (0,1,0) $(2,1,1)_{[12]}$ has the lowest AIC value. However, among all the models in Table 4, it has the highest MAPE and NRMSE values. The MAPE value was minimum in the model ARIMA (2,1,2) $(3,1,4)_{[3]}$ (Model A). However, it violated the assumption of

normality in residuals and satisfied all the assumptions of model diagnostics. Therefore, model A is the better model that can use in forecasting tourist arrivals in Case I.

Table 4: Better performed models to identify the behavior of the arrival of tourists from January 2000 to December 2019

Case I	Model	Assumptions			AIC	MAPE	NRMSE
		Heteroscedasticity	Autocorrelation	Normality			
ARIMA $(2,1,2)(3,1,4)_{[3]}$	A	Absence	Absence	Absence	–229.76	0.1941	0.8800
ARIMA $(1,1,1)(3,1,2)_{[6]}$	B	Absence	Absence	Absence	–236.53	0.2340	0.9740
ARIMA $(0,1,0)(2,1,1)_{[12]}$	C	Absence	Absence	Absence	–240.44	0.2786	1.2630

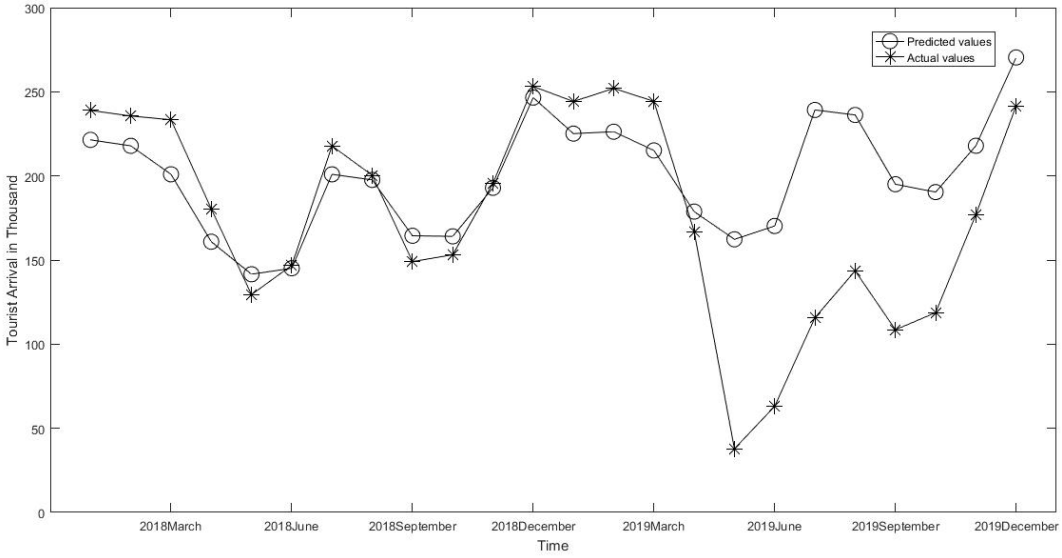


Figure 9: Actual and Predicted values of tourist arrivals from ARIMA (2,1,2) $(3,1,4)_{[3]}$

Figure 9 indicates the actual and predicted values where the asterisk marks illustrate the actual values and the point marks show the predicted values for the test set from the beginning of 2018 to the end of 2019. Until April 2019, the fitted model captured the behavior of tourist arrivals in Sri Lanka and due to the Easter bomb attack in April 2019, there was a sudden drop in May 2019. However, model A identified the actual behavior till the end of 2019 even two lines do not coincide with each other.

5.2 Window 1

Window 1 was build based on the data from January 2000 to October 2011. From Table 5, the minimum AIC is in Model F while it violated the normality of assumption of residuals and has higher MAPE and NRMSE values. The better model that can used in the prediction is ARIMA (1,1,1) $(4,1,1)_{[3]}$ (Model D). In addition, it satisfied all the assumptions of model diagnostics while all other models dissatisfied the assumption of the normality of the error terms. Compared to all the models in Table 5, the lowest performance measures were in Model D. Hence, Model D was used to forecast tourist arrivals in Window 1.

Table 5: Better performed models to identify the behavior of the arrival of tourists from January 2000 to October 2011

Window 1 SARIMA	Model	Assumptions			AIC	MAPE	NRMSE
		Heteroscedasticity	Autocorrelation	Normality			
ARIMA $(1,1,1)(4,1,1)_{[3]}$	D	Absence	Absence	Presence	–94.70	0.1727	1.1190
ARIMA $(2,1,2)(1,1,4)_{[6]}$	E	Absence	Absence	Absence	–103.39	0.2751	1.5520
ARIMA $(2,1,2)(1,1,4)_{[12]}$	F	Absence	Absence	Absence	–107.57	0.3034	1.7460

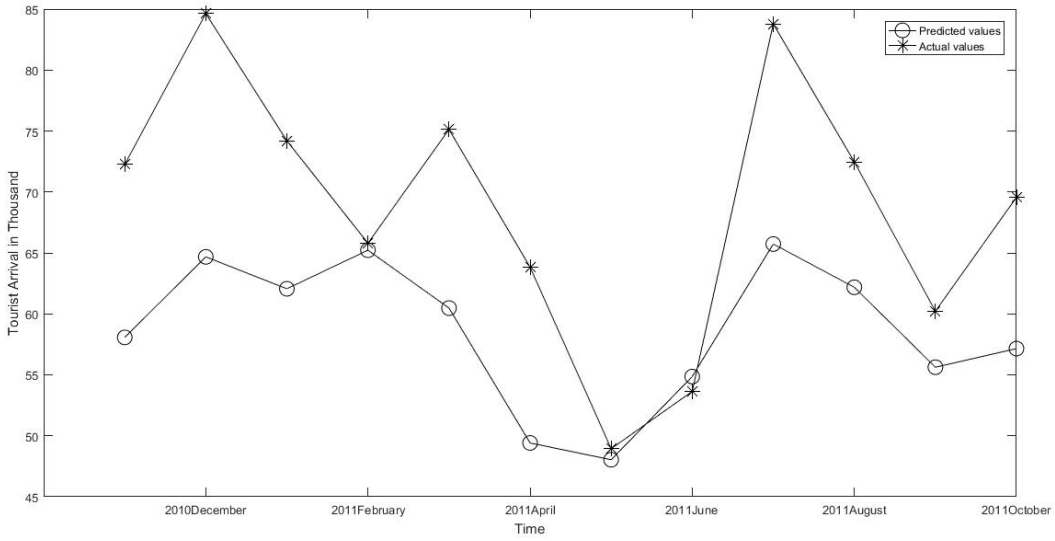


Figure 10: Actual and Predicted values of tourist arrivals from ARIMA (1,1,1) $(4,1,1)_{[3]}$

Observations from November 2010 to October 2011 were used as the testing set. Sri Lankan civil war was ended in May 2009 and as a result of that, there are fluctuations in the actual values of tourist arrivals (asterisk marks) in Figure 10. Thus, the predicted values (point marks) are not very similar to actual values.

5.3. Window 2

Window 2 was built based on the data from November 2011 to December 2019. The lowest AIC is in Model H as per Table 6 and it has higher MAPE and NRMSE values compared to model G. The better model that can be used in forecasting tourist volume is ARIMA (0,1,1) $(3,1,3)_{[3]}$ (Model G) in Window 2. Further, it consists of the lowest MAPE and NRMSE compared to all other models in Table 6 and it satisfied all the assumptions in the model diagnostics with a lower AIC value.

Table 6: Better performed models to identify the behavior of the arrival of tourists from November 2011 to the December 2019

Window 2 SARIMA	Model	Assumptions			AIC	MAPE	NRMSE
		Heterosce dasticity	Autocor relation	Normalit y			
ARIMA $(0,1,1)(3,1,3)_{[3]}$	G	Absence	Absence	Presence	–146.09	0.2740	0.8700
ARIMA $(2,1,3)(1,1,6)_{[6]}$	H	Absence	Absence	Presence	–166.12	0.3447	1.0520
ARIMA $(0,1,1)(1,1,1)_{[12]}$	I	Absence	Absence	Absence	–155.55	0.3473	1.1050
ARIMA $(1,1,2)(1,1,1)_{[12]}$	J	Absence	Absence	Presence	–154.23	0.3460	1.1000

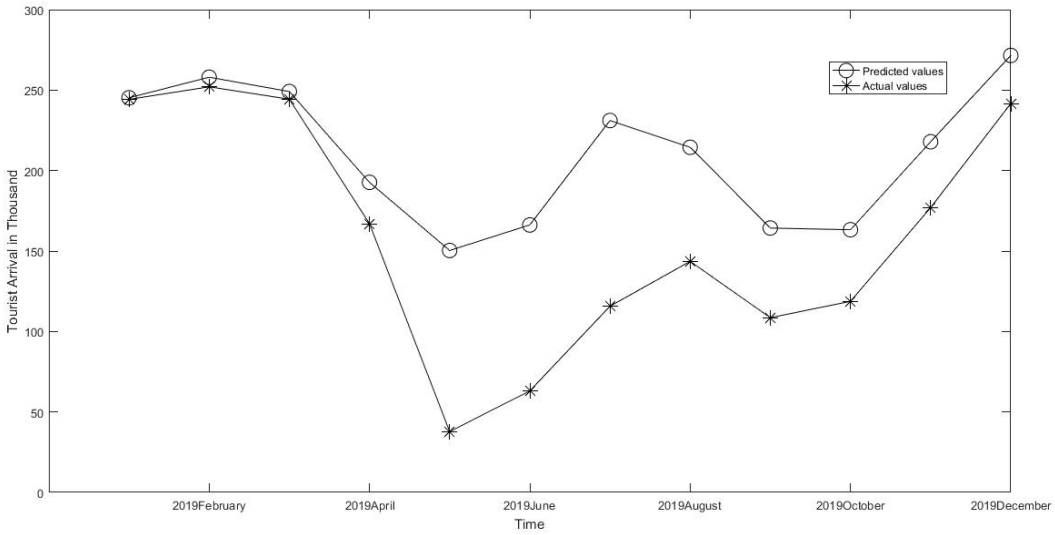


Figure 11: Actual and Predicted values of tourist arrivals from ARIMA (0,1,1) $(3,1,3)_{[3]}$

Figure 11 indicates the predicted (point marks) and actual data (asterisk marks) values in the test set from January 2019 to December 2019. The fitted model captures the behavior until April 2019. However, there was a sudden drop in the arrival of tourists in May 2019 due to the Easter bomb attack in April. Nevertheless, from this point onwards the predicted line captures the behavior of the actual values even though they do not coincide with each other. Again in December 2019, the predicted and Actual values are very close to each other.

6. Conclusions

The findings of the study are important to make the major decisions relevant to tourism to achieve sustainable development of the sector. Tourist arrivals in Sri Lanka indicate a seasonality pattern and a clear upward trend after 2010. According to this study, there were models built which were relevant to seasonal lags 3, 6 and 12. The outperformed model that can be used in forecasting tourist arrivals from January 2000 to December 2019 was the ARIMA (2,1,2) $(3,1,4)_{[3]}$ model which exhibits the lowest MAPE and NRMSE values with the satisfaction of all model diagnostics assumptions except the normality of the residuals. With the application of CPA, from January 2000 to October 2011 (Window 1) the better performed model was ARIMA (1,1,1) $(4,1,1)_{[3]}$. However, the model did not capture the actual behavior of tourist arrivals due to the fluctuations in values of tourist arrivals after the end of the civil

war in May 2009. From November 2011 to December 2019 (Window 2), the better model that can be used in forecasting was ARIMA (0,1,1) $(3,1,3)_{[3]}$ and it satisfied all the model diagnostics assumptions. Thus, this study is a benefit for both the private and public sectors as tourist arrivals have a prominent impact on the economy of the country. Moreover, future forecasting information is vital in the decision making for industries related to tourism. For further implications, supervised machine learning algorithms can be used to build forecasting models to forecast the tourist arrivals in Sri Lanka with higher accuracy.

References

Aryani, S., Aidi, M. N. and Syafitri, U. D. (2018). Analysis of the profitability of islamic banking using ARIMAX model and regression with ARIMA errors model. *International Journal of Scientific Research in Science, Engineering and Technology,* **4**(**8**), 49-53.

Aslanargun, A., Mammadov, M., Yazici, B. and Yolacan, S. (2007). Comparison of ARIMA, neural networks and hybrid models in time series: tourist arrival forecasting. *Journal of Statistical Computation and Simulation*, **77**(**1**), 29-53.

Bakka, K. B. (2018). *Changepoint Model Selection in Gaussian Data by Maximization of Approximate Bayes Factors with the Pruned Exact Linear Time Algorithm.* Master's Thesis, Norwegian University of Science and Technology.

Chan, Y. M. (1993). Forecasting tourism: A sine wave time series regression approach. *Journal of Travel Research*, **32**(**2**), 58-60.

Chapman, J. L. and Killick, R. (2020). An assessment of practitioners approaches to forecasting in the presence of changepoints. *Quality and Reliability Engineering International*, **36**(**8**), 2676-2687.

Chhorn, T. and Chaiboonsri, C. (2018). Modelling and Forecasting Tourist Arrivals to Cambodia: An Application of ARIMA-GARCH Approach. *Journal of Management, Economics and Industrial Organization*, **2**(**2**), 1-19.

Cho, V. (2003). A comparison of three different approaches to tourist arrival forecasting. *Tourism Management*, **24**(**3**), 323-330.

Eckley, I. A., Fearnhead, P. and Killick, R. (2011). Analysis of changepoint models. In *Bayesian Time Series Models*, Eds. D. Barber, A. T. Cemgil and S. Chiappa. Cambridge University Press, 205-224.

Fernando, S., Bandara, J. S. and Smith, C. (2017). Tourism in Sri Lanka. In *The Routledge Handbook of Tourism in Asia*, Eds. M. C. Hall and S. J. Page. Abingdon, Oxon, UK: Routledge, 251-264

Goh, C. and Law, R. (2002). Modeling and forecasting tourism demand for arrivals with stochastic nonstationary seasonality and intervention. *Tourism Management*, **23(5)**, 499-510.

Gonzalez-Estrada, E., Villasenor-Alva, J. A. and Gonzalez-Estrada, M. E. (2020). Package 'goft'.

Hackl, P. (Ed.). (2013). *Statistical Analysis and Forecasting of Economic Structural Change*. Springer Science and Business Media.

Jozef, W. M. (1984). Tourism forecasting and the policymaker: Criteria of usefulness. *Tourism Management*, **5**(**1**), 24-39.

Killick, R. and Eckley, I. (2014). Changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, **58**(**3**), 1-19.

Konarasinghe, U. (2016, December). *Decomposition Techniques on Forecasting Tourist Arrivals from Western European Countries to Sri Lanka*. University of Sri Jayewardenepura, Sri Lanka, 13th International Conference on Business Management (ICBM).

Kumar, M. and Sharma, S. (2016). Forecasting tourist in-flow in South-East Asia: A case of Singapore. *Tourism and Management Studies*, **12**(**1**), 107-119.

Law, R. and Au, N. (1999). A neural network model to forecast Japanese demand for travel to Hong Kong. *Tourism Management*, **20**(**1**), 89-97.

Lim, C. and McAleer, M. (2002). Time series forecasts of international travel demand for Australia. *Tourism Management*, **23**(**4**), 389-396.

Lund, R., Wang, X. L., Lu, Q. Q., Reeves, J., Gallagher, C. and Feng, Y. (2007). Changepoint detection in periodic and autocorrelated time series. *Journal of Climate*, **20**(**20**), 5178-5190.

Neusser, K., 2016. *Time Series Econometrics*. Springer.

Nyoni, T. (2019). Sri Lanka–the wonder of Asia: analyzing monthly tourist arrivals in the post-war era. *University Library of Munich*, MPRA Paper No. 96790, Germany.

Pai, P. F., Wei-Chiang, H., Ping-Teng, C. and Chen-Tung, C. (2006). The application of support vector machines to forecast tourist arrivals in Barbados: An empirical study. *International Journal of Management*, **23**(**2**), 375.

Peiris, H. (2016). A seasonal ARIMA model of tourism forecasting: The case of Sri Lanka. *Journal of Tourism, Hospitality and Sports*, **22**(**1**), 98-109.

Saayman, A. and Saayman, M. (2010). Forecasting tourist arrivals in South Africa. *Professional Accountant*, **10**(**1**), 281-293.

Singh, E. H. (2013). Forecasting tourist inflow in Bhutan using seasonal ARIMA. *International Journal of Science and Research*, **2**(**9**), 242-245.

Truong, C., Oudre, L. and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, **167**, 107299.

Villasenor, J. A. and Gonzalez-Estrada, E. (2015). A variance ratio test of fit for Gamma distributions. *Statistics and Probability Letters*, **96**(**1**), 281-286. http://dx.doi.org/10.1016/j.spl.2014.10.001

Wang, X. L. (2006). *A Recursive Testing Algorithm for Detecting and Adjusting for Multiple Artificial Changepoints in a Time Series*. Report of the Fifth Seminar for Homogenization and Quality Control in Climatological Databases, Budapest, Hungary, World Climate Data and Monitoring Programme, WMO.

Wambui, G. D., Waititu, G. A. and Wanjoya, A. (2015). The power of the pruned exact linear time (PELT) test in multiple changepoint detection. *American Journal of Theoretical and Applied Statistics*, **4**(**6**), 581.

Yao, Y. C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics and Probability Letters*, **6**(**3**), 181-189.

ANNEXURE

Table 1: SARIMA Models with seasonal difference of 3

Model	AIC	Model	AIC
ARIMA $(1,1,1)(1,1,1)_{[3]}$	–99.66	ARIMA $(0,1,1)(1,1,3)_{[3]}$	–117.70
ARIMA $(1,1,1)(1,1,3)_{[3]}$	–122.20	ARIMA $(0,1,1)(1,1,4)_{[3]}$	–142.60
ARIMA $(1,1,1)(1,1,4)_{[3]}$	–149.80	ARIMA $(0,1,1)(2,1,1)_{[3]}$	–109.30
ARIMA $(1,1,1)(2,1,1)_{[3]}$	–115.00	ARIMA $(0,1,1)(2,1,3)_{[3]}$	–162.90
ARIMA $(1,1,1)(2,1,3)_{[3]}$	–161.40	ARIMA $(0,1,1)(2,1,4)_{[3]}$	–151.30
ARIMA $(1,1,1)(2,1,4)_{[3]}$	155.96	ARIMA $(0,1,1)(3,1,1)_{[3]}$	–180.70
ARIMA $(1,1,1)(3,1,1)_{[3]}$	–185.90	ARIMA $(0,1,1)(3,1,3)_{[3]}$	–195.70
ARIMA $(1,1,1)(3,1,3)_{[3]}$	–196.70	ARIMA $(0,1,1)(3,1,4)_{[3]}$	–223.10
ARIMA $(1,1,1)(3,1,4)_{[3]}$	–226.80	ARIMA $(0,1,2)(1,1,1)_{[3]}$	–111.30
ARIMA $(1,1,2)(1,1,1)_{[3]}$	–115.40	ARIMA $(0,1,2)(1,1,3)_{[3]}$	–135.40
ARIMA $(1,1,2)(1,1,3)_{[3]}$	–139.00	ARIMA $(0,1,2)(1,1,4)_{[3]}$	–158.30
ARIMA $(1,1,2)(1,1,4)_{[3]}$	158.51	ARIMA $(0,1,2)(2,1,1)_{[3]}$	–129.70
ARIMA $(1,1,2)(2,1,1)_{[3]}$	–131.90	ARIMA $(0,1,2)(2,1,3)_{[3]}$	–166.00
ARIMA $(1,1,2)(2,1,3)_{[3]}$	–181.20	ARIMA $(0,1,2)(2,1,4)_{[3]}$	–167.70
ARIMA $(1,1,2)(2,1,4)_{[3]}$	–168.10	ARIMA $(0,1,2)(3,1,1)_{[3]}$	–184.70
ARIMA $(1,1,2)(3,1,1)_{[3]}$	–183.80	ARIMA $(0,1,2)(3,1,3)_{[3]}$	–203.50
ARIMA $(1,1,2)(3,1,3)_{[3]}$	–203.10	ARIMA $(0,1,2)(3,1,4)_{[3]}$	–223.20
ARIMA $(1,1,2)(3,1,4)_{[3]}$	–227.80	ARIMA $(1,1,0)(1,1,1)_{[3]}$	–87.43
ARIMA $(2,1,1)(1,1,1)_{[3]}$	–113.30	ARIMA $(1,1,0)(1,1,3)_{[3]}$	–117.60
ARIMA $(2,1,1)(1,1,3)_{[3]}$	–142.00	ARIMA $(1,1,0)(1,1,4)_{[3]}$	–142.00
ARIMA $(2,1,1)(1,1,4)_{[3]}$	–160.70	ARIMA $(1,1,0)(2,1,1)_{[3]}$	–106.00
ARIMA $(2,1,1)(2,1,1)_{[3]}$	–144.50	ARIMA $(1,1,0)(2,1,3)_{[3]}$	–159.70
ARIMA $(2,1,1)(2,1,3)_{[3]}$	–190.70	ARIMA $(1,1,0)(2,1,4)_{[3]}$	–150.30
ARIMA $(2,1,1)(2,1,4)_{[3]}$	–176.00	ARIMA $(1,1,0)(3,1,1)_{[3]}$	180.67
ARIMA $(2,1,1)(3,1,1)_{[3]}$	–184.60	ARIMA $(1,1,0)(3,1,3)_{[3]}$	–195.70
ARIMA $(2,1,1)(3,1,3)_{[3]}$	–204.50	ARIMA $(1,1,0)(3,1,4)_{[3]}$	–223.10
ARIMA $(2,1,1)(3,1,4)_{[3]}$	–228.20	ARIMA $(2,1,0)(1,1,1)_{[3]}$	–109.00
ARIMA $(2,1,2)(1,1,1)_{[3]}$	–147.50	ARIMA $(2,1,0)(1,1,3)_{[3]}$	–125.00
ARIMA $(2,1,2)(1,1,3)_{[3]}$	–140.00	ARIMA $(2,1,0)(1,1,4)_{[3]}$	–156.80
ARIMA $(2,1,2)(1,1,4)_{[3]}$	–158.70	ARIMA $(2,1,0)(2,1,1)_{[3]}$	–128.40
ARIMA $(2,1,2)(3,1,1)_{[3]}$	–194.70	ARIMA $(2,1,0)(2,1,3)_{[3]}$	–171.10
ARIMA $(2,1,2)(3,1,3)_{[3]}$	–211.00	ARIMA $(2,1,0)(2,1,4)_{[3]}$	–158.80
ARIMA $(2,1,2)(3,1,4)_{[3]}$	–229.76	ARIMA $(2,1,0)(3,1,1)_{[3]}$	–183.20
ARIMA $(0,1,1)(1,1,1)_{[3]}$	–89.47	ARIMA $(2,1,0)(3,1,3)_{[3]}$	–200.60
		ARIMA $(2,1,0)(3,1,4)_{[3]}$	–222.80

Statistics and Applications {ISSN 2454–7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 123–133

A Survey on Cyclic Solution of Block Designs

Shyam Saurabh[1] **and Kishore Sinha**[2]
[1]*Ranchi University, Ranchi, India*
[2]*Formerly at Birsa Agricultural University, Ranchi, India and*
#201 Maitry Residency, Kalkere Main Road, Bangalore – 560043

Received: 07 September 2021; Revised: 17 September 2021; Accepted: 21 September 2021

Abstract

Cyclic designs are incomplete block designs based on cyclic development of one or more initial blocks. John *et al.* (1972) described the advantages of cyclic designs as calibration designs and experimental designs and tabulated these designs in the useful range of parameters which was published by National Bureau of Standards, Washington, DC. The cyclic designs may have up to $v/2$ associate classes. The purpose of this survey is to present cyclic solutions of balanced incomplete block designs, group divisible designs, Latin square type designs and cyclic designs, wherever possible, which have at most two associate classes and higher efficiencies.

Key words: Balanced incomplete block (BIB) designs; Semi–regular and regular group divisible designs; Latin square type designs; Cyclic Designs.

1. Introduction

Cyclic designs are incomplete block designs based on cyclic development of one or more initial blocks. Their flexibility, ease in conduct of experiment and natural groupings for one-way elimination of heterogeneity, make them worthy of attention in their own right. All cyclic designs are partially balanced incomplete block (PBIB) designs with up to $v/2$ associate classes. Among the class of cyclic designs, cyclic balanced incomplete block (BIB) designs are obviously best in the sense that all the pair-wise treatment comparisons are measured with same and maximum efficiency. When no cyclic BIB design exists, then we look for cyclic solution of two associate class PBIB design with same (v, b, r, k). These designs are used as calibration designs and experimental designs [see John *et al.* (1972), Clatworthy (1973), John and Williams (1995)]. Cyclic designs were catalogued by John *et al.* (1972). The cyclic solutions of BIB designs were given by Hall (1998), wherever possible. Clatworthy (1973) tabulated two associate classes PBIB designs. The purpose of this paper is to present a survey on cyclic solutions of BIB designs, group divisible designs, Latin square type designs and cyclic designs in the range of $r, k \leq 10$.

The concept of cyclic designs is extended to generalized cyclic designs which are useful as factorial experiments [see Jarrett and Hall (1978), Lamacraft and Hall (1982), Nigam *et al.* (1988), Dean and Lewis (1990), Bailey (1990)].

A *Group divisible design* (*GDD*) is an arrangement of v (= mn; m, $n \geq 2$) treatments into b blocks such that each block contains k ($<v$) distinct treatments, each treatment occurs r times and any pair of distinct treatments which are first associates occur together in λ_1 blocks and in λ_2 blocks if they are second associates. Furthermore, if $r-\lambda_1 = 0$ then the GD design is

Corresponding Author: Kishore Sinha
Email: kishore.sinha@gmail.com

singular (S); if $r-\lambda_1 > 0$ and $rk-v\lambda_2 = 0$ then it is semi–regular (SR); and if $r-\lambda_1 > 0$ and $rk-v\lambda_2 > 0$, the design is regular (R). For definitions and terminologies, we refer to Dey (1986, 2010), Raghavarao (1971), Raghavarao and Padgett (2005).

2. Cyclic Solutions of Block Designs

Table 1: Cyclic Solutions: BIBD/ GD design/ Cyclic Design / Latin square type design

No.	BIBD/ GDD/ CD/ LSD: (v, r, k, b); Overall Efficiency	John No.; Overall Efficiency	Cyclic Solutions
1^{M}	*SR*1: (4, 2, 2, 4); 0.60	-	*G*: (1, 4) mod 4
2^{*}	*C*1: (5, 2, 2, 5); 0.50	-	*C*: (1, 3) mod 5
3^{*}	*C*6: (5, 6, 2, 15); 0.61	-	*C*: (1, 3); (1, 3); (1, 2) mod 5
4^{*}	*C*7: (5, 8, 2, 20); 0.59	-	*C*: (1, 3); (1, 3); (1, 3); (1, 2) mod 5
5^{*}	*C*8: (5, 10, 2, 25); 0.58	-	*C*: (1, 3); (1, 3); (1, 3); (1,3); (1, 2) mod 5
6*	*C*9: (5, 10, 2, 25); 0.62	-	*C*: (1, 3); (1, 3); (1, 3); (1, 2); (1, 2) mod 5
7^{M}	*SR*7: (6, 6, 2, 18); 0.56	2×*A*2; 0.55	*G*: (0, 1); (0, 3); (0, 5) mod 6
8^{M}	*SR*13: (12, 6, 2, 36); 0.52	2×*A*26; 0.39	*G*: (0,1); (0, 3); (0, 5) mod 12
9^{*}	*C*10: (13, 6, 2, 39); 0.50	*A*36; 0.50	(1, 3); (1, 6); (1, 7) mod 13
10^{M}	*SR*15: (16, 8, 2, 64); 0.52	*A*57; 0.52	*G*: (0, 1); (0, 3); (0, 5); (0, 7) mod 16
11^{*}	*C*11: (17, 8, 2, 68); 0.50	*A*62; 0.51	(1, 4); (1, 6); (1, 7); (1, 8) mod 17
12^{M}	*SR*17: (20, 10, 2, 100); 0.51	*A*81; 0.51	*G*: (0, 1); (0, 3); (0, 5); (0, 7); (0, 9) mod 10
13^{*}	*C*12: (5, 3, 3, 5); 0.81	-	*C*: (1, 2, 4) mod 5
14^{*}	*C*15: (5, 9, 3, 15); 0.83	-	(1, 3, 5); (1, 3, 5); (1, 2, 5) mod 5
15^{*}	*R*42: (6, 3, 3, 6); 0.78	*B*1, 0.78	*G*: (1, 2, 4) mod 6
16	*H*1: (7, 3, 3, 7); 0.78	*B*2; 0.78	*B*: (1, 2, 4) mod 7
17^{*}	*R*54: (8, 3, 3, 8); 0.75	*B*3; 0.75	*G*: (1, 2, 4) mod 8
18^{DN}	*R*55: (8, 6, 3, 16); 0.75	*B*5; 0.75	*G*: (1, 2, 3); (1, 3, 6) mod 8
19^{*}	*R*58: (8, 9, 3, 24); 0.76	3× *B*3; 0.75	*G*: (1, 2, 3); (1, 2, 5); (1, 3, 6) mod 8
20^{MD}	*SR*23: (9, 3, 3, 9); 0.73	*B*9; 0.72	*G*: (3, 5, 8); (2, 6, 8); (2, 5, 9); 1↔3, 4↔6, 7↔9 (*PC*)
21	*H*2: (9, 4, 3, 12); 0.75	-	Add the blocks: (1+3x, 2+3x, 3+3x); $0 \le x \le 2$ to the solution in Serial No. 20
22^{M}	*SR*25: (9, 9, 3, 27); 0.73	3× *B*9; 0.72	*G*: (0, 1, 2); (0, 4, 8); (0, 5, 7) mod 9
23^{MD}	*R*68: (9, 10, 3, 30); 0.74	-	*G*:(1, 2, 3); (1, 2, 6); (1, 3, 5); (1, 4, 7) mod 9
24	*H*26: (10, 9, 3, 30); 0.74	*B*14; 0.70	*B*: (∞, 0, 5); (0, 1, 4); (0, 2, 3); (0, 2, 7) mod 9
25^{*}	*C*16: (13, 3, 3, 13); 0.67	*B*50; 0.67	*C*: (1, 3, 9) mod 13
26	*H*9: (13, 6, 3, 26); 0.72	-	*B*: (1, 3, 9); (2, 5, 6) mod13
27^{*}	*C*19: (13, 9, 3, 39); 0.72	*B*54; 0.72	*C*: (1, 12, 13); (3, 10, 13); (4, 9, 13) mod 13
28^{*}	*R*80: (14, 9, 3, 42); 0.67	*B*64; 0.67	*G*: (1, 2, 8); (1, 8, 9); (1, 3, 8); (1, 8, 10); (1, 4, 8); (1, 8, 11); 1↔7, 8↔14 (*PC*)
29^{*}	*R*81: (15, 6, 3, 30); 0.71	*B*75; 0.71	*G*: (1, 4, 15); (2, 8, 15) mod 15
30^{*}	*R*83: (15, 9, 3, 45); 0.71	*B*77; 0.71	*G*: (1, 7, 13); (1, 4, 5); (1, 3, 8) mod 15 *G*: (1, 2, 5); (1, 3, 8); (1, 4, 10) mod15
31	*H*14: (15, 7, 3, 35); 0.71	*B*76; 0.71	*B*: (1_1, 4_1, 0_2); (2_1, 3_1, 0_2); (1_2, 4_2, 0_3); (2_2, 3_2, 0_3); (1_3, 4_3, 0_1); (2_3, 3_3, 0_1);

			$(0_1, 0_2, 0_3)$ mod 5
32*	*LS*18: (16, 3, 3, 16); 0.63	*C1*; 0.63	*L*: (7, 10, 16); (4, 6, 13); (4, 7, 9); (2, 9, 16); 1↔4, 5↔8, 9↔12, 13↔16 (PC)
33*	*R*86: (16, 6, 3, 32); 0.70	2× *C*1; 0.63	*G*: (1, 2, 11); (1, 3, 6) mod 16
34*	*R*87: (16, 9, 3, 48); 0.71	*C*1; 0.63	*G*: (1, 5, 13); (1, 2, 11); (1, 3, 6) mod 16
35*	*R*89: (18, 9, 3, 54); 0.70	*C*11; 0.61	*G*: (1, 11, 13); (1, 10, 14); (1, 15, 18); (1, 16, 17); (1, 2, 5); (1, 3, 12); 1↔9, 10↔18 (*PC*)
36*F*	*R*89*a*: (18, 10, 3, 60); 0.69	-	*G*: (*A*1, *A*2, *B*1); (*B*1, *B*2, *A*1); (*A*1, *A*8, *B*1); (*B*1, *B*8, *A*4); (*A*1, *A*6, *B*4); (*B*1, *B*6, *A*1); $\frac{1}{3}${(*A*1, *A*4, *A*7), (*B*1, *B*4, *B*7)} mod 9
37*	*R*91: (21, 9, 3, 63); 0.70	3× *C*32; 0.60	*G*: (1, 2, 11); (1, 3, 7); (1, 4, 9) mod 21
38	*H*38: (21, 10, 3, 70); 0.70	-	*B*: $(1_1, 6_1, 0_2)$; $(2_1, 5_1, 0_2)$; $(3_1, 4_1, 0_2)$; $(1_2, 6_2, 0_3)$;$(2_2, 5_2, 0_3)$; $(3_2, 4_2, 0_3)$; $(1_3, 6_3, 0_1)$; $(2_3, 5_3, 0_1)$; $(3_3, 4_3, 0_1)$; $(0_1, 0_2, 0_3)$ mod 7
39*	*R*92: (24, 9, 3, 72); 0.69	3× *C*52; 0.58	*G*: (1, 2, 12); (1, 3, 8); (1, 4, 10) mod 24
40*	*LS*22: (25, 6, 3, 50); 0.67	2× *C*60; 0.57	*L*: (1, 5, 25); (9, 14, 15); (18, 23, 24); (2, 7, 8); (11, 16, 17); (1, 3, 13); (12, 15, 22); (6, 21, 24); (8, 10, 20); (4, 17, 19); 1↔5, 6↔10, 11↔15, 16↔20, 21↔25 (PC)
41*	*C*20: (37, 9, 3, 111); 0.67	-	(1, 10, 26); (1, 31, 34) (1, 11, 37) mod 37
42*	*R*94: (6, 4, 4, 6); 0.89	-	*G*: (1, 2, 4, 6) mod 6
43*DB*	*SR*36: (8, 4, 4, 8); 0.84	*B*6; 0.85	*G*: (2, 3, 4, 5); (1, 6, 7, 8); 1↔4, 5↔8 (*PC*)
44*	*R*98: (8, 8, 4, 16); 0.85	2× *B*6; 0.85	*G*: (1, 2, 3, 5); (1, 2, 4, 6) mod 8
45*MD*	*SR*39: (8, 8, 4, 16); 0.84	2× *B*6; 0.85	*G*: (1, 4, 6, 7); (1, 2, 3, 4) mod 8
46*	*R*104: (9, 4, 4, 9); 0.80	*B*12; 0.83	*G*: (1, 2, 4, 7) mod 9 *J*: (1, 2, 4, 5) mod 9
47*	*R*105: (9, 8, 4, 18); 0.80	2× *B*12; 0.83	*G*: (1, 2, 4, 7); (1, 2, 5, 8) mod 9 2 copies of *J*: (1, 2, 4, 5) mod 9
48	*H*20: (9, 8, 4, 18); 0.84	2× *B*12; 0.83	*B*: (0, 1, 2, 4); (0, 1, 4, 6) mod 9
49*DB*	*R*106: (10, 8, 4, 20); 0.82	2× *B*18; 0.83	*G*: (3, 4, 5, 6); (1, 8, 9, 10); (2, 4, 5, 6); (1, 7, 9, 10); 1↔5, 6↔10 (*PC*)
50*	*R*109: (12, 4, 4, 12); 0.81	-	*G*: (1, 2, 5, 7) mod 12
51*F*	*R*109*a*: (12, 7, 4, 21); 0.82	-	*G*: (*A*1, *A*2, *A*3, *B*4); (*A*1, *A*3, *B*1, *B*6); (*A*1, *A*4, *B*2, *B*6); $\frac{1}{2}$(*B*1, *B*2, *B*4, *B*5) mod 6
52*MD*	*R*110: (12, 8, 4, 24); 0.81	*B*39; 0.82	*G*: (1, 2, 5, 7); (1, 2, 8, 10) mod 12
53*F*	*R*110*b*: (12, 10, 4, 30); 0.81	2× *B*37; 0.81	*G*: (*A*1, *A*2, *A*3, *A*6); (*A*1, *A*3, *B*4, *B*6); (*B*1, *B*2, *B*3, *A*6); (*B*1, *B*3, *A*4, *B*6); (*A*1, *A*2, *B*3, *B*5); (*B*1, *B*2, *A*3, *A*5); (mod 5) and 6 invariant
54	*H*3: (13, 4, 4, 13); 0.81	*B*55; 0.81	*B*: (0, 1, 3, 9) mod 13
55*	*C*21: (13, 8, 4, 26); 0.80	*B*56; 0.81	*C*: (1, 4, 12, 13); (1, 4, 10, 13) mod 13
56*	*R*112: (14, 4, 4, 14); 0.80	*B*65; 0.80	*G*: (1, 2, 5, 7) mod 14
57*MD*	*R*113: (14, 8, 4, 28); 0.80	*B*67; 0.80	*G*: (1, 2, 5, 7); (1, 2, 10, 12) mod 14
58*F*	*R*113*a*: (14, 10, 4, 35); 0.80	*B*69; 0.80	*G*: (*A*1, *A*2, *A*4, *B*7); (*B*1, *B*2, *B*7, *A*7); (*A*1, *A*2, *B*1, *B*2); (*A*1, *A*3, *B*1, *B*3); (*A*1, *A*4, *B*1, *B*4) mod 7
59*	*R*114: (15, 4, 4, 15); 0.80	*B*79; 0.80	*G*: (1, 3, 4, 12) mod 15
60*	*R*115: (15, 8, 4, 30); 0.73	*B*81; 0.80	*G*: (1, 2, 6, 11); (1, 6, 7, 11); (1, 6, 11, 12); (1, 6, 8, 11); (1, 6, 11, 13); (1, 3, 6, 11); 1↔-

			5, 6↔10, 11↔15 (*PC*) *J*: (1, 2, 5, 6); (1, 3, 9, 11) mod 15
61^{MD}	*R*116: (15, 8, 4, 30); 0.80	*B*81; 0.80	*G*: (0, 1, 3, 7); (1, 3, 4, 12) mod15
62*	*R*117: (15, 8, 4, 30); 0.80	*B*81; 0.80	*G*: (1, 3, 11, 15); (1, 5, 7, 15) mod15
63*	*LS*38: (16, 8, 4, 32); 0.80	2× *C*2; 0.79	*L*: (5, 6, 8, 11); (1, 5, 9, 13); (1, 4, 10, 15); (7, 13, 14, 16); (9, 10, 12, 15); (1, 2, 7, 12); (1, 5, 9, 13); (1, 3, 6, 16); 1↔4, 5↔8, 9↔12, 13↔16 (PC)
64*	*C*22: (17, 8, 4, 34); 0.79	2× *C*7; 0.78	*C*: (2, 9, 11, 17); (1, 4, 5, 17) mod 17
65^{A}	*C*22*A*: (17, 10, 5, 34); 0.85	2× *C*8; 0.84	*C*: (0, 5, 12, 14, 3); (0, 7, 10, 11, 6) mod 17
66^{F}	*R*123*a*: (18, 10, 4, 45); 0.79	-	*G*: (*A*1, *A*2, *A*3, *B*4); (*A*1, *A*3, *B*5, *C*4); $\frac{1}{2}$(*A*1, *A*4, *B*2, *B*5) perm *A*, *B*, *C* mod 6
67	*SR*46: (20, 5, 4, 25); 0.78	-	By deleting the treatments 21, 22, 23, 24, 25 from *SR*60
68^{F}	*R*124*a*: (22, 8, 4, 44); 0.77	2× *C*41; 0.76	*G*: (*A*1, *A*3, *A*4, *B*1); (*A*1, *A*7, *B*1, *B*8) perm *A*, *B* and mod 11
69^{F}	*R*126*a*: (24, 9, 4, 54); 0.77	-	*G*: (*A*1, *A*2, *A*9, *B*1); (*B*1, *B*2, *B*9, *A*1); (*A*1, *A*4, *B*1, *B*11); (*B*1, *B*4, *A*1, *A*11); $\frac{1}{2}$(*A*1, *A*7, *B*1, *B*7) mod 12
70	*H*22: (25, 8, 4, 50); 0.78	2× *C*61; 0.75	*B*: [(0, 0); (1, 0); (0, 1); (4, 4)] mod (5, 5); [(0, 0); (2, 0); (0, 2); (3, 3)] mod (5, 5)
71*	*R*128: (26, 8, 4, 52); 0.78	2× *C*68; 0.74	*G*: (2, 4, 10, 14); (1, 16, 19, 20); (3, 6, 7, 14); (1, 15, 17, 23); 1↔13, 14↔26 (*PC*)
72^{F}	*R*128*a*: (26, 10, 4, 65); 0.76	-	*G*: (*A*1, *A*6, *A*8, *B*1); (*B*1, *B*6, *B*8, *A*1); (*A*1, *A*2, *B*1, *B*4); (*B*1, *B*2, *A*1, *A*4); (*A*1, *A*5, *B*1, *B*5) mod 13
73*	*R*132: (30, 10, 4, 75); 0.78	-	*G*: (1, 3, 15, 20); (5, 16, 18, 30); (1, 5, 11, 17); (2, 16, 20, 26); (1, 9, 16, 24); 1↔15, 16↔30 (*PC*)
74*	*R*133: (8, 5, 5, 8); 0.90	-	*G*: (1, 2, 3, 5, 7) mod 8
75*	*R*134: (8, 5, 5, 8); 0.91	-	*G*: (1, 3, 4, 5, 6) mod 8
76^{DN}	*R*136: (8, 10, 5, 16); 0.91	-	*G*: (1, 5, 6, 7, 8); (1, 3, 5, 6, 8) mod 8
77*	*R*137: (9, 5, 5, 9); 0.89	-	*G*: (1, 3, 4, 6, 7) mod 9
78*	*R*138: (9, 10, 5, 18); 0.89	-	*G*: (1, 3, 4, 6, 7); (1, 3, 4, 6, 9) mod 9
79*	*R*139: (10, 5, 5, 10); 0.88	*B*21; 0.88	*G*: (1, 2, 3, 6, 8) mod 10
80*	*R*141: (10, 10, 5, 20); 0.89	2× *B*21; 0.88	*G*: (1, 2, 3, 4, 7); (1, 2, 4, 6, 8) mod 10
81	*H*5: (11, 5, 5, 11); 0.88	*B*27; 0.88	*B*: (1, 3, 4, 5, 9) mod 11
82*	*R*143: (12, 5, 5, 12); 0.81	*B*43; 0.87	*G*: (1, 2, 4, 7, 10) mod 12 *J*: (1, 2, 3, 5, 8) mod 12
83*	*R*144: (12, 5, 5, 12); 0.87	*B*43; 0.87	*G*: (1, 2, 4, 9, 12) mod 12
84*	*R*145: (12, 5, 5, 12); 0.87	*B*43; 0.87	*G*: (1, 2, 4, 6, 7) mod 12
85*	*R*146: (12, 10, 5, 24); 0.81	2× *B*43; 0.87	*G*: (1, 2, 4, 7, 10); (1, 3, 4, 7, 10) mod 12 *J*: 2 copies of (1, 2, 3, 5, 8) mod 12
86^{MD}	*R*147: (12, 10, 5, 24); 0.87	2× *B*43; 0.87	*G*: (0, 1, 2, 4, 9); (0, 1, 2, 5, 10) mod 12 *J*: 2 copies of (1, 2, 3, 5, 8) mod 12
87*	*R*148: (12, 10, 5, 24); 0.87	2× *B*43; 0.87	*G*: (1, 2, 3, 6, 12); (1, 3, 6, 8, 12) mod 12 *J*: 2 copies of (1, 2, 3, 5, 8) mod 12
88*	*R*149: (15, 10, 5, 30); 0.82	2× *B*82; 0.85	*G*: (1, 2, 6, 7, 11); (1, 3, 6, 8, 11) mod 15

			J: 2 copies of (1, 2, 3, 5, 11) mod 15
89*	*R*150: (15, 10, 5, 30); 0.86	*B*82; 0.85	*G*: (1, 2, 3, 5, 8); (1, 2, 5, 9, 11) mod 15
90*S*	*R*150*a*: (15, 10, 5, 30); 0.84	*B*82; 0.85	*G*: (1, 2, 4, 7, 11); (1, 2, 4, 10, 13) mod 15
91*	*R*152: (20, 10, 5, 40); 0.74	-	*G*: (1, 2, 6, 11, 16); (1, 6, 7, 11, 16); (1, 6, 11, 12, 16); (1, 6, 11, 16, 17); (1, 6, 8, 11, 16); (1, 6, 11, 13, 16); (1, 6, 11, 16, 18); (1, 3, 6, 11, 16); 1↔5, 6↔10, 11↔15, 16↔20 (*PC*)
92	*H*7: (21, 5, 5, 21); 0.84	*C*34; 0.84	*B*: (3, 6, 7, 12, 14) mod 21
93*F*	*R*152*a*: (22, 10, 5, 44); 0.84		*G*: (A1, *A*2, *A*3, *A*6, *B*9) (*A*1, *A*3, *A*8, *B*2, *B*10); perm *A*, *B* and mod 11
94*	*R*153: (24, 5, 5, 24); 0.83	-	*G*: (1, 2, 5, 10, 12) mod 24
95*MD*	*R*154: (24, 10, 5, 48); 0.83	2× *C*54; 0.83	*G*: (1, 2, 5, 10, 12); (1, 2, 4, 12, 21) mod 24
96	*SR*60: (25, 5, 5, 25); 0.83	*C*62; 0.83	*G*: (1, 6, 11, 16, 21); (1, 7, 13, 19, 25); (1, 10, 14, 18, 22); (1, 9, 12, 20, 23); (1, 8, 15, 17, 24); 1↔5, 6↔10, 11↔15, 16↔20, 21 ↔25 (*PC*)
97	*H*11: (25, 6, 5, 30); 0.83	-	Add the blocks: (1+5*x*, 2+5*x*, 3+5*x*, 4+5*x*, 5+5*x*); 0≤ *x* ≤4 to the solution in Serial No. 96
98*	*R*159: (35, 10, 5, 70); 0.82	-	*G*: (2, 5, 6, 11, 21); (7, 10, 11, 16, 26); (12, 15, 16, 21, 31); (1, 17, 20, 21, 26); (6, 22, 25, 26, 31); (1, 11, 27, 30, 31); (1, 6, 16, 32, 35); (3, 4, 6, 11, 21); (8, 9, 11, 16, 26); (13, 14, 16, 21, 31); (1, 18, 19, 21, 26); (6, 23, 24, 26, 31); (1, 11, 28, 29, 31); (1, 6, 16, 33, 34); 1↔5, 6↔10, 11↔15 16↔20, 21↔25, 26↔30, 31↔35 (*PC*)
99*	*R*160: (39, 10, 5, 78); 0.82	-	*G*: (2, 4, 10, 14, 27); (1, 15, 17, 23, 27); (1, 14, 28, 30, 36); (1, 14, 29, 32, 33); (1, 16, 19, 20, 27); (3, 6, 7, 14, 27); 1↔13, 14↔26, 27↔39 (*PC*)
100	*H*42: (41, 10, 5, 82); 0.82	-	*B*: (1, 10, 16, 18, 37); (5, 8, 9, 21, 39) mod 41
101*	*R*166: (10, 6, 6, 10); 0.90	-	*G*: (1, 2, 3, 5, 7, 9) mod 10
102*F*	*R*167*a*: (12, 9, 6, 18); 0.91	3× *D*5; 0.89	*G*: (*A*1, *A*2, *A*4, *A*6, *B*2, *B*3); (*B*1, *B*2, *B*4, *B*6, *A*2, *A*3); (*A*1, *A*2, *A*4, *B*1, *B*2, *B*4) mod 6
103*	*C*23: (13, 6, 6, 13); 0.90	-	*C*: (1, 2, 4, 7, 9, 13) mod 13
104*	*R*168: (15, 6, 6, 15); 0.82	-	*G*: (1, 2, 4, 7, 10, 13) mod 15
105	*H*10: (16, 6, 6, 16); 0.89	-	*B*: (1, 0, 0, 0); (0, 1, 0, 0); (0, 0, 1, 0); (0, 0, 0, 1); (1, 1, 0, 0); (0, 0, 1, 1) mod (2, 2, 2, 2)
106	*SR*72: (18, 6, 6, 18); 0.90	*C*14; 0.88	*G*: (1, 4, 7, 10, 13, 16); (1, 4, 8, 11, 15, 18); (1, 6, 8, 12, 14, 16); (1, 6, 9, 11, 13, 17); (1, 5, 7, 12, 15, 17); (1, 5, 9, 10, 14, 18); 1↔3, 4↔6, 7↔9, 10↔12, 13↔15, 16↔18,19↔21 (*PC*)
107*MD*	*R*170: (27, 6, 6, 27); 0.86	*C*77; 0.86	*G*: (0, 9, 12, 13, 16, 18) mod 27
108*MD*	*R*171: (28, 6, 6, 28); 0.86	*C*85; 0.86	*G*: (0, 1, 4, 15, 20, 22) mod 28
109	*SR*76: (30, 10, 6, 50);	5× *D*59; 0.55	By deleting the treatments 31, 32, 33, 34, 35

	0.86		from *SR*86*a*
110	*H*12: (31, 6, 6, 31); 0.86	-	*B*: (1, 5, 11, 24, 25, 27) mod 31
111	*SR*77: (42, 7, 6, 49); 0.85	-	By deleting the treatments 43, 44, 45, 46, 47, 48, 49 from *SR*87
112*	*LS*82: (49, 6, 6, 49); 0.84	-	*L*: (9, 19, 28, 32, 38, 43); (2, 13, 19, 24, 39, 49); (7, 9, 20, 26, 31, 49); (7, 15, 25, 34, 38, 44);(2, 12, 21, 25, 31, 36); (7, 13, 18, 33, 36, 45);(3, 8, 23, 33, 42, 46); 1↔7, 8↔14, 15↔21, 22↔28, 23↔29, 29↔35, 36↔42, 43↔49 (PC)
113*	*R*172: (9, 7, 7, 9); 0.96	-	*G*: (1, 2, 3, 5, 6, 8, 9) mod 9
114*	*R*173: (12, 7, 7, 12); 0.90	-	*G*: (1, 2, 3, 5, 7, 9, 11) mod 12
115*	*R*174: (12, 7, 7, 12); 0.92	-	*G*: (1, 2, 4, 5, 7, 8, 11) mod 12
116*	*R*175: (12, 7, 7, 12); 0.93	-	*G*: (1, 2, 3, 4, 6, 7, 11) mod 12
117*	*R*176: (12, 7, 7, 12); 0.93	-	*G*: (1, 4, 5, 6, 7, 8, 11) mod 12
118*MD*	*R*177: (14, 7, 7, 14); 0.92	-	*G*: (0, 1, 2, 5, 7, 8, 12) mod 14
119	*H*16: (15, 7, 7, 15); 0.92	*B*188; 0.92	*B*: (0, 1, 2, 4, 5, 8, 10) mod 15
120*	*LS*83: (16, 7, 7, 16); 0.91	*C*16; 0.92	*L*: (4, 8, 12, 13, 14, 15, 16); (4, 8, 9, 10, 11, 12, 16); (4, 5, 6, 7, 8, 12, 16); (1, 2, 3, 4, 8, 12, 16); 1↔4, 5↔8, 9↔12, 13↔16 (PC)
121*	*R*178: (18, 7, 7, 18); 0.82	*C*15; 0.90	*G*: (1, 2, 4, 7, 10, 13, 16) mod 18 *J*: (1, 2, 3, 4, 6, 9, 13) mod 18
122*MD*	*R*179: (20, 7, 7, 20); 0.90	*C*29; 0.90	*G*: (0, 1, 2, 4, 8, 11, 16) mod 20
123*DN*	*R*180*a*: (21, 7, 7, 21); 0.90	*C*36; 0.90	*G*: (1, 2, 5, 7, 11, 12, 14) mod 24
124*F*	*R*180*b*: (24, 7, 7, 24); 0.89	*C*56; 0.89	*G*: (*A*1, *A*2, *A*4, *A*5, *B*6, *B*8, *C*7) perm *A*, *B*, *C* and mod 8
125*	*C*25: (29, 7, 7, 29); 0.88	-	*C*: (1, 7, 16, 20, 23, 24, 25) mod 29
126*MD*	*R*182: (33, 7, 7, 33); 0.88	-	*G*: (2, 4, 5, 6, 10, 12, 23); (1, 13, 15, 16, 17, 21, 23); (1, 12, 24, 26, 27, 28, 32); 1↔11, 12↔22, 23↔33 (*PC*)
127*F*	*R*182*a*: (35, 7, 7, 35); 0.87	-	*G*: (*A*1, *A*2, *A*4, *B*7, *C*7, *D*7, *E*7) perm *A*, *B*, *C*, *D*, *E* and mod 7
128	*SR*86*a*: (35, 10, 7, 50); 0.88	-	By deleting the treatments 36, 37, 38, 39, 40 from *SR*95*a*
129*	*R*183: (48, 7, 7, 48); 0.87	-	*G*: (1, 2, 5, 11, 31, 36, 38) mod48
130	*SR*87: (49, 7, 7, 49); 0.87	-	*G*: (1, 8, 15, 22, 29, 36, 43); (1, 9, 17, 25, 33, 41, 49); (1, 14, 20, 26, 32, 38, 44); (1, 13, 18, 23, 35, 40, 45); (1, 12, 16, 27, 31, 42, 46); (1, 11, 21, 24, 34, 37, 47); (1, 10, 19, 28, 30, 39, 48); 1↔7, 8↔14, 15↔21, 22↔28, 29↔35, 36↔42, 43 ↔49 (*PC*)
131	*H*24: (49, 8, 7, 56); 0.87	-	Add the blocks: (1+7*x*, 2+7*x*, 3+7*x*, 4+7*x*, 5+7*x*, 6+7*x*, 7+7*x*); 0≤ *x* ≤6 to the solution in Serial No. 130
132*	*R*186: (12, 8, 8, 12); 0.95	2× *D*7; 0.95	*G*: (1, 3, 4, 5, 6, 7, 10, 11) mod 12
133*	*R*187: (14, 8, 8, 14); 0.90	2× *D*11; 0.94	*G*: (1, 2, 3, 5, 7, 9, 11, 13) mod 14 *J*: 2 copies of (1, 2, 3, 5, 8, 9, 10, 12) mod 14
134*	*C*26: (17, 8, 8, 17); 0.93		*C*: (1, 2, 4, 8, 9, 13, 15, 16) mod 17

135*	*R*188: (21, 8, 8, 21); 0.82	*C*37; 0.92	*G*: (1, 3, 6, 9, 12, 15, 18, 21) mod 21 *J*: (1, 2, 3, 5, 6, 9, 15, 17) mod 21
136	*R*189: (24, 8, 8, 24); 0.91	*C*57, 0.91	*G*: (2, 3, 4, 5, 6, 7, 13, 19); (1, 8, 9, 10, 11, 12, 13, 19); (1, 7, 14, 15, 16, 17, 18, 19); (1, 7, 13, 20, 21, 22, 23, 24); 1↔6, 7↔12, 13↔18, 19 ↔24 (*PC*)
137*	*LS*101: (25, 8, 8, 25); 0.91	*C*65; 0.91	*L*: (1, 6, 11, 16, 22, 23, 24, 25); (1, 6, 11, 17, 18, 19, 20, 21); (1, 6, 12, 13, 14, 15, 16, 21); (1, 7, 8, 9, 10, 11, 16, 21); (2, 3, 4, 5, 6, 11, 16, 21); 1↔5, 6↔10, 11↔15 16↔20, 21↔25 (*PC*)
138*	*C*27: (29, 8, 8, 29); 0.90	-	*C*: (1, 2, 8, 17, 21, 24, 25, 26) mod 29
139	*SR*95: (32, 8, 8, 32); 0.90	2× *D*66; 0.88	*G*: (1, 5, 9, 13, 17, 21, 25, 29); (1, 8, 11, 13, 18, 23, 26, 32); (1, 7, 9, 14, 19, 22, 28, 32); (1, 5, 10, 15, 18, 24, 28, 31); (1, 6, 11, 14, 20, 24, 27, 29); (1, 7, 10, 16, 20, 23, 25, 30); (1, 6, 12, 16, 19, 21, 26, 31); (1, 8, 12, 15, 17, 22, 27, 30); 1↔4, 5↔8, 9↔12, 13↔16, 17↔20, 21↔24, 25↔28, 29↔32 (*PC*)
140	*SR*95*a*: (40, 10, 8, 50); 0.89	5× *D*84; 0.62	By deleting the treatments 41, 42, 43, 44, 45 from *SR*103*a*
141*F*	*R*189*a*: (42, 8, 8, 42); 0.88	2× *D*89; 0.89	*G*: (*A*1, *A*2, *A*4, *B*7, *C*7, *D*7, *E*7, *F*7) perm *A*, *B*, *C*, *D*, *E*, *F* and mod 7
142	*H*25: (57, 8, 8, 57); 0.89	-	*B*: (1, 6, 7, 9, 19, 38, 42, 49) mod 57
143*	*R*191: (63, 8, 8, 63); 0.89	-	*G*: (1, 6, 8, 14, 38, 48, 49, 52) mod 63
144*	*R*193: (12, 9, 9, 12); 0.97	3× *D*8; 0.97	*G*: (1, 2, 3, 5, 6, 8, 9, 11, 12) mod 12
145*	*R*194: (15, 9, 9, 15); 0.94	3× *D*14; 0.94	*G*: (1, 2, 4, 5, 7, 8, 11, 13, 14) mod 15
146*	*R*195: (16, 9, 9, 16); 0.90	-	*G*: (1, 2, 4, 6, 8, 10, 12, 14, 16) mod 16
147	*H*30: (19, 9, 9, 19); 0.94	*C*24; 0.94	*B*: (1, 4, 5, 6, 7, 9, 11, 16, 17) mod 19
148*DN*	*R*197*a*: (20, 9, 9, 20); 0.93	*C*30; 0.93	*G*: (1, 2, 3, 4, 6, 10, 15, 17, 18) mod 20
149*	*R*198: (24, 9, 9, 24); 0.82	*C*58; 0.93	*G*: (1, 2, 4, 7, 10, 13, 16, 19, 22) mod 24 *J*: (1, 2, 3, 4, 7, 12, 15, 19, 21) mod 24
150*	*LS*117: (25, 9, 9, 25); 0.92	-	*L*: (1, 2, 3, 4, 5, 6, 11, 16, 21); (1, 6, 7, 8, 9, 10, 11, 16, 21); (1, 6, 11, 12, 13, 14, 15, 16, 21); (1, 6, 11, 16, 17, 18, 19, 20, 21); (1, 6, 11, 16, 21, 22, 23, 24, 25); 1↔5, 6↔10, 11↔15 16↔20, 21↔25 (PC)
151	*SR*102: (27, 9, 9, 27); 0.92	3× *D*52; 0.89	*G*: (1, 4, 7, 10, 13, 16, 19, 22, 25); (1, 6, 8, 10, 15, 17, 19, 24, 26); (1, 5, 9, 10, 14, 18, 19, 23, 27); (1, 4, 7, 12, 15, 18, 20, 23, 26); (1, 6, 8, 12, 14, 16, 20, 22, 27); (1, 5, 9, 12, 13, 17, 20, 24, 25); (1, 4, 7, 11, 14, 17, 21, 24, 27); (1, 6, 8, 11, 13, 18, 21, 23, 25); (1, 5, 9, 11, 15, 16, 21, 22, 26); 1↔3, 4↔6, 7↔9, 10↔12, 13↔15 16↔18, 19↔21, 22↔24, 25↔27 (*PC*)
152	*R*200: (28, 9, 9, 28); 0.91	*C*87; 0.92	*G*: (2, 3, 4, 5, 6, 7, 8, 15, 22); (1, 9, 10, 11, 12, 13, 14, 15, 22);

			(1, 8, 16, 17, 18, 19, 20, 21, 22); (1, 8, 15, 23, 24, 25, 26, 27, 28) 1↔7, 8↔14, 15↔21, 22 ↔28 (PC)
153	*H*34: (37, 9, 9, 37); 0.91	-	*B*: (1, 7, 9, 10, 12, 16, 26, 33, 34) mod 37
154^{DN}	*R*200*c*: (40, 9, 9, 40); 0.91	-	*G*: (1, 3, 4, 6, 10, 17, 18, 22, 35) mod 40
155	*SR*103*a*: (45, 10, 9, 50), 0.91	-	By deleting the treatments 46, 47, 48, 49, 50 from *SR*109*a*
156^{F}	*R*200*e*: (49, 9, 9, 49); 0.89	-	*G*: (*A*1, *A*2, *A*4, *B*7, *C*7, *D*7, *E*7, *F*7, *G*7); perm *A*, *B*, *C*, *D*, *E*, *F*, *G* and mod 7
157	*H*37: (73, 9, 9, 73); 0.90	-	*B*: (1, 2, 4, 8, 16, 32, 37, 55, 64) mod 73
158*	*R*202: (80, 9, 9, 80); 0.90	-	*G*: (1, 3, 6, 10, 22, 44, 57, 58, 75) mod 80
159*	*LS*134: (100, 9, 9, 100); 0.89	-	*L*: (63, 95, 59, 11, 42, 78, 87, 24, 36); (90, 29, 77, 43, 51, 8, 34, 92, 15); (37, 85, 16, 51, 69, 23, 42, 10, 98); (4, 62, 47, 21, 99, 13, 78, 85, 60); (93, 18, 31, 6, 77, 60, 24, 69, 45); (55, 39, 21, 68, 86, 93, 7, 12, 80); (72, 7, 100, 84, 11, 35, 69, 43, 26); (1, 26, 49, 68, 77, 32, 85, 14, 53); (16, 57, 84, 32, 8, 45, 99, 80, 63); (47, 74, 6, 98, 22, 70, 53, 35, 89); 1↔10, 11↔20, 21↔30, …, 91↔100
160*	*R*203: (12, 10, 10, 12); 0.98	-	*G*: (1, 2, 3, 4, 6, 7, 8, 10, 11, 12) mod 12
161*	*R*204: (14, 10, 10, 14); 0.97	2× *D*12; 0.97	*G*: (1, 2, 3, 4, 6, 7, 8, 10, 12, 14) mod 14
162^{MD}	*R*205: (14, 10, 10, 14); 0.97	2× *D*12; 0.97	*G*: (0, 1, 3, 5, 6, 7, 8, 9, 10, 11) mod 14
163*	*R*206: (18, 10, 10, 18); 0.90	2× *D*26; 0.95	*G*: (1, 2, 4, 6, 8, 10, 12, 14, 16, 18) mod 18 *J*: 2 copies of (1, 2, 3, 4, 7, 10, 11, 12, 13, 16) mod 18
164^{F}	*R*206*a*: (21, 10, 10, 21); 0.94	*C*39; 0.94	*G*: (*A*1, *A*2, *A*4, *A*7, *B*1, *B*2, *B*4, *C*1, *C*2, *C*4); perm *A*, *B*, *C* and mod 7
165^{MD}	*R*206*b*: (21, 10, 10, 21); 0.93	*C*39; 0.94	*G*: (0, 1, 3, 4, 6, 9, 10, 12, 15, 18) mod 21
166*	*R*207: (27, 10, 10, 27); 0.82	*C*81; 0.93	*G*: (1, 2, 4, 7, 10, 13, 16, 19, 22, 25) mod 27 *J*: (1, 2, 3, 4, 5, 8, 13, 17, 21, 23) mod 27
167^{F}	*R*207*a*: (28, 10, 10, 28); 0.93	*C*88; 0.93	*G*: (*A*1, *A*2, *A*4, *B*1, *B*2, *B*4, *C*1, *C*2, *C*4, *D*7); perm *A*, *B*, *C*, *D* and mod 7
168	*R*208: (32, 10, 10, 32); 0.92	2× *D*67; 0.92	*G*: (2, 3, 4, 5, 6, 7, 8, 9, 17, 25); (1, 10, 11, 12, 13, 14, 15, 16, 17, 25); (1, 9, 18, 19, 20, 21, 22, 23, 24, 25); (1, 9, 17, 26, 27, 28, 29, 30, 31, 32) 1↔8, 9↔16, 17↔24, 25 ↔32 (*PC*)
169*	*LS*136: (36, 10, 10, 36); 0.92	2× *D*79; 0.92	*L*: (2, 3, 4, 5, 6, 7, 13, 19, 25, 31); (1, 8, 9, 10, 11, 12, 13, 19, 25, 31); (1, 7, 14, 15, 16, 17, 18, 19, 25, 31); (1, 7, 13, 20, 21, 22, 23, 24, 25, 31); (1, 7, 13, 19, 26, 27, 28, 29, 30, 31); (1, 7, 13, 19, 25, 32, 33, 34, 35, 36); 1↔6, 7↔12, 13↔18, 19↔24, 25↔30, 31↔36

170^{A}	*C*30: (37, 10, 10, 37); 0.92	-	*C*: (0, 1, 16, 34, 26, 9, 33, 10, 12, 7) mod 37
171	*SR*109*a*: (50, 10, 10, 50); 0.92	5×*D*108;0.67	*G*: (1, 10, 12, 17, 25, 26, 33, 39, 44, 48);(1, 9, 15, 19, 21, 28, 32, 40, 42, 48); (1, 8, 13, 16, 22, 29, 35, 40, 44, 47); (1, 7, 11, 18, 23, 29, 32, 39, 45, 50); (1, 6, 14, 20, 24, 28, 33, 37, 45, 47); (1, 6, 11, 16, 21, 26, 31, 36, 41, 46); (1, 8, 15, 17, 24, 27, 34, 36, 43, 50); (1, 10, 14, 18, 22, 30, 34, 38, 42, 46); (1, 7, 13, 19, 25, 30, 31, 37, 43, 49); (1, 9, 12, 20, 23, 27, 35, 38, 41, 49) 1↔5, 6↔10, 11↔15, 16↔20, 21↔25, 26↔30, 31↔35, 36↔40, 41↔45, 46↔50 (PC)
172^{F}	*R*208*a*: (56, 10, 10, 56); 0.89	2× *D*125; 0.9	*G*: (*A*1, *A*2, *A*4, *B*7, *C*7, *D*7, *E*7, *F*7, *G*7, *H*7); perm *A*, *B*, *C*, *D*, *E*, *F*, *G*, *H* and mod 7
173	*H*46: (91, 10, 10, 91); 0.91	-	*B*: (0, 1, 3, 9, 27, 49, 56, 61, 77, 81) mod 91

*The cyclic solutions are reported in Clatworthy (1973). John numbers are from John *et al.* (1972). A part cycle, such as $\frac{1}{2}$(*B*1, *B*2, *B*4, *B*5) for *R*109*a*, means that only half the six blocks are needed, since the same treatments would then recur [see Freeman (1976)]. *m* × No. *X* denotes design obtained by taking *m* copies of the design No. *X*.

PC: the initial blocks are developed in partial cycles, *B*: BIBD, *G*: Group divisible design, *C*: the cyclic design from Clatworthy and Agrawal (1987), *L*: Latin square type designs; *J*: the cyclic design from John *et al.* (1972).

HX numbers are from Hall (1998) and *SRX*, *RX*, *LSX* and *CX* (in the second column of the table) numbers are from Clatworthy (1973) and Agrawal (1987).

The abbreviations *A*, *M*, *F*, *MD*, *S*, *DN* and *DB* stand for Agrawal (1987), Mukerjee *et al.* (1987), Freeman (1976), Midha and Dey (1995), Sinha (1989), Dey and Nigam (1985) and Dey and Balasubramanian (1991) respectively.

The overall efficiency of a partially balanced design is defined as the ratio of the average variance of a treatment comparison to the variance in a randomized block experiment with the same replication, assuming that the standard errors of individual plots are the same. The overall efficiency of a BIB design is obtained using $\lambda v/rk$ and the overall efficiency *E* of a two associate class PBIB design is calculated as [see Clatworthy (1973)]:

$$E = \frac{(k-1)(v-1)}{n_1(k-c_1)+n_2(k-c_2)}$$

where the computational constants c_1, c_2 are obtained by means of the following relations:

$k^2 \Delta = (rk - r + \lambda_1)(rk - r + \lambda_2) + (\lambda_1 - \lambda_2)\{(r(k-1)(p_{12}^1 - p_{12}^2) + \lambda_2 p_{12}^1 - \lambda_1 p_{12}^2\}$,
$k \Delta c_1 = \lambda_1(rk - r + \lambda_2) + (\lambda_1 - \lambda_2)(\lambda_2 p_{12}^1 - \lambda_1 p_{12}^2)$
$k \Delta c_2 = \lambda_2(rk - r + \lambda_1) + (\lambda_1 - \lambda_2)(\lambda_2 p_{12}^1 - \lambda_1 p_{12}^2)$.
In case of GD designs, the expression for overall efficiency is given as [see Freeman (1976)]:

$$E = \frac{v(v-1)\lambda_2\{\lambda_1 + (m-1)\lambda_2\}}{rk\{(m-1)\lambda_1 + (mv-2m+1)\lambda_2\}}.$$

The cyclic solutions of BIB designs: *H*2, *H*11, *H*24 and GD designs: *SR*60, *SR*72, *SR*87, *SR*95, *SR*102, *R*189, *R*200, *R*208 are new. Clatworthy (1973) reported six, seven and eight initial blocks for the cyclic solution of *R*189, *R*200 and *R*208 respectively whereas we have used four initial blocks only. For the design *R*68: the first three initial blocks give 9 blocks each and the fourth initial block gives three distinct blocks. Clatworthy (1973) did not report the solutions in cyclic form for Latin square designs except four. We have reported cyclic solutions for such designs. The cyclic block designs which are *m* – multiple of smaller block designs and are with non – repeated initial blocks are included in the above table. For each of these designs, Clatworthy (1973) reported a solution that is obtained by repeating the blocks of a smaller block design *m* times. A resolvable solution of *SR*109*a* may be found in Saurabh and Sinha (2022).

The GD scheme for the cyclic semi – regular GD designs: *SR*60, *SR*72, *SR*87, *SR*95, *SR*102, *SR*109*a* and cyclic regular GD designs: *R*189, *R*200, *R*208 is given as:

$$\begin{array}{ccccc} 1 & 2 & 3 & \cdots & n \\ n+1 & n+2 & n+3 & \cdots & 2n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (m-1)n+1 & (m-1)n+2 & (m-1)n+3 & \cdots & mn \end{array}$$

for suitable choices of *m* and *n*.

Acknowledgement

The authors are thankful to the Editor–in–Chief and an anonymous referee for nice comments leading to improvement in the readability of the paper.

References

Agrawal, H. C. (1987). Construction of two–class cyclic partially balanced incomplete block designs. *Journal of Statistical Planning and Inference*, **16**, 127–132.

Bailey, R. A. (1990). *Cyclic Designs and Factorial Designs*. Proceedings of the R.C. Bose symposium on Probability, Statistics and Design of Experiments, New Delhi, 27–30 December,1988, Wiley Eastern, 51–74.

Clatworthy, W. H. (1973). *Tables of Two–Associate–Class Partially Balanced Designs*. National Bureau of Standards (U.S.), Applied Mathematics, Series **63**.

Dean, A. M. and Lewis, S. M. (1980). A unified approach to generalized cyclic designs. *Journal of Statistical Planning and Inference*, **4**, 13–23.

Dey, A. (1986). *Theory of Block Designs*. Wiley Eastern, New Delhi.

Dey, A. (2010). *Incomplete Block Designs*. Hindustan Book Agency, New Delhi.

Dey, A. and Balasubramanian, K. (1991). Construction of some families of group divisible designs. *Utilitas Mathematica*, **40**, 283–290.

Dey, A. and Nigam, A. K. (1985). Constructions of group divisible designs. *Journal of the Indian Society of Agricultural Statistics*, **37**, 163–166.

Freeman, G. H. (1976). A cyclic method of constructing regular group divisible incomplete block designs. *Biometrika*, **63**, 555–558.

Hall, Marshal Jr. (1998). *Combinatorial Theory*. John Wiley, New York.

Jarrett, R. G. and Hall, W. B. (1978). Generalized cyclic incomplete block designs. *Biometrika*, **65**, 397–401.

John, J. A., Wolock, F. W. and David, H. A. (1972). *Cyclic Designs*. Washington D. C.: National Bureau of Standards. Appl. Math. Ser. No. **62**.

John, J. A. and Williams, E. R. (1995). *Cyclic and Computer-Generated Designs*. Chapman and Hall, London.

Lamacraft, R. R. and Hall, W. B. (1982). Tables of cyclic incomplete block designs: $r = k$. *Australian Journal of Statistics*, **24**, 350–360.

Midha, C. K. and Dey, A. (1995). Cyclic group divisible designs. *Calcutta Statistical Association Bulletin*, **45**, 179–180.

Mukerjee, R., Jimbo, M. and Kageyama, S. (1987). On cyclic semi–regular group divisible designs. *Osaka Journal of Mathematics*, **24**, 395–407.

Nigam, A. K., Puri, P. D. and Gupta, V. K. (1988). *Characterizations and Analysis of Block Designs*. Wiley Eastern Limited, New Delhi.

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

Raghavarao, D. and Padgett, Lakshmi V. (2005). Block designs. Analysis, combinatorics and applications. *Series on Applied Mathematics*, **17**, World Scientific, Singapore.

Saurabh, S. and Sinha, K. (2022). Some new resolvable group divisible designs. *Communications in Statistics – Theory and Methods*, **51**(**13**), 4509-4514. doi.org/10.1080/ 03610926. 2020. 1817487.

Sinha, K. (1989). A method of construction of PBIB designs. *Journal of the Indian Society of Agricultural Statistics*, **XLI**, 313–315.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 135-146

Robustness of Bayes Estimation of Coefficient of Variation for Normal Distribution for a Class of Moderately Non-Gamma Prior Distributions

Priyanka Aggarwal[1] and Samridhi Mehta[2]
[1]*Department of Statistics, Hindu College, University of Delhi*
[2]*Department of Mathematics, Hindu College, University of Delhi*

Received: 03 April 2021; Revised: 27 September 2021; Accepted: 03 October 2021

Abstract

In this paper, we propose to examine sensitivity of the Bayes estimate of normal coefficient of variation to a moderately non-gamma prior distribution of the unknown precision. Non-negativity and unimodality region of the considered K-prior distributions are computed for illustration purpose. Kullback-Leibler Divergence measure is employed to study the effect as the K-prior becomes much different from the conjugate gamma prior.

Key words: Positive and unimodal region; Kullback-Leibler divergence; Bayes estimate; Coefficient of variation; K-prior; MELO approach.

AMS Subject Classification Code: 62F15, 62F10

1. Introduction

The concept of coefficient of variation (CV) has been intriguing researchers for many years because of its use in assessing the variability of a series since it is independent of the unit of measurement. It has applications in various areas ranging from medical sciences to finance. Here, we study Bayesian estimation of CV for Normal distribution, with mean and precision both unknown, using Zellner's Minimum Posterior Expected Loss (MELO) approach. Zellner (1978) addressed the problem of estimating the reciprocals and the ratios of the population mean and the regression coefficients. He pointed out the situations in which maximum likelihood and other estimators of these problems do not possess finite moments and have infinite risk relative to quadratic and other loss functions, whereas MELO estimators using relative squared error loss function (RSELF) have finite moments and risk, and are hence, admissible.

In Bayes estimation for normal distribution with unknown precision, a conjugate gamma prior is used to obtain the posterior distribution. However, subjectivity involved in choosing a single prior distribution, as observed by Berger (1984), has drawn severe criticism of Bayesian methodology. A reasonable approach is to consider a family of plausible priors that are in the neighbourhood of a specific assessed approximation to the 'true' prior. Not much attention has been paid by the investigators to study the problem of sensitivity to a possible misspecification of the gamma distribution as the conjugate prior distribution in Bayesian analysis. In this paper, we follow Bansal and Singh (1999) and Aggarwal and

Corresponding Author: Samridhi Mehta
E-mail: mehta.samridhi@gmail.com

Bansal (2017) to use Khamis' class (K-class) of moderately non-gamma prior distributions for the unknown precision of the normal distribution and study the robustness of the Bayes estimate with respect to the prior.

Many researchers including Barton and Dennis (1952), and Draper and Tierney (1972) exhibited the importance of deriving the conditions under which Gram-Charlier and Edgeworth curves are positive definite and unimodal. Spiring (2011) determined the regions where Edgeworth expansion and Gram-Charlier series upto the 4th moment is positive and unimodal. Till now, no attempt has been made in this direction for K-class of moderately non-gamma densities. In this paper, the boundaries of positive and unimodal regions are obtained for K-class of moderately non-gamma densities. The corresponding plot of the region is also displayed.

In Section 2, Bayes estimate of the CV of the normal distribution using MELO approach is derived. In Section 3, we discuss the positive definite and unimodal region for K-class of non-gamma densities. In Section 4, the distance between gamma density and some non-gamma densities are computed using KLD for arbitrarily chosen values of parameters. The derived results are further illustrated using hypothetical data in Section 5.

2. Bayes Estimate of Coefficient of Variation of the Normal Distribution

In this Section, the Bayes estimate of Coefficient of Variation (CV) using MELO approach is obtained for Normal distribution with mean and precision both unknown. The conditional normal prior for unknown mean and K-prior for the unknown precision of the normal distribution are used. The posterior distribution is derived below which shall be further used to obtain Bayes estimate of CV.

2.1. Likelihood function

Let us suppose that $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from $N(\theta, r)$ with mean θ and precision r, both unknown. The likelihood function of θ and r, given observed sample $\mathbf{X} = \mathbf{x}$, is

$$\ell(\theta, r|\mathbf{x}) = \left(\frac{r}{2\pi}\right)^{\frac{n}{2}} exp\left(-\frac{r}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right)$$

$$= \left(\frac{r}{2\pi}\right)^{\frac{n}{2}} exp\left(-\frac{r}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 - \frac{nr}{2}(\bar{x} - \theta)^2\right); \theta \in (-\infty, \infty), r > 0. \quad (1)$$

2.2. Prior distributions

2.2.1. Conditional normal prior for unknown mean

The prior distribution of unknown mean θ, given r, is $N(\mu, \tau r)$, both μ and τ known, given by

$$g(\theta|r) = \sqrt{\frac{\tau r}{2\pi}} exp\left(-\frac{\tau r}{2}(\theta - \mu)^2\right); -\infty < (\theta, \mu) < \infty, (r, \tau) > 0.$$

2.2.2. Khamis' class of moderately non-gamma distributions as a prior for unknown precision (K-prior)

To study the sensitivity of Bayes estimator with respect to the prior when the 'true' prior is not the conventional natural conjugate gamma prior, we consider a class of K-prior for the unknown precision of the normal distribution. Khamis (1960), in his pioneering work, obtained a class of non-gamma densities using Laguerre expansion with Gamma function as the weight function. The application of such series expansion was discussed in Tiku and Tan (1999). Recently, Aggarwal and Bansal (2017) used K-class of moderately non-gamma distributions as a prior (K-prior) for the unknown mean of the Poisson regression super population model.

Consider density $h(r)$ (may be unknown) with first k moments about origin known for $r \in (0, \infty)$ and the Laguerre expansion

$$h_m(r) = \sum_{j=0}^{m} C_j L_j(r) p(r|\alpha, \beta) \text{ with } m \leq k$$

where $p(r|\alpha, \beta) = Gamma(\alpha, \beta)$, and

$$L_j(r) = \sum_{i=0}^{j} (-1)^i \binom{j}{i} \frac{\Gamma(\alpha+j)}{\Gamma(\alpha+j-i)} (\beta r)^{j-i} \quad , L_0(r) = 1, j = 1,2, \dots, m$$

is the Laguerre polynomial of degree j and C_j are arbitrary constants. Using Khamis (1960)'s expression for $C_j = \frac{\int_0^\infty L_j(r) h(r) dr}{\int_0^\infty \left(L_j(r)\right)^2 h(r) dr}, j = 0,1, \dots, m$, the expansion $h_m(r)$ can be used to approximate $h(r)$ for appropriate values of α. Bansal and Singh (1999) considered a particular case of Khamis' class of non-gamma distributions wherein only the first four moments $(m = 4)$ were used. This particular case was referred to as K-class of moderately non-gamma densities, given by

$$h_4(r) \approx g(r) = K(r) p(r|\alpha, \beta), \; r, \alpha, \beta > 0 \tag{2}$$

with

$$K(r) = \left(1 + \frac{\delta_3 \sqrt{\alpha}}{6(\alpha+1)(\alpha+2)} \left(L_3(r) - \frac{3}{\alpha+3} L_4(r)\right) + \frac{\delta_4 \alpha}{24(\alpha+1)(\alpha+2)(\alpha+3)} L_4(r)\right).$$

The excess of skewness and kurtosis of K-class of non-gamma densities $g(r)$ over gamma density $p(r|\alpha, \beta)$ are measured by the parameters δ_3 and δ_4, respectively.

Remark 1: In particular, if we take $\alpha = 4, \beta = 1, \delta_3 = 0.15, \delta_4 = 2$, then skewness of gamma $p(r|\alpha, \beta) = \frac{4}{\sqrt{(\alpha)}} = 2$ and kurtosis of gamma $p(r|\alpha, \beta) = 3 + \frac{6}{\alpha} = 4.5$. Hence, skewness and kurtosis of K-prior $g(r)$ are 2.15 and 6.5, respectively.

2.3. Posterior distribution

The joint prior for θ and r is

$$g(\theta, r) = g(\theta|r) g(r)$$

where $g(\theta|r)$ is $N(\mu, \tau r)$, and $g(r)$ is K-prior given in (2).

Using Bayes Theorem, the posterior distribution of θ and r, given observed sample $\mathbf{X} = \mathbf{x}$, is

$$g(\theta, r|\mathbf{x}) = g(\theta|r, \mathbf{x}) g(r|\mathbf{x})$$

where

$$g(\theta|r,\mathbf{x}) = \left(\frac{(n+\tau)r}{2\pi}\right)^{n/2} exp\left(-\frac{r}{2}(n+\tau)(\theta-\mu^*)^2\right) \equiv N(\mu^*,(n+\tau)r), \quad (3)$$

and $$g(r|\mathbf{x}) = \left(\frac{K(r)g(r|\alpha^*,\beta^*,\mathbf{x})}{G(\delta_3,\delta_4)}\right), \quad (4)$$

with

$$\alpha^* = \alpha + \frac{n}{2}, \beta^* = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{1}{2}\frac{n\tau}{n+\tau}(\mu-\bar{x})^2, \mu^* = \frac{n\bar{x}+\tau\mu}{n+\tau},$$

$$G(\delta_3,\delta_4) = 1 - \delta_3\frac{\alpha^{\frac{3}{2}}}{6}C_1(\alpha^*) + \delta_4\frac{\alpha^2}{24}C_2(\alpha^*),$$

$$C_1(\alpha^*) = 3R_4 - 13R_3 + 21R_2 - 15R_1 + 4R_0,$$

$$C_2(\alpha^*) = R_4 - 4R_3 + 6R_2 - 4R_1 + R_0,$$

and

$$R_j = \left(\frac{\Gamma(\alpha^*+j)}{\Gamma(\alpha^*)\beta^{*j}}\right)\Big/\left(\frac{\Gamma(\alpha+j)}{\Gamma(\alpha)\beta^j}\right), j = 0,1,\dots,4.$$

(See Appendix A.1 for the details of derivation)

2.4. Bayes estimate using Zellner's MELO approach

Zellner (1978) pointed out that the usual Bayes estimate of the reciprocal of normal mean often fails to exist. He recommended MELO estimate as a solution to overcome the problem of non-existence. Following him, consider $\hat{a}$ as the estimate of CV $a = \sigma/\theta$, $\left(\sigma^2 = \frac{1}{r}\right)$. Upon minimizing posterior expected loss $E((\hat{a}\theta-\sigma)^2|\mathbf{x}) = E(\theta^2(\hat{a}-a)^2|\mathbf{x})$, the MELO estimate is given by

$$\hat{a}_{MELO} = \frac{E\left(\theta^2 a|\mathbf{x}\right)}{E\left(\theta^2|\mathbf{x}\right)} = \frac{E\left(\frac{\theta}{\sqrt{r}}\middle|\mathbf{x}\right)}{E\left(\theta^2|\mathbf{x}\right)} \quad (5)$$

where the expectations are with respect to posterior distribution and are given by

$$E(\theta^2|\mathbf{x}) = \mu^{*2} + \frac{\beta^*}{(\alpha^*-1)(\tau+n)}\frac{G_1(\delta_3,\delta_4)}{G(\delta_3,\delta_4)}, \quad (6)$$

$$E\left(\frac{\theta}{\sqrt{r}}\middle|\mathbf{x}\right) = \mu^*\frac{\sqrt{\beta^*}\Gamma\left(\alpha^*-\frac{1}{2}\right)}{\Gamma(\alpha^*)}\frac{G_2(\delta_3,\delta_4)}{G(\delta_3,\delta_4)}, \quad (7)$$

with

$$G_1(\delta_3,\delta_4) = 1 - \delta_3\frac{\alpha^{\frac{3}{2}}}{6}C_1(\alpha^*-1) + \delta_4\frac{\alpha^2}{24}C_2(\alpha^*-1),$$

and

$$G_2(\delta_3,\delta_4) = 1 - \delta_3\frac{\alpha^{\frac{3}{2}}}{6}C_1\left(\alpha^*-\frac{1}{2}\right) + \delta_4\frac{\alpha^2}{24}C_2\left(\alpha^*-\frac{1}{2}\right).$$

(See Appendix A.2 for the details of derivation of the posterior expectations (6) and (7))

Remark 2: The value of $\hat{a}_{MELO}$ in (5) depends on the observed sample values.

Remark 3: If we consider gamma prior for r, that is $\delta_3 = \delta_4 = 0$, then the MELO estimate reduces to $\mu^*\frac{\sqrt{\beta^*}\Gamma\left(\alpha^*-\frac{1}{2}\right)}{\Gamma(\alpha^*)}\Big/\left(\mu^{*2} + \frac{\beta^*}{(\alpha^*-1)(\tau+n)}\right)$.

Remark 4: For non-informative prior, that is $g(\theta,r) \propto \frac{1}{r}$, the MELO estimate can be obtained by letting $\alpha \to -\frac{1}{2}, \beta \to 0, \tau \to 0$ (See De Groot (1970), page 195) and is given by

$$\frac{\bar{x}\sqrt{\frac{\sum(x_i-\bar{x})^2}{2}\frac{\Gamma\left(\frac{n-1}{2}-\frac{1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}}}{\bar{x}^2+\frac{\sum(x_i-\bar{x})^2}{2n\left(\frac{n-1}{2}-1\right)}} = \left(\sqrt{\frac{n}{2}}\frac{\Gamma\left(\frac{n}{2}-1\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right)\frac{\hat{\sigma}}{\bar{x}}\left(1+\frac{\hat{\sigma}^2}{\bar{x}^2}\frac{1}{n-3}\right)^{-1}, \tag{8}$$

where $\hat{\sigma}^2 = \frac{\sum(x_i-\bar{x})^2}{n}$. This conforms with the result obtained by Bansal (2007).

Remark 5: If we further use the result $\lim_{m\to\infty} m^{b-a}\frac{\Gamma(m+a)}{\Gamma(m+b)} = 1$, (see Abramowitz and Stegun (1964), formula 6.1.46, page 257), then on taking $m = \frac{n}{2}, a = -1$ and $b = -\frac{1}{2}$, the first factor on the right-hand side of (8) tends to one for large samples. Hence, it is seen that the MELO Bayes estimate of CV reduces to the product of the usual estimate, $\hat{\sigma}/\bar{x}$, of CV and the shrinkage factor $\left(1+\frac{\hat{\sigma}^2}{\bar{x}^2}\frac{1}{n-3}\right)^{-1}$ which has a value between zero and one. Thus, we may expect that the MELO Bayes estimate of CV to be smaller than the corresponding classical estimate for large samples and moderately non-gamma prior densities of the precision.

In the next Section, we obtain the regions in which $g(r)$ is non-negative and unimodal so that the above obtained results can be illustrated numerically using hypothetical data.

3. Positive Definite and Unimodal Region for Khamis' Class of Non-gamma Distributions

Figure 1 below exhibits the graphs of $g(r)$ for various combinations of δ_3 and δ_4 with $\alpha = 4, \beta = 1$. The Graph 1 of Figure 1 represents Gamma Distribution. Graphs 2, 3 and 4 of Figure 1 shows that the graphs change in shape and peakedness with change in δ_3 and δ_4. It may be noticed that there are combinations of δ_3 and δ_4 for which $g(r)$ is negative and multimodal. For example, for $(\delta_3, \delta_4) = (3, 4)$ and $(\delta_3, \delta_4) = (0.1, 15)$, $g(r)$ is negative and multimodal respectively as shown in Graph 5 and 6 of Figure 1 below. Thus, there is a need to obtain the regions in which $g(r)$ is non-negative and unimodal.

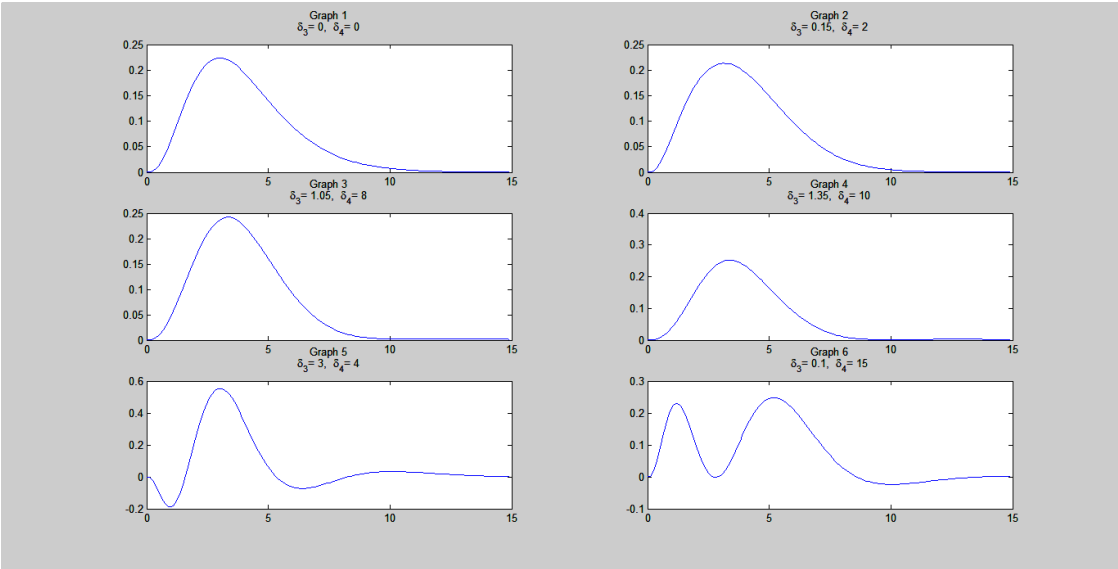


Figure 1: Graphs of $g(r)$ for various combinations of δ_3 and δ_4 with $\alpha = 4, \beta = 1$

We now determine the region where $g(r)$ is non-negative for a specific range of r. We have tabulated the combinations of (δ_3, δ_4) where $g(r)$ is non-negative for $\alpha = 4, \beta = 1$ and the boundaries of the positive regions are provided in Table 1. Figure 2 exhibits the plot of the boundary points in Table 1. This positive region is then checked for unimodality using second derivative test. It is found that the unimodality exists throughout in this positive region. For the region beyond the boundary values given in Table 1, $g(r)$ may be unimodal but not positive. Thus, such regions are not considered. It may be noted that we are providing regions only for $\alpha = 4, \beta = 1$. The entire work is done using Mathematica. The same procedure may be employed as discussed above, to obtain positive and unimodal regions for other choices of α and β.

Table 1: Positive and unimodal boundary points (δ_4, L < δ_3< U) for $g(r|4,1)$

δ_4	L	U
0	0	0.18
0.1	0	0.2
0.2	0	0.21
0.3	0	0.22
0.4	0	0.23
0.5	0	0.25
0.6	0	0.26
0.7	0	0.27
0.8	0	0.28
0.9	0	0.3
1	0	0.31
1.1	0	0.32
1.2	0	0.33
1.3	0	0.35
1.4	0	0.36
1.5	0	0.37
1.6	0	0.38
1.7	0	0.4
1.8	0	0.41
1.9	0	0.42
2	0	0.43
2.1	0.01	0.45
2.2	0.02	0.46
2.3	0.03	0.47
2.4	0.04	0.48
2.5	0.06	0.5
2.6	0.07	0.51
2.7	0.08	0.52
2.8	0.1	0.53
2.9	0.11	0.55
3	0.12	0.56
3.1	0.14	0.57
3.2	0.15	0.58
3.3	0.16	0.6
3.4	0.18	0.61
3.5	0.19	0.62
3.6	0.2	0.63
3.7	0.22	0.65
3.8	0.23	0.66
3.9	0.25	0.67
4	0.26	0.68
4.1	0.27	0.7
4.2	0.29	0.71
4.3	0.3	0.72
4.4	0.32	0.73
4.5	0.33	0.75
4.6	0.35	0.76
4.7	0.36	0.77
4.8	0.38	0.78
4.9	0.39	0.8
5	0.41	0.81
5.1	0.42	0.82
5.2	0.44	0.83
5.3	0.45	0.85
5.4	0.47	0.86
5.5	0.48	0.87
5.6	0.5	0.88
5.7	0.5	0.9
5.8	0.53	0.91
5.9	0.54	0.92
6	0.56	0.93
6.1	0.57	0.95
6.2	0.59	0.96
6.3	0.61	0.97
6.4	0.62	0.98
6.5	0.64	1
6.6	0.65	1.01
6.7	0.67	1.02
6.8	0.69	1.03
6.9	0.7	1.05
7	0.72	1.06
7.1	0.74	1.07
7.2	0.75	1.08
7.3	0.77	1.1
7.4	0.78	1.11
7.5	0.8	1.12
7.6	0.82	1.13
7.7	0.84	1.15
7.8	0.85	1.16
7.9	0.87	1.17
8	0.89	1.18
8.1	0.9	1.2
8.2	0.92	1.21
8.3	0.94	1.22
8.4	0.96	1.23
8.5	0.97	1.25
8.6	0.99	1.26
8.7	1.01	1.27
8.8	1.03	1.28
8.9	1.04	1.3
9	1.06	1.31
9.1	1.08	1.32
9.2	1.1	1.33
9.3	1.12	1.35
9.4	1.14	1.36
9.5	1.15	1.37
9.6	1.17	1.38
9.7	1.19	1.4
9.8	1.21	1.41
9.9	1.23	1.42
10	1.25	1.43
10.1	1.27	1.45
10.2	1.29	1.46
10.3	1.31	1.47

10.4	1.33	1.48
10.5	1.35	1.5
10.6	1.37	1.51
10.7	1.39	1.52
10.8	1.41	1.53
10.9	1.43	1.55

11	1.45	1.56
11.1	1.47	1.57
11.2	1.49	1.58
11.3	1.52	1.6
11.4	1.54	1.61
11.5	1.56	1.62

11.6	1.58	1.63
11.7	1.61	1.65
11.8	1.63	1.66
11.9	1.65	1.67
12	1.68	1.68
12.1	1.7	1.7

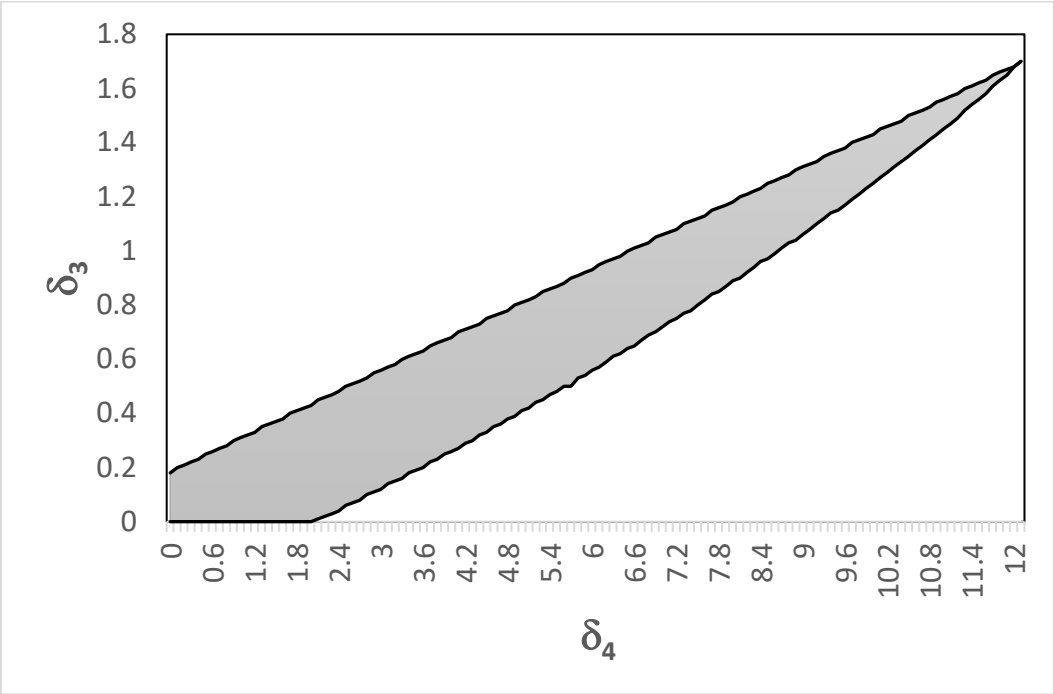


Figure 2: Plot of Positive and unimodal region for $g(r)$ $with$ $\alpha = 4, \beta = 1$

4. Quantitative Robustness using Kullback-Leibler Divergence (KLD) Measure

By virtue of the significance of Gamma distribution in problem of statistical estimation, it is deemed necessary to study the sensitivity of the estimates to its possible misspecification. In this direction, we make an effort to study the quantitative robustness employing Kullback-Leibler divergence (KLD) measure.

To examine quantitative robustness with respect to the K-class of moderately non-gamma densities $g(r)$, we compute its distance from gamma $p(r|\alpha,\beta)$ using KLD as

$$I(p,g) = \int_0^\infty \log\left(\frac{p(r|\alpha,\beta)}{g(r)}\right) p(r|\alpha,\beta)dr =$$

$$E\left(\log\left(\frac{p(r|\alpha,\beta)}{g(r)}\right)\right). \qquad (9)$$

The expectation is taken with respect to $p(r|\alpha,\beta)$. Observe that $I(p,g)$ is not a symmetric distance.

Aggarwal and Bansal (2010) used KLD to evaluate the distance between Normal and Edgeworth distributions for some selected values of $\lambda_3\ (=\delta_3)$ and $\lambda_4\ (=\delta_4)$ lying in region given by Barton and Dennis (1952). Aggarwal and Bansal (2017) computed $I(p,g)$ and it is found that there is an error in its computation. Thus, we extend the study on

quantitative robustness using corrected $I(p,g)$ while considering $KL_{min} = \min\{I(p,g), I(g,p)\}$ as a measure to find distance between $p(r|\alpha,\beta)$ and $g(r)$. It may be observed that the distance KL_{min} is a symmetric distance as specified in Bernardo and Rueda (2002).

Table 2 provides computed values of KL_{min} for arbitrarily chosen $\alpha = 4, \beta = 1$ and some selected values of δ_3 and δ_4. The chosen values of δ_3 and δ_4 are those in which $g(r)$ is unimodal and non-negative.

Table 2: Values of KL_{min} for $\alpha = 4, \beta = 1$ and some selected values of δ_3 and δ_4

δ_3	δ_4	KL_{min}	δ_3	δ_4	KL_{min}	δ_3	δ_4	KL_{min}
0	0	0	0.4	2	0.0061	0.9	6	0.016
	2	0.0144		4	0.005		8	0.0209
0.15	0	0.0049	0.6	4	0.0064	1.05	7	0.0226
	2	0.0023		6	0.0139		8	0.0192
0.3	2	0.0011	0.75	5	0.0107	1.35	9.5	0.0358
	4	0.0118		7	0.0167		10.3	0.0354

From Table 2, it may be observed that

(1) Out of the chosen combinations of (δ_3, δ_4), KL_{min} is minimum for (0, 0) as it corresponds to Gamma distribution, and is maximum for (1.35, 9.5).

(2)KL_{min} could be approximately same for different choices for (δ_3, δ_4). In particular, for the combinations (0, 2) and (0.6, 6), KL_{min} is approximately 0.014. However, the graphs of $g(r)$ for these values of (δ_3, δ_4) are different as shown in Figure 3.

(3)For (0.6, 4), $KL_{min} = 0.0064$, and for (0.15, 2), $KL_{min} = 0.0023$. So, $g(r)$ corresponding to (0.6, 4) is more non-gamma than gamma distribution as compared to the $g(r)$ corresponding to (0.15, 2).

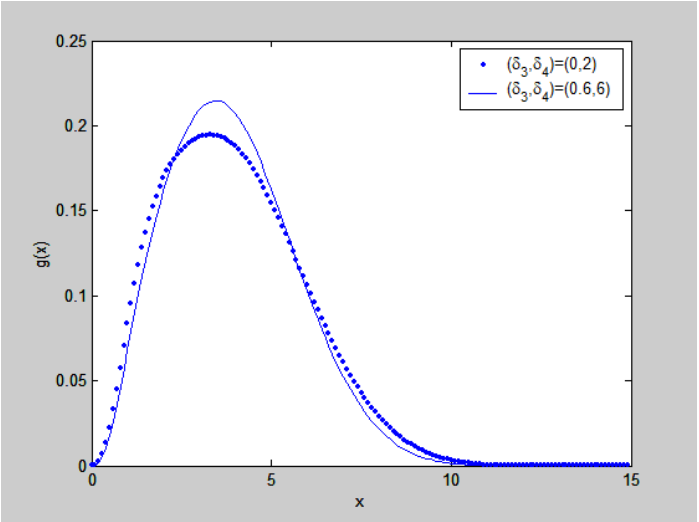


Figure 3: Graph of $g(r)$

In the next section, Bayes estimates of CV, obtained in Section (2.4), will now be calculated for hypothetical data using some values of δ_3 and δ_4 selected based on KL_{min} discussed in this section.

5. Numerical Illustration

To study the effect of non-gamma prior, we generate a hypothetical data of size $n = 10$ from $N(4,4)$ distribution given by 0.0660, 5.2140, 3.7548, 5.4743, 6.2490, 2.0363, 4.8134, 9.4950, 6.6342, 4.4920. It is clear that the true CV is 0.5 whereas the classical estimate of CV, the ratio of observed standard deviation and observed mean, is 0.505187. The MELO estimate under non-informative prior is 0.5622.

The Bayes estimates of CV, with $\alpha = 4, \beta = 1, \mu = 0, \tau = 1$, and various values of δ_3 and δ_4 selected using Table 2, are tabulated in Table 3.

Table 3: Bayes estimate of CV for various values of δ_3 and δ_4

δ_3	δ_4	KL_{min}	$\hat{a}_{MELO}$
0	0	0	0.4978
0.15	2	0.0023	0.4998
0.40	4	0.0050	0.5001
0.75	5	0.0107	0.4924
1.05	8	0.0192	0.4980
1.35	9.5	0.0358	0.4871

From Table 3, one may observe that the Bayes estimates $\hat{a}_{MELO}$ of CV are close to the Bayes estimate of CV under gamma prior for all chosen combination of δ_3 and δ_4. The difference in the maximum and minimum value of $\hat{a}_{MELO}$ is 0.013 which is insignificant and hence, we may say that the moderate deviation from gamma prior may not significantly affect Bayes estimate of coefficient of variation under MELO. We may, therefore, conclude that the Bayes estimate is robust with respect to misspecification of the prior distribution for precision in our illustration.

6. Conclusion

In this paper, Bayes estimate of coefficient of variation is derived for normal model with both mean and variance unknown. The normal conditional prior for unknown mean and K-prior for the unknown precision of the normal distribution are considered. The positive and unimodal regions for K-class of non-gamma densities are obtained for $\alpha = 4$ and $\beta = 1$. The boundary values of δ_3 and δ_4 where the pdf of non-gamma distribution changes from the positive definite to non-positive definite are provided. It is seen that in the region bounded by the above values, pdf is unimodal as well. For other values of α and β, one may find region where pdf is positive and unimodal using the same procedure. It is found that for two or more members of K-class of non-gamma distributions, KL_{min} could

be approximately same which means that these members are equally non-gamma as compared to the gamma distribution. A numerical illustration is also discussed and therein, it is observed that Bayes estimate of coefficient of variation under K-prior distributions are very close to that based on gamma prior distribution for all chosen combinations of δ_3 and δ_4. We may also conclude that the Bayes estimate of CV under MELO is reasonably insensitive to moderate deviation from generally assumed gamma prior distribution.

Acknowledgement

The authors are grateful to Dr. Ashok K. Bansal for his valuable suggestions and innumerable discussions that led to the improvement of the paper.

References

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions*. Dover Publications, Tenth printing (1972), New York.

Aggarwal, P. and Bansal, A. K. (2010). Robustness of Bayes prediction under error-in-variables superpopulation model. *Communications in Statistics: Theory and Methods*, **39**(**2**), 327-339.

Aggarwal, P. and Bansal, A. K. (2017). Bayes prediction of Poisson regression superpopulation mean with a non-gamma prior. *Communications in Statistics: Theory and Methods*, **46**(**11**), 5531-5543.

Bansal, A. K. (2007). *Bayesian Parametric Inference*. Narosa Publishing House, India.

Bansal, A. K. and Singh, R. C. (1999). Bayes estimation with a non-gamma prior. *Ganita*, **50**(**2**), 93-102.

Barton, D. E. and Dennis, K. E. (1952). The conditions under which Gram-Charlier and Edgeworth curves are positive definite and unimodal. *Biometrika*, **39**, 425-427.

Berger, J. O. (1984). The robust Bayesian view-point. In: Kadane, J. B., Ed. *Robustness of Bayesian Statistics*. Amsterdam: North-Holland, 63-124.

Bernardo, J. M. and Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *International Statistical Review*, **70**, 351-372.

De Groot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

Draper, N. and Tierney, D. (1972). Regions of positive and unimodal series expansion of the Edgeworth and Gram-Charlier approximation. *Biometrika*, **59**, 463-465.

Khamis, S. H. (1960). Incomplete gamma functions expansions of statistical distribution functions. *Bulletin of International Statistical Institute*, **37**, 385-396.

Tiku, M. L. and Tan W. Y. (1999). *Sampling Distributions in Terms of Laguerre Polynomials With Applications*. Wiley Eastern, New Age International, New Delhi.

Spiring, F. (2011). The refined positive definite and unimodal regions for the Gram-Charlier and Edgeworth Series Expansion. *Advances in Decision Sciences*, Article ID 463097, 18 pages.

Zellner, A. (1978). Estimation of functions of population means and regression coefficients including structural coefficients (A minimum expected loss (MELO) approach). *Journal of Econometrics*, **8**, 127-158.

Appendix

A.1. (Derivation of posterior distribution given in Section 4.1)

It is known that

$$g(\theta, r|\mathbf{x}) = \frac{\ell(\theta, r \,|\, \mathbf{x}) g(\theta, r)}{\int_{-\infty}^{\infty}\int_{0}^{\infty} \ell(\theta, r|\mathbf{x}) g(\theta, r) d\theta dr}$$

Using

$$\ell(\theta, r|\mathbf{x}) = \left(\frac{r}{2\pi}\right)^{n/2} exp\left(-\frac{r}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 - \frac{nr}{2}(\bar{x} - \theta)^2\right),$$

and

$$g(\theta, r) = g(\theta|r)g(r) = \left(\sqrt{\frac{r}{2\pi}} exp\left(-\frac{\tau r}{2}(\theta - \mu)^2\right)\right)(K(r)\frac{\beta^{\alpha}}{\Gamma(\alpha)}\exp(-\beta r)\, r^{\alpha-1},$$

we get

$$g(\theta, r|\mathbf{x}) = \frac{r^{\frac{n}{2}+\frac{1}{2}+\alpha-1}K(r)\exp\left(-r\left(\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\right)\exp\left(-\frac{r}{2}(n(\theta - \bar{x})^2 + \tau(\theta - \mu)^2)\right)}{\int_{-\infty}^{\infty}\int_{0}^{\infty} r^{\frac{n}{2}+\frac{1}{2}+\alpha-1}K(r)\exp\left(-r\left(\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\right)\exp\left(-\frac{r}{2}(n(\theta - \bar{x})^2 + \tau(\theta - \mu)^2)\right)d\theta dr}.$$

Using a result that

$$A(z-a)^2 + B(z-b)^2 = (A+B)(z-c)^2 + \frac{AB}{A+B}(a-b)^2, c = \frac{Aa+Bb}{A+B},$$

we can write

$$n(\theta - \bar{x})^2 + \tau(\theta - \mu)^2 = (n+\tau)(\theta - \mu^*)^2 + \frac{n\tau}{n+\tau}(\mu - \bar{x})^2, \mu^* = \frac{n\bar{x} + \tau\mu}{n+\tau}.$$

Thus,

$$g(\theta, r|\mathbf{x}) = \frac{r^{\frac{n}{2}+\frac{1}{2}+\alpha-1}K(r)\exp\left(-r\left(\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{2}\frac{n\tau}{n+\tau}(\mu - \bar{x})^2\right)\right)\exp\left(-\frac{r}{2}((n+\tau)(\theta - \mu^*)^2)\right)}{\int_{-\infty}^{\infty}\int_{0}^{\infty} r^{\frac{n}{2}+\frac{1}{2}+\alpha-1}K(r)\exp\left(-r\left(\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{2}\frac{n\tau}{n+\tau}(\mu - \bar{x})^2\right)\right)\exp\left(-\frac{r}{2}((n+\tau)(\theta - \mu^*)^2)\right)d\theta dr}$$

$$= \frac{r^{\frac{n}{2}+\frac{1}{2}+\alpha-1}K(r)\exp\left(-r\left(\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{2}\frac{n\tau}{n+\tau}(\mu - \bar{x})^2\right)\right)\exp\left(-\frac{r}{2}((n+\tau)(\theta - \mu^*)^2)\right)}{\int_{0}^{\infty} r^{\frac{n}{2}+\alpha-1}K(r)\exp\left(-r\left(\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{2}\frac{n\tau}{n+\tau}(\mu - \bar{x})^2\right)\right)\sqrt{\frac{2\pi}{n+\tau}}dr}.$$

Writing $\alpha^* = \alpha + \frac{n}{2}, \beta^* = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{2}\frac{n\tau}{n+\tau}(\mu - \bar{x})^2$, we can write

$$g(\theta, r|\mathbf{x}) = \left(\sqrt{\frac{(n+\tau)r}{2\pi}}\exp\left(-\frac{r}{2}\left((n+\tau)(\theta - \mu^*)^2\right)\right)\right)\frac{\left(r^{\alpha^*-1}K(r)\exp(-r\beta^*)\right)}{\int_{0}^{\infty} r^{\alpha^*-1}K(r)\exp(-r\beta^*)\, dr}$$

$$= \left(\sqrt{\frac{(n+\tau)r}{2\pi}}\exp\left(-\frac{r}{2}\left((n+\tau)(\theta - \mu^*)^2\right)\right)\right)\frac{K(r)p(r|\alpha^*, \beta^*, \mathbf{x})}{G(\delta_3, \delta_4)}$$

where

$$G(\delta_3, \delta_4) = 1 + A_1\frac{\Gamma(\alpha+3)}{\Gamma(\alpha)}\sum_{i=0}^{3}(-1)^i\binom{3}{i}R_{3-i} + A_2\frac{\Gamma(\alpha+4)}{\Gamma(\alpha)}\sum_{i=0}^{4}(-1)^i\binom{4}{i}R_{4-i}$$

$$= 1 - \delta_3 \frac{\alpha^{\frac{3}{2}}}{6} C_1(\alpha^*) + \delta_4 \frac{\alpha^2}{24} C_2(\alpha^*),$$
$$C_1(\alpha^*) = 3R_4 - 13R_3 + 21R_2 - 15R_1 + 4R_0,$$
$$C_2(\alpha^*) = R_4 - 4R_3 + 6R_2 - 4R_1 + R_0,$$

and

$$R_j = \left(\frac{\Gamma(\alpha^* + j)}{\Gamma(\alpha^*)\beta^{*j}}\right) \Big/ \left(\frac{\Gamma(\alpha + j)}{\Gamma(\alpha)\beta^j}\right).$$

ν_j and μ_j are the moments about origin of order j of gamma prior and posterior gamma, respectively.

A.2. (Derivation of the posterior expectations given in Section 4.2)

$$(1)\ E(\theta^2|\mathbf{x}) = \int_0^\infty \frac{K(r)p(r|\alpha^*,\beta^*,\mathbf{x})}{G(\delta_3,\delta_4)} \left(\int_{-\infty}^{\infty} \theta^2 \sqrt{\frac{(n+\tau)r}{2\pi}} \exp\left(-\frac{r}{2}\left((n+\tau)(\theta-\mu^*)^2\right)\right) d\theta \right) dr$$
$$= \int_0^\infty \frac{K(r)p(r|\alpha^*,\beta^*,\mathbf{x})}{G(\delta_3,\delta_4)} \left(\mu^{*2} + \frac{1}{r(\tau+n)}\right) dr$$
$$= \mu^{*2} + \frac{G_1(\delta_3,\delta_4)}{G(\delta_3,\delta_4)} \frac{\beta^{*\alpha^*}}{\Gamma(\alpha^*)} \frac{\Gamma(\alpha^*-1)}{\beta^{*\alpha^*-1}} \frac{1}{\tau+n}$$

where

$$G_1(\delta_3,\delta_4) = 1 - \delta_3 \frac{\alpha^{\frac{3}{2}}}{6} C_1(\alpha^*-1) + \delta_4 \frac{\alpha^2}{24} C_2(\alpha^*-1),$$
$$\therefore E\left(\theta^2\middle|\mathbf{x}\right) = \mu^{*2} + \frac{\beta^*}{(\alpha^*-1)(\tau+n)} \frac{G_1(\delta_3,\delta_4)}{G(\delta_3,\delta_4)}.$$

$$(2)\ E\left(\frac{\theta}{\sqrt{r}}\middle|\mathbf{x}\right) = \int_0^\infty r^{-\frac{1}{2}} \frac{K(r)p(r|\alpha^*,\beta^*,\mathbf{x})}{G(\delta_3,\delta_4)} \left(\int_{-\infty}^{\infty} \theta \sqrt{\frac{(n+\tau)r}{2\pi}} \exp\left(-\frac{r}{2}\left((n+\tau)(\theta-\mu^*)^2\right)\right) d\theta \right) dr$$
$$= \int_0^\infty r^{-\frac{1}{2}} \frac{K(r)p(r|\alpha^*,\beta^*,\mathbf{x})}{G(\delta_3,\delta_4)} \mu^* dr$$
$$= \mu^* \frac{G_2(\delta_3,\delta_4)}{G(\delta_3,\delta_4)} \frac{\beta^{*\alpha^*}}{\Gamma(\alpha^*)} \frac{\Gamma\left(\alpha^*-\frac{1}{2}\right)}{\beta^{*\alpha^*-\frac{1}{2}}}$$

where

$$G_2(\delta_3,\delta_4) = 1 - \delta_3 \frac{\alpha^{\frac{3}{2}}}{6} C_1\left(\alpha^*-\frac{1}{2}\right) + \delta_4 \frac{\alpha^2}{24} C_2\left(\alpha^*-\frac{1}{2}\right)$$
$$\therefore E\left(\frac{\theta}{\sqrt{r}}\middle|\mathbf{x}\right) = \mu^* \frac{\sqrt{\beta^*}\,\Gamma\left(\alpha^*-\frac{1}{2}\right)}{\Gamma(\alpha^*)} \frac{G_2(\delta_3,\delta_4)}{G(\delta_3,\delta_4)}.$$

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp. 147-156

Almost Unbiased Dual Exponential Type Estimators of Population Mean Using Auxiliary Information

Sajad Hussain[1], Manish Sharma[2], Banti Kumar[3] and Vilayat Ali Bhat[4]
[1]*University School of Business, Chandigrah University, Mohali, Gharuan - 140413, INDIA.*
[2]*Division of Statistics and Computer Science, FBSc, SKUAST-J, Chatha-180009, INDIA.*
[3]*Department of Physical Sciences and Languages, COBS CSK HPKV H.P, Palampur-176062, INDIA.*
[4]*Department of Statistics, Pondicherry University, R. V. Nagar, Kalapet, Puducherry-605014, INDIA.*

Received: 02 July 2021; Revised: 13 October 2021; Accepted: 17 October 2021

Abstract

In this paper dual exponential type estimators of population mean have been proposed. The large sample properties of the proposed estimators have been studied by obtaining the bias and mean square error (MSE) expressions. The proposed estimators under optimum conditions were found to be unbiased and more efficient than sample mean, ratio estimator of Cochran (1940), product estimator of Robson (1957), exponential ratio and product estimators of Bahl and Tuteja (1991), exponential ratio estimators of Singh *et al.* (2009) and the exponential product type estimators of Onyeka (2013). A numerical study has also been carried out to support the theoretical findings of the paper.

Key words: Dual estimator; Exponential estimator; Auxiliary variable; Unbiased estimator; Mean square error.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Whenever a researcher intends to get the precise estimates, the best choice is to make the wise use of auxiliary information. The auxiliary information can be used either at the design stage or at the estimation stage or at both stages. The ratio, product, difference and regression estimators are defined by using the available auxiliary information at the estimation stage. At this stage the auxiliary information may be available in the form of correlation coefficient, mean, median, coefficient of variation, skewness, kurtosis etc. The pioneer work for estimation of population mean using auxiliary information was done by Cochran (1940) while proposing classical ratio estimator, used when there is a high positive correlation between study variable (Y) and auxiliary variable (X) with the regression line passing through origin. If the correlation between Y and X is negative high, product method of estimation proposed by Robson (1957) can be used. While as the linear regression estimator is preferred when there is a very high (positive or negative) correlation between X and Y and the regression line of Y on X has intercept on y-axis. Many researchers such as

Corresponding Author: Sajad Hussain
E-mail: sajad.stat321@gmail.com

Sisodia and Dwivedi (1981), Singh and Tailor (2003), Khoshnevisan *et al.* (2007), Sharma and Bhatnagar (2008), Sharma *et al.* (2010), Yadav and Kadilar (2013), Kumar *et al.* (2018) and others proposed modified ratio or product type estimators by utilizing different known values of the parameters of auxiliary variable and these estimators have gained relevance in estimation theory because of their improved precision than conventional ratio and product estimators. The modified ratio and product estimators can work well only when correlation between Y and X is high, therefore Bahl and Tuteja (1991) proposed exponential ratio and product type estimators and these estimators can be employed even when there is not a high degree of correlation between X and Y. Later Singh *et al.* (2007, 2009), Onyeka (2013), Yasmeen *et al.* (2016), Panigrahi and Mishra (2017) and Hussain *et al.* (2021) proposed some improved versions of the exponential ratio and product type estimators.

On taking a note of the above discussion, it was observed that most of the estimators available in the literature are biased. They lack the very first property of a good estimator which may lead to over or under estimation of the population mean. Therefore, the authors Singh and Singh (1993), Yadav *et al.* (2012), Singh *et al.* (2016) and others worked in this direction and proposed almost unbiased estimators of population mean. Further, it is also observed that for positively correlated variables ratio estimators are used and for negatively correlated variables product estimators are used. So the authors Singh *et al.* (2009*a*), Tailor and Sharma (2009), Sharma and Tailor (2010), Tailor *et al.* (2012) and others proposed ratio cum product estimators which can be employed for both positively and negatively correlated variables. It is also observed that the exponential estimators can also be employed for low degree of correlation. By keeping the stated points in view, two almost unbiased dual exponential type estimators of population mean have been proposed in the paper.

Consider a finite population containing N number of units in total and draw a random sample of size n ($n < N$) by simple random sampling without replacement (SRSWOR) sampling scheme. Associated with every unit, there are two variables Y and X, the population mean of X is assumed to be known. The sample mean of Y and X i.e $\bar{y} = \frac{1}{n}\sum_i^n y_i$ and $\bar{x} = \frac{1}{n}\sum_i^n x_i$ are the unbiased estimates of $\bar{X} = \frac{1}{N}\sum_i^N X_i$ and $\bar{Y} = \frac{1}{N}\sum_i^N Y_i$ respectively. Other formula and notations that are used in the paper (John and Inyang (2015)) are as

Study Variable	**Auxiliary Variable**
$C_y = \frac{S_y}{\bar{Y}}$ is the coefficient of variation .	: $C_x = \frac{S_x}{\bar{X}}$ is the coefficient of variation.
$S_y^2 = \frac{1}{N-1}\sum_{i=1}^N (Y_i - \bar{Y})^2$ is the population mean square.	: $S_x^2 = \frac{1}{N-1}\sum_{i=1}^N (X_i - \bar{X})^2$ is the population mean square.
$s_y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2$ is the sample mean square.	: $s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$ is the sample mean square.

Further,
$S_{yx} = \frac{1}{N-1}\sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X})$ is the population covariance between Y and X.
$s_{yx} = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})$ is the sample covariance between y and x.
$\rho = \frac{S_{xy}}{\sqrt{S_x^2 S_y^2}}$ is the population correlation coefficient between X and Y.
$\theta = \frac{a\bar{X}}{2(a\bar{X}+b)}$ and $\gamma = \frac{1-f}{n}$, where the sampling fraction $f = \frac{n}{N}$.

The Percent relative efficiency (PRE) of the estimators is obtained using the formula as $\text{PRE} = \frac{MSE\ of\ existing\ estimator}{MSE\ of\ proposed\ estimator} \times 100$

2. Review of Some Existing Ratio and Product Type Estimators

The sample mean estimator is

$$t_1 = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

With the Bias and MSE are as

$$Bias(t_1) = 0. \tag{1}$$

$$MSE(t_1) = \gamma \bar{Y}^2 C_y^2. \tag{2}$$

The ratio estimator proposed by Cochran (1940) which is more efficient than the estimator t_1, if $\frac{C_x}{2C_y} < \rho \leq +1$ is

$$t_2 = \bar{y}\frac{\bar{X}}{\bar{x}}.$$

The expressions of Bias and MSE for the estimator t_2 are as

$$Bias(t_2) = \gamma \bar{Y}(C_x^2 - C_{yx}). \tag{3}$$

$$MSE(t_2) = \gamma \bar{Y}^2(C_y^2 + C_x^2 - 2C_{yx}). \tag{4}$$

When the variables X and Y are negatively correlated and $-1 \leq \rho < -\frac{C_x}{2C_y}$, Robson (1957) proposed product estimator of population. The main advantage of this estimator is that the exact expressions of Bias and mean squared error were obtained. The proposed estimator is given as

$$t_3 = \bar{y}\frac{\bar{x}}{\bar{X}}.$$

The Bias and MSE expressions are as

$$Bias(t_3) = \gamma \bar{Y} C_{yx}. \tag{5}$$

$$MSE(t_3) = \gamma \bar{Y}^2(C_y^2 + C_x^2 + 2C_{yx}). \tag{6}$$

The exponential ratio (t_4) and product (t_5) type estimators proposed by Bahl and Tuteja (1991) which are efficient even when there is a low degree of correlation between X and Y are as

$$t_4 = \bar{y}\exp\left(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}\right),$$

with the Bias and MSE as

$$Bias(t_4) = \gamma \bar{Y}\left(\frac{3}{8}C_x^2 - \frac{1}{2}C_{yx}\right), \tag{7}$$

$$MSE(t_4) = \gamma \bar{Y}^2\left(C_y^2 + \frac{C_x^2}{4} - C_{yx}\right), \tag{8}$$

and

$$t_5 = \bar{y} \exp\left(\frac{\bar{x} - \bar{X}}{\bar{X} + \bar{x}}\right),$$

with the Bias and MSE as

$$Bias(t_5) = \gamma \bar{Y}\left(\frac{1}{2}C_{yx} - \frac{1}{8}C_x^2\right), \tag{9}$$

$$MSE(t_5) = \gamma \bar{Y}^2\left(C_y^2 + \frac{C_x^2}{4} + C_{yx}\right). \tag{10}$$

A class of modified exponential ratio estimators proposed by Singh *et al.* (2009) using the constants a($\neq$ 0) and b, where a and b are either the real number or functions of some known parameters of auxiliary variable such as coefficient of variation, skewness, correlation etc. The proposed estimators are as

$$t_6 = \bar{y} \exp\left[\frac{(a\bar{X} + b) - (a\bar{x} - b)}{(a\bar{X} + b) + (a\bar{x} - b)}\right].$$

The Bias and MSE expressions of t_6 are as

$$Bias(t_6) = \gamma \bar{Y}(\theta^2 C_x^2 - \theta C_{yx}). \tag{11}$$

$$MSE(t_6) = \gamma \bar{Y}^2(C_y^2 + \theta^2 C_x^2 - 2\theta C_{yx}). \tag{12}$$

Onyeka (2013) proposed a class of product type estimators as

$$t_7 = \bar{y} \exp\left[\frac{(a\bar{X} + b) - (a\bar{x} + b)}{(a\bar{X} + b) + (a\bar{x} + b)}\right].$$

The expressions of Bias and MSE for the estimator t_7 are as

$$Bias(t_7) = \gamma \bar{Y}\left(\frac{1}{2}\theta C_{yx} - \frac{1}{8}\theta^2 C_x^2\right). \tag{13}$$

$$MSE(t_7) = \gamma \bar{Y}^2\left(C_y^2 + \frac{1}{4}\theta^2 C_x^2 + \theta C_{yx}\right). \tag{14}$$

3. Proposed Estimators

The proposed dual type exponential estimators are as

$$t_{de1} = \bar{y}\left[\alpha \exp\left(\frac{\bar{X} - \bar{x}}{p\bar{X}}\right) + (1 - \alpha)\exp\left(\frac{\bar{x} - \bar{X}}{p\bar{X}}\right)\right] \text{ and}$$

$$t_{de2} = \bar{y}\left[\beta \exp\left(\frac{\bar{X} - \bar{x}}{q\bar{x}}\right) + (1 - \beta)\exp\left(\frac{\bar{x} - \bar{X}}{q\bar{x}}\right)\right].$$

Where p and q are non zero constants whose value is chosen such that the estimators t_{de1} and t_{de2} should be unbiased. The value of constants α and β are chosen such that the MSE of t_{de1} and t_{de2} is minimum. The expressions of Bias and MSE are obtained through the

following methodology.
Let

$$e_0 = \frac{\bar{y} - \bar{Y}}{\bar{Y}} \qquad ; \qquad e_1 = \frac{\bar{x} - \bar{X}}{\bar{X}}.$$

Therefore, the following expected values are obtained

$$E(e_0) = E(e_1) = 0.$$
$$E(e_0^2) = \gamma C_y^2; \qquad E(e_1^2) = \gamma C_x^2; \qquad E(e_0 e_1) = \gamma \rho C_y C_x.$$

On writing the estimator t_{de1} and t_{de2} in terms of $e_i (i = 0, 1)$, the following equations are obtained as

$$t_{de1} = \bar{Y}(1 + e_0)\left[\alpha \exp\left(\frac{-e_1}{p}\right) + (1 - \alpha)\exp\left(\frac{e_1}{p}\right)\right]. \tag{15}$$

$$t_{de2} = \bar{Y}(1 + e_0)\left[\beta \exp\left(\frac{-e_1}{q}(1 + e_1)^{-1}\right) + (1 - \beta)\exp\left(\frac{e_1}{q}(1 + e_1)^{-1}\right)\right]. \tag{16}$$

After solving equation (15) & (16) and keeping the terms up 2^{nd} degree, the equations reduce to

$$t_{de1} = \bar{Y}\left[1 + e_0 + (1 - 2\alpha)\frac{e_1}{p} + \frac{e_1^2}{2p^2} + (1 - 2\alpha)\frac{e_0 e_1}{p}\right]$$

$$\Rightarrow t_{de1} - \bar{Y} = \left[e_0 + (1 - 2\alpha)\frac{e_1}{p} + \frac{e_1^2}{2p^2} + (1 - 2\alpha)\frac{e_0 e_1}{p}\right]. \tag{17}$$

$$t_{de2} = \bar{Y}\left[1 + e_0 + (1 - 2\beta)\frac{e_1}{q} + \left(\frac{1}{2q} + 2\beta - 1\right)\frac{e_1^2}{q} + (1 - 2\beta)\frac{e_0 e_1}{q}\right]$$

$$\Rightarrow t_{de2} - \bar{Y} = \left[e_0 + (1 - 2\beta)\frac{e_1}{q} + \left(\frac{1}{2q} + 2\beta - 1\right)\frac{e_1^2}{q} + (1 - 2\beta)\frac{e_0 e_1}{q}\right]. \tag{18}$$

Now taking expectation on both sides of (17) and (18), the bias of the estimators t_{de1} and t_{de2} is obtained as

$$Bias(t_{de1}) = \gamma \bar{Y}\left[\frac{1}{2p^2}C_x^2 + \frac{1}{p}(1 - 2\alpha)\rho C_x C_y\right] \text{ and} \tag{19}$$

$$Bias(t_{de2}) = \gamma \bar{Y}\frac{1}{q}\left[\left(\frac{1}{2q} + 2\beta - 1\right)C_x^2 + (1 - 2\beta)\rho C_x C_y\right] \text{ respectively.} \tag{20}$$

On squaring the equations (17) & (18) and taking expectation on both sides. After solving and retaining the terms up to 2^{nd} degree only, the mean square error of t_{de1} and t_{de2} is obtained as

$$MSE(t_{de1}) = \gamma \bar{Y}^2\left[C_y^2 + (1 - 2\alpha)^2\frac{C_x^2}{p^2} + \frac{2}{p}(1 - 2\alpha)\rho C_x C_y\right] \text{ and} \tag{21}$$

$$MSE(t_{de2}) = \gamma \bar{Y}^2\left[C_y^2 + (1 - 2\beta)^2\frac{C_x^2}{q^2} + \frac{2}{q}(1 - 2\beta)\rho C_x C_y\right] \text{ respectively.} \tag{22}$$

The estimator t_{de1} is unbiased, if

$$p = \frac{C_x}{2(2\alpha - 1)\rho C_y}, \tag{23}$$

and the estimator t_{de2} is unbiased, if

$$q = \frac{C_x}{2(2\beta - 1)(\rho C_y - C_x)}. \tag{24}$$

Substituting the values of (23) and (24) in equations (21) and (22) respectively, the following equations are obtained as

$$MSE(t_{de1}) = \gamma \bar{Y}^2 \left[C_y^2 + 4(1-2\alpha)^4 \rho^2 C_y^2 - 4(1-2\alpha)^2 \rho^2 C_y^2 \right]. \tag{25}$$

$$MSE(t_{de2}) = \gamma \bar{Y}^2 \left[C_y^2 + 4(1-2\beta)^4 (\rho C_y - C_x)^2 - 4(1-2\beta)^2 (\rho C_y - C_x)\rho C_y \right]. \tag{26}$$

For obtaining optimum value of α and β differentiate equations (25) and (26) with respect to α and β respectively and equating to zero.

$$\text{Therefore, } \alpha = 0.146, 0.854 \qquad \text{and} \qquad \beta = \frac{1}{2} \pm \frac{1}{2}\sqrt{\frac{\rho C_y}{2(\rho C_y - C_x)}}.$$

The value of unknown quantity (C_y) used to find the values of p, q and β can be obtained quite accurately from some previous survey or from the experience of the researcher (See Reddy (1974), Singh and Vishwakarma (2008), Singh and Kapre (2010)). Now by using the values of α (0.146 or 0.854) and β $\left(\frac{1}{2} + \frac{1}{2}\sqrt{\frac{\rho C_y}{2(\rho C_y - C_x)}} \text{ or } \frac{1}{2} - \frac{1}{2}\sqrt{\frac{\rho C_y}{2(\rho C_y - C_x)}} \right)$ in equations (25) and (26) respectively, the minimum value of MSE of the estimators t_{de1} and t_{de2} is obtained as

$$MSE_{min}(t_{dei}) = \gamma \bar{Y}^2 C_y^2 (1 - \rho^2). \qquad i = 1, 2 \tag{27}$$

4. Efficiency Comparisons

From equations (2), (4), (6), (8), (10), (12), (14) and (27), the conditions under which the proposed estimators will be preferred for better precision are obtained as

$$MSE_{min}(t_{dei}) < V(t_1)$$

$$\Rightarrow \gamma \bar{Y}^2 C_y^2 (1 - \rho^2) < \gamma \bar{Y}^2 C_y^2, \; if \; \rho^2 \bar{Y}^2 > 0. \tag{28}$$

$$MSE_{min}(t_{dei}) < MSE(t_2)$$

$$\Rightarrow \gamma \bar{Y}^2 C_y^2 (1 - \rho^2) < \gamma \bar{Y}^2 (C_y^2 + C_x^2 - 2C_{yx}), \; if \; (\rho C_y - C_x)^2 > 0. \tag{29}$$

$$MSE_{min}(t_{dei}) < MSE(t_3)$$

$$\Rightarrow \gamma \bar{Y}^2 C_y^2 (1 - \rho^2) < \gamma \bar{Y}^2 (C_y^2 + C_x^2 + 2C_{yx}), \; if \; (\rho C_y + C_x)^2 > 0. \tag{30}$$

$$MSE_{min}(t_{dei}) < MSE(t_4)$$

$$\Rightarrow \gamma \bar{Y}^2 C_y^2 (1 - \rho^2) < \gamma \bar{Y}^2 \left(C_y^2 + \frac{C_x^2}{4} - C_{yx} \right), \; if \; (2\rho C_y - C_x)^2 > 0. \tag{31}$$

$$MSE_{min}(t_{dei}) < MSE(t_5)$$

$$\Rightarrow \gamma\bar{Y}^2C_y^2(1-\rho^2) < \gamma\bar{Y}^2\left(C_y^2+\frac{C_x^2}{4}+C_{yx}\right),\ if\ (2\rho C_y+C_x)^2>0. \tag{32}$$

$$MSE_{min}(t_{dei}) < MSE(t_6)$$

$$\Rightarrow \gamma\bar{Y}^2C_y^2(1-\rho^2) < \gamma\bar{Y}^2(C_y^2+\theta^2C_x^2-2\theta C_{yx}),\ if\ (\rho C_y-\theta C_x)^2>0. \tag{33}$$

$$MSE_{min}(t_{dei}) < MSE(t_7)$$

$$\Rightarrow \gamma\bar{Y}^2C_y^2(1-\rho^2) < \gamma\bar{Y}^2\left(C_y^2+\frac{1}{4}\theta^2C_x^2+\theta C_{yx}\right),\ if\ (2\rho C_y-\theta C_x)^2>0. \tag{34}$$

The conditions (28) to (34) hold, therefore the estimators $t_{dei}(i=1,2)$ are more efficient than $t_1, t_2, t_3, t_4, t_5, t_6$ and t_7.

5. Numerical Illustration

The performance of the estimators proposed and considered for comparison in the paper have been evaluated by using the data of four populations P1, P2, P3 and P4 (See Table 1). In the populations P1 and P2, the variables X and Y are positively correlated while as for P3 and P4 are negatively correlated. The source of the population P1 is Sukhatme and Chand (1977) where the variable Y represents the apple trees of bearing age in 1964 and the variable X represents bushels harvested in 1964. The population P2 is from Murthy (1967), where the variable Y is fixed capital and the variable X is the output of 80 factories. The source of population P3 is Onyeka (2013) where the variable Y represents percentage of hives affected by disease and X the date of flowering of a particular summer species (no. of days from Jan. 1). The population P4 is from Gujarati (2004) where the variable Y is average miles per gallon and the variable X is top speed, miles per hour.

Table 1: Summary statistics of the population data sets.

Population	N	n	$\bar{Y}$	$\bar{X}$	C_y	C_x	ρ
P1	200	20	1031.82	2934.58	1.598	2.006	0.93
P2	80	20	11.264	51.826	0.750	0.354	0.94
P3	10	4	52	200	0.156	0.046	−0.94
P4	81	13	33.835	112.457	0.297	0.126	−0.69

Table 2: MSE, Bias and PRE of the estimators t_1, t_2, t_4, t_6 and t_{dei}.

Estimator	Population					
	P1			P2		
	MSE	$\mid Bias \mid$	PRE	MSE	$\mid Bias \mid$	PRE
t_1	122341.540	0.000	100.000	2.676	0.000	100.000
t_2	29476.060	48.421	415.054	0.898	0.052	297.995
t_4	27711.550	0.855	441.482	1.638	0.033	163.379
t_6	27745.680	22.518	440.939	1.644	0.039	162.773
t_{dei}	16528.340	0.000	740.192	0.311	0.000	860.450

It can be observed from Table-2 that the proposed estimators have minimum MSE amongst sample mean estimator (t_1) and the ratio estimators t_2, t_4, t_6 considered. The percent relative efficiency (PRE) of the proposed estimators is highest among all other estimators considered.

Table 3: MSE, Bias and PRE of the estimators t_1, t_3, t_5, t_7 and t_{dei}.

Estimator	Population					
	P3			P4		
	MSE	$\lvert Bias \rvert$	PRE	MSE	$\lvert Bias \rvert$	PRE
t_1	9.871	0.000	100.000	6.521	0.000	100.000
t_3	5.257	0.053	187.769	3.877	0.056	168.197
t_5	7.349	0.028	134.318	4.906	0.033	132.919
t_7	8.556	0.014	115.369	5.639	0.015	115.641
t_{dei}	1.149	0.000	859.095	3.416	0.000	190.896

Table-3 depicts that the proposed estimators have minimum MSE amongst sample mean estimator (t_1) and product estimators t_3, t_5, t_7 considered and are also unbiased. The percent relative efficiency (PRE) of proposed estimators is highest among all other estimators considered.

Thus, it can be concluded from Table-2 and Table-3 that the proposed estimators are unbiased and work efficiently in estimating the population mean irrespective of negative or positive correlation between the study and auxiliary variable.

6. Discussion

The population mean can be estimated using the proposed dual exponential type estimators of population mean by plugging in the values of $p = \frac{C_x}{2(2\alpha-1)\rho C_y}$, $q = \frac{C_x}{2(2\beta-1)(\rho C_y - C_x)}$ and the optimum values of α (0.146 or 0.854) and β $\left(\frac{1}{2} + \frac{1}{2}\sqrt{\frac{\rho C_y}{2(\rho C_y - C_x)}} \text{ or } \frac{1}{2} - \frac{1}{2}\sqrt{\frac{\rho C_y}{2(\rho C_y - C_x)}} \right)$ in the respective estimators t_{ue1} and t_{ue2}. The constants p, q, β are dynamic in nature and therefore depend upon the parameters of population data whereas the value of constant α is static and results an estimator as

$$t_{de1} = \bar{y}\left[0.854\exp\left(\frac{1.414\rho C_y(\bar{X}-\bar{x})}{C_x\bar{X}}\right) + 0.146\exp\left(\frac{1.414\rho C_y(\bar{x}-\bar{X})}{C_x\bar{X}}\right)\right].$$

7. Conclusion

- The proposed almost unbiased dual exponential type estimators of population mean are as

$$t_{de1} = \bar{y}\left[\alpha\exp\left(\frac{\bar{X}-\bar{x}}{p\bar{X}}\right) + (1-\alpha)\exp\left(\frac{\bar{x}-\bar{X}}{p\bar{X}}\right)\right].$$

$$t_{de2} = \bar{y}\left[\beta\exp\left(\frac{\bar{X}-\bar{x}}{q\bar{x}}\right) + (1-\beta)\exp\left(\frac{\bar{x}-\bar{X}}{q\bar{x}}\right)\right].$$

- The proposed estimators are always more efficient than sample mean, ratio estimator of Cochran (1970), product estimator of Robson (1957), exponential ratio and product estimators of Bahl and Tuteja (1991), exponential ratio estimators of Singh *et al.* (2009) and the exponential product type estimators of Onyeka (2013).

- The proposed estimators t_{de1} and t_{de2} are unbiased and can be used for both positively and negatively correlated variables equal efficiently.

References

Bahl, S. and Tuteja, R. K. (1991). Ratio and product type exponential estimators. *Information and Optimization Sciences*, **1**, 159-163.

Cochran, W. G. (1940). The estimation of the yields of the cereal experiments by sampling for the ratio of grain to total produce. *The Journal of Agricultural Science*, **30**, 262-275.

Gujarati, D. N. (2004). *Basic Econometrics*. McGraw-Hill Companies, New York.

Hussain, S., Sharma, M. and Bhat, M. I. J. (2021). Optimum exponential ratio type estimators for estimating the population mean. *Journal of Statistics Applications and Probability Letters*, **8**(**2**), 73-82.

John, E. E and Inyang, E. E. (2015). Efficient exponential ratio estimators for estimating the population mean in simple random sampling. *Hacettepe Journal of Mathematics and Statistics*, **44**(**3**), 689-705.

Khoshnevisan, M., Singh, R., Chauhan, P., Sawan, N. and Smarandache, F. (2007). A general family of estimators for estimating population mean using known value of some population parameters. *Far East Journal of Theoretical Statistics*, **22**, 181–191.

Kumar, B., Kumar, M., Rizvi, S. E. H. and Bhat, M. I. J. (2018). Estimation of population mean through improved class of ratio type estimators using auxiliary information. *Journal of Relability and Statistical Studies*, **11**(**1**), 01-07.

Murthy, M. N. (1967). *Sampling Theory and Methods*. Calcutta India, Statistical Publishing Society.

Onyeka, A. C. (2013). A class of product-type exponential estimators of the population mean in simple random sampling scheme. *Statistics in Transition-New Series*, **14**(**2**), 189-200.

Panigrahi, A. and Mishra, G. (2017). Improved exponential product type estimators of finite population mean. *Open Acess International Journal of Science and Engineering*, **2**(**7**), 25-28.

Robson, D. S. (1957). Applications of multivariate polykays to the theory of unbiased ratio type estimation. *Journal of the American Statistical Association*, **52**, 511–522.

Reddy, V. N. (1974). On a transformed ratio method of estimation. *Sankhya*, **C36**, 59–70.

Sharma, B. and Tailor, R. (2010). A new ratio-cum-dual to ratio estimator of finite population mean in simple random sampling. *Global Journal of Science Frontier Research*, **10**(**1**), 27–31.

Sukhatme, B. and Chand, L. (1977), Multivariate ratio-type estimators in proceedings of the social statistics section. *American Statistical Association, Michigan*, 927–931.

Singh, R., Chauhan, P., Sawan, N. and Smarandache, F. (2007). Improvement in estimating the population mean using exponential estimator in simple random sampling. *Renaissance High Press Ann Arbor USA*, 33-40.

Sisodia, B. V. S. and Dwivedi, V. K. (1981). A modified ratio estimator using coefficient of variation of auxiliary variable. *Journal of the Indian Society of Agricultural Statistics*, **33**(**2**), 13-18.

Singh, H. P. and Tailor, R. (2003). Use of known correlation coefficient in estimating the finite population mean. *Statistics in Transition*, **6**(**4**), 555-560.

Singh, R., Chauhan, P., Sawan, N. and Smarandache, F. (2009). Improvement in estimating the population mean using exponential estimators in simple random sampling. *Bulletin of Statistics and Economics*, **3**(**A09**), 13-18.

Singh, H. P., Upahayaya, L. N. and Tailor, R. (2009a). Ratio-cum-product type exponential estimator. *Statistica*, **69**(**4**), 299-310.

Singh, R., Gupta, S. B. and Malik, S. (2016). Almost unbiased estimator using known value of population parameter(s) in sample surveys. *Journal of Modern Applied Statistical Methods*, **15**(**1**), 600-615.

Singh, S. and Singh, R. (1993). A new method: Almost seperation of bias precipitates in sample survey. *Journal of the Indian Statistical Association*, **31**, 99-105.

Sharma, M. and Bhatnagar, S. (2008). A general class of estimators for estimating population mean through auxiliary information. *Journal of Indian Statistical Association*, **46**(**2**), 155-162.

Sharma, M., Bhatnagar, S. and Imran, K., (2010), A general class of improved ratio type estimators through auxiliary information. *International Journal of Agricultural and Statistical Sciences*, **6**(**1**), 115-118.

Singh, H. P. and Karpe, N. (2010). Estimation of mean, ratio and product using auxiliary information in the presence of measurement errors in sample surveys. *Journal of Statistical Theory and Practice*, **4**(**1**), 111–136.

Singh, H. P. and Vishwakarma, G. K. (2008). Some families of estimators of variance of stratified random sample mean using auxiliary information. *Journal of Statistical Theory and Practice*, **2(1)**, 21–43.

Tailor, R. and Sharma, B. K. (2009). A modified ratio-cum-product estimator of finite population mean using known coefficient of variation and coefficient of kurtosis. *Statistics in Transition-New Series*, **10**(**1**), 15–24.

Tailor, R., Tailor, R., Parmar, R. and Kumar, M. (2012). Dual to ratio-cum-product estimator using known parameters of auxiliary variables. *Journal of Reliability and Statistical Studies*, **5**(**1**), 65–71.

Yasmeen, U., Noor ul Amin, M. and Hanif, M. (2016). Exponential ratio and product type estimators of population mean. *Journal of Statistics and Management System*, **19**(**1**), 55-71.

Yadav, R., Upadhayaya, L. N., Singh, H. P. and Chatterjee, S. (2012). Almost unbiased ratio and product type exponential estimators. *Statistics in Transition-New Series*, **13**(**3**), 537-550.

Yadav, S. K. and Kadilar, C. (2013). Improved class of ratio and product estimators. *Applied Mathematics and Computation*, **219**(**22**), 10726-10731.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 157–175

Theory and Applicability of the Weighted Modified Lindley Distribution

Christophe Chesneau[1], Lishamol Tomy[2] and Jiju Gillariose[3]
[1]*Department of Statistics*
Université de Caen, LMNO, Campus II, Science 3, 14032, Caen, France
[2]*Department of Statistics*
Deva Matha College, Kuravilangad, Kerala, 686633, India
[3]*Department of Statistics*
CHRIST (Deemed to be University), Hosur Road, Bengaluru-560029, India

Received: 06 August 2021; Revised: 07 September 2021; Accepted: 19 October 2021

Abstract

As a bridge between the exponential and Lindley distributions, the modified Lindley distribution was created. It has been used successfully in a variety of fields related to survival analysis. In this study, we present a novel distribution that extends the modified Lindley distribution using the traditional weighted (or length/size-biased) approach. It is named as weighted modified Lindley distribution. This idea is mainly used to flexibilize the former modified Lindley distribution through the use of a one-parameter polynomial weight. This weight is intended to modulate the functionalities of the new distribution, well beyond those of the former modified Lindley distribution. The related probability density function, cumulative density function, hazard rate function, moments, moment generating function and characteristic function are analysed from a theoretical and practical point of view. Estimation of the parameters is done by the classical method of maximum likelihood and a simulation study is carried out to check the consistency of the maximum likelihood estimates. A data set is used to illustrate the application of the proposed distribution.

Key words: Data analysis; Lindley distribution; Estimation; Modified Lindley distribution; Moments; Weighted distributions.

AMS Subject Classifications: 60E05, 62E15

1. Introduction

Lindley is the inventor of the Lindley (L) distribution (see Lindley (1958)). For many statistical settings, the L distribution is established as an alternative to the exponential distribution. It is governed by the following one-parameter cumulative density function (cdf):

$$F_L(x;\upsilon) = 1 - \left[1 + \frac{\upsilon x}{1+\upsilon}\right] e^{-\upsilon x}, \quad x > 0,$$

Corresponding Author: Christophe Chesneau
Email: christophe.chesneau@unicaen.fr

where $\upsilon > 0$, and $F_L(x;\upsilon) = 0$ for $x \leq 0$. Then its probability density function (pdf) is derived as

$$f_L(x;\upsilon) = \frac{\upsilon^2}{1+\upsilon}(1+x)e^{-\upsilon x}, \quad x > 0,$$

and $f_L(x;\upsilon) = 0$ for $x \leq 0$.

Several authors have researched and generalized this distribution during the last few decades. There is a vast literature in this area. Some examples of such distributions include the three-parameter L distribution by Zakerzadeh and Dolati (2009), generalized L distribution by Nadarajah *et al.* (2011), generalized Poisson-L distribution by Mahmoudi and Zakerzadeh (2010), power L distribution by Ghitany *et al.* (2013), two parameter-L distribution by Shanker and Mishra (2013a), quasi L distribution by Shanker and Mishra (2013b), transmuted L distribution by Merovci (2013), transmuted L-geometric distribution by Merovci and Elbatal (2014), beta-L distribution by Merovci and Sharma (2014), negative binomial-L distribution Zamani and Ismail (2010) and gamma-L distribution by Zeghdoudi and Nedjar (2016). For more details, see a comprehensive review study of the L distribution by Tomy (2018).

Among its generalizations, Ghitany *et al.* (2011) introduced the weighted L (WL) distribution, with pdf determined as

$$f_{WL}(x;\alpha,\upsilon) = \Psi_\alpha^{-1} x^{\alpha-1} f_L(x;\upsilon),$$

where $\alpha > 0$, Ψ_α represents the normalizing constant corresponding to the expectation of $X^{\alpha-1}$, X being a random variable with the L distribution with parameter υ. The pdf of the WL distribution can also be expressed as

$$f_{WL}(x;\alpha,\upsilon) = \frac{\upsilon^{\alpha+1}}{(\upsilon+\alpha)\Gamma(\alpha)} x^{\alpha-1}(1+x)e^{-\upsilon x}, \quad x > 0,$$

where $\Gamma(\alpha)$ denotes the Euler gamma function at α, and $f_{WL}(x;\alpha,\upsilon) = 0$ for $x \leq 0$. It is proved that the polynomial weight $x^{\alpha-1}$ modulates the shape properties of the functions of the former L distribution, increasing their capabilities in terms of modeling. As a consequence, the hazard rate function (hrf) of the WL distribution exhibits bathtub or increasing shapes. Furthermore, for some non-grouped or grouped survival data, the WL model is better than several well-known two-parameter survival models.

Recently, an intermediary distribution between the classical exponential and the L distribution has been proposed by Chesneau *et al.* (2019), called the modified L (ML) distribution. Its cdf is specified by

$$F_{ML}(x;\upsilon) = 1 - \left[1 + \frac{\upsilon x}{1+\upsilon} e^{-\upsilon x}\right] e^{-\upsilon x}, \quad x > 0,$$

with $\upsilon > 0$ and $F_{ML}(x;\upsilon) = 0$ for $x \leq 0$, and the related pdf is obtained as

$$f_{ML}(x;\upsilon) = \frac{\upsilon}{1+\upsilon} \left[(1+\upsilon)e^{\upsilon x} + 2\upsilon x - 1\right] e^{-2\upsilon x}, \quad x > 0,$$

and $f_{ML}(x;\upsilon) = 0$ for $x \leq 0$. In Chesneau *et al.* (2019), it is proved that a strong first order stochastic ordering property relates the exponential, L and ML distributions. In this precise mathematical sense, the ML distribution is "sandwiched" between the exponential and L distributions. Also, the hrf of the ML distribution is non-monotonic, contrary to the hrf of the exponential distribution, which is constant, and the one of the L distribution, which is increasing. In addition, an important structural property of the ML distribution is that $f_{ML}(x;\upsilon)$ can be expressed as a linear combination of exponential and gamma pdfs. Furthermore, in Chesneau *et al.* (2019), it is discussed the applicability of the ML model and illustrated its workability via several relevant practical data sets. More recently, Chesneau *et al.* (2020a,c) introduced two extensions of the ML distribution, namely the inverse ML distribution and the wrapped ML distribution, respectively.

The aim of this study is to offer an extension of the ML model that allows for more flexibility in modeling lifetime data. Following the idea of Ghitany *et al.* (2011), we propose the weighted ML (WML) distribution by considering the following weighted pdf:

$$f_{WML}(x;\alpha,\upsilon) = \Phi_\alpha^{-1} x^{\alpha-1} f_{ML}(x;\upsilon),$$

where $\alpha > 0$, Φ_α represents the normalizing constant corresponding to the expectation of $X^{\alpha-1}$, X being a random variable with the ML distribution with parameter υ. After simplifications, we arrive at the following analytical expression:

$$f_{WML}(x;\alpha,\upsilon) = \frac{(2\upsilon)^\alpha}{[(\upsilon+1)2^\alpha+\alpha-1]\Gamma(\alpha)} x^{\alpha-1}\left[(1+\upsilon)e^{\upsilon x}+2\upsilon x-1\right]e^{-2\upsilon x}, \quad x>0, \quad (1)$$

and $f_{WML}(x;\alpha,\upsilon) = 0$ for $x \leq 0$. Thus, the WML distribution is to the ML distribution, what the WL distribution is to the L distribution, with the hope of the same additional benefit from the statistical modelling point of view. This study develops all these aspects, respecting the rules of the art in the field.

The sections of this article are arranged as follows: Section 2 concerns some characteristics and properties of the WML distribution. Section 3 is devoted to the estimation of model parameters as well as real data applications. Section 4 ends the paper with conclusions.

2. Theoretical Work

Some relevant theoretical results on the WML distribution are presented in this section.

2.1. Analysis of the pdf

The pdf of the WML distribution as defined by Equation (1) satisfies the following asymptotic properties. In the case where x tends to be in the neighborhood of 0; an equivalent function is described below:

$$f_{WML}(x;\alpha,\upsilon) \sim \frac{(2\upsilon)^\alpha \upsilon}{[(\upsilon+1)2^\alpha+\alpha-1]\Gamma(\alpha)} x^{\alpha-1}.$$

Hence, we see the importance of the new parameter α is the behavior of this function in 0; When $\alpha < 1$, $f_{WML}(x;\alpha,\upsilon)$ diverges to $+\infty$, when $\alpha = 1$, $f_{WML}(x;\alpha,\upsilon)$ tends to $\upsilon^2/(\upsilon+1)$,

and when $\alpha > 1$, $f_{WML}(x; \alpha, \upsilon)$ tends to 0.

For the behavior at $x \to +\infty$, the following result holds:

$$f_{WML}(x; \alpha, \upsilon) \sim \frac{(2\upsilon)^{\alpha}(1+\upsilon)}{[(\upsilon+1)2^{\alpha}+\alpha-1]\Gamma(\alpha)} x^{\alpha-1} e^{-\upsilon x} \to 0.$$

In this case, the dominant term in the convergence is $e^{-\upsilon x}$; α plays a secondary role. The critical points of $f_{WML}(x; \alpha, \upsilon)$ are the solutions to the following equation: $d \log f_{WML}(x; \alpha, \upsilon)/dx = 0$, which is equivalent to the following analytical equation:

$$(\alpha-1)\frac{1}{x} + \upsilon \frac{(\upsilon+1)e^{\upsilon x}+2}{(1+\upsilon)e^{\upsilon x}+2\upsilon x-1} = 2\upsilon.$$

We see that α only modulates the term $(\alpha-1)/x$, which can be of great impact on the small values of x. The described critical points contain the possible mode of the WML distribution. They are not expressible in the strict mathematical sense, but can be determined numerically via any scientific software.

In order to provide a comprehensive study of the characteristics of $f_{WML}(x; \alpha, \upsilon)$, we end this part with a graphical analysis in Figure 1; it shows the panel of its possible shapes, depending on the conjoint values of the parameters α and υ. From Figure 1, one can observe various kinds of non-monotonic or monotonic shapes, such as reverse J-shaped, right-skewed and unimodal shapes.

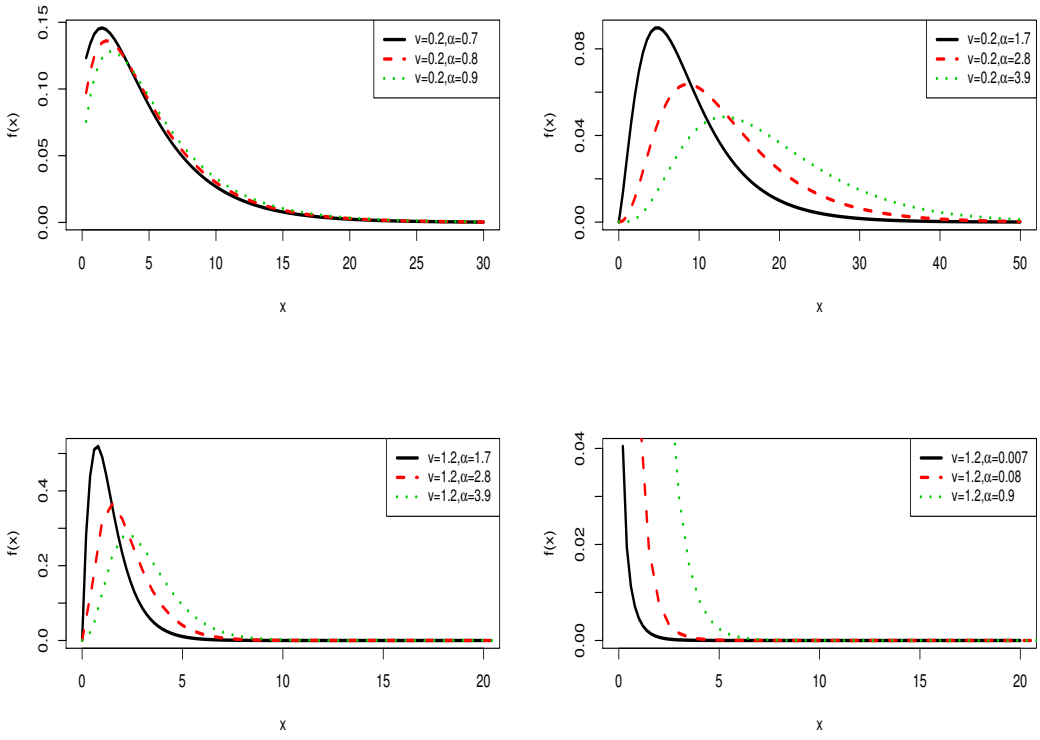


Figure 1: Examples of graphs of the pdf of the WML distribution

2.2. Expression of the cdf

Based on Equation (1), the cdf of the WML distribution can be determined; it can be expressed according to the lower incomplete Euler gamma function defined as $\gamma(s,x) = \int_0^x t^{s-1}e^{-t}dt$ with $s>0$ and $x>0$. Concretely, for any $x>0$, we have

$$\begin{aligned}
F_{WML}(x;\alpha,\upsilon) &= \int_{-\infty}^{x} f_{WML}(t;\alpha,\upsilon)dt \\
&= \frac{(2\upsilon)^\alpha}{[(\upsilon+1)2^\alpha+\alpha-1]\Gamma(\alpha)}\left[(1+\upsilon)\int_0^x t^{\alpha-1}e^{-\upsilon t}dt + 2\upsilon\int_0^x t^{\alpha}e^{-2\upsilon t}dt - \int_0^x t^{\alpha-1}e^{-2\upsilon t}dt\right] \\
&= \frac{(2\upsilon)^\alpha}{[(\upsilon+1)2^\alpha+\alpha-1]\Gamma(\alpha)}\left[\frac{1+\upsilon}{\upsilon^\alpha}\gamma(\alpha,\upsilon x) + \frac{1}{(2\upsilon)^\alpha}\gamma(\alpha+1,2\upsilon x) - \frac{1}{(2\upsilon)^\alpha}\gamma(\alpha,2\upsilon x)\right].
\end{aligned}$$

By using the relation: $\gamma(s+1,x) = s\gamma(s,x) - x^s e^{-x}$, we arrive at the simple expression:

$$\begin{aligned}
&F_{WML}(x;\alpha,\upsilon) = \\
&\frac{1}{[(\upsilon+1)2^\alpha+\alpha-1]\Gamma(\alpha)}\left[(1+\upsilon)2^\alpha\gamma(\alpha,\upsilon x) + (\alpha-1)\gamma(\alpha,2\upsilon x) - (2\upsilon x)^\alpha e^{-2\upsilon x}\right]. \qquad (2)
\end{aligned}$$

For $x \le 0$, we put $F_{WML}(x;\alpha,\upsilon) = 0$.

Some technical comments on this cdf are now given. As expected, by taking $\alpha = 1$, we get

$$F_{WML}(x;\alpha,\upsilon) = \frac{1}{2(\upsilon+1)}\left[2(1+\upsilon)(1-e^{-\upsilon x}) - 2\upsilon x e^{-2\upsilon x}\right] = F_{ML}(x;\upsilon).$$

Moreover, since $-(2\upsilon x)^\alpha e^{-2\upsilon x} < 0$, the following first-order stochastic dominance holds: $F_{WML}(x;\alpha,\upsilon) \le F_{MixG}(x;\alpha,\upsilon)$ for all $x \in \mathbb{R}$, where $F_{MixG}(x;\alpha,\upsilon)$ denotes the following generalized mixture cdf:

$$F_{MixG}(x;\alpha) = \lambda F_G(x;\alpha,\upsilon) + (1-\lambda)F_G(x;\alpha,2\upsilon),$$

where $\lambda = (1+\upsilon)2^\alpha/[(\upsilon+1)2^\alpha+\alpha-1]$ and $F_G(x;\alpha,\upsilon) = \gamma(\alpha,\upsilon x)/\Gamma(\alpha)$ corresponds to the cdf of the classical gamma distribution with parameters α and υ. Note that λ is always positive, but $1-\lambda$ can be negative if $\alpha < 1$. In this case, since $F_G(x;\alpha,\upsilon) \le F_G(x;\alpha,2\upsilon)$, for all $x \in \mathbb{R}$, we have

$$F_{WML}(x;\alpha,\upsilon) \le F_{MixG}(x;\alpha,\upsilon) \le F_G(x;\alpha,\upsilon).$$

Thus, in this case, the WML distribution first-order stochastically dominates the gamma distribution. For $\alpha > 1$, there is no such dominance; the situation is more complex. For illustrative purposes, Figure 2 shows the variation of $F_{WML}(x;\alpha,\upsilon)$ for varying α and υ.

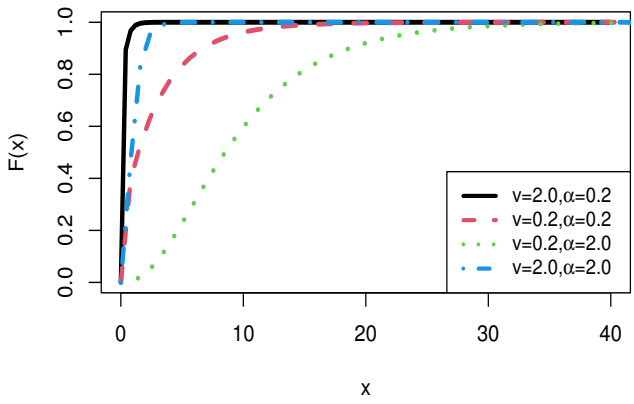


Figure 2: Examples of graphs of the cdf of the WML distribution

Last but not least, the cdf is essential for defining other distributional functions, such as the quantile function (qf) and hrf, which will be the subject of two coming subsections.

2.3. Quantile function

The qf is defined by the inverse function of $F_{WML}(x;\alpha,\upsilon)$, say $F^{-1}_{WML}(u;\alpha,\upsilon)$ with $u \in (0,1)$. In view of Equation (2), it is not possible to express it in an analytical way. However, it is always possible to do a numerical evaluation by giving values for the first quartile (when $u = 1/4$), the median (when $u = 1/2$) and the third quartile (when $u = 3/4$). In addition, this qf has a simple functional lower bound; the following inequality holds: For all $u \in (0,1)$ and $\alpha < 1$, $F^{-1}_{WML}(x;\alpha,\upsilon) \geq F^{-1}_{G}(x;\alpha)$, where $F^{-1}_{G}(x;\alpha,\upsilon)$ denotes the qf of the classical gamma distribution with parameters α and υ defined by $F^{-1}_{G}(x;\alpha,,\upsilon) = \upsilon^{-1}\gamma^{-1}(\alpha, u\Gamma(\alpha))$ with $u \in (0,1)$, $\gamma^{-1}(\alpha,y)$ being the inverse function of $\gamma(\alpha,x)$.

2.4. On the hrf

From $f_{WML}(x;\alpha,\upsilon)$ and $F_{WML}(x;\alpha,\upsilon)$ as given by Equations (1) and (2), respectively, we can present the hrf of the WML distribution by the following ratio function: $h_{WML}(x;\alpha,\upsilon) = f_{WML}(x;\alpha,\upsilon)/[1 - F_{WML}(x;\alpha,\upsilon)]$. When $x > 0$, it is given as

$$h_{WML}(x;\alpha,\upsilon) = \frac{(2\upsilon)^{\alpha}x^{\alpha-1}\left[(1+\upsilon)e^{\upsilon x} + 2\upsilon x - 1\right]e^{-2\upsilon x}}{[(\upsilon+1)2^{\alpha}+\alpha-1]\Gamma(\alpha) - (1+\upsilon)2^{\alpha}\gamma(\alpha,\upsilon x) - (\alpha-1)\gamma(\alpha,2\upsilon x) + (2\upsilon x)^{\alpha}e^{-2\upsilon x}}.$$

and $h_{WML}(x;\alpha,\upsilon) = 0$ for $x \leq 0$. The possible shapes of $h_{WML}(x;\alpha,\upsilon)$ are of great interest in understanding the modelling capability of the WML model (see, Aarset (1987)). Since the expression of $h_{WML}(x;\alpha,\upsilon)$ is mathematically complex, we conduct a visual analysis in Figure 3, showing the diverse shapes possessed by this model. From Figure 3, it is clear that hrf has various kinds of non-monotonic shapes, such as reverse J-shaped, reversed N-shaped, right-skewed and unimodal shapes, which makes the proposed distribution more flexible to

fit different data sets. As we know, the L and ML distributions have only unimodal hrf. Hence, the WML distribution is more flexible than its parent distributions, such as the L and ML distributions.

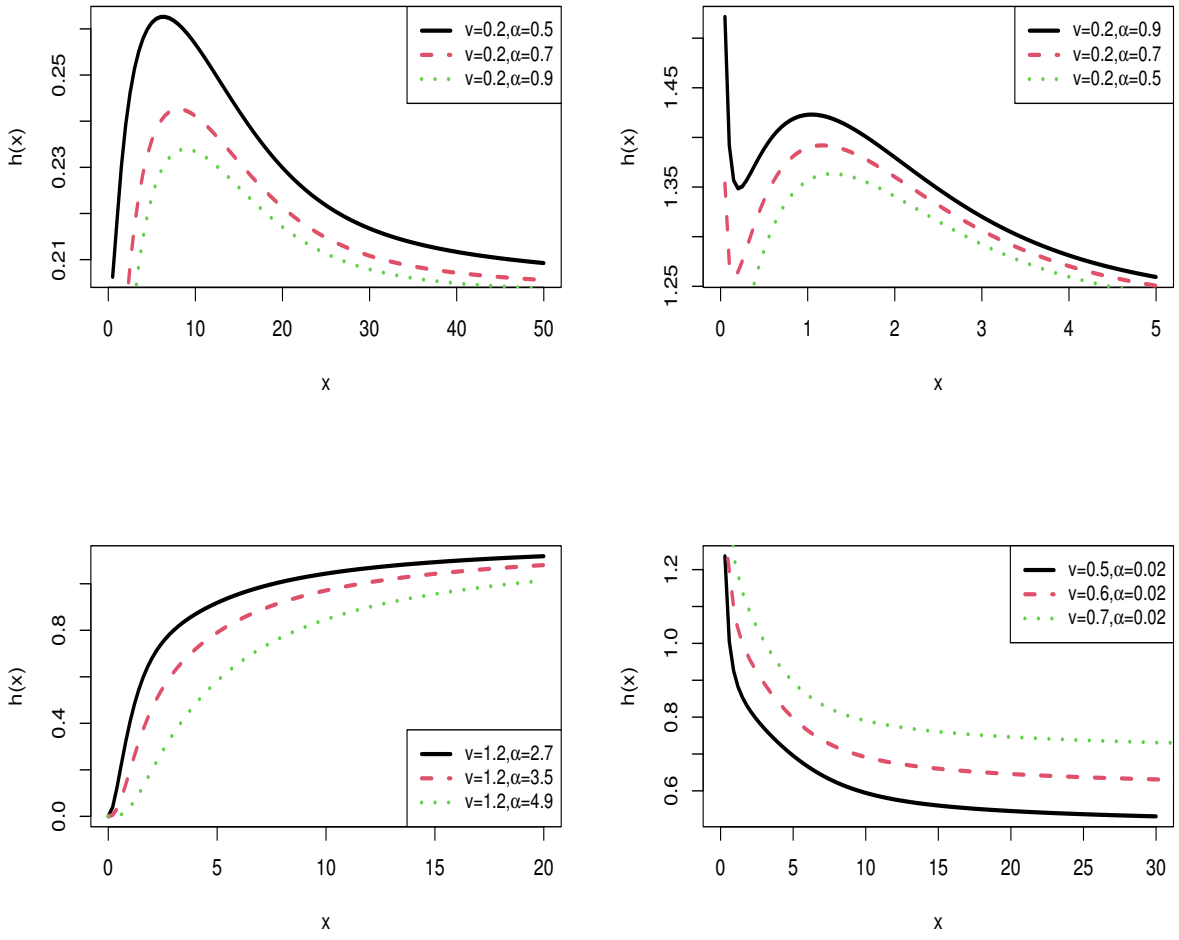


Figure 3: Examples of graphs of the hrf of the WML distribution

2.5. Mathematical moments

In this section, we study the useful moment characteristics and measures of the WML distribution. Let X be a random variable that follows the WML distribution. Besides, we discuss the incomplete moments of X, from which we derive moments and discuss some related quantities. The moment generating and characteristic functions are also expressed.

2.5.1. Incomplete moments

As a first information, the sth incomplete moment of X exists, and it is classically defined by $m_s(x) = E(X^s I(X \leq x))$, where E denotes the mathematical expectation operator.

Therefore, by taking into account the definition of $f_{WML}(x;\alpha,\upsilon)$, the integral definition of $m_s(x)$ becomes

$$\begin{aligned}
m_s(x) &= \int_0^x t^s f_{WML}(t;\alpha,\upsilon)dt \\
&= \frac{(2\upsilon)^\alpha}{[(\upsilon+1)2^\alpha+\alpha-1]\Gamma(\alpha)}\left[(1+\upsilon)\int_0^x t^{s+\alpha-1}e^{-\upsilon t}dt + 2\upsilon\int_0^x t^{s+\alpha}e^{-2\upsilon t}dt - \int_0^x t^{s+\alpha-1}e^{-2\upsilon t}dt\right] \\
&= \frac{(2\upsilon)^\alpha}{[(\upsilon+1)2^\alpha+\alpha-1]\Gamma(\alpha)}\times \\
&\left[\frac{1+\upsilon}{\upsilon^{s+\alpha}}\gamma(s+\alpha,\upsilon x) + \frac{1}{(2\upsilon)^{s+\alpha}}\gamma(s+\alpha+1,2\upsilon x) - \frac{1}{(2\upsilon)^{s+\alpha}}\gamma(s+\alpha,2\upsilon x)\right].
\end{aligned}$$

Since $\gamma(s+\alpha+1,2\upsilon x) = (s+\alpha)\gamma(s+\alpha,2\upsilon x) - (2\upsilon x)^{s+\alpha}e^{-2\upsilon x}$, the sth incomplete moment is reduced to

$$\begin{aligned}
m_s(x) = &\frac{1}{(2\upsilon)^s[(\upsilon+1)2^\alpha+\alpha-1]\Gamma(\alpha)}\times \\
&\left[(1+\upsilon)2^{s+\alpha}\gamma(s+\alpha,\upsilon x) + (s+\alpha-1)\gamma(s+\alpha,2\upsilon x) - (2\upsilon x)^{s+\alpha}e^{-2\upsilon x}\right]. \qquad (3)
\end{aligned}$$

From this expression, by taking $s=0$, we logically obtain the expression of $F_{WML}(x;\alpha,\upsilon)$. Furthermore, some uses of this manageable expression are described below. Also, by taking $\alpha=1$, we rediscover the sth incomplete moment of a random variable with the former ML distribution.

2.5.2. Ordinary moments and related measures

Also, the ordinary moments of X can be easily obtained by applying $x\to+\infty$ in Equation (3). That is, the sth ordinary moment of X is given as

$$m_s = m_s(+\infty) = \frac{1}{(2\upsilon)^s[(\upsilon+1)2^\alpha+\alpha-1]\Gamma(\alpha)}\left[(1+\upsilon)2^{s+\alpha}+s+\alpha-1\right]\Gamma(s+\alpha).$$

In particular, by using the relation: $\Gamma(x+1)=x\Gamma(x)$ for $x>0$, the four first ordinary moments of X are

$$m_1 = \frac{\alpha\left[(1+\upsilon)2^{1+\alpha}+\alpha\right]}{2\upsilon[(\upsilon+1)2^\alpha+\alpha-1]}, \quad m_2 = \frac{\alpha(\alpha+1)\left[(1+\upsilon)2^{2+\alpha}+\alpha+1\right]}{(2\upsilon)^2[(\upsilon+1)2^\alpha+\alpha-1]},$$

$$m_3 = \frac{\alpha(\alpha+1)(\alpha+2)\left[(1+\upsilon)2^{3+\alpha}+\alpha+2\right]}{(2\upsilon)^3[(\upsilon+1)2^\alpha+\alpha-1]}$$

and

$$m_4 = \frac{\alpha(\alpha+1)(\alpha+2)(\alpha+3)\left[(1+\upsilon)2^{4+\alpha}+\alpha+3\right]}{(2\upsilon)^4[(\upsilon+1)2^\alpha+\alpha-1]}.$$

The classical central and dispersion moment parameters of X follow immediately, including the mean given as $m = m_1$, variance given as $V = m_2 - m_1^2$, coefficient of variation $\sqrt{V}/m$, as well as the skewness and kurtosis coefficients obtained as

$$S = \frac{m_3 - 3m_2m + 2m^3}{V^{3/2}}$$

and

$$K = \frac{m_4 - 4m_3m + 6m_2m^2 - 3m^4}{V^2},$$

respectively. The numerical pliancy of these important probabilistic measures is shown in Table 1. Since the skewness has positive values, the WML distribution is skewed to the right. In addition, the WML distribution can be platykurtic (when $K < 3$) and leptokurtic (when $K > 3$). Furthermore, the mean of the proposed distribution can be smaller or greater than its variance.

2.5.3. Some related functions

Incomplete and ordinary moments are the main ingredients of various functions or indexes that are useful in various applied areas. For instance, from the first incomplete and ordinary moments, one can express the mean residual function given as

$$\begin{aligned} M_{WML}(x) &= E(X - x \mid X > x) \\ &= \frac{m_1 - m_1(x)}{1 - F_{WML}(x;\alpha,\upsilon)} - x \\ &= \frac{\left[(1+\upsilon)2^{1+\alpha} + \alpha\right]\Gamma(\alpha+1) - (1+\upsilon)2^{1+\alpha}\gamma(\alpha+1,\upsilon x) - \alpha\gamma(1+\alpha,2\upsilon x) + (2\upsilon x)^{1+\alpha}e^{-2\upsilon x}}{2\upsilon[(\upsilon+1)2^{\alpha} + \alpha - 1]\Gamma(\alpha) - (1+\upsilon)2^{\alpha}\gamma(\alpha,\upsilon x) - (\alpha-1)\gamma(\alpha,2\upsilon x) + (2\upsilon x)^{\alpha}e^{-2\upsilon x}} - x \end{aligned}$$

and the mean reversed residual function defined as

$$\begin{aligned} M_{WML}^{rev}(x) &= E(x - X \mid X \leq x) \\ &= x - \frac{m_1(x)}{F_{WML}(x;\alpha,\upsilon)} \\ &= x - \frac{1}{2\upsilon}\frac{(1+\upsilon)2^{1+\alpha}\gamma(1+\alpha,\upsilon x) + \alpha\gamma(1+\alpha,2\upsilon x) - (2\upsilon x)^{1+\alpha}e^{-2\upsilon x}}{(1+\upsilon)2^{\alpha}\gamma(\alpha,\upsilon x) + (\alpha-1)\gamma(\alpha,2\upsilon x) - (2\upsilon x)^{\alpha}e^{-2\upsilon x}}. \end{aligned}$$

In terms of reliability and life testing, these functions play a crucial role. See, for instance, Barlow and Proschan (1975) and Nanda *et al.* (2003). They can be used in the setting of the WML distribution for further purposes in this direction.

2.5.4. Moment functions

Based on Equation (1), the moment generating and characteristic functions of the WML distribution can be obtained using the lower incomplete Euler gamma function. Indeed, for

any $x < \upsilon$, we have

$$\begin{aligned}
R_{WML}(x) &= E(e^{xX}) = \int_0^{+\infty} e^{xt} f_{WML}(t;\alpha,\upsilon)dt \\
&= \frac{(2\upsilon)^\alpha}{[(\upsilon+1)2^\alpha+\alpha-1]\Gamma(\alpha)} \times \\
&\left[(1+\upsilon)\int_0^{+\infty} t^{\alpha-1}e^{-(\upsilon-x)t}dt + 2\upsilon\int_0^{+\infty} t^{\alpha}e^{-(2\upsilon-x)t}dt - \int_0^{+\infty} t^{\alpha-1}e^{-(2\upsilon-x)t}dt\right] \\
&= \frac{(2\upsilon)^\alpha}{[(\upsilon+1)2^\alpha+\alpha-1]\Gamma(\alpha)}\left[\frac{1+\upsilon}{(\upsilon-x)^\alpha}\Gamma(\alpha) + \frac{2\upsilon}{(2\upsilon-x)^{\alpha+1}}\Gamma(\alpha+1) - \frac{1}{(2\upsilon-x)^\alpha}\Gamma(\alpha)\right] \\
&= \frac{(2\upsilon)^\alpha}{(\upsilon+1)2^\alpha+\alpha-1}\left[\frac{1+\upsilon}{(\upsilon-x)^\alpha} + \frac{2\upsilon(\alpha-1)+x}{(2\upsilon-x)^{\alpha+1}}\right].
\end{aligned}$$

By taking $\alpha = 1$, we rewrite the moment generating function of the former ML distribution. By using the standard formula, we have $m_s = R_{WML}(x)^{(s)} \mid_{x=0}$. Also, the rth cumulant of X can be obtained through the following equation: $\kappa_s = \{\log R_{WML}(x)\}^{(s)} \mid_{x=0}$. Also, the characteristic function of X is immediately deduced from $R_{WML}(x)$; it is given as

$$\Psi_{WML}(x) = E(e^{ixX}) = \frac{(2\upsilon)^\alpha}{(\upsilon+1)2^\alpha+\alpha-1}\left[\frac{1+\upsilon}{(\upsilon-ix)^\alpha} + \frac{2\upsilon(\alpha-1)+ix}{(2\upsilon-ix)^{\alpha+1}}\right], \quad x \in \mathbb{R},$$

where $i^2 = -1$. This function fully characterizes the WML distribution, and can be used for further results in distribution involving the WML distribution.

3. Estimation and Application

In this section, we will discuss estimation and its applicability in a concrete data analysis scenario.

3.1. Parametric estimation

The maximum likelihood (MaxLik) method can be applied to obtain efficient estimates of the WML model parameters. In this context, what is necessary is specified below. Let $x_1, \ldots, x_n$ be realizations of n independent random variables, all distributed following the WML distribution with parameters α and υ. Then, the estimates suggested by the MaxLik method are given by the arguments of the maxima of the likelihood function, or the log-likelihood function defined by

$$\begin{aligned}
\ell(\alpha,\upsilon) = \sum_{i=1}^n \log f_{WML}(x_i;\alpha,\upsilon) &= n\alpha\log 2 + n\alpha\log\upsilon - n\log\Gamma(\alpha) - n\log[(\upsilon+1)2^\alpha+\alpha-1] \\
&+ (\alpha-1)\sum_{i=1}^n \log x_i - 2\upsilon\sum_{i=1}^n x_i + \sum_{i=1}^n \log\left[(1+\upsilon)e^{\upsilon x_i} + 2\upsilon x_i - 1\right].
\end{aligned}$$

The maximum likelihood estimates (MaxLikEs) are denoted by $\hat{\alpha}$ and $\hat{\upsilon}$, satisfying $\ell(\alpha,\upsilon) \leq \ell(\hat{\alpha},\hat{\upsilon})$ for any $\alpha > 0$ and $\upsilon > 0$, by construction. They are also the solutions of the two

following equations with respect to the parameters:

$$\frac{\partial}{\partial\alpha}\ell(\alpha,\upsilon) = n\log 2 + n\log\upsilon - n\frac{\partial\Gamma(\alpha)/\partial\alpha}{\Gamma(\alpha)} - n\frac{(\upsilon+1)2^{\alpha}\log(2)+1}{(\upsilon+1)2^{\alpha}+\alpha-1} + \sum_{i=1}^{n}\log x_i = 0$$

and

$$\frac{\partial}{\partial\upsilon}\ell(\alpha,\upsilon) = n\frac{\alpha}{\upsilon} - n\frac{2^{\alpha}}{(\upsilon+1)2^{\alpha}+\alpha-1} - 2\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\frac{x_i[(\upsilon+1)e^{\upsilon x_i}+2]+e^{\upsilon x_i}}{(1+\upsilon)e^{\upsilon x_i}+2\upsilon x_i-1} = 0.$$

Explicit formulations for $\hat{\alpha}$ and $\hat{\upsilon}$ are not possible due to the intricacy of these equations. As a result, numerical methods involving Newton-type algorithms must be used to solve them. Alternatively, one can investigate the maximization of $\ell(\alpha,\upsilon)$ numerically through specific functions in the `R` package, such as the `constrOptim` function, `optim` function or `maxLik` function.

The theory of MaxLikEs ensures that $\hat{\alpha}$ and $\hat{\upsilon}$ are efficient in several senses, including their fast numerical convergence to the underlying true values of the parameters. Other important properties are described in Casella and Berger (1990).

Using the asymptotic normal distribution of the MaxLikEs, we can evaluate the confidence intervals (CIs) of unknown parameters. In this regard, the observed Fisher information matrix $I(\alpha,\upsilon)$ formed of the negative second derivatives of the log-likelihood function must be determined. In this asymptotic framework, the $100(1-\gamma)\%$ CI for α is defined by the interval with the following lower bound (LB) and upper bound (UB):

$$LB = \hat{\alpha} \pm z_{\gamma/2}\sqrt{I_{\hat{\alpha}\hat{\alpha}}}, \quad UB = \hat{\alpha} \pm z_{\gamma/2}\sqrt{I_{\hat{\alpha}\hat{\alpha}}},$$

where $z_{\gamma/2}$ is the percentile of the standard normal distribution with right tail probability $\gamma/2$, and $I_{\hat{\alpha}\hat{\alpha}}$ is the first diagonal component of $I^{-1}(\hat{\alpha},\hat{\upsilon})$. The same holds for the parameter υ by the consideration of $\hat{\upsilon}$ instead of $\hat{\alpha}$, and $I_{\hat{\upsilon}\hat{\upsilon}}$ instead of $I_{\hat{\alpha}\hat{\alpha}}$.

3.2. Simulation study

We are now conducting a simulation research to assess the performance of the MaxLikEs of the parameters of the WML distribution. The Newton formula is used because the qf of this distribution is not available in closed form. The simulation experiment was repeated 1000 times with sample sizes of 25, 80 and 150 from the WML distribution. The assessment was based on the following steps of simulation study:

1. Generate 1000 samples of size N.
2. Assign the sample size of n and the values of the parameters.
3. Assign the initial value for the random start y_0.
4. For $j = 1,\ldots,n$, generate u_j from a random variable U_j following the unit uniform distribution.

5. Change y_0 by y^* by using the Newton formula as follows:

$$y^* = y_0 - \left\{ \frac{F_{WML}(y_0; \alpha, \upsilon) - u_j}{f_{WML}(y_0; \alpha, \upsilon)} \right\}.$$

6. If $|y_0 - y^*| \leq \epsilon$ for small $\epsilon > 0$, ϵ being considered as a tolerance limit, then $y = y^*$ is considered as a generated value from the WML distribution with parameters α and υ, else set $y_0 = y^*$ and go to the previous step.

7. Repeat steps 4 to 6 for $j = 1, \ldots, n$ to obtain n values $y_1, \ldots, y_n$.

8. Compute the MaxLikEs of α and υ from $y_1, \ldots, y_n$.

9. Repeat steps 2 to 8, N times.

10. Compute the Bias and mean square error (MSE) for each parameter, defined as

$$\text{Bias}(\alpha) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_i - \alpha), \quad \text{MSE}(\alpha) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_i - \alpha)^2,$$
$$\text{Bias}(\upsilon) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\upsilon}_i - \upsilon), \quad \text{MSE}(\upsilon) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\upsilon}_i - \upsilon)^2,$$

where $\hat{\alpha}_i$ and $\hat{\upsilon}_i$ are the MaxLikEs of α and υ, respectively, obtained at the i^{th} replication.

The parameter combinations are given below:

1. $\alpha = 1.5$, $\upsilon = 1.5$
2. $\alpha = 2$, $\upsilon = 3.5$
3. $\alpha = 3.5$, $\upsilon = 2.5$
4. $\alpha = 4$, $\upsilon = 2.5$

Table 2 presents the Bias, MSE, LB and UB related to the CIs of the parameters for different sample sizes. The Bias and MSE decrease as n increases. As a result, the MaxLik approach for estimating the parameters of the WML distribution using Bias and MSE works fairly well.

3.3. Application

This portion contains an application of the WML distribution to real lifetime data. To demonstrate the potential of the WML distribution, a comparison is made using two-parameter extensions or modifications of the L distribution, which are the WL distribution by (Ghitany *et al.*, 2011) and some other extended L distributions. Below is a list of the competing distributions.

1. The quasi L (QL) distribution (Shanker and Mishra, 2013b) with pdf

$$f(x;\alpha,\upsilon)=\frac{\upsilon(\alpha+x\upsilon)}{\alpha+1}e^{-\upsilon x},\quad x>0,$$

and $f(x;\alpha,\upsilon)=0$ for $x\leq 0$.

2. The two-parameter L (SL) distribution (Shanker and Mishra, 2013a) with pdf,

$$f(x;\alpha,\upsilon)=\frac{\upsilon^2}{\alpha\upsilon+1}(\alpha+x)e^{-\upsilon x},\quad x>0,$$

and $f(x;\alpha,\upsilon)=0$ for $x\leq 0$.

3. The exponentiated L (EL) distribution (see, Cordeiro *et al.*, 2013) with pdf,

$$f(x;\alpha,\upsilon)=\frac{\alpha\upsilon^2}{\upsilon+1}e^{-\upsilon x}(1+x)\left[1-\left(1+\frac{\upsilon x}{1+\upsilon}\right)e^{-\upsilon x}\right]^{\alpha-1},\quad x>0,$$

and $f(x;\alpha,\upsilon)=0$ for $x\leq 0$.

4. The power L (PL) distribution (Ghitany *et al.*, 2013) with pdf,

$$f(x;\alpha,\upsilon)=\frac{\alpha\upsilon^2}{\upsilon+1}(1+x^{\alpha})x^{\alpha-1}e^{-\upsilon x^{\alpha}},\quad x>0,$$

and $f(x;\alpha,\upsilon)=0$ for $x\leq 0$.

For the pdfs above, it is supposed that $\upsilon>0$ and $\alpha>0$.

The MaxLik method is applied to estimate the unknown parameters, along with the determination of the related standard errors (SEs). The following criteria are used to choose the best-fitting distribution: negative maximized Log-likelihood value ($-\log L$), Akaike information criterion (AIC) and Bayesian information criterion (BIC). The value of the Kolmogorov-Smirnov (K-S) statistic and the p-value are also provided.

The real data set corresponds to the life of a fatigue fracture of Kevlar 373/epoxy that was subjected to steady pressure (at 90% stress) until it failed. Therefore, we have comprehensive data with accurate failure periods. The data set has been obtained from Barlow *et al.* (1984) and Andrews and Herzberg (1985). For previous studies on the data, see Chesneau *et al.* (2020a).

The values of this data set are: 0.0251, 0.0886, 0.0891, 0.2501, 0.3113, 0.3451, 0.4763, 0.5650, 0.5671, 0.6566, 0.6748, 0.6751, 0.6753, 0.7696, 0.8375, 0.8391, 0.8425, 0.8645, 0.8851, 0.9113, 0.9120, 0.9836, 1.0483, 1.0596, 1.0773, 1.1733, 1.2570, 1.2766, 1.2985, 1.3211, 1.3503, 1.3551, 1.4595, 1.4880, 1.5728, 1.5733, 1.7083, 1.7263, 1.7460, 1.7630, 1.7746, 1.8275, 1.8375, 1.8503, 1.8808, 1.8878, 1.8881, 1.9316, 1.9558, 2.0048, 2.0408, 2.0903, 2.1093, 2.1330, 2.2100, 2.2460, 2.2878, 2.3203, 2.3470, 2.3513, 2.4951, 2.5260, 2.9911, 3.0256, 3.2678, 3.4045, 3.4846, 3.7433, 3.7455, 3.9143, 4.8073, 5.4005, 5.4435, 5.5295, 6.5541, 9.0960

The findings of a descriptive evaluation of the fitted models for the data set are shown in Table 3. The R program is used to perform the necessary calculations.

Based on the goodness-of-fit measures, the smallest $-\log L$, AIC, BIC, K-S statistics and the highest p-values are obtained for the WML distribution. These observations indicate that the WML model provides the best fit for the data set. Moreover, from the study, the competing distributions can be ranked in the following order (best to the least): EL distribution, SL distribution, WL distribution, QL distribution, and PL distribution.

As a graphical approach, in Figure 4, we present the estimated pdfs against the fitted pdfs. In addition, the empirical cdf against the fitted cdfs is also given in Figure 5. From these figures, we see that the two fits of the estimated functions of the WML model have well captured the forms or curvatures of the empirical objects, confirming the previous numerical analysis.

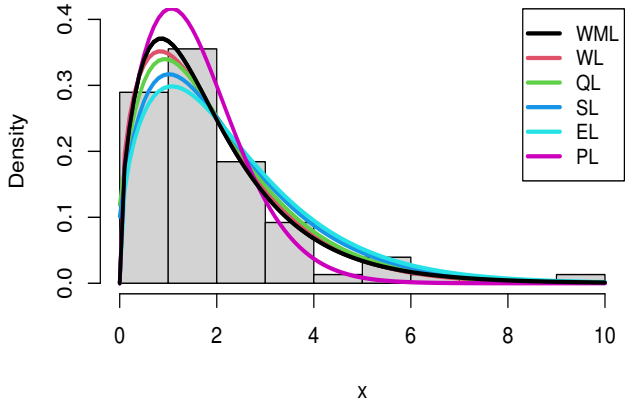


Figure 4: Graphs of the estimated pdfs of the considered distributions

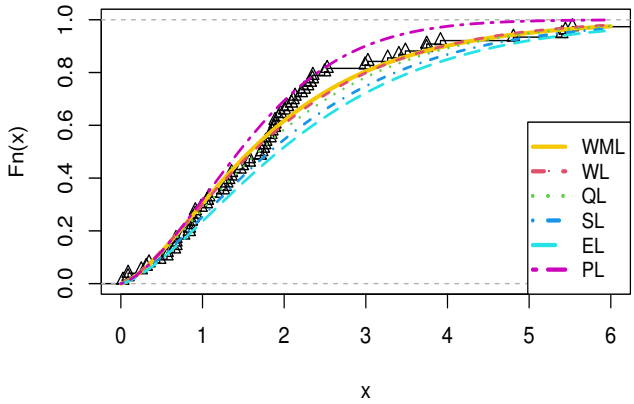


Figure 5: Graphs of the estimated cdfs of the considered distributions

4. Conclusions

In this paper, we introduced a weighted scheme for the modified L distribution, referred to as the weighted modified Lindley distribution. The main motivations for introducing this new distribution are provided. Various shapes of pdf and hrf, which are attractive for statistical modeling, are highlighted. In particular, we have exhibited that the pdf and hrf can be unimodal and monotonically decreasing. In addition, detailed and elegant discussions of incomplete moments, ordinary moments with their related measures, moment generating function and characteristic function are given. Parameter estimation is approached by the use of the maximum likelihood function in a simulation study. The usefulness of the new distribution is illustrated in an analysis of real data. Thus, the proposed model can be used quite effectively for analysing lifetime data.

Acknowledgements

We thank our students for participating in the classroom activity which eventually led to this paper. We are grateful to an anonymous referee who suggested many improvements and furthermore generously listed many useful references. Finally, we thank the Chair Editor for his encouragement, guidance and counsel.

References

Aarset, M. V. (1987). How to identify bathtub hazard rate. *IEEE Transactions and Reliability*, **36**, 106-108.

Andrews, D. F. and Herzberg, A. M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. Springer Series in Statistics, New York.

Barlow, R. E. and Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing: Probability Models*. Holt, Rinehart, and Winston, New York.

Barlow, R. E., Toland, R. H. and Freeman, T. (1984). *A Bayesian analysis of stress-rupture life of kevlar 49/epoxy spherical pressure vessels*. In: Proceedings of Canadian Conference Applied Statistics, Marcel Dekker, New York.

Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Brooks/Cole Publishing Company: Bel Air, CA, USA.

Chesneau, C., Tomy, L. and Gillariose, J. (2019). A new modified Lindley distribution with properties and applications. *Journal of Statistics and Management Systems*, DOI: 10.1080/09720510.2020.1824727.

Chesneau, C., Tomy, L., Gillariose, J. and Jamal, F. (2020a). The inverted modified Lindley distribution. *Journal of Statistical Theory and Practice*, **14**, 1-17.

Chesneau, C., Tomy, L. and Gillariose, J. (2020b). On a sum and difference of two Lindley distributions: theory and applications. *REVSTAT- Statistical Journal*, **18**, 673-695.

Chesneau, C., Tomy, L. and Jose, M. (2020c). Wrapped modified Lindley distribution. *Journal of Statistics and Management Systems*, DOI: 10.1080/09720510.2020.1796313.

Cordeiro, G. M., Ortega, E. M. and Cunha, D. C. C. (2013). The exponentiated generalized class of distributions. *Journal of Data Science*, **11**, 1-27.

Ghitany, M. E., Alqallaf, F., Al-Mutairi, D. K. and Husain, H. A. (2011). A two-parameter weighted Lindley distribution and its applications to survival data. *Mathematics and Computers in Simulation*, **81**, 1190-1201.

Ghitany, M. E., Al-Mutairi, D. K., Balakrishnan, N. and Al-Enezia, L. J. (2013). Power Lindley distribution and associated inference. *Computational Statistics and Data Analysis*, **6**, 20-33.

Lindley, D. V. (1958). Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society*, **20**, 102-107.

Mahmoudi, E. and Zakerzadeh, H. (2010). Generalized Poisson-Lindley distribution. *Communications in Statistics - Theory and Methods*, **39**, 1785-1798.

Merovci, F. (2013). Transmuted Lindley distribution. *International Journal of Open Problems in Computer Science and Mathematics*, **6**, 63-72.

Merovci, F. and Elbatal, I. (2014). Transmuted Lindley-geometric distribution and its applications. *Journal of Statistics Applications & Probability*, **3,** 77-91.

Merovci, F. and Sharma, V. K. (2014). The beta Lindley distribution: Properties and applications. *Journal of Applied Mathematics*, **2014**, 1-10.

Nadarajah, S., Bakouch, H. and Tahmasbi, R. (2011). A generalized Lindley distribution. *Sankhya B-Applied and Interdisciplinary Statistics*, **73**, 331-359.

Nanda, A. K., Singh, H., Misra, N. and Paul, P. (2003). Reliability properties of reversed residual lifetime. *Communications in Statistics-Theory and Methods*, **32**, 2031-2041.

Shanker, R. and Mishra, A. (2013a). A two parameter Lindley distribution. *Statistics in Transition New Series*, **14**, 45-56.

Shanker, R. and Mishra, A. (2013b). A quasi Lindley distribution. *African Journal of Mathematics and Computer Science Research*, **6**, 64-71.

Tomy, L. (2018). A retrospective study on Lindley distribution. *Biometrics and Biostatistics International Journal*, **7**, 163-169.

Zakerzadeh, H. and Dolati, A. (2009). Generalized Lindley distribution. *Journal of Mathematical Extension*, **3**, 13-25.

Zamani, H. and Ismail, N. (2010). Negative binomial-Lindley distribution and its application. *Journal of Mathematics and Statistics*, **6**, 4-9.

Zeghdoudi, H. and Nedjar, S. (2016). Gamma-Lindley distribution and its application. *Journal of Applied Probability and Statistics*, **11**, 129-138.

ANNEXURE

Table 1: Some numerical values of moment measures of the WML distribution

$(\alpha, v) \rightarrow$	$(2, 0.02)$	$(2, 2)$	$(0.2, 0.2)$	$(0.2, 2)$	$(0.75, 70)$
m	0.0015	1.0000	0.5182	9.9910	64.7160
m_2	0.0008	1.4711	3.1704	143.5341	5751.618
m_3	0.0008	2.8846	31.9664	2741.3372	514672.4
m_4	0.0012	7.1034	461.7491	66105.4505	46346481
V	7.9357	0.4712	2.9019	43.5341	1563.458
S	11.1390	1.4569	5.5259	1.5155	0.9691
K	185.9511	6.2699	47.5460	6.6337	2.0563

Table 2: Simulation results related to the parameters of the WML distribution

Combinations	n		Bias	MSE	LB	UB
$\alpha = 1.5$, $\upsilon = 1.5$	100	α	0.0585	0.0751	1.4990	1.6110
		υ	0.0644	0.0780	1.4981	1.6177
	200	α	0.0295	0.0346	1.4941	1.5550
		υ	0.0336	0.0348	1.4962	1.5591
	500	α	0.0169	0.0138	1.49067	1.5271
		υ	0.0173	0.0132	1.4973	1.5273
$\alpha = 2$, $\upsilon = 3.5$	100	α	0.0635	0.1055	1.9999	2.1259
		υ	0.1309	0.3794	3.4981	3.7489
	200	α	0.0352	0.0477	1.9988	2.0651
		υ	0.0677	0.1700	3.4911	3.6241
	500	α	0.0120	0.0207	1.9994	2.0246
		υ	0.0215	0.0704	3.4913	3.5447
$\alpha = 3.5$, $\upsilon = 2.5$	100	α	0.1240	0.3002	3.4993	3.7287
		υ	0.0947	0.1818	2.4992	2.6762
	200	α	0.0453	0.1397	3.4930	3.5968
		υ	0.0450	0.0887	2.4988	2.5858
	500	α	0.0103	0.0555	3.4897	3.5310
		υ	0.0146	0.0325	2.4942	2.5303
$\alpha = 4$, $\upsilon = 2.5$	100	α	0.1163	0.3703	3.9992	4.2334
		υ	0.0826	0.1625	2.4905	2.6600
	200	α	0.0556	0.1778	3.9976	4.1135
		υ	0.0368	0.0839	2.4970	2.5767
	500	α	0.0162	0.0667	3.9936	4.0388
		υ	0.0135	0.0307	2.4982	2.5288

Table 3: Descriptive evaluation of the fitted models for the data set

Model	MaxLikE (SE)	−logL	AIC	BIC	K-S	p-value
WML	$\hat{\upsilon} = 0.7020$ (0.1303) $\hat{\alpha} = 1.2723$ (0.2657)	121.4213	246.8426	251.5041	0.0931	0.4965
WL	$\hat{\upsilon} = 1.0007$ (0.1469) $\hat{\alpha} = 1.3809$ (0.2339)	122.0275	248.055	252.7164	0.10413	0.3573
QL	$\hat{\upsilon} = 0.9543$ (0.0954) $\hat{\alpha} = 0.1498$ (0.1437)	121.6503	247.3006	251.962	0.13049	0.1374
SL	$\hat{\upsilon} = 0.9544$ (0.0954) $\hat{\alpha} = 6.3676$ (6.4571)	121.6503	247.3006	251.962	0.10247	0.3765
EL	$\hat{\upsilon} = 9364$ (0.1047) $\hat{\alpha} = 1.3905$ (0.2376)	121.8991	247.7981	252.4596	0.10221	0.3796
PL	$\hat{\upsilon} = 0.7046$ (0.0819) $\hat{\alpha} = 1.1425$ (0.0908)	122.4001	248.8001	253.4616	0.11233	0.2719

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 177-188

Bayesian Credible Intervals for Generalized Inverse Weibull Distribution

Kamaljit Kaur[1], Sangeeta Arora[2] and Kalpana K. Mahajan[3]
[1]Assistant Professor, SGGS College, Sector-26, Chandigarh -160019, India
[2,3]Professor, Department of Statistics, Panjab University, Chandigarh-160014, India

Received: 15 May 2021; Revised: 24 July 2021; Accepted: 24 October 2021

Abstract

Bayesian credible intervals are obtained for Generalized Inverse Weibull distribution using different priors. Gibbs sampling procedure is used to draw Markov Chain Monte Carlo (MCMC) samples which are used to construct the Bayesian estimates and corresponding credible intervals. Simulation study is conducted by taking different configurations of parameter points and sample sizes to highlight the properties and comparison of the credible intervals. Illustrative example based on a real data set is also provided.

Key words: Generalized inverse Weibull distribution; Credible interval; MCMC algorithm; Posterior distribution.

1. Introduction

The three-parameter Generalized Inverse Weibull distribution (GIWD), introduced by Gusmao *et al.* (2011), is a positively skewed distribution used to model the income data and because of its ability of possessing decreasing and unimodal failure rate, is also useful in reliability and biological studies. Generalized inverse Weibull distribution is the generalization of various well-known and useful distributions, including inverse Weibull, inverse exponential, inverse Rayleigh and Fréchet distributions as special sub-models.

These distributions play an important role in many applications, including the dynamic components of diesel engines, several data sets such as the times to breakdown of an insulating fluid subject to the action of a constant tension, failure characteristics such as infant mortality, useful life and wear-out periods, analyzing the wind speed data (Drapella (1993), Jiang *et al.* (2001), Nelson (1982), Khan (2008), Zaharim *et al.* (2009)). Most of the sub-cases of generalized inverse Weibull distribution are families of inverse distributions, which can be easily fitted to income related data. These distributions have two parameters but in order to fit better at the tails, three parameters distribution (GIWD) is used in the present study.

The cdf of generalized inverse Weibull distribution is

$$F(x) = e^{\left(-\gamma\left(\frac{\alpha}{x}\right)^{\beta}\right)}, x > 0; \ \alpha, \beta, \gamma > 0,$$

where α is scale parameter and β, γ are shape parameters.

The pdf of generalized inverse Weibull distribution is

Corresponding Author: Kamaljit Kaur
Email: kamaljitk010@gmail.com

$$f(x) = \gamma\beta\alpha^{\beta}x^{-(\beta+1)}e^{\left(-\gamma\left(\frac{\alpha}{x}\right)^{\beta}\right)}.$$

Sub-models:

- For $\alpha = 1$, it reduces to inverse Weibull (IW) distribution.
- For $\gamma = \alpha = 1$, it reduces to Fréchet (F) distribution.
- For $\beta = 2, \alpha = 1$, it reduces to inverse Rayleigh (IR) distribution.
- For $\gamma = \beta = 1$, it reduces to inverse exponential (IE) distribution.

In this paper, credible intervals for the parameters of the generalized inverse Weibull distributions are obtained. Some work using Generalized Pareto Distribution (Hosking (1987)), Weibull distribution (Kundu (2008)), Generalized Exponential Distribution (Kundu *et. al.* (2009)) and Generalized Inverted Exponential Distribution (Dey *et. al.* (2014) is already available in the literature in the case of credible interval, however, in the context of Bayesian and income inequality measure is already available in the literature for GIWD and some other distributions (Bhattacharya *et al*. (1999), Mahajan *et. al.* (2015), Arora *et. al.* (2017), Kaur *et. al.* (2018), Kaur *et. al.* (2021)). In the context of Credible interval, no work has been done for Generalized inverse Weibull distribution.

Credible interval is an interval in the domain of a posterior probability distribution or predictive distribution in Bayesian statistics. The Bayesian equivalent of the confidence interval in the classical inference is the credible interval. Bayesian interval estimators have a clearer and more direct interpretation than classical confidence intervals. Like classical confidence interval, the 95% Bayesian credible interval contains the true value with approximately 95% confidence. Bayesian intervals treat their bounds as fixed and the estimated parameter as a random variable, whereas frequentist confidence intervals treat their bounds as random variables and the parameter as a fixed value. 95% credible interval is any interval which contains a 95% percent of the posterior probability. Because the posterior density is a true probability density, we can compute quantiles and percentiles of the parameter. The simplest 95% credible interval is bounded by the 2.5th and 97.5th percentiles. This interval is called a symmetric credible interval because it removes equal probability (2.5%) from both tails of the distribution.

According to Eberly and Casella (2003) the $100\ (1-\alpha)\%$ equal tail credible interval for exact posterior distribution can be defined as

$$P(\theta < L) = \int_{-\infty}^{L} \pi(\theta|x)\, d\theta = \frac{\alpha}{2}\ ,\ \ P(\theta > U) = \int_{U}^{\infty} \pi(\theta|x)\, d\theta = \frac{\alpha}{2} \qquad (1)$$

where,

$\pi(\theta|x)$is posterior density of θ and
$(L\ , U)$are the lower and upper limits of the credible interval respectively for specified value of α (level of significance).

The posterior distribution is always available, although in realistically complex problems it cannot be represented analytically and becomes difficult in generation of random samples.

There are two types of algorithms used to draw samples from the true posterior. The first type is a direct method, when we draw a sample from an easily sampled density and reshape

this sample by only accepting some of the values into the final sample in such a way that the accepted values constitute a random sample from the posterior. This method is inefficient as the number of parameters increases in the posterior distribution.

Secondly, the simulation method for sampling from posterior distribution is called the Markov Chain Monte Carlo (MCMC) method (Metropolis *et. al.* (1949)). The advantage of MCMC is that it gives not only a point estimator of the parameter, but also gives an interval estimation based on the final simulated empirical distribution. MCMC is essentially an iterative sampling algorithm, drawing values from the posterior distribution of the parameter in the model concerned. The simulation method for sampling from posterior distribution which computes posterior quantities of interest is called the Markov Chain Monte
Carlo (MCMC) method. A Markov chain is a well-known stochastic process model that can be used to characterize the probability of moving from one state to another. Numerous algorithms have been developed that will simulate samples from a discrete-time continuous-space Markov chain such that, after reaching a steady-state, the sequence of samples constitutes a sample from the desired joint posterior distribution. These simulated samples estimate the mean and especially the quantiles (used to compute credible intervals) of marginal posterior distributions for the parameters of interest. MCMC involves two methods, Metropolis–Hastings' algorithm and Gibbs sampling for generating samples from the posterior distribution (Metropolis *et al.* (1953), Hastings (1970)). For more details about MCMC and the related methodologies, one can refer to Gentle (1998), Chen *et al.* (2000) and Robert and Casella (2004). Gibbs sampling procedure and Metropolis-Hastings (M-H) method are used to generate samples from the posterior density function to compute the Bayesian point estimates and credible intervals. When using a Markov Chain Monte Carlo algorithm such as the Gibbs sampler to generate a sample from the posterior distribution (marginal) of interest, calculations are often easier.

1.1. Metropolis-Hastings (Bolstad, 2010) algorithm

The algorithm of Metropolis-Hastings (Bolstad, 2010) is as follows:

Let the proposed density using the Metropolis-Hastings algorithm is denoted by $q(\theta, \theta')$, which is close to target density $g(\theta|x)$,

where

$\theta$is starting value,
$\theta'$is the next generating value of θ and
$g(\theta|x)$is the posterior target density from which we need to generate θ.

1. Start at an initial value$\theta^{(0)}$.
2. Do for $n = 1,2,\ldots\ldots,n$ (n is the number of iterations)
   (a) Draw a sample from $q(\theta^{(n-1)}, \theta')$.
   (b) Calculate r = probability of acceptance = $\alpha(\theta^{(n-1)}, \theta')$.
   (c)Draw u from the uniform distribution $U(0,1)$.
   (d) If $u < r$, then let $\theta^{(n)} = \theta'$, else let $\theta^{(n)} = \theta^{(n-1)}$.

The density $q(\theta, \theta')$ close to the target density $g(\theta|x)$ leads to more points being accepted. In fact, proposed density has the same shape as the target density.

$q(\theta, \theta') = kg(\theta'|x)$

the acceptance probability

$$\alpha(\theta, \theta') = min\left[1, \frac{g(\theta'|x)q(\theta', \theta)}{g(\theta|x)q(\theta, \theta')}\right]$$
$$= min\left[1, \frac{g(\theta'|x)g(\theta|x)}{g(\theta|x)g(\theta'|x)}\right]$$
$$= 1$$

i.e in this case, all points will be accepted.

After generating the sample from the posterior distribution using MCMC simulation method, one important question is: how many samples are needed to accurately approximate the characteristics of the posterior distribution? This question is difficult to answer because samples generated on successive iterations are not independent of one another. Frequently, the values from one iteration and the next will be highly correlated, and a very large number of iterations will be necessary to make sure that the sample covers the entire range of the distribution. We would like our Markov chain to move about the space covered by the distribution freely. When outcome of one iteration has little effect on the next iteration, we say that the chain is mixing quickly. If the outcomes on successive iterations are highly linked, then we say that the chain is mixing slowly. If the chain is mixing slowly then it will have to be run for a long time until we can be sure that our sample properly represents the posterior distribution.

1.2. Trace plots

The simplest tool for visualizing the convergence of a Markov chain is the **trace plot**, the plot of the values generated from the Markov chain versus the iteration number. This plot shows that the chain is mixing well, moving back and forth over the space and suggests how much sample values are enough to produce accurate approximation of the posterior summaries.

It may be noted that if the chain does not converge to its stationary distribution, then there will be long burn-in period. One can observe from a trace plot that there is a relatively constant mean and variance in case of stationarity.

1.3. Burn-in period

To discard the initial portion of a Markov chain, so that the effect of initial values on the posterior inference is minimized, we use Burn-in procedure. The initial samples are not completely valid because the Markov Chain has not stabilized to the stationary distribution or at beginning of sequence, we need to run MCMC for a while to achieve convergence to target pdf. The burn in samples allows us to discard these initial samples that are not yet at the stationary distribution.

This study focuses on the generation of samples using MCMC algorithm from the posterior distribution. Then the generated samples using Metropolis–Hastings' algorithm and Gibbs sampling are used to compute the credible intervals for the parameters of interest using different prior and squared error loss function (SELF) in case of generalized inverse Weibull distribution.

The outline of the paper is: the Posterior distributions of generalized inverse Weibull distribution using different priors are given in Section 2. In Section 3, algorithms are given to compute credible intervals for the above said distributions. The convergence and mixing of Markov chain through graphical method are presented in Section 4. In this section, simulation study along with real life illustration is also carried out to compute credible intervals using different priors in case of generalized inverse Weibull distribution. Finally, Section 5 gives the conclusion of the study.

2. Posterior Distributions for Parameters of Generalized Inverse Weibull Distribution

The pdf of generalized inverse Weibull distribution is

$$f(x) = \gamma\beta\alpha^{\beta}x^{-(\beta+1)}e^{\left(-\gamma\left(\frac{\alpha}{x}\right)^{\beta}\right)},\ \alpha,\beta,\gamma > 0.$$

The likelihood function of generalized inverse Weibull distribution is given by

$$L(x|\gamma,\alpha,\beta) = \gamma^{n}\beta^{n}\alpha^{n\beta}\left(\prod_{i=1}^{n}x_i^{-(\beta+1)}\right)exp\left(-\gamma\alpha^{\beta}\sum_{i=1}^{n}x_i^{-\beta}\right).$$

2.1. Posterior densities of parameters of GIWD using informative prior

Informative prior depends on the elicitation of prior distribution based on pre-existing scientific knowledge in the area of investigation. This information may be available from the previous investigation or from non-statistician experts. Assuming parameters α,β,γ have independent Gamma priors with the pdf's

$$g(\alpha; a_2, b_2) = \frac{{b_2}^{a_2}\alpha^{a_2-1}exp\,(-\alpha b_2)}{\Gamma(a_2)},$$

$$g(\beta; a_3, b_3) = \frac{{b_3}^{a_3}\beta^{a_1-1}exp\,(-\beta b_3)}{\Gamma(a_3)},$$

$$g(\gamma; a_1, b_1) = \frac{{b_1}^{a_1}\gamma^{a_1-1}exp\,(-\gamma b_1)}{\Gamma(a_1)},$$

where a_i, b_i for $i = 1,2,3$ are hyperparameters.

Assuming that the parameters are mutually independent, the posterior distribution is proportional to the product of the prior and the likelihood function given by

$$g^{*}(\alpha,\beta,\gamma|x) \propto \gamma^{n}\beta^{n}\alpha^{n\beta}\prod_{i=1}^{n}x_i^{-(\beta+1)}\,exp\left(-\gamma\alpha^{\beta}\sum_{i=1}^{n}x_i^{-\beta}\right)\gamma^{a_1-1}\,exp(-\gamma b_1)\,\alpha^{a_2-1}$$

$$exp\,(-\alpha b_2)\beta^{a_3-1}exp(-\beta b_3)$$

The full conditional posterior density of α is

$$g^{*}(\alpha|\beta,\gamma,x) \propto \alpha^{n\beta+a_2-1}exp\,(-\gamma\alpha^{\beta}\textstyle\sum_{i=1}^{n}x_i^{-\beta} - b_2\alpha).$$

The full conditional posterior density of β is

$$g^*(\beta|\alpha,\gamma,x) \propto \beta^{n+a_3-1}\alpha^{n\beta+a_3-1}\prod_{i=1}^{n} x_i^{-(\beta+1)} exp\,(-\gamma\alpha^{\beta}\sum_{i=1}^{n} x_i^{-\beta} - b_3\beta).$$

The full conditional posterior density of γ is

$$g^*(\gamma|\alpha,\beta,x) \propto \gamma^{n+a_1-1} ex\,p\left(-\left(\alpha^{\beta}\sum_{i=1}^{n} x_i^{-\beta} - b_1\right)\gamma\right) \sim Gamma(n + a_1,\ \alpha^{\beta}\sum_{i=1}^{n} x_i^{-\beta} - b_1).$$

2.2. Posterior densities of parameters of GIWD using Jeffreys' prior

Jeffreys' (1946) prior based on the Fisher's information, is defined as

$$\pi(\theta) \propto \sqrt{I(\theta)},$$

where $I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2} lnL(\theta|x)\right]$ is Fisher's information based on likelihood function $L(\theta|x)$.

The expected value of double derivatives is not in the closed form, hence the explicit experrsion for the Jeffreys' prior is not obtained. For simplicity it is assumed that all the three parameters are independent, therefore joint prior in case of Jeffreys' prior (Guure, 2012), Singh (2011) is written as

$$g(\alpha,\beta,\gamma) \propto \frac{1}{\gamma\alpha\beta}.$$

The full conditional posterior density of α is

$$g^*(\alpha|\beta,\gamma,x) \propto \alpha^{n\beta-1} exp\,(-\gamma\alpha^{\beta}\sum_{i=1}^{n} x_i^{-\beta}).$$

The full conditional posterior density of β is

$$g^*(\beta|\alpha,\gamma,x) \propto \beta^{n-1}\alpha^{n\beta-1}\prod_{i=1}^{n} x_i^{-(\beta+1)} exp\,(-\gamma\alpha^{\beta}\sum_{i=1}^{n} x_i^{-\beta}).$$

The full conditional posterior density of γ is

$$g^*(\gamma|\alpha,\beta,x) \propto \gamma^{n-1} ex\,p\left(-\left(\alpha^{\beta}\sum_{i=1}^{n} x_i^{-\beta}\right)\gamma\right) \sim Gamma(n, \alpha^{\beta}\sum_{i=1}^{n} x_i^{-\beta}).$$

Note: The full conditional posterior densities of α, β and γ using Jeffreys' prior are obtained by taking hyperparametres as zero ($a_1 = b_1 = a_2 = b_2 = a_3 = b_3 = 0$).

3. Algorithms to Compute Credible Intervals for Generalized Inverse Weibull Distribution

The posterior densities using different priors cannot be solved directly to compute lower limit (L) and upper limit (U) of credible interval as stated in equation 1. MCMC simulation techniques allow us to generate a sample from these posterior densities using Metropolis-Hastings (M-H) method and Gibbs sampling method.

The conditional posterior distributions of α and β cannot be reduced analytically to well-known distributions and therefore it is not possible to simplify it directly by standard methods, but their graphs indicate that they are like the Gamma and Weibull distributions, respectively.

So, to generate random numbers from these distributions, use the Metropolis-Hastings (M-H) method with Gamma and Weibull as the proposed distributions. To generate γ from the posterior density, Gibbs sampling method is used.

The following algorithm is given to generate α, β and γ from their posterior density functions and in turn to obtain the Bayes estimates and the corresponding credible intervals.

- Start with $\alpha_0 = \hat{\alpha}$ and $\beta_0 = \hat{\beta}$ as their initial approximation.
- Set $j = 1$, using Metropolis – Hasting generate α_j from conditional posterior density of α with the Gamma $(\alpha_{j-1}, 2)$ as the proposal distribution and also generate β_j from conditional posterior density of β with the Weibull $(\beta_{j-1}, 2)$ as the proposal distribution. Generate γ_j from Gamma $(n + a_1, (\alpha^{\beta} \sum_{i=1}^{n} x_i^{-\beta} + b_1))$ using Gibbs sampling.
- Set $j = j + 1$
- Repeat step 2, N times.
- Obtain the Bayes estimates of α, β and γ using SELF as
  $\hat{\alpha} = \frac{\sum_{i=M+1}^{N} \alpha_i}{N-M}$, where M is the burn-in period.
  $\hat{\beta} = \frac{\sum_{i=M+1}^{N} \beta_i}{N-M}$, where M is the burn-in period.
  $\hat{\gamma} = \frac{\sum_{i=M+1}^{N} \gamma_i}{N-M}$, where M is the burn-in period.
- To compute the credible intervals of α, β and γ, order $\alpha_{M+1}, \alpha_{M+2} \ldots\ldots, \alpha_N$; $\beta_{M+1}, \beta_{M+2} \ldots\ldots, \beta_N$ and $\gamma_{M+1}, \gamma_{M+2} \ldots\ldots, \gamma_N$ in ascending order as $\alpha_{(1)}, \alpha_{(2)} \ldots\ldots, \alpha_{(N-M)}$; $\beta_{(1)}, \beta_{(2)} \ldots\ldots, \beta_{(N-M)}$; $\gamma_{(1)}, \gamma_{(2)} \ldots\ldots, \gamma_{(N-M)}$. Then the $100(1-\eta)\%$ credible intervals for α, β and γ are
- $\left(\alpha_{\left(\frac{(N-M)\eta}{2}\right)}, \alpha_{\left(1-\frac{\eta}{2}\right)(N-M)}\right)$, $\left(\beta_{\left(\frac{(N-M)\eta}{2}\right)}, \beta_{\left(1-\frac{\eta}{2}\right)(N-M)}\right)$ and $\left(\gamma_{\left(\frac{(N-M)\eta}{2}\right)}, \gamma_{\left(1-\frac{\eta}{2}\right)(N-M)}\right)$,

where η is the level of significance.

In the next section, credible intervals are computed using R-software by the above algorithm.

4. Simulation Study

Statistical simulation study is carried out to compute the 95% and 99% credible intervals using different priors for generalized inverse Weibull distribution. The comparisons of priors are also done based on the width of the credible intervals; smaller the width better is the interval. According to distributions, combinations of parameters, hyper parameters and sample size should be chosen, and these are discussed below for all the three parameters. The credible intervals are computed based on 10,000 MCMC samples and first 500 values are discarded as burn-in. We plot the trace plots of the chains to determine whether the chain is exploring the parametric space well for all the parameters of GIWD. The monitoring MCMC convergence and mixing is also checked using trace and autocorrelation plots. The autocorrelation shows the mixing rate, and it is measured by autocorrelations of different lags.

The credible interval are computed for Jeffreys' and informative prior for the parametres of GIWD obtained using squared error loss function. These intervals are computed for different sample sizes $n = 30, 50$ with parameters combinations $\alpha = \beta = \gamma = 2$. The combinations of hyperparameters are taken as $a_1 = b_1 = 6, a_2 = b_2 = a_3 = b_3 = 4$ (Kaur *et. al*. (2018)) according to misfit measure. The trace, posterior density and autocorrelation plots of α, β and γ are plotted in case of informative prior.

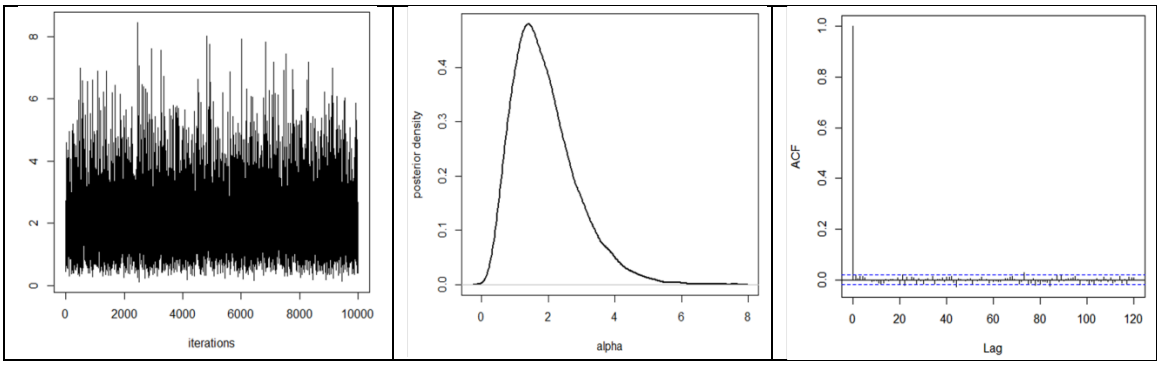


Figure 1: Trace, posterior density and autocorrelation plots of α

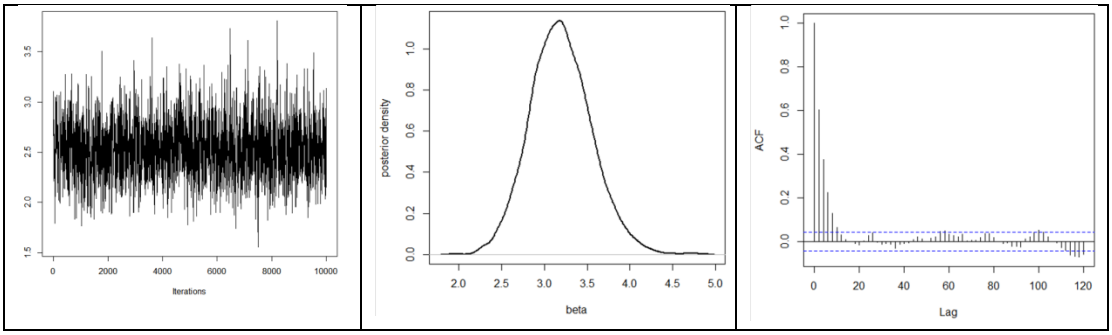


Figure 2: Trace, posterior density and autocorrelation plots of β

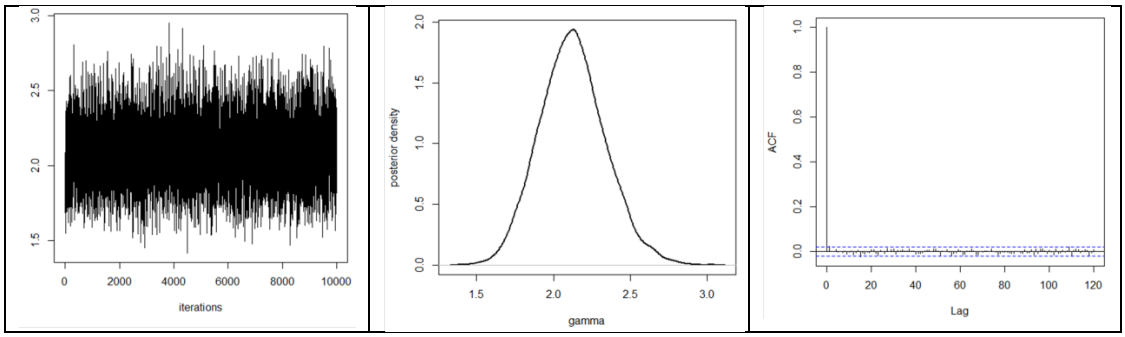


Figure 3: Trace, posterior density and autocorrelation plots of γ

Based on trace, autocorrelation and posterior plots (Figure 1-3), we conclude that

- Markov Chain (MC) has reached convergence,
- trace plot is perfect and the centre of the chain having small fluctuations indicates that the MC has reached the right distribution,
- all autocorrelations are close to zero for α and γ *i.e.*, MCMC sampling is done in independent manner and stationarity is reached. The autocorrelation plots for β shows low mixing at the starting lags and good mixing after 10^{th} lag.

The credible intervals are reported in the following Tables 1-3. From the Tables, it may be seen that

(i) Credible intervals using informative priors lead to smaller width of the interval as compared to non-informative prior for all the three parameters both for 95% and 99% C.I.
(ii) As the sample size increases, the width of the credible intervals decreases

Table 1: Credible intervals for α

n	Prior	Estimate	95% C.I. (width)	99% C.I. (width)
30	Jeffrey	1.93205	(0.56643,4.87645) 4.31002	(0.35410,6.02160) 5.6675
	Informative	2.17695	(0.51799,4.46347) 3.94548	(0.36721,5.68138) 5.31417
50	Jeffrey	2.01670	(0.57516,4.48867) 3.91351	(0.34895, 5.91940) 5.57045
	Informative	2.05906	(0.54841,4.37218) 3.82377	(0.33537,5.60383) 5.26846

Table 2: Credible intervals for β

N	Prior	Estimate	95% C.I (width)	99% C.I (width)
30	Jeffrey	2.17912	(1.73348, 2.67003) 0.93655	(1.60958, 2.86279) 1.25321
	Informative	2.05582	(1.62910, 2.53639) 0.90729	(1.49765, 2.71638) 1.21873
50	Jeffrey	1.82986	(1.55504, 2.13338) 0.57834	(1.48468, 2.25778) 0.7731
	Informative	1.90897	(1.59993, 2.15669) 0.55676	(1.51976, 2.27387) 0.75411

Table 3: Credible intervals for γ

n	Prior	Estimate	95% C.I. (width)	99% C.I. (width)
30	Jeffrey	2.01406	(1.49211, 2.60510) 1.11299	(1.46391, 2.92407) 1.46016
	Informative	2.13966	(1.61030, 2.71203) 1.10173	(1.36377, 2.81584) 1.45207
50	Jeffrey	2.15621	(1.75397, 2.59709) 0.84312	(1.65186, 2.76850) 1.11664
	Informative	2.00077	(1.64707, 2.39050) 0.74343	(1.55070, 2.51419) 0.96349

Table 4: Credible intervals for α, β and γ

Parameters	Prior	estimate	95% C.I (width)	99% C.I. (width)
α	Jeffrey	1.01693	(0.29426, 2.72271) 2.42845	(0.18097, 3.62492) 3.44395
	Informative	1.14379	(0.26616, 2.34805) 2.08189	(0.17551, 2.97161) 2.7961
β	Jeffrey	2.20413	(1.81032, 2.65222) 0.8419	(1.69353, 2.83398) 1.14045
	Informative	2.04210	(1.67144, 2.46173) 0.79029	(1.56256, 2.63407) 1.07151
γ	Jeffrey	5.12292	(3.53113, 6.96405) 3.43292	(3.12454, 7.53388) 4.40934
	Informative	4.83125	(3.36143, 6.57182) 3.21039	(2.95502, 7.17773) 4.22271

Real Life Example

The real-life data of percentage of GDP of different countries is taken from Dataset: Central Government Dept of 2009. The GIWD is used to fit this data set. To check the validity of the model, we compute of Kolmogorov-Smirnov test and p-value for this test is 0.1859, suggesting thereby the appropriateness of the GIWD. The credible intervals are computed based on 10,000 MCMC samples and first 500 values are discarded as burn-in. The trace plots are also plotted to determine whether the chain is exploring the parametric space well for all the parameters of three distributions in case of real-life example.

It is seen from the above tables, for all three parameters of GIWD the informative prior performs better as compared with non-informative prior (Jeffreys' prior).

5. Conclusion

The informative prior performs better as compared to non-informative prior and findings from the analysis of real life example are in accordance with those of simulation study in case of generalized inverse Weibull distribution. One can further infer that as the sample sizes

increases, the width of the credible interval decreases for both 95% and 99% credible intervals in case of Generalized inverse Weibull distribution.

References

Arora, S., Kaur, K. and Mahajan, K. K. (2017). Generalized maximum likelihood Estimators for Gamma distribution: Semi-Bayesian approach. *The Aligarh Journal of Statistics,* **37**, 41-76.

Bhattacharya, S. K., Chaturvedi, A. and Singh, N. K. (1999). Bayesian estimation for Pareto income distribution. *Statistical Papers,* **40**, 247-262.

Bolstad, W. M. (2010). *Understanding Computational Bayesian Statistics*. John Wiley & Sons, Inc., Publication.

Chen, M. H., Shao, Q. M. and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York.

Dey, S. and Dey, T. (2014). Generalized inverted exponential distribution: Different methods of estimation. *American Journal of Mathematical and Management Sciences*, **33(3)**, 194-215.

Drapella, A. (1993). The complementary Weibull distribution, unknown or just forgotten? *Quality and Reliability Engineering International*, **9**, 383-385.

Eberly, L. E. and Casella, G. (2003). Estimating Bayesian credible intervals. *Journal of Statistical Planning and Inference*, **112**,115-132.

Gentle, J. E. (1998). *Random Number Generation and Monte Carlo Methods*. Springer, New York.

Gusmão, F. R. S., Ortega, E. M. M. and Cordeiro, G. M. (2011). The generalized inverse Weibull distribution. *Statistical Papers,* **52**, 591–619.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika,* **57**, 97-109.

Hosking, J. R. and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, **29(3)**, 339-349.

Jiang, R., Murthy, D. N. P. and Ji, P. (2001). Models involving two inverse Weibull distributions. *Reliability Engineering and System Safety,* **73,** 73–81.

Kaur, K., Mahajan, K. K. and Arora, S. (2018). Bayesian and Semi-Bayesian estimation of the parameters of generalized inverse Weibull distribution. *Journal of Modern Applied Statistical Methods*, **17(1)**, 22.

Kaur, K., Arora, S. and Mahajan, K. K. (2021). Bayesian vs semi-Bayesian estimation for income inequality measures in case of GIWD. *Journal of Statistics and Management Systems*, 1-32.

Khan, M. S., Pasha, G. R. and Pasha, A. H. (2008). Theoretical analysis of inverse Weibull distribution. *Wseas Transactions on Mathematics,* **2**(**7**), 30-38.

Kundu, D. (2008). Bayesian inference and life testing plan for the Weibull distribution in presence of progressive censoring. *Technometrics*, **50**(**2**), 144-154.

Kundu, D. and Pradhan, B. (2009). Estimating the parameters of the generalized exponential distribution in presence of hybrid censoring. *Communications in Statistics: Theory and Methods*, **38**(**12**), 2030-2041.

Mahajan, K. K., Arora, S. and Kaur, K. (2015). Bayesian estimation for Gini index and a poverty measure in case of Pareto distribution using Jeffreys' prior. *Model Assisted Statistics and Applications*, **10(1)**, 63-72.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953). Equations of state calculations by fast computing machines. The *Journal of Chemical Physics*, **21**(**6**),1087–1092.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, **44**, 335-341.

Nelson, W. B. (1982). *Applied Life Data Analysis*. John Wiley & Sons, New York, NY, USA.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Second Edition, Springer, New York.

Zaharim, A., Najid, S. K., Razali, A. M. and Sopian, K. (2009). Analyzing Malaysian wind speed data using statistical distribution. In Proceedings of the 4th IASME / WSEAS *Internatio*nal *Conference on Energy and Environment*, University of Cambridge, February 24-26.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 189-201

Robust Parameter Design Using 20 Run Plackett-Burman Design

Renu Kaul and Sanjoy Roy Chowdhury
Department of Statistics, Lady Shri Ram College for Women (University of Delhi), New Delhi 110024

Received: 29 July 2021; Revised: 01 November 2021: Accepted: 07 November 2021

Abstract

Taguchi's parameter design technique for improving product quality has aroused a great deal of interest among statisticians and quality practitioners. He proposed the use of product array for reducing variation and improving product quality. However, in some applications, his approach results in an exorbitant number of runs. As an alternative to the product array approach, Welch *et al.* (1990), Shoemaker *et al.* (1991) and Montgomery (1991a), proposed the use of combined arrays wherein control and noise factors are combined in a single array. Further, Shoemaker *et al.* (1991) used an optimal design algorithm to reduce the size of the combined array.

In this paper, we have exploited the non-orthogonal column structure of the 20-run Plackett-Burman Design. It is shown that by making use of the columns of the 20-run Plackett-Burman design, the size of the experiment can further be reduced. The results have been shown for designs with six factors and the results for three, four and five factors are given in the annexure.

Key words: Robust parameter design; Orthogonal arrays; Fractional factorial designs; Combined array; Plackett-Burman designs; D-efficiency; Projective rationale.

1. Introduction

Taguchi (1959, 1987) introduced an off-line quality control technique known as robust parameter design for reducing variation and improving product quality. The root of this idea is the notion that products lack in quality because of inconsistency in performance produced by factors that are controllable in the design of the product. He thus classifies the factors into two groups: Control factors and Noise factors.

The overall objective of Taguchi's approach is to determine the levels of control factors at which the effect of the noise factors on the performance characteristics is minimized. To achieve this objective he made use of product arrays by taking the Kronecker product of two orthogonal arrays, one involving only the control factors (inner array) and the other involving only the noise factors (outer array). Direct products of orthogonal arrays are themselves orthogonal arrays but the product operation greatly increases the number of observations in the array without generally increasing its strength. Several different methods of construction have been suggested, with the underlying idea of choosing levels for the controllable factors so that the uncontrollable factors have least influence on the response. Welch *et al*. (1990), Borkowski and Lucas (1991, 1997), Montgomery (1991 a, b), Myers (1991), Shoemaker *et al.* (1991), Welch and Sacks (1991), Box and Jones (1992) and Lucas (1994) suggested the

Corresponding Author: Sanjoy Roy Chowdhury
E-mail:sanjoy.lsrc@gmail.com

use of combined arrays, wherein the control factors and noise factors are combined in a single array. A combined array lets the experimenter choose the interactions to be estimated. This provides more flexibility so that the experimental budget can be used to fit models more refined than the main effects only models frequently used in Taguchi's loss model approach. An excellent review of the robust parameter technique is made by Nair (1992) and Myers and Montgomery (1995). Kunert *et al*. (2007) compared Taguchi's product array with a combined array.

In this paper we have exploited the non-orthogonal column structure of the 20-run Plackett-Burman (PB) design to generate non-orthogonal combined arrays.

2. The Role of Interactions

In parameter design, one is interested in choosing the levels of control factors so that the product's performance is insensitive to noise factors and can be adjusted on target as appropriate. The control × noise (C × N) interactions are exploited to accomplish this. The structure of these interactions provides special insights in the combined array/response model approach because they are the effects that can be exploited to reduce response variability. The noise × noise (N × N) interactions play little role in making a product's performance insensitive to noise factors. The presence of large C × C interactions is considered highly undesirable; thus, every attempt is made to reduce the number of C × C interactions through judicious choice of the quality characteristics.

3. Objectives and the Supportive Models

Keeping in view the above justification for the inclusion of various terms in the models we now specify our objectives:

Let there be *r* control factors, say, $x_1, x_2, \ldots\ldots, x_r$ and *s* noise factors *viz.* $z_1, z_2, \ldots\ldots, z_s$.

Then our objective is:

1. To estimate the main effects of all the control factors and noise factors.
2. To estimate C × N interactions.
3. To estimate if possible, (depending on the degrees of freedom) the C × C interactions.

The above objectives can be explained more precisely with the help of regression models. Let *y* denote a quality characteristic associated with a product. We can then express:

$$y = f(x, z) \tag{1}$$

If the response is well modelled by a linear function of the independent variables, then the approximating function is the first order model:

$$y = \beta_0 + \sum \beta_i x_i + \sum \gamma_j z_j + \epsilon \tag{2}$$

But, in model (2), the settings of *x* have no influence on variability. For robust parameter design to be successful, the functional relationship between control factors and

noise variables should be such that they interact. Thus, a second order model will be more appropriate:

$$y = \beta_0 + \sum \beta_i x_i + \sum \beta_{ii} x_i^2 + \sum\sum \beta_{ii'} x_i x_{i'} + \sum \gamma_j z_j + \sum \gamma_{jj} z_j^2 + \sum\sum \gamma_{jj'} z_j z_{j'} + \sum\sum \delta_{ij} x_i z_j + \epsilon$$

where $i \neq i' = 1,2,....,r;\ j \neq j' = 1,2,....,s$ (3)

In order to meet the first two objectives mentioned above the reduced model, by keeping the origin at (0, 0), would be:

$$y = \sum \beta_i x_i + \sum \gamma_j z_j + \sum\sum \delta_{ij} x_i z_j + \epsilon. \quad (4)$$

Whereas, when one is also interested in estimating the C × C interactions (the third objective), the corresponding model would be:

$$y = \sum \beta_i x_i + \sum \gamma_j z_j + \sum\sum \delta_{ij} x_i z_j + \sum\sum \beta_{ii'} x_i x_{i'} + \epsilon \quad (5)$$

4. Efficiency Criterion

We have used the following D-criterion for measuring the overall efficiency for estimating a collection of effects:

$$\text{D-efficiency} = |X'X|^{1/k} \quad (6)$$

where, $X = [x_1/||x_1||, ..., x_k/||x_k||]$; and x_i is the coefficient vector of the i^{th} effect. To find the efficiency of each individual effect, we have used the following D_s criterion:

$$\frac{\left\{x_i'x_i - x_i'X_{(i)}\left(X_{(i)}'X_{(i)}\right)^{-1}X_{(i)}'X_{(i)}\right\}}{x_i'x_i} \quad (7)$$

where, $X_{(i)}$ is obtained from X by deleting x_i.

5. Steps Used for Combined Array Approach

We give below the steps used in the combined array approach:

i. Choose p columns from the totality of n–1 columns and consider all the non-equivalent designs.
ii. For each design allocate the control factors and noise factors to p columns.
iii. Write the appropriate model by considering the required set of C × N interactions and C × C interactions (depending upon run-size).
iv. For all possible choices of the control and noise factors find the D value for the whole design and D_s values for the various effects.
v. Compare the D value of all the designs obtained and take the one with maximum D value. If there are more designs with the maximum D value, consider all of them.
vi. Sort the D_s values of these designs on the basis of C × N interactions and take the design for which it is maximum. If there are more than one designs with the same values of D_s for C × N interactions, consider all of them.

vii. Sort the D_s values of these designs on the basis of C × C interactions and take the design for which it is minimum. If there are more than one designs with the same values of D_s for C × C interactions, take all of them.

viii. Among the designs chosen by step (vii), finally sort these designs on the basis of the D_s values for control factors and noise factors and select the design for which it is maximum.

ix. Once a design has been selected by following the aforesaid steps, the D_s values of the various effects are reported according to the order of column allocations of respective control factors, noise factors and their interactions in the tables.

6. Plackett-Burman Designs

Plackett and Burman (1946) provided a series of two-level fractional factorial designs, for examining (n–1) factors in n runs, where n is a multiple of 4 and $n \leq 100$. These are non-orthogonal designs in which the aliasing coefficient between any two effects lies between –1 and +1. They gave the following design for 20-runs:

Table 1: 20-Run Plackett-Burman design

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
–	+	–	–	+	+	+	+	–	+	–	+	–	–	–	–	+	+	–
–	–	+	–	–	+	+	+	+	–	+	–	+	–	–	–	–	+	+
+	–	–	+	–	–	+	+	+	+	–	+	–	+	–	–	–	–	+
+	+	–	–	+	–	–	+	+	+	+	–	+	–	+	–	–	–	–
–	+	+	–	–	+	–	–	+	+	+	+	–	+	–	+	–	–	–
–	–	+	+	–	–	+	–	–	+	+	+	+	–	+	–	+	–	–
–	–	–	+	+	–	–	+	–	–	+	+	+	+	–	+	–	+	–
–	–	–	–	+	+	–	–	+	–	–	+	+	+	+	–	+	–	+
+	–	–	–	–	+	+	–	–	+	–	–	+	+	+	+	–	+	–
–	+	–	–	–	–	+	+	–	–	+	–	–	+	+	+	+	–	+
+	–	+	–	–	–	–	+	+	–	–	+	–	–	+	+	+	+	–
–	+	–	+	–	–	–	–	+	+	–	–	+	–	–	+	+	+	+
+	–	+	–	+	–	–	–	–	+	+	–	–	+	–	–	+	+	+
+	+	–	+	–	+	–	–	–	–	+	+	–	–	+	–	–	+	+
+	+	+	–	+	–	+	–	–	–	–	+	+	–	–	+	–	–	+
+	+	+	+	–	+	–	+	–	–	–	–	+	+	–	–	+	–	–
–	+	+	+	+	–	+	–	+	–	–	–	–	+	+	–	–	+	–
–	–	+	+	+	+	–	+	–	+	–	–	–	–	+	+	–	–	+
+	–	–	+	+	+	+	–	+	–	+	–	–	–	–	+	+	–	–
–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

We shall now discuss the projection properties of this design. The choice of p columns, where $p \leq (n–1)$ may result in a number of designs for given n and p, not all of which may be equivalent. Two such designs are said to be equivalent if one can be obtained from the other by permutations of rows, columns and sign changes. Draper and Lin (1990) have given detailed tables giving the number of distinct designs for 12-, 20- and 24-run PB designs for different values of p. Each design is characterized by the number of repeat runs, mirror images or distinct runs it has. For p = 2, Draper and Lin (1990, Table 3B), Lin and Draper

(1992, 1995) found that projection is a 2^2 design, $n/4$ times over. For $p = 3$, they found that there are two different projections, each one consisting of at least a full 2^3 factorial. For $p = 4$, there are three non-isomorphic 20×4 submatrices: designs 4.1, 4.2 and 4.3. The 20 points in design 4.1 have one treatment combination omitted and five duplicated. Both designs 4.2 and 4.3 have each 4 points missing, 5 points appear once, 6 points appear twice and one point appears three times. For $p = 5$, Draper and Lin (1990, Table 3B) found that there are nine non-isomorphic 20×5 submatrices *viz.* designs 5.1, 5.2, …, and 5.9. Design 5.1 has no run with repeats, design 5.4 has one run with 2 repeats and the remaining designs have at least two runs with repeats. For $p = 6$, there are 50 non-isomorphic 20×6 submatrices. To save the enormity of calculations, we consider only the 17 designs considered by Draper and Lin (1990, Table 3B) *viz.* designs 6.1, 6.2,…, and 6.17 based on their mirror image patterns or repeat run pairs. Designs 6.1, 6.2, 6.4, 6.9 and 6.13 have no runs with repeats while rest of the designs have at least one run with a repeat. We now discuss the combined array concept for this design.

7. Combined Array Results for the 20-Run PB Design

There are 19 independent columns for studying the factor effects and 20 design points. We shall discuss here only one case as others can be obtained in a similar manner. Suppose we have six factors we then need to choose six columns from the 19 columns. Now 6 factors can be divided into control and noise factors in five different ways:

(a) $r = 5, s = 1$ (b) $r = 1, s = 5$ (c) $r = 4, s = 2$ (d) $r = 2, s = 4$ (e) $r = 3, s = 3$

Consider the first possibility:

(a) $r = 5, s = 1$

Allocate five columns to the control factors and one to the noise factor. There are 21 parameters to be estimated including the C × C interactions. Out of 17 designs given by Draper and Lin (1990, Table 3B), 5 designs have no repeats and thus enable us to estimate 19 parameters in 20 runs. Out of these, design 6.1 is the best having maximum D-efficiency. The following Table gives the allocation of control and noise factors which have come out to be the best for this design:

Table 2: $r = 5, s = 1$

Design 6.1 (1,2,3,4,5,6), (20)						
D. No.	C	N	C × N	C × C	D	D_s
1	1,3,4,5,6	2	12,32,42,52,62	14,15,34,35, 36,45,46,56	.73	.31,.31,.56,.44,.64,.58,.28,.57,.64,.51,.56, .58,.41, .57,.43,.32,.51,.28,.44

In the Table after giving the design number we give the column allocation of the selected design in the first parenthesis and the number in the second parenthesis gives the number of distinct runs in the design.

Five out of 17 designs have one run with a repeat and thus enable us to estimate 18 parameters. Design 6.5 is the best. We call a design to be good if it has the highest D-efficiency and provides maximum flexibility in the allocation of control and noise factors.

We give below the allocations of control and noise factors that have come out to be the best for this design:

Table 3: $r = 5, s = 1$

Design 6.5 (1,2,4,5,6,7), (19)						
D. No.	C	N	C × N	C × C	D	D_s
1	1,2,4,6,7	5	15,25,45,65,75	12,14,16,17, 26, 27,46	.74	.68,.49,.36,.64,.64,.68,.66,.59,.56, .38,.33,.59,.56,.38,.33,.53,.33,.49
2	2,4,5,6,7	1	21,41,51,61,71	25.26,27,45, 46,56,57	.74	.49,.36,.68,.64,.64,.68,.59,.56,.66, .38,.33,.59,.53,.33,.56,.49,.38,.33

There are 4 designs with 2 repeats, out of which, design 6.10 is the best having the highest D-efficiency, which enables us to estimate 17 parameters in 18 runs. The following Table gives the allocation of control and noise factors that has come out to be the best for this design:

Table 4: $r = 5, s = 1$

Design 6.10 (1,4,5,6,7,9), (18)						
D. No.	C	N	C × N	C × C	D	D_s
1	1,4,5,6,9	7	17,47,57,67,97	14,16,45,46, 56,59	.71	.57,.77,.51,.67,.28,.67,.4,.54,.4, .67, .41,.31,.23,.49,.22,.31,.22

There are 3 designs with 3 repeats, out of which design 6.16 is the best having highest D-efficiency. This design enables us to estimate 16 parameters in 17 runs. The following table gives the allocation of control and noise factors that have come out to be the best for this design:

Table 5: $r = 5, s = 1$

Design 6.16 (1,2,3,6,9,12), (17)						
D. No.	C	N	C × N	C × C	D	D_s
1	1,2,3,6,9	12	112,212,312, 612,912	12,13,16,23, 26	.69	.28,.77,.18,.18,.10,.29,.77,.31, .54, .54,.10, .24,.05,.28,.16,.54

(b) $r = 1, s = 5$

Allocate one column to the control factor and five to the noise factors. In this case 11 parameters are to be estimated as there are no C × C interactions. Design 6.17 is the best having highest D-efficiency, which estimates all the 11 parameters in minimum number of runs. We give below the allocations of control and noise factors which have come out to be the best for this design:

Table 6: $r = 1, s = 5$

Design 6.17 (1,2,3,5,8,13), (17)						
D. No.	C	N	C × N	C × C	D	D_s
1	1	2,3,5,8,13	12,13,15,18,113	-	0.93	.93,.83,.86,.80,.86,.86,.86, .86,.81, .86,.86
2	2	1,3,5,8,13	21,23,25,28,213	-	0.93	.92,.86,.86,.86,.80,.86,.86, .86,.86, .81,.86
3	13	1,2,3,5,8	131,132,133,135,138	-	0.93	.93,.83,.86,.80,.86,.86,.86, .86,.81, .86,.86

(c) $r = 4, s = 2$

Allocate four columns to the control factors and two to the noise factors. There are in all 20 parameters to be estimated. However, in 20 runs one can estimate at the most 19 parameters. Out of 5 designs, having no repeats, design 6.1 is the best having maximum D-efficiency. The following table gives the allocation of control and noise factors which has come out to be the best for this design:

Table 7: $r = 4, s = 2$

Design 6.1 (1,2,3,4,5,6), (20)						
D. No.	C	N	C × N	C × C	D	D_s
1	1,2,3,6	4,5	14,15,24,25,34, 35,64,65	12,13,23,26,36	.68	.25,.56,.31,.13,.39,.3,.56,.46, .64, .2,.56,.47, .32, .29, .23, .13, .39, .56, .37

Out of 5 designs having one repeat, design 6.6 performs the best. We give below the allocation of control and noise factors that has come out to be the best for this design:

Table 8: $r = 4, s = 2$

Design 6.6 (1,2,4,5,7,8), (19)						
D. No.	C	N	C × N	C × C	D	D_s
1	1,2,5,8	4,7	14,17,24,27,54, 57, 84,87	12,15,25,28	.74	.53,.66,.66,.53,.38,.38,.64,.64,.36, .36,.49,.49,.56,.56,.53,.36,.66,.33

There are 4 designs with 2 repeats, out of which design 6.10 is the best having the highest D-efficiency, which enables us to estimate 17 parameters in 18 runs. The following table gives the allocation of control and noise factors that has come out to be the best for this design:

Table 9: $r = 4, s = 2$

Design 6.10 (1,4,5,6,7,9), (18)						
D. No.	C	N	C × N	C × C	D	D_s
1	1,4,5,9	6,7	16,17,46,47,56, 57,96,97	14,15,19	.71	.55,.29,.55,.42,.67,.67,.31,.32,.27, .58, .23,.32,.22,.55,.49,.67,.49

There are 3 designs with 3 repeats, out of which design 6.16 is the best having the highest D-efficiency. This design enables us to estimate 16 parameters in 17 runs. The following Table gives the allocation of control and noise factors that has come out to be the best for this design:

Table 10: $r = 4, s = 2$

Design 6.16 (1,2,3,6,9,12), (17)						
D. No.	C	N	C × N	C × C	D	D_s
1	1,3,6,9	2,12	12,112,32,312,62,612, 92,912	13,16	.63	.14,.31,.07,.23,.77,.47,.33,.24, .33,.54,.33,.54, .08,.24,.05,.28

Consider the following example discussed by Shoemaker *et al.* (1991) to illustrate the flexibility afforded by a combined array:

7.1. Example 1

Suppose there are 4 two-level control factors, A, B, C and D, and 2 two-level noise factors, r and s. Assume that the control × control interactions – AB, AC and AD are potentially important and that we wish to estimate them. If we use the product array approach, we first construct a control array (CA) that estimates all main effects – A, B, C and D and the three important interactions – AB, AC and AD. We then construct a noise array (NA) that estimates the two main effects – r and s, and the interaction – rs. The defining relation of this plan is $I = ABCD$. According to the general result concerning estimation capacity of CA × NA designs, the resulting 32-run product array allows us to estimate six main effects – A, B, C, D, r, and s, the 12 two-factor interactions – *AB, AC, AD, rs, Ar, Br, Cr, Dr, As, Bs, Cs*, and *Ds* and 13 higher-order interactions. On the other hand, a combined array 2^{6-1} with resolution VI using

$$I = x_1x_2x_3x_4z_1z_2$$

is much more appropriate. This design allows the estimation of all the six main effects and all 15 two-factor interactions.

As yet a better approach, Shoemaker *et al*. (1991) used an optimal design algorithm to reduce the size of experiment further. As the 13 higher order interactions are less likely to be important, they constructed a linear model consisting of six main effects and 12 two-factor interactions mentioned above. Three combined arrays of size 20, 22, and 24 were generated from an optimal design algorithm DETMAX (Mitchell 1974), used in the software system RS/ DISCOVER (1988). All the three designs are approximately two-third the size of the product /combined array but allow efficient estimation of all the main effects and two-factor interactions mentioned earlier.

For the above example, we exploited the non-orthogonal column structure of the 20-run PB design. Also, as the role of noise × noise interactions in making a product's performance insensitive to noise factors is almost negligible, we therefore exclude them from our model. We are now left with 17 parameters to be estimated. There are 4 designs with 2 repeats each, *viz*, 6.8, 6.10, 6.11, and 6.14 given by Draper and Lin (1990, Table 3B). As a result, they have only seventeen degrees of freedom for estimating factor effects. Out of these four designs, design 6.10 estimates the 17 parameters with highest D-efficiency. Table 9 gives the allocation of control and noise factors that has come out to be the best for this design.

Thus, if we allocate the four control factors to columns 1, 4, 5, and 9 and noise factors to columns 6 and 7 of design 6.10, this design allows us to estimate all the 17 parameters in 18 runs only as compared to the design given by Shoemaker *et al.* (1991).

(c) $r = 2, s = 4$

Allocate two columns to the control factors and four to the noise factors. There are in all 15 parameters to be estimated. Out of 17 designs, designs 6.15, 6.16, and 6.17 enable us to estimate 15 parameters in 17 runs. However, as design 6.17 provides more flexibility for the allocation of control and noise factors, we give below the results for this design only:

Table 11: $r = 2, s = 4$

Design 6.17 (1,2,3,5,8,13), (17)						
D. No.	C	N	C × N	C × C	D	D_s
1	2,3	1,5,8,13	21, 25, 28, 213, 31, 35, 38, 313	23	.66	.35,.52,.35,.12,.25,.30,.46,.20,.24, .20, .28,.37, .06, .16, .48
2	8,13	1,2,3,5	81, 82, 83, 85, 131, 132, 133, 135	813	.66	.52,.35,.35,.30,.25,.12,.28,.16,.06, .37,.46,.20,.24,.20,.48

(d) $r = 3, s = 3$

Allocate three columns to the control factors and three to noise factors. Out of 17 designs, 5 designs have one repeat and thus enable us to estimate 18 parameters in 19 runs. Out of these 5 designs, design 6.5 is the best having maximum D value. The following table gives the allocation of control and noise factors that has come out to be the best for this design:

Table 12: $r = 3, s = 3$

Design 6.5 (1,2,4,5,6,7), (19)						
D. No.	C	N	C × N	C × C	D	D_s
1	1,2,5	4,6,7	14,16,17,24,26,27, 54,56,57	12,15,25	.68	.66,.42,.66,.36,.64,.12,.48,.19, .36,.12,.36,.26,.48,.19,.36,.42, .66,.42

Out of 4 designs having 2 repeats, design 6.10 performs the best and enables us to estimate 17 parameters in 18 runs. The following Table gives the allocation of control and noise factors that has come out to be the best for this design:

Table 13: r = 3, s = 3

Design 6.10 (1,4,5,6,7,9), (18)						
D. No.	C	N	C × N	C × C	D	D_s
1	1,6,7	4,5,9	14,15,19,64,65,69, 74,75,79	16,17	.71	.55,.67,.67,.29,.55,.42,.48,.67,.48, .27,.23,.22,.58,.32,.55,.31,.32

Out of 3 designs having 3 repeats, design 6.17 performs the best and enables us to estimate 16 parameters in 17 runs. The following Table gives the allocation of control and noise factors which have come out to be the best for this design:

Table 14: $r = 3, s = 3$

Design 6.17 (1,2,3,5,8,13), (17)						
D. No.	C	N	C × N	C × C	D	D_s
1	1,3,13	2,5,8	12,15,18,32,35,38, 132, 135,138	13	.63	.32,.53,.53,.28,.14,.28,.77,.28, .30,.28, .20,.05,.10,.12,.17,.34
2	2,3,13	1,5,8	21,25,28,31,35,38, 131, 135,138	23	.63	.32,.53,.53,.28,.28,.14,.77,.30, .28, .28,.05,.20,.10,.17,.12,.34

In the presence of 3 two-level control factors and 3 two-level noise factors, Shoemaker *et al.* (1991) have shown with the help of an example, the flexibility offered by a combined array *vis-a-vis* a product array.

8. Concluding Remarks

Many authors have advocated the use of combined arrays as an alternative to Taguchi's product arrays by modelling the response itself as a function of control and noise factors. These combined arrays are based on orthogonal fractional factorial designs, which do not exist for all values of *n*. Also, a major concern of most of the industries is to reduce the number of runs or minimize it. In this paper, we have exploited the non-orthogonal column structure of the 20-run Plackett-Burman design, giving a systematic method for choosing columns of a PB design for the allocation of control and noise factors. It has been shown that most of the designs using this approach, though not orthogonal, result in the reduction of the size of the experiment, a major benefit to the industry.

Acknowledgements

The authors would like to acknowledge the valuable comments and suggestions of the Reviewer and the Chair Editor, which have led to the improved version of the paper.

References

Borkowski, J. and Lucas, J. M. (1991). The analysis of mixed resolution design. Paper presented at the *Joint Statistical Meeting*, Atlanta, GA, August.

Borkowski, J. and Lucas, J. M. (1997). Designs of mixed resolution for process robustness studies. *Techonometrics*, **39**, 63-70.

Box, G. E. P. and Jones, S. (1992). Designing products that are robust to the environment. *Total Quality Management*, **3**, 265-282.

Draper, N. R. and Lin, D. K. J. (1990). Using Plackett and Burman designs with fewer than N–1 Factors. University of Tennessee College of Business Administration.

Kunert, J., Auer, C., Erdbrügge, M. and Ewers, R. (2007). An experiment to compare Taguchi's product array and the combined array. *Journal of Quality Technology*, **39**(**1**), 17-34.

Lin, D. K. J. and Draper, N. R. (1992. Projection properties of Plackett and Burman designs. *Technometrics*, **3**. 423-428.

Lin, D. K. J. and Draper N. R. (1995). Screening properties of certain two-level designs. *Metrika*, **42**, 99-118.

Lucas, J. M. (1994). How to achieve a robust process using response surface methodology. *Journal of Quality Technology*, **26**, 248-260.

Mitchell, T. J. (1974). An algorithm for the construction of 'D-optimal' experimental designs. *Technometrics*, **16**, 203-210.

Montgomery, D. C. (1991a). Using fractional factorial designs for robust design process development. *Quality Engineering*, **3**, 193-205.

Montgomery, D. C. (1991b). *Design and Analysis of Experiments*. 3rd Edition, John Wiley and Sons, New York.

Myers, R. H. (1991). Response surface methodology in quality improvement. *Communications in Statistics- Theory and Methods*, **20**, 457-476.

Myers, R. H. and Montgomery, D. C. (1995). *Response Surface Methodology Process and Product Optimization Using Designed Experiments.* John Wiley and Sons, Inc.

Nair, V. N. (1992). Taguchi's parameter design: A panel discussion. *Technometrics*, **34**, 127-161.

Plackett, R. L. and Burman, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika*, **33**, 305-325.

Shoemaker, A. C., Tsui, K. L. and Wu, C. F. J. (1991). Economical experimentation methods for robust design. *Technometrics*, **33**, 415-427.

Taguchi, G. (1959). *Linear Graphs for Orthogonal Arrays and Their Application to Experimental Design, with the Aid of Various Techniques*. Report of Statistical Application Research JUSE, **6**, 1-43.

Taguchi, G. (1987). *System of Experimental Design*. Vol I and II. UNIPUB, New York.

Welch, W. J., Yu, T. K., Kang, S. M. and Sacks, J. (1990). Computer experiments for quality control by parameter design. *Journal of Quality Technology*, **22**, 15-22.

Welch, W. J. and Sacks, J. (1991). A system for quality improvement *via* computer experiments. *Communications in Statistics- Theory and Methods*, **20**, 477-495.

RS/DISCOVER (1988). Cambridge, MA: BBN Software Products Corporation.

ANNEXURE

Combined Array Designs for Three, Four and Five Factors

For $p = 3$

(a) $r = 2, s = 1$

Design 3.1, (1,2,3), (8)						
D. No.	C	N	C × N	C × C	D	D_S
1	1,2	3	13, 23	12	1	1,1,1,1,1,1

Design 3.2, (1,3,6), (8)						
D. No.	C	N	C × N	C × C	D	D_S
1	1,3	6	16,36	13	1	1,1,1,1,1,1

The other designs can be obtained by renaming the control and noise factors.

(b) $r = 1, s = 2$

Design 3.1, (1,2,3), (8)						
D. No.	C	N	C × N	C × C	D	D_S
1	1	2,3	12,13	-	1	1,1,1,1,1

The other designs can be obtained by renaming the control and noise factors.

Design 3.2, (1,3,6), (8)						
D. No.	C	N	C × N	C × C	D	D_S
1	1	3,6	13,16	-	1	1,1,1,1,1

For $p = 4$

(a) $r = 3, s = 1$

Design 4.3, (1,5,6,7), (12)						
D. No.	C	N	C × N	C × C	D	D_S
1	1,5,6	7	17,57,67	15,16,56	0.86	.73,.73,.73,.73,.67,.67,.67,.67,.67,.67

The other designs can be obtained by renaming the control and noise factors.

(b) $r = 1, s = 3$

Design 4.3, (1,5,6,7), (12)						
D. No.	C	N	C × N	C × C	D	D_S
1	1	5,6,7	15,16,17	-	0.95	.91,.91,.91,.91,.89,.89,.89

The other designs can be obtained by renaming the control and noise factors.

(c) $r = 2, s = 2$

Design 4.3, (1,5,6,7), (12)						
D. No.	C	N	C × N	C × C	D	D_S
1	1,5	6,7	16,17,56,57	15	0.88	.78,.78,.78,.78,.67,.67,.67,.67,.89

The other designs can be obtained by renaming the control and noise factors.

For $p = 5$

(a) $r = 4, s = 1$

Design 5.3, (1,2,3,5,6), (18)						
D. No.	C	N	C × N	C × C	D	D_S
1	1,2,3,6	5	15,25,35,65	12,13,16,23, 26,36	0.76	.44,.77,.44,.21,.45,.77,.44, .77,.45,.45, .21,.44,.45,.77,.44

Design 5.5, (1,2,5,6,7), (18)						
D. No.	C	N	C × N	C × C	D	D_S
1	1,5,6,7	2	12,52,62,72	15,16,17,56, 57,67	0.76	.77,.77,.45,.45,.44,.45,.45,.77, .77,.21, .44,.44,.44,.44,.21

(b) $r = 1, s = 4$

Design 5.9, (1,2,3,6,9), (14)						
D. No.	C	N	C × N	C × C	D	D_S
1	3	1,2,6,9	31,32,36,39	-	0.86	.91,.61,.79,.61,.76,.61, .76,.61,.79
2	9	1,2,3,6	91,92,93,96	-	0.86	.91,.61,.61,.76,.79,.61, .61,.79,.76

(c) $r = 3, s = 2$

Design 5.8 (1,3,5,6,8), (16)						
D. No.	C	N	C × N	C × C	D	Ds
1	1,3,5	6,8	16,18,36,38,56,58	13,15,35	0.68	.32,.17,.39,.32,.39,.39,.29,.39, .32,.29,.39,.39, .50,.32
2	1,3,8	5,6	15,16,35,36,85,86	13,18,38	0.68	.39,.17,.32,.32,.39,.29,.39,.39, .32,.39,.29,.32, .50,.39
3	3,5,6	1,8	31,38,51,58,61,68	35,36,56	0.68	.17,.32,.39,.39,.32,.32,.39,.29, .39,.39,.29,.39, .32,.50

(d) $r = 2, s = 3$

Design 5.9, (1,2,3,6,9), (14)						
D. No.	C	N	C × N	C × C	D	Ds
1	1,2	3,6,9	13,16,19,23,26,29	12	0.72	.46,.46,.43,.29,.43,.30,.43,.46, .57,.61,.33,.43
2	1,6	2,3,9	12,13,19,62,63,69	16	0.72	.46,.46,.29,.43,.43,.43,.46,.30, .61,.33,.57,.43

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 203–218

Inference on $P(X < Y)$ for Morgenstern Type Bivariate Exponential Distribution Based on Record Values

Manoj Chacko and Shiny Mathew
Department of Statistics
University of Kerala, Kerala, India

Received: 26 June 2021; Revised: 24 September 2021; Accepted: 18 November 2021

Abstract

In this paper, we consider the problem of estimation of $R = P(X < Y)$, when X and Y are dependent. The maximum likelihood estimates and Bayes estimates of R are obtained based on record values when (X, Y) follows Morgenstern type bivariate exponential distribution. The percentile bootstrap and HPD confidence intervals for R are also obtained. Monte Carlo simulations are carried out to study the accuracy of the proposed estimators.

Key words: Morgenstern type bivariate exponential distribution; Record values; Maximum likelihood estimation; Bayes estimation.

AMS Subject Classifications: 62N05, 62F15

1. Introduction

Record value data arise in a wide variety of practical situations. Examples include destructive stress testing, mateorological analysis, hydrology, seismology, sporting and athletic events and oil and mining surveys. Interest in records has increased steadily over the years since Chandler (1952) formulation. Let $\{X_i, i \geq 1\}$ be a sequence of independent and identically distributed (iid) random variables having an absolutely continuous cumulative distribution function (cdf) $F(x)$ and probability density function (pdf) $f(x)$. An observation X_j is called an upper record if $X_j > X_i$, for every $i < j$ (see Arnold *et al.* 1998, p.8). An analogous definition deals with lower record values. Let (X_1, Y_1), (X_2, Y_2),... be a sequence of iid random variables with common continuous joint cdf $F(x, y), (x, y) \in R \times R$. Let $F_X(x)$ and $F_Y(y)$ be the marginal cdfs of X and Y respectively. Let $R_n, n \geq 1$ be the sequence of upper record values arising from the sequence of X's. Then the Y-variate associated with the X-value, which qualified as the nth record will be called the concomitant of the nth record and will be denoted by $R_{[n]}$. Suppose in an experiment, individuals are measured based on an inexpensive test, and only those individuals whose measurement breaks the previous records are retained for the measurement based on an expensive test; then the resulting data involves record values and concomitants of record values. For a detailed discussion on the distribution theory of concomitants of record values see, Arnold *et al.* (1998), Ahsanullah and Nevzorov (2000), Barakat *et al.* (2013) and Ahsanullah and Shakil (2013). Chacko and Thomas (2006,2008) considered the problem of estimation of parameters of Morgenstern

Corresponding Author: Shiny Mathew
Email: mshiny15@gmail.com

type bivariate logistic distribution and bivariate normal distribution based on concomitants of record values.

The joint pdf of first n upper record values and its concomitants $(\mathbf{R_{(n)}}, \mathbf{R_{[n]}}) = ((R_{(1)}, R_{[1]}), (R_{(2)}, R_{[2]}), \ldots, (R_{(n)}, R_{[n]}))$ is given by

$$f_{(\mathbf{R_{(n)}},\mathbf{R_{[n]}})}(\mathbf{r_{(n)}}, \mathbf{r_{[n]}}) = \prod_{i=1}^{n} f(r_{[i]}|r_{(i)}) f_{1,2,\ldots,n}(r_{(1)}, r_{(2)}, \ldots, r_{(n)}), \tag{1}$$

where $f_{1,2,\ldots,n}(r_{(1)}, r_{(2)}, \ldots, r_{(n)})$ is the joint pdf of first n upper record values and is given by

$$f_{1,2,\ldots,n}(r_{(1)}, r_{(2)}, \ldots, r_{(n)}) = f(r_{(n)}) \prod_{i=1}^{n-1} \frac{f(r_{(i)})}{1 - F(r_{(i)})}. \tag{2}$$

Now a days the inference on $R = P(X < Y)$ is studied in many branches of sciences and social sciences such as psychology, medicine, pedagogy, pharmaceutics and engineering. In the context of reliability the stress-strength model describes the life of a component which has a random strength Y and is subjected to a random stress X. The component fails at the instant that the stress applied to it exceeds the strength and the component will function satisfactorily whenever $X < Y$. Thus $R = P(X < Y)$ is a measure of component reliability. It has found applications in many life testing problems and engineering. The application of R in engineering includes deterioration of rocket motors, static fatigue of ceramic components, fatigue failure of aircraft structures etc. For example, if X and Y are future observations on the stability of an engineering design, then R would be predictive probability that X is less than Y. Similarly, if X and Y represents life times of two electronic devices, then R is the probability that one fails before the other. For more details on applications of R in engineering see, Nadarajah and Kotz (2006).

The estimation of R has been extensively investigated in the literature when X and Y are independent random variables belonging to the same bivariate family of distributions. However, there is a relative little work when X and Y are dependent random variables. The problem of estimating R when the X and Y are dependent was considered by Abu-Salih and Shamseldin (1988), Awad *et al.* (1981), Jana and Roy (1994) and Cramer (2001). Estimation of R when (X, Y) follows bivariate normal distribution has been discussed by Enis and Geisser (1971) and Mukherjee and Saran (1985). Jana(1994) and Hanagal (1995) discussed the estimation of R when (X, Y) follows Marshall-Olkin bivariate exponential distribution. Hanagal (1997) discussed the estimation of R when (X, Y) has a bivariate Pareto distribution. Chacko and Mathew (2019) considered the estimation of $R = P(X < Y)$ for bivariate normal distribution based on ranked set sample. Chacko and Mathew (2020) considered the estimation of $R = P(X < Y)$ for bivariate normal distribution based on record values. In this paper, we focus on estimation of $R = P(X < Y)$ based on upper record values and its concomitants, corresponding to a bivariate random variable (X, Y) which follows a Morgenstern Type Bivariate Exponential distribution (MTBED) with pdf given by (see, Kotz *et al.*, 2000, P.353)

$$f(x,y) = \begin{cases} \theta_1\theta_2 exp(-\theta_1 x - \theta_2 y)[1 + \alpha(1 - 2exp(-\theta_1 x))(1 - 2exp(-\theta_2 y))], \\ \qquad\qquad x > 0, y > 0; -1 \leq \alpha \leq 1; \theta_1 > 0, \theta_2 > 0 \\ 0, \qquad\qquad \text{otherwise} \end{cases} \tag{3}$$

It may be noted that if (X, Y) has a MTBED as defined in (3) then the marginal distributions of both X and Y have exponential distributions. The correlation between X and Y is $\alpha/4$. As α lies between -1 and 1, MTBED accomodates correlation in the range of (-1/4,1/4). Exponential distributions are the most popular and the most applied life time models in many areas, including life testing and reliability studies. Let T_1 and T_2 be two dependent components of a system with lifetimes X and Y respectively. Then $R = P(X < Y)$ is the probability that the first component T_1 fails before second component T_2. If (X,Y) follows a bivariate exponential distribution and the data available are in the form of upper record values and its concomitants then the methods describe in this paper can easily be used to estimate $R = P(X < Y)$.

The organization of the paper is as follows. In section 2, we consider maximum likelihood estimation of R and also obtain the bootstrap confidence interval (CI) based on the maximum likelihood estimator (MLE). In section 3, we consider the Bayes estimation of R using importance sampling method under both symmetric and assymetric loss functions. Section 4 is devoted to some simulation studies and in section 5, we give concluding remarks.

2. Maximum Likelihood Estimation

Let (X,Y) follows MTBED with pdf defined in (3), then $R = P(X < Y)$ is given by

$$\begin{aligned} R &= P(X < Y) \\ &= \frac{\theta_1}{\theta_1 + \theta_2}[1 + \alpha \frac{\theta_1(\theta_1 - \theta_2)}{(2\theta_1 + \theta_2)(2\theta_2 + \theta_1)}]. \end{aligned} \tag{4}$$

If we denote $\theta = (\theta_1, \theta_2, \alpha)$ then we can write R as

$$R = R(\theta).$$

In this section, we obtain the MLE of R for MTBED using record values and its concomitants. Let $(R_{(i)}, R_{[i]}), i = 1, 2, \ldots, n$ be the upper record values and its concomitants arising from MTBED. Then from (1), the likelihood function is given by

$$\begin{aligned} L(\theta) &= (\theta_1\theta_2)^n \prod_{i=1}^{n} exp(-\theta_1 r_{(i)} - \theta_2 r_{[i]})[1 + \alpha(1 - 2exp(-\theta_1 r_{(i)})) \\ &\quad \times (1 - 2exp(-\theta_2 r_{[i]}))] \prod_{i=1}^{n-1} \frac{1}{exp(-\theta_1 r_{(i)})}. \end{aligned}$$

Then the log-likelihood function is given by

$$\begin{aligned} \log L(\theta) &= n \log \theta_1 + n \log \theta_2 - \theta_1 r_{(n)} - \theta_2 \sum_{i=1}^{n} r_{[i]} \\ &\quad + \sum_{i=1}^{n} \log[1 + \alpha(1 - 2exp(-\theta_1 r_{(i)}))(1 - 2exp(-\theta_2 r_{[i]}))]. \end{aligned}$$

Thus we have

$$\frac{\partial \log L}{\partial \theta_1} = \frac{n}{\theta_1} - r_{(n)} + \sum_{i=1}^{n} \frac{2\alpha r_{(i)}(1 - 2exp(-\theta_2 r_{[i]}))exp(-\theta_1 r_{(i)})}{[1 + \alpha(1 - 2exp(-\theta_1 r_{(i)}))(1 - 2exp(-\theta_2 r_{[i]}))]},$$

$$\frac{\partial \log L}{\partial \theta_2} = \frac{n}{\theta_2} - \sum_{i=1}^{n} r_{[i]} + \sum_{i=1}^{n} \frac{2\alpha r_{[i]}(1-2exp(-\theta_1 r_{(i)}))exp(-\theta_2 r_{[i]})}{[1+\alpha(1-2exp(-\theta_1 r_{(i)}))(1-2exp(-\theta_2 r_{[i]}))]}$$

and

$$\frac{\partial \log L}{\partial \alpha} = \sum_{i=1}^{n} \frac{(1-2exp(-\theta_1 r_{(i)}))(1-2exp(-\theta_2 r_{[i]}))}{[1+\alpha(1-2exp(-\theta_1 r_{(i)}))(1-2exp(-\theta_2 r_{[i]}))]}.$$

The MLEs of θ_1, θ_2 and α can be obtained as the solutions of the following non-linear equations

$$\frac{n}{\theta_1} - r_{(n)} + \sum_{i=1}^{n} \frac{2\alpha r_{(i)}(1-2exp(-\theta_2 r_{[i]}))exp(-\theta_1 r_{(i)})}{[1+\alpha(1-2exp(-\theta_1 r_{(i)}))(1-2exp(-\theta_2 r_{[i]}))]} = 0,$$

$$\frac{n}{\theta_2} - \sum_{i=1}^{n} r_{[i]} + \sum_{i=1}^{n} \frac{2\alpha r_{[i]}(1-2exp(-\theta_1 r_{(i)}))exp(-\theta_2 r_{[i]})}{[1+\alpha(1-2exp(-\theta_1 r_{(i)}))(1-2exp(-\theta_2 r_{[i]}))]} = 0$$

and

$$\sum_{i=1}^{n} \frac{(1-2exp(-\theta_1 r_{(i)}))(1-2exp(-\theta_2 r_{[i]}))}{[1+\alpha(1-2exp(-\theta_1 r_{(i)}))(1-2exp(-\theta_2 r_{[i]}))]} = 0.$$

If $\hat{\theta} = (\hat{\theta_1}, \hat{\theta_2}, \hat{\alpha})$ is the MLE of θ obtained by solving the above nonlinear equations, then the MLE of R is given by

$$\hat{R}_{ML} = \frac{\hat{\theta_1}}{\hat{\theta_1}+\hat{\theta_2}} \left[1 + \hat{\alpha}\frac{\hat{\theta_1}(\hat{\theta_1}-\hat{\theta_2})}{(2\hat{\theta_1}+\hat{\theta_2})(2\hat{\theta_2}+\hat{\theta_1})}\right]. \tag{5}$$

2.1. Asymptotic confidence interval

In this subsection, the asymptotic confidence interval of R is obtained. Towards this, we consider the observed information matrix of θ. Let

$$I(\theta) = \begin{bmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{bmatrix},$$

where

$$I_{11} = \frac{\partial^2 \log L}{\partial \theta_1^2} = \frac{-n}{\theta_1^2} - \sum_{i=1}^{n} 2\alpha r_{(i)}(1-2exp(-\theta_2 r_{[i]})) \times \left(\frac{1-\alpha r_{(i)} exp(-\theta_1 r_{(i)})(1-2exp(-\theta_2 r_{[i]})}{[1+\alpha(1-2exp(-\theta_1 r_{(i)})(1-2exp(-\theta_2 r_{[i]})]^2}\right),$$

$$I_{12} = \frac{\partial^2 \log L}{\partial \theta_1 \partial \theta_2} = \sum_{i=1}^{n} \frac{2\alpha r_{(i)} exp(-\theta_1 r_{(i)})}{[1+\alpha(1-2exp(-\theta_1 r_{(i)})(1-2exp(-\theta_2 r_{[i]})]^2},$$

$$I_{13} = \frac{\partial^2 \log L}{\partial\theta_1\partial\alpha} = \sum_{i=1}^{n} \frac{2r_{(i)}exp(-\theta_1 r_{(i)})(1-2exp(-\theta_2 r_{[i]})}{[1+\alpha(1-2exp(-\theta_1 r_{(i)})(1-2exp(-\theta_2 r_{[i]})]^2},$$

$$I_{21} = \frac{\partial^2 \log L}{\partial\theta_2\partial\theta_1} = \sum_{i=1}^{n} \frac{2\alpha r_{[i]}exp(-\theta_2 r_{[i]})}{[1+\alpha(1-2exp(-\theta_1 r_{(i)})(1-2exp(-\theta_2 r_{[i]})]^2},$$

$$I_{22} = \frac{\partial^2 \log L}{\partial\theta_2^2} = \frac{-n}{\theta_2^2} + \sum_{i=1}^{n} 2\alpha r_{[i]}(1-2exp(-\theta_1 r_{(i)}) \times \left(\frac{1-\alpha r_{[i]}exp(-\theta_2 r_{[i]})(1-2exp(-\theta_1 r_{(i)})}{[1+\alpha(1-2exp(-\theta_1 r_{(i)})(1-2exp(-\theta_2 r_{[i]})]^2}\right),$$

$$I_{23} = \frac{\partial^2 \log L}{\partial\theta_2\partial\alpha} = \sum_{i=1}^{n} \frac{2r_{[i]}exp(-\theta_2 r_{[i]})(1-2exp(-\theta_1 r_{(i)}))}{[1+\alpha(1-2exp(-\theta_1 r_{(i)})(1-2exp(-\theta_2 r_{[i]})]^2},$$

$$I_{31} = \frac{\partial^2 \log L}{\partial\alpha\partial\theta_1} = \sum_{i=1}^{n} \frac{(1-2exp(-\theta_2 r_{[i]}))}{[1+\alpha(1-2exp(-\theta_1 r_{(i)})(1-2exp(-\theta_2 r_{[i]})]^2},$$

$$I_{32} = \frac{\partial^2 \log L}{\partial\alpha\partial\theta_2} = \sum_{i=1}^{n} \frac{(1-2exp(-\theta_1 r_{(i)}))}{[1+\alpha(1-2exp(-\theta_1 r_{(i)})(1-2exp(-\theta_2 r_{[i]})]^2}$$

and

$$I_{33} = \frac{\partial^2 \log L}{\partial\alpha^2} = -\sum_{i=1}^{n} \frac{(1-2exp(-\theta_1 r_{(i)})^2(1-2exp(-\theta_2 r_{[i]})^2}{[1+\alpha(1-2exp(-\theta_1 r_{(i)})(1-2exp(-\theta_2 r_{[i]})]^2}.$$

Let $\hat{\theta} = (\hat{\theta_1}, \hat{\theta_2}, \hat{\alpha})$ be the MLE of θ. Then the observed information matrix is given by $I(\hat{\theta})$. Thus by using delta method, we obtain the asymptotic distribution of $\hat{R}$. For that we have

$$\begin{aligned} \hat{Var}(\hat{R}_{ML}) &= \hat{Var}(R(\hat{\theta})) \\ &\approx h(\hat{\theta})[I(\hat{\theta})]^{-1}h(\hat{\theta})^{\top}. \end{aligned}$$

where

$$h(\hat{\theta}) = \left(\frac{\partial R}{\partial\theta_1}, \frac{\partial R}{\partial\theta_2}, \frac{\partial R}{\partial\alpha}\right)\Big|_{\theta=\hat{\theta}}$$

with

$$\begin{aligned} \frac{\partial R}{\partial\theta_1} &= \frac{\theta_2}{(\theta_1+\theta_2)^2} + \alpha\frac{\theta_1\theta_2(4\theta_2^3+7\theta_1\theta_2^2-2\theta_1^3)}{(2\theta_1^3+2\theta_2^3+7\theta_1\theta_2^2+7\theta_1^2\theta_2)^2}, \\ \frac{\partial R}{\partial\theta_2} &= \frac{-\theta_1}{(\theta_1+\theta_2)^2} + \alpha\frac{\theta_1^2(2\theta_1^3-4\theta_2^3-7\theta_1\theta_2^2)}{(2\theta_1^3+2\theta_2^3+7\theta_1\theta_2^2+7\theta_1^2\theta_2)^2} \end{aligned}$$

and

$$\frac{\partial R}{\partial\alpha} = \frac{\theta_1^2\theta_2}{(2\theta_1^3+2\theta_2^3+7\theta_1\theta_2^2+7\theta_1^2\theta_2)}.$$

Thus $\frac{\hat{R}-R}{\sqrt{\hat{var}(\hat{R})}}$ is aymptotically distributed as $N(0,1)$. Thus a $(1-\nu)100\%$ confidence interval for R based on the MLE is $(\hat{R} - z_{\nu/2}\sqrt{\hat{Var}(\hat{R})}, \hat{R} + z_{\nu/2}\sqrt{\hat{Var}(\hat{R})})$, where $z_{\nu/2}$ is the $(1-\nu/2)100th$ percentile of N(0,1).

2.2. Bootstrap confidence interval

In this subsection, we consider percentile bootstrap CI for R based on MLEs. For that we do the following.

1. Compute the MLEs $\hat{\theta_1}^{(0)}$, $\hat{\theta_2}^{(0)}$ and $\hat{\alpha}^{(0)}$ of θ_1, θ_2 and α using original record values and its concomitants and set k=1.
2. Generate a bootstrap sample using $\hat{\theta_1}^{(0)}$, $\hat{\theta_2}^{(0)}$ and $\hat{\alpha}^{(0)}$ from MTBED and obtain the MLEs $\hat{\theta_1}^{(k)}$, $\hat{\theta_2}^{(k)}$ and $\hat{\alpha}^{(k)}$ using the bootstrap sample.
3. Obtain the MLE $\hat{R}_k = R(\hat{\theta_1}^{(k)}, \hat{\theta_2}^{(k)}, \hat{\alpha}^{(k)})$.
4. Set $k = k + 1$.
5. Repeat steps (2)to(4) B times to obtain the MLEs $\hat{R}_1, \hat{R}_2, \cdots, \hat{R}_B$, for sufficiently large B .
6. Arrange $\hat{R}_1, \hat{R}_2, \cdots, \hat{R}_B$ in ascending order as $\hat{R}_{(1)} \leq \hat{R}_{(2)}, \ldots, \leq \hat{R}_{(B)}$. Then the $100(1-\nu)$ percentile bootstrap CI for R is given by $\left(\hat{R}_{([B(\nu/2)])}, \hat{R}_{([B(1-\nu/2)])}\right)$, [.] is the greatest integer function.

3. Bayesian Estimation

In this section, we consider Bayesian estimation of R for MTBED under symmetric as well as asymmetric loss functions. For symmetric loss function we consider squared error loss (SEL) function and for asymmetric loss function we consider both LINEX loss (LL) and the general entropy loss (EL) function. The Bayes estimate of any parameter μ under SEL function is the posterior mean of μ. The Bayes estimate of any parameter μ under LL is given by

$$\hat{d}_{LB}(\mu) = \frac{-1}{h} \log\{E_\mu(e^{-h\mu}|\underline{x})\}, h \neq 0, \tag{6}$$

provided E_μ exists. The Bayes estimate of any parameter μ under EL function is given by

$$\hat{d}_{EB}(\mu) = (E_\mu(\mu^{-q}|\underline{x}))^{\frac{-1}{q}}, q \neq 0, \tag{7}$$

provided E_μ exists.
Let $(R_{(i)}, R_{[i]}), i = 1, 2, \ldots, n$ be the vector of record value and its concomitants arising from MTBED $(\theta_1, \theta_2, \alpha)$. Then from (1) the likelihood function is given by

$$\begin{aligned} L(\theta) &= (\theta_1\theta_2)^n \prod_{i=1}^{n} exp(-\theta_1 r_{(i)} - \theta_2 r_{[i]})[1 + \alpha(1 - 2exp(-\theta_1 r_{(i)})) \\ &\quad \times(1 - 2exp(-\theta_2 r_{[i]}))] \prod_{i=1}^{n-1} \frac{1}{exp(-\theta_1 r_{(i)})}. \end{aligned}$$

Assume that the prior distributions of $\theta_1 \sim$ Gamma(a, b), $\theta_2 \sim$ Gamma(c, d) and $\alpha \sim U[-1, 1]$. Thus the prior density functions of θ_1, θ_2 and α are respectively given by

$$\pi_1(\theta_1|a, b) = \frac{b^a}{\Gamma(a)} \theta_1^{a-1} e^{-b\theta_1}; a > 0, b > 0, \tag{8}$$

$$\pi_2(\theta_2|c,d) = \frac{d^c}{\Gamma(c)}\theta_2^{c-1}e^{-d\theta_2}; c > 0, d > 0 \tag{9}$$

and

$$\pi_3(\alpha) = \frac{1}{2}, -1 \leq \alpha \leq 1. \tag{10}$$

Then the joint prior distribution of θ is given by

$$\pi(\theta) = \frac{1}{2}\frac{b^a}{\Gamma(a)}\frac{d^c}{\Gamma(c)}\theta_1^{a-1}\theta_2^{c-1}e^{-b\theta_1}e^{-d\theta_2} \tag{11}$$

Then the joint posterior density of θ is given by

$$\pi^*(\theta) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)d\theta}. \tag{12}$$

Therefore the Bayes estimate of $R(\theta)$ under SEL, LL and EL are respectively given by

$$\hat{R}_S = \frac{\int R(\theta)L(\theta)\pi(\theta)d\theta}{\int L(\theta)\pi(\theta)d\theta}, \tag{13}$$

$$\hat{R}_L = \frac{-1}{h}\log\frac{\int e^{-hR(\theta)}L(\theta)\pi(\theta)d\theta}{\int L(\theta)\pi(\theta)d\theta} \tag{14}$$

and

$$\hat{R}_E = \left[\frac{\int R(\theta)^{-q}L(\theta)\pi(\theta)d\theta}{\int L(\theta)\pi(\theta)d\theta}\right]^{\frac{-1}{q}}. \tag{15}$$

It is not possible to compute (13)-(15) explicitly. The popular approach to perform the integrals (13) to (15) is the Markov Chain Monte Carlo (MCMC) method which replace the expectation values of the parameters with the average values over Monte Carlo (posterior) samples obtained through the Markov Chain. A drawback of the MCMC method is that the time series of the Monte Carlo samples obtained through the Markov Chain are usually correlated. The importance sampling method introduces an importance sampling density which should be handled easily and can generate Monte Carlo data randomly. The Monte Carlo data generated randomly by the importance sampling method can be autocorrelation-free. The autocorrelation-free nature of the importance sampling could be considered to be an advantage over the MCMC method. Thus we consider importance sampling method to find the Bayes estimates for R.

3.1. Importance sampling method

In this subsection, we consider the importance sampling method to generate samples from the posterior distributions and then find the Bayes estimate of R. The numerator in the posterior distribution given in (12) can be written as

$$L(\theta)\pi(\theta) \propto Q(\theta)f_1(\theta_1)f_2(\theta_2)f_3(\alpha),$$

where

$$Q(\theta) = \prod_{i=1}^{n}[1 + \alpha(1 - 2exp(-\theta_1 r_{(i)}))(1 - 2exp(-\theta_2 r_{[i]}))], \tag{16}$$

$$f_1(\theta_1) \propto \theta_1^{n+a-1} exp[-\theta_1(r_{(n)}+b)] \tag{17}$$

$$f_2(\theta_2) \propto \theta_2^{m+c-1} exp\left[-\theta_2\left(\sum_{i=1}^{n} r_{[i]}+d\right)\right] \tag{18}$$

and

$$f_3(\alpha) = \frac{1}{2}. \tag{19}$$

Thus from (17) we can see that distribution of θ_1 follows Gamma distribution with parameters $(n+a)$ and $(r_{(n)}+b)$. Again from (18) one can see that distribution of θ_2 follows gamma distribution with parameters $(m+c)$ and $(\sum_{i=1}^{n} r_{[i]}+d)$. From (19) we can see that $\alpha \sim U(-1,1)$. Let $\theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \alpha^{(t)})$, $t = 1, 2, \ldots, N$ be the observations generated from (17),(18) and (19) respectively. Then by importance sampling method the Bayes estimators under SEL, LL and EL given by (13)-(15) can be respectively written as

$$\hat{R}_S = \frac{\sum_{t=1}^{N} R(\theta^{(t)})Q(\theta^{(t)})}{\sum_{t=1}^{N} Q(\theta^{(t)})}, \tag{20}$$

$$\hat{R}_L = \frac{-1}{h}\log\left[\frac{\sum_{t=1}^{N} exp(-hR(\theta^{(t)})Q(\theta^{(t)})}{\sum_{t=1}^{N} Q(\theta^{(t)})}\right] \tag{21}$$

and

$$\hat{R}_E = \left[\frac{\sum_{t=1}^{N} (R(\theta^{(t)}))^{-q} Q(\theta^{(t)})}{\sum_{t=1}^{N} Q(\theta^{(t)})}\right]^{-1/q}. \tag{22}$$

3.2. HPD interval

In this subsection, we construct HPD intervals for R as described in Chen and Shao (1999). In this method a Monte Carlo approach is used to approximate the pth quantile of R and then obtain an estimate of Bayesian credible or HPD interval. Define $R_t = R(\theta^{(t)})$, where $\theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \alpha^{(t)})$ for $t = 1, 2, \ldots, M$ are posterior samples generated respectively from (17), (18) and (19) for θ_1, θ_2 and α. Let $R_{(t)}$ be the ordered values of R_t. Define

$$w_t = \frac{Q(\theta^{(t)})}{\sum_{t=1}^{M} Q(\theta^{(t)})}.$$

Then the pth quantile of R can be estimated as

$$\hat{R}^{(p)} = \begin{cases} R_1 & \text{if} \quad p = 0 \\ R_{(i)} & if \quad \sum_{j=1}^{i-1} w_{(j)} < p < \sum_{j=1}^{i} w_{(j)}, \end{cases}$$

where $w_{(j)}$ is the weight associated with jth ordered value $R_{(j)}$. Then the $100(1-\nu)\%$, $0 < \nu < 1$, confidence interval for R is given by $(\hat{R}^{(j/M)}, \hat{R}^{(j+[(1-\nu)M])/M)})$, $j = 1, 2, \ldots, M$, where [.] is the greatest integer function. Then the required HPD interval for R is the interval with smallest width.

4. Simulation Study

In this section, we carry out a simulation study for illustrating the estimation procedures developed in the previous sections. First we obtain the MLE of R using (5). We have obtained the bias and MSE of MLEs for different combinations of θ_1, θ_2 and α and are given in Table 1. The bootstrap CI for R are also obtained. The average interval length (AIL) and coverage probability (CP) are also obtained and are included in Table 1. We consider four sets of true parameter values, (θ_1, θ_2)= (5,1), (3,2), (2,4) and (0.5,5). Since prior distribution of θ_1 follows gamma distribution with mean $\frac{a}{b}$, we take the hyperparameters for θ_1 =5, 3, 2, and 0.5 as (a, b)=(5,1), (3,1), (2,1) and (0.5,1) respectively. Similarly we take the hyperparamerters of θ_2=1,2,4, and 5 as (c, d) = (1,1), (2,1), (4,1) and (5,1). We have obtained the Bayes estimators for R of MTBED under SEL, LL and EL functions using importance sampling method and are given in Table 2. For importance sampling method we use the following algorithm.

1. Generate n upper record values and its concomiants from MTBED distribution with parameters θ_1, θ_2 and α.
2. Calculate the Bayes estimators of R as described below.
   (a) Set t=1
   (b) Generate $\theta_1^{(t)}$ from Gamma distribution with parameters $n + a$ and $r_{(n)} + b$.
   (c) Generate $\theta_2^{(t)}$ from Gamma distribution with parameters $m + c$ and $\sum_{i=1}^{n} r_{[i]} + d$.
   (d) Generate $\alpha^{(t)}$ from $Uniform(-1, 1)$ distribution.
   (e) Calculate $\hat{R}(\theta^{(t)})$ using (5) and $Q(\theta^{(t)})$ using (16).
   (f) Set t=t+1.
   (g) Repeat steps (b) to (f) 50,000 times.
   (h) Calculate the Bayes estimators for R using (20)-(22)
3. Repeat steps 1 and 2 for 500 times to obtain the estimators $\hat{R}_1, \hat{R}_2, \cdots, \hat{R}_{500}$.
4. Calculate the average bias= $\frac{1}{500}\sum_i^{500}(\hat{R}_i - R)$ and $MSE = \frac{1}{500}\sum_i^{500}(\hat{R}_i - \bar{R}) + bias^2$ of the estimators.

We repeat the simulation study for different values of α and n. From the tables we can see that the bias and MSE of all estimators decrease when the number of records increase. We can also see that among different estimators Bayes estimator under SEL have minimum bias and MSE. From Table 1 we can see that the AILs of HPD intervals are smaller than that of bootstrap CIs and the CPs of HPD intervals are higher than that of bootstrap CIs.

Table 1: The AIL and CP for bootstrap CIs and HPD intervals

α	n	θ_1	θ_2	R	Bootstrap		HPD	
					AIL	CP	AIL	CP
-0.75	6	5	1	0.67100	0.22830	0.85	0.11755	0.94
		3	2	0.57589	0.21336	0.85	0.13603	0.92
		2	4	0.34583	0.18995	0.87	0.12971	0.93
		0.5	5	0.09334	0.18229	0.86	0.12178	0.92
	8	5	1	0.67100	0.17457	0.87	0.14590	0.95
		3	2	0.57589	0.16680	0.87	0.11195	0.94
		2	4	0.34583	0.15912	0.88	0.12494	0.95
		0.5	5	0.09334	0.15215	0.88	0.11086	0.94
	10	5	1	0.67100	0.14993	0.88	0.11234	0.96
		3	2	0.57589	0.13375	0.87	0.08454	0.96
		2	4	0.34583	0.16985	0.89	0.12985	0.95
		0.5	5	0.09334	0.21065	0.85	0.14707	0.95
-0.5	6	5	1	0.72511	0.23566	0.85	0.13176	0.94
		3	2	0.58393	0.17582	0.84	0.11582	0.93
		2	4	0.34167	0.17737	0.84	0.13245	0.93
		0.5	5	0.09253	0.16639	0.86	0.11629	0.94
	8	5	1	0.72511	0.18512	0.86	0.12621	0.95
		3	2	0.58393	0.15397	0.86	0.13670	0.95
		2	4	0.34167	0.14288	0.88	0.12190	0.94
		0.5	5	0.09253	0.15278	0.87	0.13116	0.95
	10	5	1	0.72511	0.14377	0.88	0.11098	0.94
		3	2	0.58393	0.16415	0.87	0.18154	0.95
		2	4	0.34167	0.16366	0.88	0.08478	0.96
		0.5	5	0.09253	0.23831	0.85	0.12197	0.95
-0.25	6	5	1	0.77922	0.22563	0.85	0.13355	0.93
		3	2	0.59196	0.14293	0.84	0.11844	0.94
		2	4	0.33750	0.14189	0.85	0.12333	0.93
		0.5	5	0.09172	0.16277	0.86	0.16214	0.94
	8	5	1	0.77922	0.15815	0.86	0.12882	0.95
		3	2	0.59196	0.14140	0.87	0.11425	0.95
		2	4	0.33750	0.13703	0.85	0.12929	0.95
		0.5	5	0.09172	0.14633	0.88	0.11552	0.94
	10	5	1	0.77922	0.12808	0.87	0.12552	0.95
		3	2	0.59196	0.13565	0.89	0.12982	0.96
		2	4	0.33750	0.12893	0.88	0.11873	0.95
		0.5	5	0.09172	1.12536	0.89	0.11320	0.96

Table 1: Continued

α	n	θ_1	θ_2	R	Bootstrap		HPD	
					AIL	CP	AIL	CP
0.25	6	5	1	0.88745	0.15200	0.84	0.09624	0.93
		3	2	0.60804	0.17320	0.84	0.11650	0.93
		2	4	0.32917	0.18756	0.85	0.08888	0.92
		0.5	5	0.09010	0.16752	0.85	0.09039	0.92
	8	5	1	0.88745	0.13493	0.85	0.09131	0.95
		3	2	0.60804	0.16607	0.85	0.11750	0.94
		2	4	0.32917	0.16677	0.86	0.08159	0.95
		0.5	5	0.09010	0.15563	0.86	0.08744	0.95
	10	5	1	0.88745	0.12890	0.86	0.08205	0.93
		3	2	0.60804	0.13831	0.87	0.10826	0.95
		2	4	0.32917	0.12724	0.88	0.07175	0.95
		0.5	5	0.09010	0.13398	0.88	0.07624	0.96
0.5	6	5	1	0.94156	0.16974	0.85	0.11738	0.94
		3	2	0.61607	0.15253	0.86	0.12961	0.93
		2	4	0.32500	0.14018	0.83	0.11854	0.94
		0.5	5	0.08929	0.14557	0.84	0.12974	0.95
	8	5	1	0.94156	0.13592	0.85	0.11546	0.94
		3	2	0.61607	0.13528	0.85	0.10860	0.95
		2	4	0.32500	0.12432	0.86	0.10255	0.96
		0.5	5	0.08929	0.12819	0.88	0.11897	0.95
	10	5	1	0.94156	0.12177	0.88	0.12423	0.94
		3	2	0.61607	0.12387	0.86	0.12557	0.96
		2	4	0.32500	0.11362	0.88	0.12771	0.96
		0.5	5	0.08929	0.11695	0.89	0.08317	0.95
0.75	6	5	1	0.99567	0.18762	0.84	0.11731	0.93
		3	2	0.62411	0.17517	0.85	0.12341	0.92
		2	4	0.32083	0.16081	0.83	0.09304	0.94
		0.5	5	0.08847	0.16947	0.84	0.09253	0.94
	8	5	1	0.99567	0.14872	0.85	0.12235	0.93
		3	2	0.62411	0.13736	0.87	0.13994	0.95
		2	4	0.32083	0.12521	0.87	0.09638	0.95
		0.5	5	0.08847	0.13757	0.88	0.09070	0.96
	10	5	1	0.99567	0.12422	0.86	0.10111	0.95
		3	2	0.62411	0.12523	0.88	0.11063	0.96
		2	4	0.32083	0.12777	0.87	0.08842	0.96
		0.5	5	0.08847	0.12388	0.89	0.09357	0.96

Table 2: The bias and MSE for MLE and Bayes estimators for R

α	n	θ_1	θ_2	MLE		SE		LL		EL	
				Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
-0.75	6	5	1	0.15841	0.12508	0.12893	0.02014	0.13763	0.02264	0.14868	0.02623
		3	2	-0.12726	0.12173	0.02384	0.00752	0.00917	0.00907	0.01595	0.00935
		2	4	0.12681	0.15718	-0.01435	0.00854	0.00669	0.00825	0.01677	0.00864
		0.5	5	0.19078	0.16283	0.00522	0.00162	0.02135	0.00240	0.02290	0.00256
	8	5	1	-0.14307	0.11006	0.12063	0.01724	0.12942	0.01956	0.14035	0.02282
		3	2	0.11088	0.11727	-0.00596	0.00775	0.00451	0.00815	0.01270	0.00837
		2	4	0.11998	0.09947	-0.01343	0.00837	0.00377	0.00653	0.01565	0.00675
		0.5	5	0.16437	0.14880	0.00417	0.00158	0.01192	0.00195	0.01295	0.00216
	10	5	1	0.12832	0.10176	0.10952	0.01594	0.11912	0.01818	0.13063	0.02139
		3	2	-0.10925	0.10386	-0.00592	0.00612	0.00388	0.00769	0.00961	0.00783
		2	4	0.11148	0.09322	0.01237	0.00576	0.00108	0.00549	0.00530	0.00564
		0.5	5	0.13270	0.08616	0.00172	0.00149	0.00901	0.00178	0.00985	0.00185
-0.5	6	5	1	-0.14634	0.13198	0.07667	0.00944	0.08515	0.00995	0.09591	0.01322
		3	2	0.17554	0.18434	0.01319	0.00853	0.01647	0.00841	0.02230	0.00877
		2	4	0.19828	0.15297	-0.00146	0.00621	0.00271	0.00981	0.00674	0.00915
		0.5	5	-0.12645	0.16493	0.00700	0.00122	0.01692	0.00178	0.01831	0.00190
	8	5	1	0.13805	0.12617	0.06490	0.00858	0.07459	0.00898	0.08654	0.01228
		3	2	0.15785	0.13086	-0.00762	0.00762	0.00343	0.00719	0.01002	0.00732
		2	4	0.11615	0.14922	-0.00123	0.00580	-0.00863	0.00826	0.00412	0.00845
		0.5	5	0.10410	0.14556	0.00433	0.00095	0.01231	0.00163	0.01315	0.00128
	10	5	1	-0.12854	0.11433	0.06238	0.00673	0.07095	0.00798	0.08148	0.00989
		3	2	0.12208	0.12936	-0.00613	0.00641	0.00849	0.00665	0.00884	0.00678
		2	4	0.12199	0.13811	0.00108	0.00476	-0.00417	0.00712	0.00353	0.00730
		0.5	5	0.08247	0.11382	-0.00365	0.00079	0.00789	0.00109	0.00884	0.00104

Table 2: Continued

α	n	θ_1	θ_2	MLE		SE		LL		EL	
				Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
-0.25	6	5	1	0.12725	0.15281	0.01976	0.00428	0.02879	0.00486	0.04018	0.00604
		3	2	-0.14357	0.12200	0.02778	0.00977	-0.01681	0.00880	-0.01049	0.00871
		2	4	0.12894	0.14564	-0.00505	0.00807	0.01759	0.00801	0.02039	0.00841
		0.5	5	0.12438	0.15906	0.00797	0.00146	0.01257	0.00193	0.01381	0.00204
	8	5	1	0.11938	0.12006	0.01673	0.00374	0.02615	0.00417	0.03892	0.00518
		3	2	-0.14834	0.10475	-0.01844	0.00855	0.01445	0.00796	0.01226	0.00819
		2	4	-0.11049	0.13814	-0.00708	0.00782	0.01541	0.00785	0.02008	0.00818
		0.5	5	0.11796	0.13364	0.00640	0.00118	0.01118	0.00149	0.01213	0.00155
	10	5	1	0.10124	0.10853	0.01262	0.00255	0.02425	0.00304	0.03439	0.00396
		3	2	0.11841	0.11926	-0.01207	0.00800	-0.01166	0.00735	-0.00628	0.00730
		2	4	0.09921	0.12727	0.00337	0.00563	0.01496	0.00534	0.00435	0.00546
		0.5	5	-0.09222	0.11548	-0.00393	0.00071	0.00190	0.00077	0.00565	0.00079
0.25	6	5	1	0.16839	0.12261	-0.07787	0.00996	-0.08808	0.01129	-0.09631	0.01265
		3	2	-0.13160	0.11563	-0.02081	0.01138	-0.02764	0.01163	-0.03948	0.01314
		2	4	0.12543	0.10017	0.01213	0.00565	0.01863	0.00848	0.01393	0.00552
		0.5	5	0.12936	0.10947	0.01771	0.00260	0.01652	0.00477	0.00358	0.00189
	8	5	1	-0.15803	0.11962	-0.07764	0.00917	-0.08717	0.01043	-0.09485	0.01169
		3	2	-0.12533	0.10829	-0.01730	0.00680	-0.02298	0.00697	-0.03243	0.00783
		2	4	0.11079	0.09383	0.01208	0.00795	0.00802	0.00771	0.00669	0.00786
		0.5	5	0.12187	0.09976	0.01282	0.00180	0.01196	0.00145	0.00282	0.00114
	10	5	1	0.13589	0.11125	-0.06289	0.00709	-0.07466	0.00835	-0.08383	0.00965
		3	2	0.11101	0.10076	-0.01277	0.00537	-0.01760	0.00547	0.02536	0.00601
		2	4	-0.10887	0.08118	0.00638	0.00149	0.00141	0.00470	-0.01720	0.00509
		0.5	5	-0.11165	0.08092	0.01127	0.00091	0.01059	0.00088	0.00209	0.00069

Table 2: Continued

α	n	θ_1	θ_2	MLE		SE		LL		EL	
				Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
0.5	6	5	1	0.15162	0.11095	0.11730	0.01587	-0.12781	0.01819	-0.13604	0.02023
		3	2	-0.13110	0.09081	-0.04190	0.00861	-0.04643	0.00895	0.05423	0.00998
		2	4	0.11021	0.09824	0.02081	0.00658	0.01599	0.00629	-0.08451	0.00617
		0.5	5	0.15137	0.10037	0.01325	0.00631	0.02208	0.00219	0.01924	0.00158
	8	5	1	-0.13262	0.09341	-0.11504	0.01456	0.12539	0.01776	-0.12349	0.01969
		3	2	0.12683	0.08046	-0.02679	0.00650	-0.03326	0.00686	0.04428	0.00810
		2	4	-0.10137	0.07384	0.01085	0.00570	0.00701	0.00552	-0.07148	0.00564
		0.5	5	0.13209	0.08589	0.00947	0.00538	0.00871	0.00133	0.00729	0.00110
	10	5	1	0.12558	0.08917	-0.11237	0.01364	-0.12415	0.01608	0.11328	0.02029
		3	2	-0.11186	0.06539	-0.01501	0.00628	-0.02041	0.00641	-0.02901	0.00709
		2	4	-0.10113	0.06314	-0.00301	0.00439	-0.00617	0.00433	-0.01820	0.00465
		0.5	5	0.11845	0.07321	0.00211	0.00089	0.00453	0.00086	0.00315	0.00076
0.75	6	5	1	0.17269	0.10671	-0.18041	0.03485	-0.19013	0.03819	-0.19789	0.05107
		3	2	0.14495	0.11425	-0.02737	0.00675	-0.05236	0.00700	0.04044	0.00786
		2	4	0.13727	0.10991	0.02759	0.00617	0.03377	0.00680	0.03682	0.00923
		0.5	5	0.13290	0.11231	0.01795	0.00222	0.01814	0.00214	0.02786	0.00728
	8	5	1	0.12592	0.09042	-0.15980	0.02955	0.16805	0.02916	0.18938	0.04232
		3	2	-0.12520	0.07051	-0.02599	0.00568	-0.04199	0.00612	0.05215	0.00734
		2	4	0.12947	0.09929	0.01740	0.00562	0.02961	0.00582	-0.02351	0.00736
		0.5	5	0.11302	0.08131	0.01433	0.00182	0.01335	0.00174	0.01885	0.00234
	10	5	1	-0.12679	0.07175	-0.11297	0.02318	-0.13501	0.02640	-0.17271	0.03915
		3	2	-0.11129	0.06732	-0.01954	0.00459	-0.03385	0.00583	-0.04074	0.00651
		2	4	0.12162	0.06734	0.00728	0.00419	0.00826	0.00508	0.00897	0.00518
		0.5	5	0.10832	0.06674	0.00415	0.00060	0.00363	0.00088	0.00858	0.00073

5. Illustration Using Simulated Data

In this section, we illustrate the estimation procedures developed in the previous sections using a simulated data. For that we have generated 10 upper record values and its concomitants from MTBED with parameters $\theta_1 = 2$, $\theta_2 = 1$ and $\alpha = 0.5$. The generated record values and its concomitants are given below.

i	1	2	3	4	5	6	7	8	9	10
$r_{(i)}$	0.201	0.383	0.868	1.433	1.589	1.7034	2.258	3.123	3.657	4.166
$r_{[i]}$	0.245	0.066	0.563	3.379	0.685	0.411	1.111	3.526	2.721	5.317

Based on the simulated data we have obtained the MLE of $R = P(X < Y)$ and also the bootstrap CL of R based on the MLE. For the Bayesian estimation we took hyperparameters as $a = 2$, $b = 1$, $c = 2$ and $d = 2$. The HPD interval of R under SEL is also obtained. The estimated values are given below.

MLE (Bootstrap CI)	Bayes estimates		
	SEL (HPD Interval)	LL	EL
0.6375 (0.4124,0.7124)	0.6587 (0.4841, 0.7124)	0.6457	0.6387

6. Conclusion

In this work, we considered the problem of estimation of $R = P(X < Y)$ for Morgenstern type bivariate exponential distribution using record values and its concomitants. The maximum likelihood and Bayesian estimators were obtained for R. For obtaining the Bayes estimates, importance sampling method was applied. Based on the simulation study we concluded that among different estimators, Bayes estimators under squared error loss function perform better in terms of bias and MSE. AILs of HPD intervals are smaller and the associated CPs are higher than that of bootstrap confidence intervals.

Acknowledgements

The authors would like to thank the editor and the referees for the constructive comments.

References

Abu-Salih, M. S. and Shamseldin, A. A. (1988). Bayesian estimation of $P(X < Y)$ for bivariate exponential distribution. *Arab Gulf Journal of Scientific Research. A, Mathematical and Physical Sciences*, **6(1)**, 17–26.

Ahsanullah, M. and Nevzorov, V. B. (2000). Some distributions of induced record values. *Biometrika,* **42**, 1069–1081.

Ahsanullah, M. and Shakil, M. (2013). Characterizations of Rayleigh distribution based on order Statistics and record Values. *Bulletin of the Malaysian Mathematical Sciences Society*, **36(3)**, 625–635.

Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1998). *Records*, Wiley, New York.

Awad, A., Azzam, M. and Hamdan, M. (1981). Some inference results on $P(Y < X)$ in the bivariate exponential model. *Communications in Statistics Theory and Methods*, **10**, 2515–2525.

Barakat, H. M., Nigm, E. M. and Elsawah, A. M. (2013). Asymptotic distributions of the generalized and the dual generalized extremal quotient. *Bulletin of the Malaysian Mathematical Sciences Society*, **36(3)**, 657–670.

Chacko, M. and Mathew, S. (2019). Inference on $P(X < Y)$ for bivariate normal distribution based on ranked set sample. *Metron*, **77**, 239–252.

Chacko, M. and Mathew, S. (2020). Inference on $P(X < Y)$ based on records for bivariate normal distribution. *Aligarh Journal of Statistics*, **40**, 97–116.

Chacko, M. and Thomas, P. Y. (2006). Concomitants of record values arising from Morgenstern type bivariate logistic distribution and some of their applications in parameter estimation. *Metrika*, **64**, 317–333.

Chacko, M. and Thomas, P. Y. (2008). Estimation of parameters of bivariate normal using concomitants of record values. *Statistical Papers*, **49**, 263–275.

Chandler, K. N. (1952). The distribution and frequency of record values. *Journal of Royal Statistical Society Ser.*, **B 14,** 220–228.

Cramer, E. (2001). Inference for stress-strength models based on wienman multivariate exponential samples. *Communications in Statistics - Theory and Methods*, **30**, 331–346.

Chen, M. H. and Shao, Q. M. (1999). Monte Carlo estimation of bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics*, **8,** 69–92.

Enis, P. and Geisser, S. (1971). Estimation of the prabability that $P(Y < X)$. *Journal of American Statistical Association,* **66,** 162–186.

Hanagal, D. D. (1995). Testing reliability in a bivariate exponential stress-strength model. *Journal of the Indian Statistical Association,* **33,** 41–45.

Hanagal, D. D. (1997). Estimation of reliability when stress is censored at strength. *Communication in Statistics - Theory and Methods,* **26(4)**, 911–919.

Jana, P. K. (1994). Estimation of $P(Y < X)$ in the bivariate exponential case due to Marshall-Olkin. *Journal of the Indian Statistical Association,* **32**, 35–37.

Jana, P. K. and Roy, D. (1994). Estimation of reliability under stress-strength model in a bivariate exponential set-up. *Calcutta Statistical Association Bulletin*, **44**, 175–181.

Kotz, S., Balakrishnan, N. and Johnson, N. L. (2000). *Distributions in Statistics: Continuous Multivariate Distributions*, New York: Wiley.

Mukherjee, S. P. and Saran, L. K. (1985). Estimation of failure probability from a bivariate normal stress-strength distribution. *Microelectronics and Reliability,* **25**, 692–702.

Nadarajah, S. and Kotz, S. (2006). Reliability for some bivariate exponential distributions. *Mathematical Problems in Engineering*, **2006**, 1–14. Article ID 41652.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 219–237

Clustering using Skewed Data via Finite Mixtures of Multivariate Lognormal Distributions

Deepana R. and Kiruthika C.
Department of Statistics, Pondicherry University, India

Received: 10 July 2021; Revised: 21 October 2021; Accepted: 26 December 2021

Abstract

Model-based clustering techniques are based on the finite mixture models. In this paper, an attempt is made to explore effect of the skewness in heterogeneous data using finite mixture models to clustering. In particular, this paper deals with model-based clustering using finite mixtures of multivariate lognormal distributions which can deal with skewness effectively. The Expectation Maximization (EM) algorithm is used for computing maximum likelihood estimates for model parameters. To examine the performance of clustering multivariate log normal mixtures models, some simulation studies are presented for heterogeneous data with asymmetric behavior. A real dataset is also used to illustrate the use of finite mixtures of multivariate lognormal distributions to clustering.

Key words: Multivariate log normal distribution; Finite mixture model; Model based clustering; EM algorithm.

AMS Subject Classifications: 62K99, 62J05

1. Introduction

Clustering is an unsupervised learning technique. Clustering is grouping of a set of data objects into several clusters so that objects within a cluster have high level of similarity, but they are dissimilar to the objects in other clusters. Clustering is also defined in a probabilistic approach, where the notion of clusters is formalized through their probability distributions. One of the main advantages of this probabilistic approach is that it can be interpreted from a statistical point of view for the obtained clusters. In the model-based clustering methods, the observations are generated from a mixture of probability distributions, in which each component represents a different cluster. An extensive review of finite mixture models and their clustering applications are given by Everitt and Hand (1981), Titterington *et al.* (1985) and McLachlan and Peel (2000). Finite mixtures of multivariate Gaussian distribution are widely used in model-based clustering. One may refer to McLachlan and Basford (1988), McNicholas and Murphy (2008), Beak and McLachlan (2010) and among others. Melnykov and Semhar (2016) have discussed about the challenges of model-based clustering such as initialization techniques, dimension reduction and variable selection. However, clustering based on Gaussian mixture models is not capable of reasonably fittings for heavy tails, asymmetric and outliers to the heterogeneous data.

Corresponding Author: Deepana R
Email: deepana192007@gmail.com

Model-based clustering using finite mixture models with non-normal distributions have received increasing attention and showed advantages in modeling heterogeneous data with heavy tails, asymmetric and outliers. Non-normal finite mixture distribution plays an important role in clustering applications when the component densities are skewed and heavy tailed. Karlis *et al.* (2002), Lin *et al.* (2007), Pyne *et al.* (2009), Soltyk and Gupta (2011) have given application of univariate and multivariate finite mixtures of skew-normal and skew-t distributions to clustering. Schnatter *et al.* (2010) have proposed Bayesian approach for finite mixture models of univariate and multivariate skew-t and skew normal distributions. The estimation of parameters in these mixture models is carried out by EM algorithm. Lee and McLachlan (2013a) have provided finite mixture models with skew normal and skew-t distributions and it has increased importance in modeling data withequal asymmetry and heavy tails simultaneously. Also, they have classified multivariate skew distributions into four types namely, 'restricted', 'unrestricted', 'extended' and 'generalized' forms. Lee and McLachlan (2013b) have compared the clustering performance of mixture in multivariate skew normal and skew-t distributions with other non-normal mixture distributions like generalized hyperbolic distributions, multivariate inverse-Gaussian distributions and shifted asymmetric Laplace distributions. Lee and McLachlan (2014) have provided some recent developments of mixtures in multivariate skew-t distributions. Also, they have discussed about various characterizations of multivariate skew-t distribution. Further, they have used existing EM algorithms for estimating the parameters of the restricted and unrestricted forms of multivariate skew-t mixture models. Sanjeena *et al.* (2014) have considered univariate and multivariate normal inverse Gaussian distribution for model-based clustering approach in finite mixture models and parameter estimation is carried out by the EM algorithm. A shifted asymmetric Laplace distribution is considered for model-based clustering by Franczak *et al.* (2014). A multivariate generalized hyperbolic mixture model was proposed by Browne and McNicholas (2015). Adrian *et al.* (2016) proposed clustering using multivariate normal inverse Gaussian distribution for heavy tails and asymmetric data. Melnykov *et al.* (2018) have developed finite mixture modeling with components that can handle skewness in matrix-valued data.

Although many non-symmetric distributions are available, model-based clustering using finite mixtures of multivariate lognormal distribution is considered in this paper. A finite mixture of multivariate lognormal distribution is useful in modeling heterogeneous data with asymmetric behaviour. In the present study, an attempt is made to obtain clusters for skewed data based on model-based clustering using finite mixtures of multivariate lognormal distribution. A parsimonious family of finite mixtures of multivariate lognormal distribution is also developed. Algorithms for model parameter estimation and initialization technique are presented in this paper. Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) are used for model selection. The clustering performance is evaluated using Adjusted Rand Index (ARI) and Misclassification Rate (MR). The performance of multivariate lognormal mixture models in clustering for real and simulated data are studied. The proposed initialization method to determine the initial value for the component parameters using EM algorithm is presented in the next section. The methodology for initialization technique considered in this paper overcomes the issue of initial values in EM algorithm by using K-means clustering with Mahalanobis distance measures.

The rest of this paper is organized as follows. Section 2 presents the initialization techniques for model-based clustering approach. Section 3 describes the multivariate lognormal

mixture models using EM algorithm. In Section 4 real and simulated datasets are applied for multivariate lognormal mixture models to clustering and are compared to some well-known existing methods in Section 5. In Section 6, some concluding remarks are given.

2. Initialization Technique for Multivariate Lognormal Mixture Models

The EM algorithm relies on the specified starting values for component parameters. However, it is difficult to specify good starting values. Several research works have been done for initialization for component parameters in EM algorithm. Mahalanobis distance measure is used to capture the covariance structures of clusters. Mahalanobis distance measure is used to identify and correctly classify non-spherical clusters for non-homogeneous data. Mahalanobis distance measure overcomes the variable standardization by yielding scale invariant classification. The proposed algorithm is presented below.

Algorithm

Input: Data X and the number of groups G

Output: Cluster Indicator $z_1, z_2, ..., z_n$

1. Randomly select the mean vector according to G groups from the dataset X.
2. Compute Euclidean distance based on the mean vectors. Assigning each observation nearest to the group mean vector. Compute the new mean vector $\underline{c}_k; k = 1, 2, ..., G$ and the covariance matrix $S_k; k = 1, 2, ..., G$ based on the assignments.
3. While for $1, 2, ..., G$ do
4. Compute the Mahalanobis distance measure based on the new mean vector $\underline{c}_k$ and the covariance matrix S_k

$$D(\underline{x}_i, \underline{c}_k) = \sqrt{(\underline{x}_i, \underline{c}_k^{(q)(t)}) S_k^{-1(q)} (\underline{x}_i, \underline{c}_k^{(q)})}$$

5. Assignment: Assign each observation nearest to cluster center $\underline{z}_{ik} = 1$ if $D(\underline{x}_i, \underline{c}_k)$
6. Update: Recalculate the mean and covariance matrix for (k = 1, 2, ..., G) based on the assignments.

$$\underline{c}_k^{(q+1)} = \frac{\sum_{i=1}^{n} z_{ik} \underline{x}_i}{\sum_{i=1}^{n} z_{ik}}$$

$$S_k^{(q+1)} = \frac{\sum_{i=1}^{n} z_{ik} (\underline{x}_i, \underline{c}_k^{(q+1)}) (\underline{x}_i, \underline{c}_k^{(q+1)})^{(t)}}{\sum_{i=1}^{n} z_{ik}}$$

where, q is the iteration number and t represents the transpose.

7. end While

Based on the cluster indicators $z_1, z_2, ..., z_n$ the initial component parameter values $\pi_k^{(0)}, \underline{m}u_k^{(0)}, \Sigma_k^{(0)}$. The initial values are used to initiate the EM algorithm for Multivariate Lognormal (MLN) mixture models to clustering. The parameter estimation procedure is derived in the following section.

3. Parameter Estimation Procedure for Multivariate Lognormal Mixture Models

Let data X be a d-dimensional random variable which follows a multivariate lognormal distribution with mean vector $\underline{\mu}_k$ and the covariance matrix Σ_k. The G-component finite mixture model of multivariate lognormal distributions is given by

$$f(\underline{x}_i|\Theta) = \sum_{k=1}^{G} \pi_k \frac{1}{(2\pi)^{1/2}|\Sigma_k|^{1/2}|\underline{x}_i} e^{-\frac{1}{2}(ln(\underline{x}_i)-\underline{\mu}_k)^t\Sigma_k^{-1}(ln(\underline{x}_i)-\underline{\mu}_k)} \tag{1}$$

where π_k represents the mixing proportion with $\sum_{k=1}^{G}\pi_k = 1, 0 < \pi_k < 1$. The unknown parameter Θ is $\{\pi_1, \pi_2, ..., \pi_{G-1}, \underline{x}_1, \underline{x}_2, ..., \underline{x}_G, \Sigma_1, \Sigma_2, ..., \Sigma_G\}$.

Consider the random sample of size n from multivariate Lognormal mixture models defined the probability density function given in (1). EM algorithm [Dempster *et al.* 1977] is used for the parameter estimation. The complete data in EM algorithm is written as (X, Z). The observed data vector $X = (x_1, x_2, ..., x_n)^T$ is viewed as incomplete. The component label vector is defined as $Z = z_1, z_2, ..., z_n$. The likelihood of complete data of multivariate lognormal mixture model is given by

$$L(\Theta; X, Z) = \prod_{i=1}^{n}\prod_{k=1}^{G}[\pi_k f(\underline{x}_i; \underline{\mu}_k, \Sigma_k)]^{z_{ik}} \tag{2}$$

$$= \prod_{i=1}^{n}\prod_{k=1}^{G}[\pi_k \frac{1}{(2\pi)^{1/2}|\Sigma_k|^{1/2}|\underline{x}_i} e^{-\frac{1}{2}(ln(\underline{x}_i)-\underline{\mu}_k)^t\Sigma_k^{-1}(ln(\underline{x}_i)-\underline{\mu}_k)}]^{z_{ik}}$$

The log-likelihood of complete data of multivariate lognormal mixture models is given by

$$l(\Theta; X, Z) = \sum_{i=1}^{n}\sum_{k=1}^{G} z_{ik}[log\pi_k + log[\frac{1}{(2\pi)^{1/2}|\Sigma_k|^{1/2}\underline{x}_i}] + -\frac{1}{2}(ln(\underline{x}_i) - \underline{\mu}_k)^t\Sigma_k^{-1}(ln(\underline{x}_i) - \underline{\mu}_k)] \tag{3}$$

The conditional expectation of the log-likelihood of multivariate lognormal mixture models is given by

$$E_{Z|X}l(\Theta; X, Z) = \sum_{i=1}^{n}\sum_{k=1}^{G}\tau_{ik}[log\pi_k + f(\underline{x}_i; \underline{\mu}_k, \Sigma_k)]$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{G}\tau_{ik}[log\pi_k - \frac{nd}{2}log(2\pi) + log(\underline{x}_i) - \frac{1}{2}log|\Sigma_k|^{-1} - \frac{1}{2}[(ln(\underline{x}_i) - \underline{\mu}_k)^t\Sigma_k^{-1}(ln(\underline{x}_i) - \underline{\mu}_k)]]$$

E-step:

The expectation of $l(\Theta; X, Z)$ over $Z|X$ based on current parameter choice Θ^s is $Q(\Theta, \Theta^{(s)})$

$$Q(\Theta, \Theta^{(s)}) = E_{Z|X}[l(\Theta; X, Z); \Theta^{(s)}] \tag{4}$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{G}\hat{\tau}_{ik}log\pi_k - \frac{nd}{2}log(2\pi) + \sum_{i=1}^{n}\sum_{k=1}^{G}\hat{\tau}_{ik}log(\underline{x}_i) - \sum_{i=1}^{n}\sum_{k=1}^{G}\frac{\hat{\tau}_{ik}}{2}log|\Sigma_k|^{-1}$$

$$- \sum_{i=1}^{n}\sum_{k=1}^{G}\frac{\hat{\tau}_{ik}}{2}[(ln(\underline{x}_i) - \underline{\mu}_k)^t\Sigma_k^{-1}(ln(\underline{x}_i) - \underline{\mu}_k)]$$

where $\hat{\tau}_{ik}$ is the probability of observation i belonging to the group k based on the current parameter choice $\Theta^{(s)}$. It can be calculated by

$$\hat{\tau}_{ik}^{(s)} = \frac{\pi_k^{(s)} f(\underline{x}_i; \underline{\mu}_k^{(s)}, \Sigma_k^{(s)})}{\sum_{k=1}^{G} \pi_k^{(s)} f(\underline{x}_i; \underline{\mu}_k^{(s)}, \Sigma_k^{(s)})} \tag{5}$$

M-step:

Find the estimate $\hat{\Theta}$, which maximizes $Q(\Theta, \Theta^{(s)})$ for fixed $\Theta^{(s)}$ subject to the equation $\sum_{k=1}^{G} \pi_k = 1$. Using Lagrangian method, we have

$$\Psi = Q(\Theta, \Theta^{(s)}) + \gamma(1 - \sum_{k=1}^{G}) \tag{6}$$

Maximizing the function Ψ with respect to π_j and equation them zero, we get

$$\hat{\pi_j} = \frac{\sum_{i=1}^{n} \hat{\tau}_{ij}}{n}; j = 1, 2, ..., G \tag{7}$$

Maximizing the function $Q(\Theta, \Theta^{(s)})$ with respect to $\underline{\mu}_j$ and equating them zero, we get

$$\frac{\partial(\Theta, \Theta^{(s)})}{\partial \underline{\mu}_j} = 0$$

$$\hat{\underline{\mu}}_j = \frac{\sum_{i=1}^{n} \hat{\tau}_{ij} \underline{ln}(x_i)}{\sum_{i=1}^{n} \hat{\tau}_{ij}} \tag{8}$$

To maximize the function $Q(\Theta, \Theta^{(s)})$ with respect to Σ_j

$$= -\frac{1}{2}[\sum_{i=1}^{n} \hat{\tau}_{ij} log|\Sigma_j| + tr\Sigma_j^{-1} \sum_{i=1}^{n} [(ln(\underline{x}_i) - \underline{\mu}_k)(ln(\underline{x}_i) - \underline{\mu}_k)^t]]$$

So, maximizing the function $Q(\Theta, \Theta^{(s)})$ with respect to Σ_j is equivalent to maximizing the above expression with respect to Σ_j. Here, $\hat{\Sigma}_j$ is obtained by using the Lemma 3.2.2 of Anderson (1984) and we get

$$\hat{\underline{\Sigma}}_j = \frac{\sum_{i=1}^{n} \hat{\tau}_{ij} [(ln(\underline{x}_i) - \underline{\mu}_k)(ln(\underline{x}_i) - \underline{\mu}_k)^t]}{\hat{\tau}_{ij}} \tag{9}$$

Another important objective of model-based clustering is to study the covariance structures. Fraley *et al.* (1998) have considered different covariance structures for Gaussian mixture models to clustering techniques. Different covariance structures for multivariate lognormal mixture models are developed in the following section.

4. Estimation via Geometric Decomposition

To provide easy and simple interpretable models, Banfield *et al.* (1993) have parameterized the covariance matrices in terms of the eigen-value decompositions for Gaussian mixture models. Fraley *et al.* (1998) considered an eigen-value decomposition of the cluster covariance matrices to provide a wide range of parsimonious covariance structures. Fraley *et al.* (2002) have provided an in-depth discussion of the eigen-value decomposition approach for finite mixture models to clustering. This work is implemented in the MCLUST package. MCLUST package consists of 14 mixture models that arise from the imposition of constraints upon the group of covariance matrix. MCLUST is the most well-established package for model-based clustering technique using Gaussian mixture models. Details of the constraints that can be imposed are summarized in Fraley *et al.* (2003, 2006) which is available in the R software. Fraley *et al.* (2012) summarized the covariance structures available in the MCLUST package, corresponding to geometric characteristics such as shape, volume and orientation. If the number of components is not specified, it assumes that the number of components lies between one to nine. Following this, EM algorithm is implemented corresponding to each initial classification and estimates for parameters are obtained. Then BIC is computed for each resulting mixture model. The model having highest BIC value is identified as the best model.

Browne *et al.* (2014) have pointed out that the covariance technique of Celeux *et al.* (1995) for the EVE and VVE models are computationally infeasible in higher dimensions. They have proposed an alternative algorithm for these two models, based on an accelerated line search on the orthogonal model. Browne *et al.* (2015) have developed another approach, using fast maximization-minimization algorithms, for the EVE and VVE models. This approach is implemented in the mixture packages for R. Several other approaches have been presented, and the excellent review of covariance structures is given by Bouveyron *et al.* (2007).

From the above existing procedures, it is observed that different covariance structures are important for multivariate non-normal mixture models. This paper considers the different covariance structures based on eigen-value decomposition techniques. Let us recall the conditional expectation of the log-likelihood for multivariate lognormal finite mixture models as given in the equation (4).

4.1. The Parsimonious MLN family of models

An eigen-value decomposition of the component covariance matrices is given by

$$\Sigma_k = \lambda_k D_k A_k D_k^t \tag{10}$$

where λ_k is a constant of proportionality, D_k is a orthogonal matrix of eigen vectors and A_k is a orthogonal matrix of eigen vectors and $\det A_k = 1$. Celeux *et al.* (1995) developed eight eigen-value decomposition of a component covariance matrix. The volume of the cluster is determined by λ_k. D_k determines the orientation of the clusters and A_k determines the shape of the density contours. d is the number of dimensions in the datasets. The parsimonious MLN mixture models, herein referred to as PMLN, whose density is given by

$$f(\underline{x}_i|\Theta) \ = \sum_{k=1}^{G} \pi_k f(\underline{x}_i; \underline{\mu}_k, \lambda_k D_k A_k D_k^t) \tag{11}$$

To fit the parsimonious MLN mixture models, EM algorithm is used. The details of parameter estimation methods are like those described in Section 3. To compute $\hat{\Sigma_k}$ To fit the parsimonious MLN mixture models, EM algorithm is used. The details of parameter estimation methods are like those described in Section 3. For the most general MLN family member (VVV model), the complete-data likelihood is given by

$$L(\Theta; X, Z) = \prod_{i=1}^{n}[\prod_{k=1}^{G}[\pi_k]^{z_{ik}}[\prod_{k=1}^{G} f(\underline{x}_i; \underline{\mu}_k, \lambda_k D_k A_k D_k^t)]^{z_{ik}}]] \tag{12}$$

where $f(\underline{x}_i; \underline{\mu}_k, \lambda_k D_k A_k D_k^t)$ is the density of multivariate lognormal distribution with mean vector $\underline{\mu}_k$ and covariance matrix $\Sigma_k = \lambda_k D_k A_k D_k^T$.The conditional expectation of the complete data log-likelihood Q is given by

$$Q(\Theta, \Theta^{(s)}) = E_{Z|X}[l(\Theta; X, Z); \Theta^{(s)}] \tag{13}$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{G} \hat{\tau}_{ik}[log\pi_k - \frac{nd}{2}log(2\pi) + log(\underline{x}_i) - \frac{1}{2}log|\Sigma_k|^{-1} - \frac{1}{2}[(ln(\underline{x}_i) - \underline{\mu}_k)^t \Sigma_k^{-1}(ln(\underline{x}_i) - \underline{\mu}_k)]]$$

The E-step of sth iteration consists of the component membership labels with their conditional expected values is given by

$$\hat{\tau}_{ik}^{(s)} = \frac{\pi_k^{(s)} f(\underline{x}_i; \underline{\mu}_k^{(s)}, \lambda_k D_k A_k D_k^{T(s)})}{\sum_{k=1}^{G} \pi_k^{(s)} f(\underline{x}_i; \underline{\mu}_k^{(s)}, \lambda_k D_k A_k D_k^{T(s)})}$$

To perform the decomposition for MLN mixture models, we follow the procedures outlined in Celeux *et al.* (1995).

Sperical Family

In spherical family, the shape of the clusters is spherical. The shape of the covariance matrix is always diag(1,1). Two spherical families are considered here.

(1) Fitting of EII model ($\Sigma = \lambda I$)

First consider the simplest structure where every component has spherical shape and equal volume. Substitute the $\Sigma_k = \Sigma = \lambda I$ in equation (13). The complete data log-likelihood for the EII model is given by

$$l(\lambda I) = \sum_{i=1}^{n}\sum_{k=1}^{G} \hat{\tau}_{ik}[log\pi_k - \frac{nd}{2}log(2\pi) + log(\underline{x}_i) - \frac{1}{2}log|\lambda I|^{-1} - \frac{1}{2}[(ln(\underline{x}_i) - \underline{\mu}_k)^t \lambda I^{-1}(ln(\underline{x}_i) - \underline{\mu}_k)]] \tag{14}$$

$$= K - \sum_{i=1}^{n}\sum_{k=1}^{G} \frac{\hat{\tau}_{ik}}{2} log\det \lambda I - \sum_{i=1}^{n}\sum_{k=1}^{G} \frac{\hat{\tau}_{ik}}{2}[(ln(\underline{x}_i) - \underline{\mu}_k)^T(\lambda I)^{-1}(ln(\underline{x}_i) - \underline{\mu}_k)]$$

$$= \lambda^{-1}\sum_{k=1}^{G} tr(W_k) + dlog\sum_{k=1}^{G}\sum_{i=1}^{n} \hat{\tau}_{ik}$$

$$= \lambda^{-1} \sum_{k=1}^{G} tr(W) + dlog\lambda$$

where K is the constant with respect to model parameters $\underline{\mu}_k$ and λ. Maximizing the equation (13) with respect to λ, we get

$$\hat{\Sigma} = \hat{\lambda} = \frac{tr(W)}{nd} = \frac{\sum_{k=1}^{G} \sum_{i=1}^{n} \hat{\tau}_{ik}[(ln(\underline{x}_i) - \underline{\mu}_k)(ln(\underline{x}_i) - \underline{\mu}_k)^t]}{nd}$$

where $n = \sum_{k=1}^{G} \sum_{i=1}^{n} \hat{\tau}_{ik}$

(2) Fitting of VII model ($\Sigma_k = \lambda_k I$)

This is the second simplest model where the component has spherical shape and different volume. Substitute in the equation (14) $\Sigma_k = \lambda_k I$ in equation (13). The complete data log-likelihood for the EII model is given by

$$l(\lambda_k I) = \sum_{i=1}^{n} \sum_{k=1}^{G} \hat{\tau}_{ik}[log\pi_k - \frac{nd}{2} log(2\pi) + log(\underline{x}_i) - \frac{1}{2} log|\lambda_k I|^{-1} - \frac{1}{2}[(ln(\underline{x}_i) - \underline{\mu}_k)^t \lambda_k I^{-1} (ln(\underline{x}_i) - \underline{\mu}_k)]] \tag{15}$$

$$= K - \sum_{i=1}^{n} \sum_{k=1}^{G} \frac{\hat{\tau}_{ik}}{2} log \det \lambda_k I - \sum_{i=1}^{n} \sum_{k=1}^{G} \frac{\hat{\tau}_{ik}}{2} [(ln(\underline{x}_i) - \underline{\mu}_k)^t (\lambda_k I)^{-1} (ln(\underline{x}_i) - \underline{\mu}_k)]$$

$$= \lambda^{-1} \sum_{k=1}^{G} tr(W_k) + dlog \sum_{k=1}^{G} \sum_{i=1}^{n} \hat{\tau}_{ik}$$

$$= \lambda^{-1} \sum_{k=1}^{G} tr(W) + d \sum_{k=1}^{G} log\lambda_k \sum_{i=1}^{n} \hat{\tau}_{ik}$$

where K is the constant with respect to model parameters $\underline{\mu}_k$ and λ_k. Maximizing the equation (15) with respect to λ_k, we get

$$\hat{\Sigma_k} = \hat{\lambda}_k = \frac{\sum_{k=1}^{G} \sum_{i=1}^{n} \hat{\tau}_{ik}[(ln(\underline{x}_i) - \underline{\mu}_k)(ln(\underline{x}_i) - \underline{\mu}_k)^t]}{\tau_k d}; k = 1, 2, ..., G$$

where $\tau_k = \sum_{i=1}^{n} \hat{\tau}_{ik}$

General Family

(3) Fitting an EVV model ($\Sigma_k = \lambda D_k A_k D_k^T$)

This is generalized model and the component has the same volume but different shape and orientation. Substitute in the equation (13) $\Sigma_k = \lambda D_k A_k D_k^T$ and $C_k = D_k A_k D_k^T; \Sigma_k = \lambda C_k$. The complete data log-likelihood for the EVV model is given by

$$l(\lambda C_k) = \sum_{i=1}^{n} \sum_{k=1}^{G} \hat{\tau}_{ik}[log\pi_k - \frac{nd}{2} log(2\pi) + log(\underline{x}_i) - \frac{1}{2} log|\lambda C_k|^{-1} - \frac{1}{2}[(ln(\underline{x}_i) - \underline{\mu}_k)^t \lambda C_k^{-1} (ln(\underline{x}_i) - \underline{\mu}_k)]] \tag{16}$$

$$= K - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{G}\hat{\tau}_{ik}2log|\lambda C_k| + \sum_{k=1}^{G} tr(W_k)(\lambda C_k)^{-1}$$

where K is the constant with respect to model parameters C_k and λ. The equation (16) is maximizing with respect to C_k and λ and equating them zero. We get,

$$\hat{C}_k = \frac{\sum_{i=1}^{n}\hat{\tau}_{ik}(ln(\underline{x}_i) - \underline{\mu}_k)(ln(\underline{x}_i) - \underline{\mu}_k)^T}{|\sum_{i=1}^{n}\hat{\tau}_{ik}(ln(\underline{x}_i) - \underline{\mu}_k)(ln(\underline{x}_i) - \underline{\mu}_k)^t|^{\frac{1}{d}}}$$

and

$$\hat{\lambda}_k = \frac{|\sum_{i=1}^{n}\hat{\tau}_{ik}(ln(\underline{x}_i) - \underline{\mu}_k)(ln(\underline{x}_i) - \underline{\mu}_k)^t|^{\frac{1}{d}}}{n}$$

$$\hat{\Sigma_k} = \hat{\lambda}\hat{C_k}$$

(4) Fitting an EEE model ($\Sigma_k = \Sigma = \lambda DAD^t$)

This model is a common model for all components and it considers same size, volume and orientation. Substitute in the equation (13) $\Sigma_k = \Sigma = \lambda DAD^t$. The complete data log-likelihood for the EEE model is given by

$$\begin{aligned} l(\lambda DAD^t) = \sum_{i=1}^{n}\sum_{k=1}^{G}\hat{\tau}_{ik}[log\pi_k - \frac{nd}{2}log(2\pi) + log(\underline{x}_i) - \frac{1}{2}log|(\lambda DAD^t)|^{-1} \\ -\frac{1}{2}[(ln(\underline{x}_i) - \underline{\mu}_k)^t(\lambda DAD^T)^{-1}(ln(\underline{x}_i) - \underline{\mu}_k)]] \end{aligned} \tag{17}$$

$$= K - \sum_{i=1}^{n}\sum_{k=1}^{G}\frac{\hat{\tau}_{ik}}{2}log|\lambda DAD^T| - \sum_{i=1}^{n}\sum_{k=1}^{G}\frac{\hat{\tau}_{ik}}{2}[(ln(\underline{x}_i) - \underline{\mu}_k)^t(\lambda DAD^t)^{-1}(ln(\underline{x}_i) - \underline{\mu}_k)]$$

$$= K - \frac{1}{2}[tr(W\Sigma^{-1}) + nlog|\Sigma|]$$

where k is the constant with respect to the model parameters $\mu_k, \lambda, DandA$.

$$W = \sum_{k=1}^{G} W_k = \sum_{i=1}^{n}\hat{\tau}_{ik}[(ln(\underline{x}_i) - \underline{\mu}_k)(ln(\underline{x}_i) - \underline{\mu}_k)^t)]$$

and

$$n = \sum_{i=1}^{n}\sum_{k=1}^{G}\hat{\tau}_{ik}$$

EEE model is unconstrained model and it's considered common covariance matrix.

$$\hat{\Sigma_k} = \frac{W}{n} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{G}\hat{\tau}_{ik}(ln(\underline{x}_i) - \underline{\mu}_k)(ln(\underline{x}_i) - \underline{\mu}_k)^t}{n}$$

(5) Fitting an VVV model ($\Sigma_k = \lambda_k D_k A_k D_k^t$)

This is the most generalized model. This is the model where every component has different shape, different volume and different orientation. VVV model is the unconstrained model. Substitute in the equation (13) $\Sigma_k = \lambda_k D_k A_k D_k^t$. The complete data log-likelihood for the VVV model is given by

$$\begin{aligned} l(\lambda_k D_k A_k D_k^t) = \sum_{i=1}^{n}\sum_{k=1}^{G} \hat{\tau}_{ik}[log\pi_k - \frac{nd}{2}log(2\pi) + log(\underline{x}_i) - \frac{1}{2}log|\lambda_k D_k A_k D_k^t|^{-1} \\ -\frac{1}{2}[(ln(\underline{x}_i) - \underline{\mu}_k)^t(\lambda_k D_k A_k D_k^t)^{-1}(ln(\underline{x}_i) - \underline{\mu}_k)]] \end{aligned} \tag{18}$$

$$= K - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{G}\hat{\tau}_{ik}2log|\lambda_k D_k A_k D_k^t| + \sum_{k=1}^{G} tr(W_k)(\lambda_k D_k A_k D_k^t)^{-1}$$

where K is the constant with respect to model parameters μ_k, D_k, A_k and λ_k.

$$\Sigma_k = \frac{\sum_{k=1}^{G} W_k}{n} = \frac{\sum_{i=1}^{n}\hat{\tau}_{ik}[(ln(\underline{x}_i) - \underline{\mu}_k)(ln(\underline{x}_i) - \underline{\mu}_k)^t)]}{\tau_k}; k = 1, 2, ..., G$$

where

$$\tau_k = \sum_{i=1}^{n}\hat{\tau}_{ik}$$

The summary of eigen-value decomposition covariance structures is given in the Table 1.

Table 1: Nomenclature, scale matrix structure and the number of free scale parameters for the eigen-decomposed family of models

Model	λ_k	A_k	D_k	Σ_k	$NumberofCovarianceParameters$
EII	Equal	Spherical	-	λI	1
VII	Variable	Spherical	-	$\lambda_k I$	G
EVV	Equal	Variable	Variable	$\lambda D_k A_k D_k^T$	$\frac{Gd(d+1)}{2} - (G-1)d$
EEE	Equal	Equal	Equal	λDAD^T	$\frac{d(d+1)}{2}$
VVV	Variable	Variable	Variable	$\lambda_k D_k A_k D_k^T$	$\frac{Gd(d+1)}{2}$

Covariance Estimation

An alternative estimation method for covariance matrix is presented in this paper. The decomposed elements of the covariance matrix are updated according to the following algorithm. τ_{ik} represents the probability that observation i belongs to group k given the current component parameters

$$n_k = \tau_{ik} = \frac{\pi_k f(\underline{x}_i; \underline{\mu}_k, \Sigma_k)}{\sum_{j=1}^{G}\pi_j f(\underline{x}_j; \underline{\mu}_j, \Sigma_j)}; j \neq k$$

M-step involves the conditionally maximizing the parameters with respect to complete log-likelihood. The estimated mixing proportion and sample cross-product matrix for the kth component is given by

$$\hat{\pi_k} = \frac{n_k}{n}; k = 1, 2, ..., G$$

$$W_k = \sum_{i=1}^{n} n_k(\underline{x}_i - \underline{\mu}_k)(\underline{x}_i - \underline{\mu}_k)^t; k = 1, 2, ..., G$$

1. Iteration $q = 1$
2. Update
$$\lambda_k = \frac{\sum_{k=1}^{G} tr(n_k.W_k)}{nd}$$
where n is the number of observations and d is the dimension.
3. Update
$$A_k = \frac{diag(n_k.W_k)}{|n_k.W_k|^{\frac{1}{d}}}$$
4. Update
$$D_k = n_k W_k a_k$$
Where a_k is the largest eigen value of W_k
5. Update A_k, D_k, λ_k in Σ_k
6. Calculate $E_q = \frac{1}{\lambda} tr(n_k \lambda_k D_k A_k D_k^T + n * dlog(\lambda))$
7. If $t > 1, E_q - E_q - 1 > \epsilon$. If true $t = t + 1$ and return step 2 , or else end.

Five types of covariance structures are considered for finite mixtures of multivariate lognormal distributions to clustering. All covariance models based on eigen-value decomposition structures are used in the M-step of the EM algorithm. The description of the EM algorithm for MLN mixture models is given below.

EM Algorithm

1. Initialization: The initial values of $\pi_k^{(0)}, \underline{m}u_k^{(0)}, \Sigma_k^{(0)}$ are obtained using the algorithm in Section 2.

2. E-step: The conditional Expectation $(\hat{\tau_{ik}}^{(q)})$ of the group membership for each observation is obtained using the equation (5).

3. Mstep: Update the parameters $\hat{\pi_j}^{(q)}$ and $\hat{\mu_j}^{(q)}$ using the formula (7) and (8). Five parsimonious covariance models for MLN mixtures which are derived in the Section 4.1 are updated in the M-step.

4. Compute the log-likelihood $l_j^{(q)}$ and $l_j^{(q+1)}$ and Compare $l_j^{(q+1)}$ and $l_j^{(q)}$. —— $l_j^{(q+1)} - l_j^{(q)}|| < \epsilon$. STOP.

5. E-step and M-step are repeated till the same log-likelihood values are met.

After the convergence is reached, the $(\hat{\tau_{ik}}^{(q)})$ is the posterior probability of component membership for each observation and it is used to cluster the observation into groups. Predicated membership is obtained through Maximum A Posterior probability (MAP).

5. Experimental Results

In this section, the clustering performance of PLMN mixture models is assessed in terms of BIC, AIC, ARI and misclassification rate through simulated as well as real datasets. Numerical comparison of PMLN mixture models have been made with Multivariate Skew Normal (MSN) and Multivariate Normal (MN) mixture models. All numerical computations have been implemented through a program developed in R.

5.1. Simulation Experiment

Here, we consider a finite mixture of multivariate Lognormal distribution with three components. Random sample of size n = 262, 270 and 268 are simulated with parameters $\mu_1 = (0.29, 0.685)$, $\mu_2 = (1.68, 0.69)$, $\mu_3 = (0.88, 1.71)$ with same covariance matrix $\Sigma = \begin{bmatrix} 0.1986 & 0.8876 \\ 0.8876 & 0.8876 \end{bmatrix}$. The mean vectors and covariance matrix are generated from the clusterGeneration package which is available in R. Figure 1 displays the scatter plot of the simulated dataset.

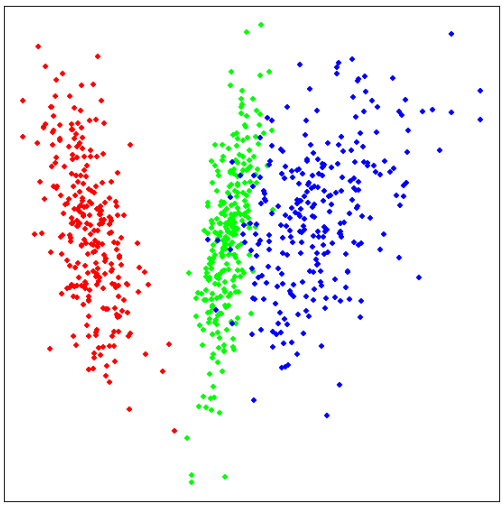


Figure 1: Scatter plot for simulated data

The initial component parameter values are obtained using the algorithm in Section 2. All the covariance models are initiated with the same initial values of the component parameters. The initial values are obtained iteratively till the same cluster membership labels are met. From the cluster membership labels, the initial mixing proportion, initial mean vector and initial covariance matrix are calculated. The initial values are presented in Table 2.

Table 2: Initial parameter values of three-component MLN mixture models

Component 1	$\pi_1 = 0.5211$	$\mu_1 = (0.29, 0.685)$	$\Sigma_1 = \begin{bmatrix} 0.1986 & 0.8876 \\ 0.8876 & 0.8876 \end{bmatrix}$
Component 2	$\pi_2 = 0.238$	$\mu_2 = (2.27, 0.57)$	$\Sigma_2 = \begin{bmatrix} 0.5392 & 1.0275 \\ 1.0275 & 6.2978 \end{bmatrix}$
Component 3	$\pi_3 = 0.2409$	$\mu_3 = (1.24, 2.65)$	$\Sigma_3 = \begin{bmatrix} 0.9786 & 2.9376 \\ 2.9376 & 9.2136 \end{bmatrix}$

Table 3: Clustering performance of various multivariate mixture models

Distributions	Model	BIC	AIC	MR	ARI	Log likelihood
MLN	EII	3380.15	3279.15	0.10	0.7169	-1523.851
MLN	VII	3256.14	3126.86	0.11	0.7328	-1503.132
MLN	EEE	3178.23	2814.25	0.09	0.8354	-1523.57
MLN	EVV	3445.76	3437.52	0.04	0.8369	-1529.57
MLN	VVV	3045.28	3012.19	0.07	0.7425	-1496.09
MSN	EEV	3389.461	3145.58	0.155	0.8269	-1467.04
MN	EEE	3193.09	3436.29	0.133	0.7932	-1498.96

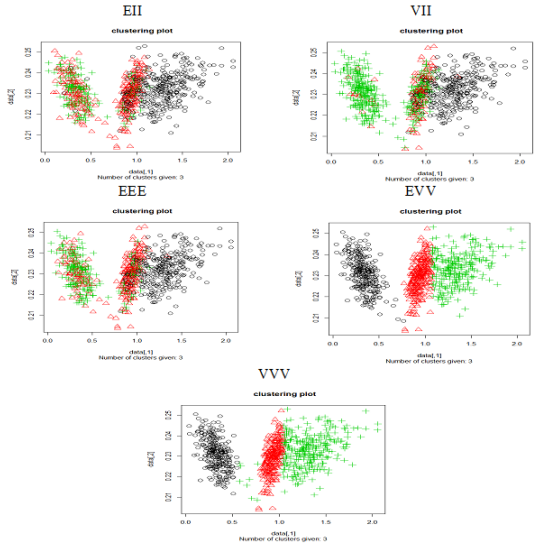


Figure 2: Scatter plot for five MLN mixture models

Different covariance structures in multivariate lognormal mixture models are considered. The clustering results of the simulated dataset are provided in Table 3. From Table 3, it is observed that EVV model gives lowest misclassification rate (0.04). The ARI is 83 % with BIC 3445.76 and AIC 3437.52. Among five covariance structures of MLN mixture models, EVV model achieved the highest ARI. The best model (EVV) is compared with other multivariate mixture models. The ARI value for MLN mixture model ranges from 0.71 to 0.83 which indicates that the dataset is classified with greater precision. EEV model gives better clustering performance for multivariate skew normal mixture models and EEE

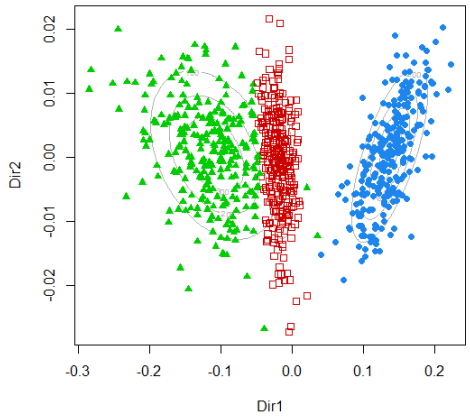


Figure 3: Contour plot for the EVV model

model provides better clustering results for multivariate normal mixture models. The results of both model are shown in Table 3. Table 4 provides the estimated parameter values of EVV model in case of MLN mixture models.

Table 4: Estimated parameter values of three-component MLN mixture (EVV) model

component 1	$\pi_1 = 0.5901$	$\mu_1 = (1.07, 0.974)^t$	$\Sigma_1 =$	$\begin{bmatrix} 0.09379 & 0.9396 \\ 0.9396 & 4.2789 \end{bmatrix}$
Component 2	$\pi_2 = 0.111$	$\mu_2 = (2.005, 0.772)^t$	$\Sigma_2 =$	$\begin{bmatrix} 0.3327 & 2.0235 \\ 2.0235 & 5.1936 \end{bmatrix}$
Component 3	$\pi_3 = 0.3989$	$\mu_3 = (1.984, 2.728)^t$	$\Sigma_3 =$	$\begin{bmatrix} 0.9726 & 2.9506 \\ 2.9506 & 8.9349 \end{bmatrix}$

From the table, correctly classified samples are presented here. That is, Almost 81% of samples are correctly classified for all models of MLN mixture models. The best model for MLN mixture gives 95% correct classification of the simulated dataset. For multivariate skew normal mixture, EEV model achieved 85 correct classification. Multivariate normal mixture models EEE model gives 87% correct classification. Figure 2 depicts the estimation of the cluster memberships into three clusters for the five models. In these figures, the clusters are indicated by three different characters (+, o and D). The volume of the five models is:

i) $\lambda I : \lambda = 0.2996$

ii) $\lambda_k I : \lambda_1 = 0.983, \lambda_2 = 2.371, \lambda_3 = 0.693$

iii) $\lambda DAD^t : \lambda = 3.2996$

iv) $\lambda D_k A_k D_k^t : \lambda = 5.2996$

v) $\lambda_k D_k A_k D_k^t : \lambda_1 = 1.283, \lambda_2 = 3.591 and \lambda_3 = 7.753$

The contour plot of the best (EVV) model is shown in Figure 3. The contour plot shows the different volume, size and orientation of the three clusters. The best fitted model is selected based on BIC and AIC value. It is also noticed that from the simulated dataset, general models perform better than spherical models.

5.2. Real Data (Old Faithful Dataset)

In this section, old faithful dataset is used for the PMLN mixture models. This dataset contains two variables (eruptions and waiting) and 275 observations. It is a bivariate dataset measuring the length of eruption and time to eruption, both variables are in millimeters. This dataset is available in R software. Many researchers have analyzed this dataset for model-based clustering approach. This dataset does not have true class labels. The original plot of the faithful dataset is shown in Figure4, where the observations are displayed into two clusters very clearly.

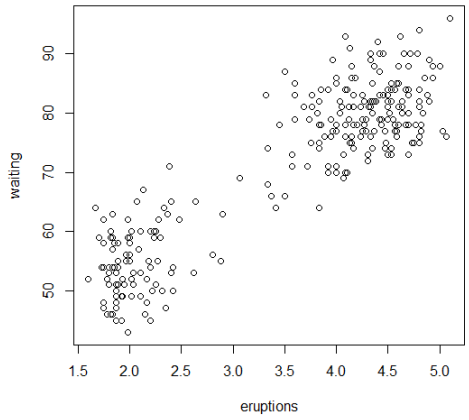


Figure 4: The bivariate Old faithful dataset

Table 5: Initial parameter values of two-component faithful dataset

Component 1	$\pi_1 = 0.5389$	$\mu_1 = (3.457, 70.794)^t$	$\Sigma_1 = \begin{bmatrix} 1.3899 & 14.3525 \\ 14.3525 & 182.461 \end{bmatrix}$
Component 2	$\pi_2 = 0.4611$	$\mu_2 = (3.518, 71)^t$	$\Sigma_2 = \begin{bmatrix} 1.2232 & 13.7003 \\ 13.7003 & 188.5333 \end{bmatrix}$

We compare the clustering performance of MLN, MSN and MN mixture models. The initial values of component parameters are calculated based on the algorithm as given in Section 2. Initial values of faithful datasets are presented in the Table 5.

For MSN and MN mixture models the best results are given in Table 6. The classification plot of each model for MLN mixture models are displayed in Figure 5. The clusters are

Table 6: Clustering performance of various multivariate mixture models

Distributions	Model	BIC	AIC	Log likelihood
MLN	EII	1876.912	1844.562	-796.53
MLN	VII	1887.072	1854.825	-769.94
MLN	EEE	1889.649	1883.544	-868.27
MLN	EVV	1825.195	1852.052	-893.82
MLN	VVV	1895.839	1869.302	-788.28
MSN	EVV	1892.361	1825.427	-834.25
MN	VVV	2371.702	2148.597	-919.29

Table 7: Estimated parameter values of two-component faithful dataset

Component 1	$\pi_1 = 0.653$	$\mu_1 = (3.093, 71.814)^t$	$\Sigma_1 =$	$\begin{bmatrix} 1.2903 & 14.1739 \\ 14.1739 & 181.281 \end{bmatrix}$
Component 2	$\pi_2 = 0.347$	$\mu_2 = (2.948, 70.542)^t$	$\Sigma_2 =$	$\begin{bmatrix} 1.2232 & 13.8103 \\ 13.8103 & 187.4933 \end{bmatrix}$

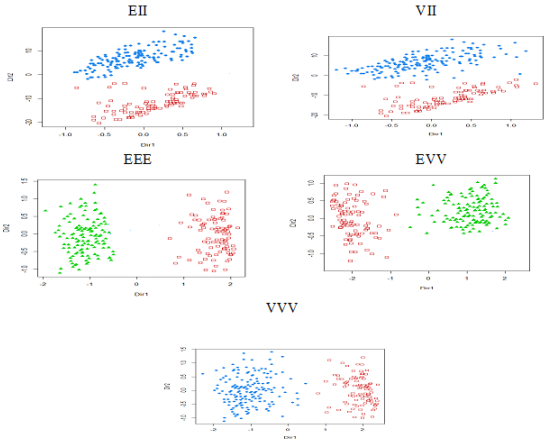


Figure 5: Scatter plot for five models using multivariate lognormal mixture models

represented by different symbols. VVV model gives good clustering results for multivariate normal mixture models. The parsimonious family of multivariate lognormal distributions shows that the clusters have different volume and size. The contour plot in Figure 6 shows different volume and size of clusters. Estimated parameters of VVV models for MLN mixtures are given in the Table 7.

The number of observations in each cluster for MLN, MSN and MN mixture models are presented in Table 8. The volume of the clusters is given below:

i) $\lambda I : \lambda = 197.17$

ii) $\lambda_k I : \lambda_1 = 180, \lambda_2 = 69$

iii) $\lambda DAD^t : \lambda = 109.26$

Table 8: Clustering table for multivariate mixture models

	MLN Mixture Model					MSN	MN
Clusters	EII	VII	EEE	EVV	VVV	EVV	VVV
Cluster 1	177	99	174	170	178	175	168
Cluster 2	95	173	98	102	94	97	104

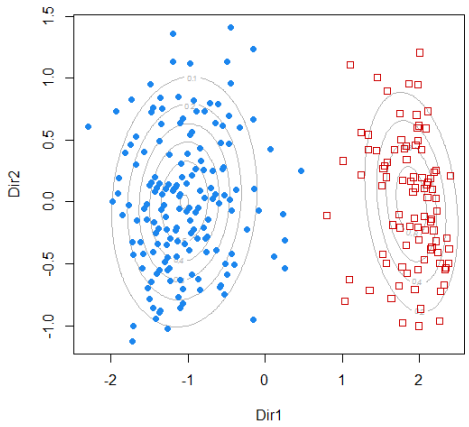


Figure 6: Contour plot for VVV model for multivariate lognormal mixture model

iv) $\lambda D_k A_k D_k^t : \lambda = 166.1296$

v) $\lambda_k D_k A_k D_k^t : \lambda_1 = 170, \lambda_2 = 110$

6. Conclusion

In this paper, a family of parsimonious MLN mixture models is introduced through an eigen-value decomposition of the components covariance matrix. From simulation experiments, the general (EVV) covariance model provides best clustering results than spherical models. The results of real dataset showed that all covariance model gives better clustering results according to BIC and AIC criteria. Proposed initialization techniques plays important role, because it gives reliable and true estimated parameter values for components. It is noticed that among general covariance models from numerical experiments, VVV gives good clustering results. VVV model allows with different size, volume, and orientation. Some parsimonious models give good clustering results, because those covariance models are close to the structure of the data.

References

Adrian, O. H., Thomas, B. M., Bredan, M., Isobel, C. G., Paul, D. M. and Dimitris, K. (2016). Clustering with the Multivariate Normal Inverse Gaussian Distribution. *Computational Statistics and Data Analysis*, **93**, 18–30.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian Clustering. *Biometrics*, **49(3)**, 803–821.

Baek, J. and McLachlan, G. J. (2010). Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualization of high dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32(7)**, 1298–1309.

Bouveyron, C., Girard, S. and Schmid, C. (2007). High-dimensional data clustering. *Computational Statistics and Data Analysis*, **52(1)**, 502–519.

Browne, R. P. and McNicholas, P. D. (2014). Orthogonal Stiefel manifold optimization for eigen-decomposed covariance parameter estimation in mixture models. *Statistics and Computing*, **24(2)**, 203–210.

Browne, R. P. and McNicholas, P. D. (2015). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics*, **43(2)**, 176–198.

Bouveyron, C. and Brunet, C. (2012). Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Statistics and Computing*, **22(1)**, 301–324.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering model. *Pattern Recognition*, **28(5)**, 781–793.

Dempster, A. P., Laird, N. M and Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (Series B)*, **39**, 1–38.

Everitt, B. S. and Hand, D. J. (1981). *Finite Mixture Distributions*. Chapman and Hall, London.

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering methods? Answers via model-based cluster analysis. *The Computer Journal*, **41(8)**, 578–588.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97(458)**, 611–631.

Fraley, C. and Raftery, A. E. (2003). Enhanced software for model-based clustering, density estimation, discriminant analysis: MCLUST. *Journal of Classification*, **20(2)**, 263–286.

Fraley, C. and Raftery, A. E. (2006). MCLUST Version 3 for R: Normal Mixture Modeling and Model-Clustering. *Technical Report 504, Department of Statistics, University of Washington*, First Published September 2006, Minor revisions January 2007 and November.

Fraley, C., Raftery, A. E., Murphy, T. B. and Scrucca, L. (2012). MCLUST Version 4 for R: Normal Mixture Modeling and Model-Clustering. *Technical Report 597*, Department of Statistics, University of Washington.

Schnatter, F. S. and Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions, *Biostatistics*, **11**, 317—336.

Karlis, D. (2002). An EM type algorithm for maximum likelihood estimation of the normal-inverse Gaussian distribution. *Statistics and Probability and Letters*, **57(1)**, 43-–52.

Lin, T. I., Lee, J. C. and Yen, S. Y. (2007). Finite mixture modeling using skew normal distribution. *Statistica Sinica*, **17(3)**, 909–927.

Lee, X. L. and McLachlan, G. J. (2013a). On mixtures of skew normal and skew t-distributions. *Advances in Data Analysis and Classification*, **7(3)**, 241–266.

Lee, X. L. and McLachlan, G. J. (2013b). Model-based clustering and classification with non-normal mixture distributions. *Statistical Methods and Applications*, **22(4)**, 427–454.

Lee, X. L. and McLachlan, G. J. (2014). Finite mixtures of multivariate skew t- distributions: some recent and new results. *Statistics and Computing*, **24(2)**, 181–202

Melnykov, V. and Xuwen, Z. (2018). On Model-based clustering of skewed matrix data. *Journal of Multivariate Analysis*, **167(c)**, 181–194.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications*. Marcel Dekker, New York.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. John Wiley and Sons, Inc, New York.

McNicholas, P. D. and Murphy, T. B. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing*, **18(3)**, 285–296.

Melnykov, V. and Semhar, M. (2016). Studying complexity of model-based clustering. *Communications in Statistics-Simulation and Computation*, **45(6)**, 2051–2069.

Pyne S., Hu X., Wang K., Rossin E., Lin, T. I., Maier et.al. (2009). *Automated High-Dimensional Flow Cytometric Data Analysis*. Proceedings of the National Academy of Sciences, USA.

Sanjeena, S. and McNicholas, P. D. (2014). Variational bayes approximations for clustering via mixtures of normal inverse gaussian distributions. *Advances in Data Analysis and Classification*, **8**, 167-193.

Soltyk, S. and Gupta, R. (2011). Application of the Multivariate Skew Normal Mixture Model with the EM Algorithm to Value-at-Risk. *MODSIM 2011 - 19th International Congress on Modelling and Simulation*, Perth, Australia, December 12-16, 2011.

Titterington, D., Smith, A. and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons, New York.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series) pp 239-249

Simultaneous Testing Procedure for the Ordered Pair-Wise Comparisons of Location Parameters of Exponential Populations under Heteroscedasticity

Jatesh Kumar[1], Amar Nath Gill[2] and Anju Goyal[1]
[1]*Department of Statistics, Panjab University, Chandigarh-160014 (India)*
[2]*School of Basic Sciences, Indian Institute of Information Technology, Una-177220 (India)*

Received: 23 June 2021; Revised: 24 October 2021; Accepted: 26 December 2021

Abstract

Zheng (2013) provided the penalized maximum likelihood estimators (PMLEs) of the location and scale parameters of two-parametric exponential distribution and proved that these estimators are uniformly minimum variance unbiased estimators (UMVUE). In this paper, a test procedure has been proposed, on the basis of the PMLEs of the location and scale parameters of the two-parametric exponential distribution. The purpose of the proposed procedure is to construct the simultaneous confidence intervals (SCIs) for the ordered pair-wise comparisons of location parameters of multi-sample two-parameter exponential distributions under the heteroscedasticity of scale parameters. A Monte Carlo simulation study has revealed that the proposed procedure is better than the existing procedure of Singh and Singh (2013) in terms of coverage probability, average volume, and power. Implementation of the proposed procedure is illustrated through real-life numerical data.

Key words: Simultaneous confidence interval (SCIs); Penalized maximum likelihood estimators (PMLEs); Heteroscedasticity; Simulated power comparison.

1. Introduction

Suppose the k (≥ 3) independent populations are such that the statistical model for the observations from the ith population is a two-parameter exponential distribution, denoted by $E_i(\mu_i, \theta_i)$, with probability density function (pdf)

$$f(x|\gamma_i, \delta_i) = \begin{cases} \frac{1}{\theta_i} e^{-(\frac{x-\mu_i}{\theta_i})}, & x \geq \mu_i, \theta_i > 0 \\ 0, & Otherwise, \end{cases}$$

where μ_i and $\theta_i$are the location and the scale parameters respectively, $i = 1, \ldots, k$.

In some of the practical situations, there is prior information of the ordering among the location parameters. For example, in dose-response experiments, the effect of a treatment may be related monotonically to the increasing levels of dose of a drug. Similarly, in against accelerated life testing, the higher stress level may lead to lowering the guaranteed lifetime. Many researchers have proposed statistical tests to test the null hypothesis $H_0: \mu_1 = \cdots =$

Corresponding author: Jatesh Kumar
Email: kumarjatesh31@gmail.com

μ_k the simple ordered alternative $H_1: \mu_1 \leq \cdots \leq \mu_k$, with at least one strict inequality, for normal and exponential probability models. This problem of simple ordered alternative is a member of the class of order restricted alternatives. A detailed discussion on order restricted statistical inferences can be found in Barlow *et al.* (1972) and Robertson and Dykstra (1988). Marcus (1976), Hayter (1990), Lee and Spurrier (1995), Liu *et al.* (2000) have also proposed tests for the simple ordered alternatives under normal probability model. Chen (1982) and Dhawan and Gill (1997) inverted the test procedures for testing homogeneity of the location parameters of k (≥ 3) two-parameter exponential distributions to construct simultaneous confidence intervals (SCIs) for the ordered pair-wise differences of location parameters under the assumption of homogeneity of scale parameters. Singh *et al*. (2006) proposed a procedure for successive comparisons of the location parameters of exponential distributions by assuming the equality of scale parameters. Maurya *et al.* (2011) came up with one-stage and two-stage multiple comparison procedures using Lam's (1987,1988) technique and obtained the conservative simultaneous confidence intervals (SCIs) for successive differences of the location parameters of several exponential distributions under the heteroscedasticity of scale parameters, *i.e*., $\theta_i \neq \theta_j, 1 \leq i < j \leq k$. Later, Singh and Singh (2013) put forward less conservative SCIs by extending Maurya *et al*. (2011) procedure. Kharrati Kopaei (2014) introduced a new lemma and used the same to provide SCIs for the successive differences of the location parameters which were less conservative than the SCIs of Maurya *et al*. (2011). It may be noted that Maurya *et al.* (2011), Singh and Singh (2013) and Kharrati Kopaei (2014) used the maximum likelihood estimator (MLE) of the location parameter. Although, the MLEs have a few desirable properties like efficiency and consistency but may not be unbiased. Zheng (2013) provided the penalized maximum likelihood estimators (PMLEs) of the location and scale parameters of two-parameter exponential distribution which are uniformly minimum variance unbiased estimators (UMVUEs). In this article, we have proposed one-stage and two-stage multiple comparison procedures to construct SCIs using the PMLEs of the location parameters for the ordered pair-wise differences of location parameters under heteroscedasticity of scale parameters. The layout of the paper is as follows.

In this paper, Sections 2 and 3 respectively contain the proposed one-stage and two-stage multiple comparison procedures to construct the simultaneous confidence intervals (SCIs) for the ordered pair-wise differences of location parameters. In Section 4, the results of Monte Carlo simulation studies conducted to compare the power, coverage probabilities (CP), and average volume (AV) of the proposed procedures with the procedure of Singh and Singh (2013), are presented. The implementation and the better performance ability of the proposed procedures over the more conservative procedure of Singh and Singh (2013), is demonstrated by taking a real-life example in Section 5. Finally, a brief conclusion is presented in Section 6.

2. One-Stage Procedure for the Simultaneous Testing of the Ordered Differences of Location Parameters

Let there be k independent exponential populations and that $X_{i1}, X_{i2}, \ldots, X_{im}$ be a random sample of size $m\ (>2)$ from the ith population $E_i(\mu_i, \theta_i)$, $i = 1, \ldots, k$. The maximum likelihood estimators (MLEs) of μ_i and θ_i are $X_i = \min(X_{i1}, X_{i2}, \ldots, X_{im})$ and $V_i = \sum_j^m (X_{ij} - X_i)/m$, respectively and these MLEs are not unbiased estimators. In literature, an approach exists in which a penalty is added to the regular likelihood function so that the new function no longer remains a monotone function of the location parameter. Let $X_{i[1]} \leq$

$X_{i[2]} \ldots \leq X_{i[m]}$ be the ordered values corresponding to the above random sample. Zheng (2013) used the penalty term $x_{i[1]} - \mu_i$ in the regular likelihood function, where $x_{i[1]}$ is the realized value of $X_{i[1]}$, and gave the penalized maximum likelihood function as follows:

$$L(\mu_i, \theta_i) = (x_{i[1]} - \mu_i) \prod_{j=1}^{m} f(x_{i[1]} | \mu_i, \theta_i) = (x_{i[1]} - \mu_i) \frac{1}{\theta_i^m} e^{-\frac{1}{\theta} \sum_{j=1}^{m} (x_{ij} - \mu_i)}, x_{i[1]} \geq \mu_i$$

The penalized maximum likelihood estimators (PMLEs) of μ_i and θ_i obtained from the above likelihood function are $Y_i = \frac{m X_{i[1]} - \bar{X}}{(m-1)}$ and $S_i = \frac{m(\bar{X} - X_{i[1]})}{(m-1)}$ respectively, where $\bar{X} = \sum_{j=1}^{m} X_{ij} / m$, is the sample mean. It is also proven that these estimators of the location and scale parameters are unique minimum variance unbiased estimators (UMVUEs). Previously, the same estimators have also been obtained by Cohen and Helm (1973) and Sarhan (1954) using different methods of estimation such as modified moment and least square, respectively.

Consider the family of hypotheses for the ordered location parameters

(i) $H_{0i}: \mu_j - \mu_i = 0$ against $H_{1i}: \mu_j - \mu_i > 0,\ 1 \leq i < j \leq k$ (One-sided problem)
(ii) $H_{0i}: \mu_j - \mu_i = 0$ against $H_{2i}: \mu_j - \mu_i \neq 0, 1 \leq i < j \leq k$ (Two-sided problem)

For the testing of these hypotheses, we can use the one-stage multiple comparison procedure given by Lam (1987, 1988) to construct simultaneous confidence intervals (SCIs) for the one-sided and two-sided sets of pair-wise differences of the ordered location parameters when the scale parameters are unknown and $\theta_i \neq \theta_j, 1 \leq i < j \leq k$, *i.e.*, heteroscedasticity of scale parameters exists. One-stage multiple comparison procedure has the merit over a two-stage procedure (as described in detail in Section 3) in practical situations where the second stage of sampling is not possible due to the shortage of time, budget, and destructive type of experiments or some other factors.

The PMLEs of the location and scale parameters have been utilized instead of the MLEs for the simultaneous testing of the ordered location parameters. It is easy to verify that the PMLEs of the location and scale parameter can be written $Y_i = X_i - S_i/m$ and $S_i = \sum_j^m (X_{ij} - X_i)/(m-1)$. Define a constant $d = \max_{1 \leq i \leq k} (S_i/m)$. The random variables $T_i = (X_i - \mu_i)/\theta_i$ and $2(m-1)S_i/\theta_i$ are stochastically independently distributed as $E(0,1)$ and Chi-square with $2(m-1)$ degree of freedom (d.f.), respectively. Hence, the statistic $W_i^* = m(X_i - \mu_i)/S_i$ is distributed as Snedecor F with $(2, 2m-2)$ degree of freedom (d.f.). Using a one-stage procedure on the similar lines of Lam's (1987, 1988), the proposed one-sided and two-sided simultaneous confidence intervals (SCIs) for the ordered pair-wise differences of location parameters under heteroscedasticity of scale parameters are given in the following theorem.

Theorem 1: Let $q_{k,m,\alpha} = F_{2,2m-2}^{-1}(1-\alpha)^{1/(k-1)} - 1$ and $r_{k,m,\alpha} = F_{2,2m-2}^{-1}(1-\alpha)^{1/k} - 1$, for given $0 < \alpha < 1$

(i) $P(\mu_j - \mu_i \geq Y_j - Y_i - d q_{k,m,\alpha},\ 1 \leq i < j \leq k) \geq 1 - \alpha$.
Then $(Y_j - Y_i - d q_{k,m,\alpha}, \infty)$ is the set of one-sided simultaneous confidence intervals for $\mu_j - \mu_i$ with confidence coefficient at least $(1-\alpha)$.
(ii) $P(Y_j - Y_i - d r_{k,m,\alpha} \leq \mu_j - \mu_i \leq Y_j - Y_i + d r_{k,m,\alpha},\ 1 \leq i < j \leq k) \geq 1 - \alpha$.

Then $(Y_j - Y_i - dr_{k,m,\alpha}, Y_j - Y_i + dr_{k,m,\alpha})$ is the set of two-sided simultaneous confidence intervals for $\mu_j - \mu_i$ with confidence coefficient at least $(1-\alpha)$.

We applied the following lemma of Lam (1987, 1988) to prove Theorem 1.

Lemma 1: Suppose X and Y are two random variables, and a and b are two positive constants; then

$$[aX \geq bY - dmax(a,b)] \supseteq [X \geq -d, Y \leq d \text{ and } X \geq Y - d].$$

The proofs of the part (i) and (ii) of Theorem 1 on the basis of Lemma 1 are as follow

Proof of part (i): $P(\mu_j - \mu_i \geq Y_j - Y_i - dq_{k,m,\alpha},\ 1 \leq i < j \leq k)$
$= P(X_i - \mu_i - S_i/m \geq X_j - \mu_j - S_j/m - dq_{k,m,\alpha},\ 1 \leq i < j \leq k)$
$= P(S_i/m(W_i^* - 1)) \geq S_j/m(W_j^* - 1) - dq_{k,m,\alpha},\ 1 \leq i < j \leq k)$
$= P(W_j^* - 1 \leq q_{k,m,\alpha},\ 1 \leq i < j \leq k)$
$= 1 - \alpha$ $(\text{Since } q_{k,m,\alpha} = F^{-1}_{2,2m-2}(1-\alpha)^{1/(k-1)} - 1)$.

Proof of part (ii): $P(Y_j - Y_i - dr_{k,m,\alpha} \leq \mu_j - \mu_i \leq Y_j - Y_i + dr_{k,m,\alpha},\ 1 \leq i < j \leq k)$
$= P(X_j - S_j/m - X_i + S_i/m - dr_{k,m,\alpha} \leq \mu_j - \mu_i \leq X_j - S_j/m - X_i + S_i/m + dr_{k,m,\alpha}, 1 \leq i < j \leq k)$
$= P(S_i/m\,(W_i^* - 1) \geq S_j/m(W_j^* - 1) - dr_{k,m,\alpha} \cap S_j/m(W_j^* - 1)$
$\geq S_i/m(W_i^* - 1) - dr_{k,m,\alpha}, 1 \leq i < j \leq k$
$= P(W_j^* - 1 \leq r_{k,m,\alpha} \cap W_i^* - 1 \leq r_{k,m,\alpha}, 1 \leq i < j \leq k)$
$= 1 - \alpha$. $(\text{Since } r_{k,m,\alpha} = F^{-1}_{2,2m-2}(1-\alpha)^{1/(k)} - 1)$.

Here $F^{-1}_{2,2m-2}(x)$ denotes the xth quantile of the snedecor F distribution with $(2, 2m-2)$ degree of freedom (d.f.).

3. Two-Stage Procedure for the Simultaneous Testing of the Ordered Differences of Location Parameters

A two-stage multiple comparison procedure has been used on the similar lines of Lam's (1987, 1988) to construct one-sided and two-sided simultaneous confidence intervals (SCIs) for the ordered pair-wise comparisons of location parameters of several exponential populations under the heteroscedasticity of scale parameters, which is explained below:

Stage 1: In the first stage, the procedure begins by taking random sample $X_{i1}, X_{i2}, \ldots, X_{im}$, of size $m\ (\geq 2)$ from the ith population $E_i(\mu_i, \theta_i)$. Let $\tilde{Y}_i = X_i - S_i/m$ and $S_i = \sum_j^m (X_{ij} - X_i)/(m-1)$ be the PMLEs of μ_i and θ_i, respectively, where $X_i = \min(X_{i1}, X_{i2}, \ldots, X_{im})$, $i = 1, \ldots, k$. The random variables $T_i = (X_i - \mu_i)/\theta_i$ and $2(m-1)S_i/\theta_i$ are independently distributed as $E(0,1)$ and Chi-square with $2(m-1)$ d.f., respectively.

Stage 2: In the second stage $(N_i - m)$ additional observations are taken, for that we defined $N_i = \max[m, [S_i/c] + 1]$, $i = 1, \ldots, k$, where c is an arbitrary positive constant to be chosen to control the width of the confidence intervals and $[x]$ denotes the greatest integer less than or equal to x. If $N_i = m$, we do not take any more sample observations from each population. If $N_i > m$, then take $(N_i - m)$ more/additional sample observations $X_{i,m+1}, \ldots, X_{iN_i}$, from the

ith population $E_i(\mu_i, \theta_i)$. This is known as the second stage of the two-stage procedure. Now, based on the combined sample observations $X_{i,1}, \dots, X_{i,m}, X_{i,m+1}, \dots, X_{i,N_i}$, let $\tilde{X}_i = \tilde{X}_{iN_i} = \min\left(X_{i,1}, \dots, X_{i,m}, X_{i,m+1}, \dots, X_{i,N_i}\right)$ and $\tilde{Y}_i = \tilde{X}_i - S_i/N_i$. It can be noted that $U_i = N_i\left(\tilde{X}_{iN_i} - \mu_i\right)/\theta_i$ and $2(m-1)S_i/\theta_i$ are stochastically independently distributed as $E(0,1)$ and Chi-square with $2(m-1)$ d.f., respectively. Hence $W_i = N_i\left(\tilde{X}_{iN_i} - \mu_i\right)/S_i$ is distributed as Snedecor F with $(2, 2m-2)$ d.f.

The following theorem will provide us the one-sided and two-sided simultaneous confidence intervals (SCIs) for the ordered pair-wise differences of location parameters under heteroscedasticity of scale parameters.

Theorem 2: Let $u_{k,m,\alpha} = F^{-1}_{2,2m-2}(1-\alpha)^{1/(k-1)} - 1$ and $v_{k,m,\alpha} = F^{-1}_{2,2m-2}(1-\alpha)^{1/k} - 1$, for given $0 < \alpha < 1$

(i) $P\left(\mu_j - \mu_i \geq \tilde{Y}_j - \tilde{Y}_i - cu_{k,m,\alpha}, 1 \leq i < j \leq k\right) \geq 1-\alpha$.
Then $(\tilde{Y}_j - \tilde{Y}_i - cu_{k,m,\alpha}, \infty)$ is the set of one-sided simultaneous confidence intervals for $\mu_j - \mu_i$ with confidence coefficient at least $(1-\alpha)$.

(ii) $P\left(\tilde{Y}_j - \tilde{Y}_i - cv_{k,m,\alpha} \leq \mu_j - \mu_i \leq \tilde{Y}_j - \tilde{Y}_i + cv_{k,m,\alpha},\ 1 \leq i < j \leq k\right) \geq 1-\alpha$.

Then $(\tilde{Y}_j - \tilde{Y}_i - cv_{k,m,\alpha}, \tilde{Y}_j - \tilde{Y}_i + cv_{k,m,\alpha})$ is the set of two-sided simultaneous confidence intervals for $\mu_j - \mu_i$ with confidence coefficient at least $(1-\alpha)$.

Proof: The proof of the Theorem 2 is based on the similar lines of Theorem 1, by replacing c with d.

4. Simulation Study

For the purpose of comparison of the proposed procedures, say Prop, with the procedure of Singh and Singh (2013), say SS, a Monte Carlo simulation study has been performed using 10^5 iterations. The simulated power, coverage probability (CP), and the average volume (AV) of SCIs under each of these procedures have been computed. In each iteration fresh random samples were generated from each of the $k = 4$ exponential distributions with location parameters $(\mu_1, \mu_2, \mu_3, \mu_4)$ and scale parameters $(\theta_1, \theta_2, \theta_3, \theta_4)$. We have used the values of sample size and parametric configuration, i.e., the value of m, $(\mu_1, \mu_2, \mu_3, \mu_4)$ and $(\theta_1, \theta_2, \theta_3, \theta_4)$, as taken by Singh and Singh (2013) so that their simulated results can be incorporated in the comparison Tables 1-4. The simulated coverage probability is the proportion of repetitions in which all the ordered differences of location parameters are contained in the respective confidence intervals among 10^5 repetitions. The volume of simultaneous confidence intervals in a repetition is the product of lengths of all the underlying confidence intervals. The average volume is the average of the volumes obtained under 10^5 repetitions. Thus, the average volume is with respect to two-sided SCIs where the lower and upper limits are finite. Simulated power is the proportion of repetitions in which at least one of the ordered differences $\mu_j - \mu_i$, $1 \leq i < j \leq k$ falls outside the corresponding confidence interval.

Table 1: Simulated powers of one-stage procedure at $1-\alpha=.95$, for varied configuration of $(\mu_1,\mu_2,\mu_3,\mu_4)$ when $(\theta_1,\theta_2,\theta_3,\theta_4)=(1,1.1,1.2,1.3)$

m	$(\mu_1,\mu_2,\mu_3,\mu_4)$	One-sided case		Two-sided case	
		SS	Prop	SS	Prop
10	(0,0,0,.4)	.069	.180	.044	.117
15		.434	.761	.313	.631
16		.666	.933	.425	.746
17		.792	.972	.548	.835
18		.885	.989	.670	.902
19		.943	.996	.772	.942
20		.926	.987	.860	.969
25		.999	1	.986	.998
30		1	1	1	1
10	(0,.2,.3,.4)	.066	.16	.039	.090
15		.369	.629	.218	.424
16		.469	.724	.357	.611
17		.572	.803	.455	.706
18		.672	.863	.559	.787
19		.764	.908	.658	.851
20		.832	.941	.666	.846
25		.981	.995	.944	.982
30		.998	1	.994	.998

Table 2: Simulated powers of one-stage procedure at $1-\alpha=.95$, for varied configuration of $(\mu_1,\mu_2,\mu_3,\mu_4)$ when $(\theta_1,\theta_2,\theta_3,\theta_4)=(1,1,1,1)$

m	$(\mu_1,\mu_2,\mu_3,\mu_4)$	One-sided case		Two-sided case	
		SS	Prop	SS	Prop
10	(0,0,0,.4)	.087	.271	.056	.172
15		.660	.937	.517	.862
16		.791	.973	.667	.933
17		.885	.989	.790	.971
18		.943	.996	.884	.989
19		.976	.999	.943	.996
20		.990	1	.975	.999
25		1	1	1	1
30		1	1	1	1
10	(0,.2,.3,.4)	.080	.225	.050	.145
15		.524	.812	.402	.718
16		.641	.878	.530	.811
17		.750	.924	.646	.877

18		.833	.955	.775	.925
19		.892	.973	.831	.954
20		.932	.984	.892	.972
25		.994	.999	.990	.998
30		1	1	.999	1

Table 3: The Coverage Probabilities (CP) and Average Volumes (AV) of two-sided SCIs under one-stage procedure for$1-\alpha=.95$

m	$(\theta_1,\theta_2,\theta_3,\theta_4)$	SS		Prop	
		CP	AV	CP	AV
10	(1,1,1,1)	.996	24.853	.987	7.66
15		.993	.724	.977	.196
16		.993	.428	.976	.114
17		.992	.263	.974	.069
18		.991	.167	.972	.063
19		.990	.108	.971	.028
20		.990	.073	.969	.019
25		.989	.013	.963	.003
30		.986	.004	.957	.001
10	(1,1.1,1.2,1.3)	.996	68.961	.987	21.255
15		.992	2.031	.978	.551
16		.992	1.195	.975	.319
17		.992	.742	.974	.195
18		.991	.469	.973	.122
19		.991	.308	.971	.079
20		.991	.327	.972	.083
25		.988	.061	.965	.015
30		.987	.016	.962	.004

Table 4: Simulated powers of two-stage procedure for varied configurations of $(\mu_1,\mu_2,\mu_3,\mu_4)$ and $(\theta_1,\theta_2,\theta_3,\theta_4)$ for $1-\alpha=.95$

L	m	$(\theta_1,\theta_2,\theta_3,\theta_4)$	$(\mu_1,\mu_2,\mu_3,\mu_4)$	One-sided case		Two-sided case	
				SS	Prop	SS	Prop
0.6	10	(1,1.1,1.2,1.3)	(0,0,0,.3)	.755	.748	.757	.749
	20			.772	.718	.767	.724
	30			.782	.691	.783	.692
	10		(0,.1,.2,.3)	.555	.546	.552	.545
	20			.580	.514	.570	.518
	30			.581	.478	.578	.475
	10	(1,1.1,1.2,1)	(0,0,0,.3)	.751	.757	.754	.756
	20			.739	.773	.742	.771
	30			.734	.784	.734	.783
	10		(0,.1,.2,.3)	.549	.550	.547	.546

	20			.537	.535	.540	.539
	30			.519	.516	.519	.517
	10	(1,1,1,1)	(0,0,0,.3)	.755	.755	.754	.754
	20			.750	.754	.749	.750
	30			.751	.752	.751	.751
	10		(0,.1,.2,.3)	.543	.544	.542	.544
	20			.530	.530	.525	.525
	30			.509	.510	.509	.509

Tables 1-2 show that the power of the proposed one-stage procedure using the PMLEs is substantially higher for small and moderate sample sizes than the power of the MLEs based procedure of Singh and Singh (2013). The analysis of Table 3 also suggests that the simulated coverage probability (CP) of the proposed procedure is closer to the nominal level .95 for moderate and large sample sizes whereas it is too high (close to .99) under the procedure of Singh and Singh (2013). Further, the average volume is also substantially smaller under the proposed procedure than the Singh and Singh (2013) procedure and it indicates that the length of the SCIs under the proposed procedure is smaller than the Singh and Singh (2013) procedure. The simulated powers under a two-stage setup are the same for the Proposed and Singh and Singh (2013) procedures.

5. Real Life Example

We have taken the same data set as illustrated in Maruya *et al.* (2011) and Singh and Singh (2013), presented in Table 5. The data is about the survival times of inoperable lung cancer patients, categorized on the basis of histological type of tumor (squamous, small, adeno and large), who were subjected to standard chemotherapeutic agents.

Singh and Singh (2013) have constructed one-sided and two-sided simultaneous confidence (SCIs) by taking $c = 11.862$. Note that the choice of c determines the size of the sample from each population. In this numerical example, the choice of $c = 11.862$, gives the same sample sizes (9, 9, 9, 9) from all the four populations under the proposed and Singh and Singh (2013) procedures so that the comparison is feasible. Therefore for $c = 11.862$, the length $l = 2cu_{k,m,\alpha}$ of SCIs under the proposed two-stage procedure are 187.656, 143.981 and 113.92 at $\alpha = .01$, $\alpha = .025$ and $\alpha = .05$, respectively. The lengths of these SCIs are smaller than those reported in Singh and Singh (2013).

Table 5: Survival time (days) of inoperable lung cancer patients

	Type of Tumor			
	Squamous	Small	Adeno	Large
Survival Days	72	30	8	177
	10	13	92	162
	81	23	35	553
	110	16	117	200
	100	21	132	156
	42	18	12	182
	8	20	162	143

	25	27	3	105
	11	31	95	103

We have constructed simultaneous confidence intervals (SCIs) using Theorem 1, since for $c = 11.862$ the sample sizes are same under both the one-stage and two-stage procedures. The estimates of the scale and location parameters respectively, for the above reported data in Table 5 are, $S_1 = 48.375, S_2 = 10.25, S_3 = 78.265, S_4 = 106.7$ and $Y_1^{'} = 8 - \frac{48.375}{8} = 2.625, Y_2^{'} = 13 - \frac{10.25}{8} = 11.861, Y_3^{'} = 3 - \frac{78.265}{8} = -5.696, Y_4^{'} = 103 - \frac{106.7}{8} = 91.144$. The required values of the critical constants for $m = 9, k = 4$ and at the level of significance $\alpha = .01, .025$ and $.05$ are $q_{k,m,.05} = 4.318, q_{k,m,.025} = 5.539, q_{k,m,.01} = 7.314$, and $r_{k,m,.05} = 4.080, r_{k,m,.025} = 6.069, r_{k,m,.01} = 7.910$. The constructed one-sided and two-sided simultaneous confidence intervals are presented in Table 6.

Table 6: Simultaneous confidence intervals (SCIs) under the proposed (Prop) and Singh and Singh (2013) (SS) procedures

	Difference	SS	Prop
		$\alpha = .01$	$\alpha = .01$
One-Sided SCI	$\mu_2 - \mu_1$	(-93.620, ∞)	(-77.522, ∞)
	$\mu_3 - \mu_2$	(-108.620, ∞)	(-104.315, ∞)
	$\mu_3 - \mu_1$	(-103.690, ∞)	(-95.079, ∞)
	$\mu_4 - \mu_3$	(1.379, ∞)	(10.076, ∞)
	$\mu_4 - \mu_2$	(-8.620, ∞)	(-7.480, ∞)
	$\mu_4 - \mu_1$	(-3.620, ∞)	(1.755, ∞)
Two-Sided SCI	$\mu_2 - \mu_1$	(-100.690,110.690)	(-84.592, 103.064)
	$\mu_3 - \mu_2$	(-115.690,95.690)	(-111.385, 76.271)
	$\mu_3 - \mu_1$	(-110.690,100.690)	(-102.149,85.507)
	$\mu_4 - \mu_3$	(-5.690,205.690)	(3.006,190.663)
	$\mu_4 - \mu_2$	(-15.690,195.690)	(-14.550,173.106)
	$\mu_4 - \mu_1$	(-10.690,200.690)	(-5.314,182.342)

A pair-wise difference is declared to be significant if the corresponding simultaneous confidence interval (SCI) does not contain zero. Accordingly, at the level $\alpha = .01$, we infer that: (i) Under one-sided SCIs the Singh and Singh (2013) procedure declares the difference $\mu_4 - \mu_3$ as significant whereas the proposed procedure declares two differences $\mu_4 - \mu_3$ and $\mu_4 - \mu_1$ as significant (the corresponding SCIs do not contain zero); (ii) Under two-sided SCIs, the proposed procedure declares the difference $\mu_4 - \mu_1$ as significant whereas the Singh and Singh (2013) procedure does not declare any difference as significant.

6. Conclusion

We have observed that lengths of SCIs of the proposed one-stage and two-stage procedures, based on the PMLEs, are significantly smaller and that their coverage probability is also close to the nominal level as compared to the MLEs based procedure of Singh and Singh (2013). Thus, the Singh and Singh (2013) procedure is too conservative than the

proposed procedure. Further, the power of the proposed one-stage procedure is higher than the one-stage procedure of Singh and Singh (2013) and both procedures have almost the same power under the two-stage setup. Keeping in view the dominance of the proposed procedures in terms of lengths of SCIs, coverage probability, and average volume, we recommend the use of proposed procedures, particularly, the one-stage procedure when there are smaller samples from the populations. In most of the practical situations we get smaller samples on life lengths and the use of the proposed one-stage procedure, based on the PMLEs, is recommended since it dominates the procedure of Singh and Singh (2013) in terms of lengths of SCIs, power, coverage probability and average volume.

Acknowledgement

The authors are thankful to the reviewer and the Chair Editor for their valuable comments, which led to substantial improvement in the presentation of the manuscript.

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions*. John Wiley, New York. (ISBN-10: 0471049700).

Chen, H. (1982). A new range statistic for comparisons of several exponential location parameters. *Biometrika*, **69**, 257-260.

Cohen, A. C. and Helm, F. R. (1973). Estimators in the exponential distribution. *Technometrics*, **15**, 415-418.

Dhawan, A. K. and Gill, A. N. (1997). Simultaneous confidence intervals for the ordered pairwise differences of exponential location parameters. *Communications in Statistics-Theory and Methods*, **26**(**1**), 247-262.

Hayter, A. (1990). A one-sided studentized range test for testing against a simple ordered alternative. *Journal of the American Statistical Association*, **85**, 778-785.

Kharrti, Kopaei M. (2014). A note on the simultaneous confidence intervals for the differences of successive differences of exponential location parameters under heteroscedasticity. *Statistical Methodology*, **22**, 1-17.

Lam, K. (1987). Subset selection of normal populations under heteroscedasticity. *In: Proceeding of the Second International Advanced Seminar/Workshop on Inference Procedures Associated with Ranking and Selection*, *Sydney, Australia.*

Lam, K. (1988). An improved two-stage selection procedure. *Communications in Statistics- Computation and Simulations*, **17**(**3**), 995-1006.

Lee, R. E. and Spurrier, J. D. (1995). Successive comparison between ordered treatments. *Journal of Statistical Planning and Inference*, **43**, 323-330.

Liu, W., Miwa, T. and Hayter, A. J. (2000). Simultaneous confidence interval estimation for successive comparisons of ordered treatment effects. *Journal of Statistical Planning and Inference*, **88**, 75-86.

Marcus, R. (1976). The power of some tests of the equality of normal means against an ordered alternative. *Biometrika*, **63**, 177-183.

Maurya, V. Goyal, A. and Gill, A. N. (2011). Simultaneous testing for successive difference of location parameters under heteroscedasticity. *Statistics and Probability Letters*, **81**(**10**), 1507-1517.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Ordered Restricted Statistical Inferences*. John Wiley, New York. (ISBN: 0-471-91787-7).

Sarhan, A. E. (1954). Estimation of the mean and standard deviation by ordered statistics. *Annals of Mathematics and Statistics*, **25**, 317-318.

Singh, P., Abebe, A. and Mishra, S. (2006). Simultaneous testing for successives differences of exponential location parameters. *Communications in Statistics-Simulation and Computations*, **35** (**3**), 547-561.

Singh, P. and Singh, N. (2013). Simultaneous confidence intervals for ordered pair wise differences of exponential location parameters under heteroscedasticity. *Statistics and Probability Letters*, **83**, 2673-2678.

Zheng, M. (2013). Penalized maximum likelihood estimation of two-parameter exponential distribution. An unpublished *Project Submitted to the Faculty of the Graduate School of the University Minnesota.*

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 251–263

Some Results of Auto-Relevation Transform in Reliability Analysis

Dileepkumar M.[1] **and P. G. Sankaran**[2]
[1]*Department of Statistics, University of Calicut, Thenhipalam, 673635, Kerala, India.*
[2]*Department of Statistics, Cochin University of Science and Technology, Cochin 682022, Kerala, India.*

Received: 21 September 2021; Revised: 27 November 2021; Accepted: 28 December 2021

Abstract

In this paper, we study some important reliability characteristics of auto-relevation transform. Various ageing and ordering concepts are discussed. Important results in terms of reliability and information measures are studied. Some characterizations are presented. A new lifetime distribution called auto-relevated Lomax (ARL) is introduced and its practical applicability is illustrated with a real dataset.

Key words: Relevation transform; Hazard rate; Ageing properties; Stochastic orders.

AMS Subject Classifications: 90B25, 60E05.

1. Introduction

Let X and Y be two absolutely continuous non-negative random variables, with survival functions $\bar{F}(.)$ and $\bar{G}(.)$ respectively. Consider an item from a population with survival function $\bar{F}(x)$, which is being replaced at the time of its failure at age x, by another item of the same age x from another population with survival function $\bar{G}(x)$. Then the survival function

$$\bar{T}(x) = \bar{F}\#\bar{G}(x) = \bar{F}(x) - \bar{G}(x)\int_0^x \frac{1}{\bar{G}(t)} d\bar{F}t. \tag{1}$$

is called the relevation transform of X and Y introduced by Krakowski (1973). Let $Y(X)$ denote the total lifetime of the random variable Y given it exceeds a random time X, (i.e $Y(X) \stackrel{d}{=} \{Y|Y > X\}$). Then (1) is the survival function of the random variable $Y(X)$. The probability density function (p.d.f.) of the relevation random variable is obtained as

$$t(x) = T'(x) = g(x)\int_0^x \frac{f(t)}{\bar{G}(t)} dt. \tag{2}$$

Grosswald *et al.* (1980) presented two characterizations of the exponential distribution based on relevation transform. The concept of dependent relevation transform and its importance in reliability analysis is given in Johnson and Kotz (1981). Baxter (1982) discussed

Corresponding Author: Dileepkumar M.
Email: drdileepkumar@uoc.ac.in

certain reliability applications of the relevation transform. Shanthikumar and Baxter (1985) provided closure properties of certain ageing concepts in the context of relevation transforms. Improved versions of the results in Grosswald *et al.* (1980) are given by Lau and Rao (1990). Chukova *et al.* (1993) established characterizations of the class of distributions with almost lack of memory property based on the relevation transform. Sankaran and Dileepkumar (2019) studied important reliability properties of the relevation transform in the context of proportional hazards model.

When the random variables X and Y are identically distributed, the tail distribution of the random variable $Y(X)$ can be simplified to

$$\bar{T}^*(x) = \bar{F}(x)(1 - \log(\bar{F}(x))). \tag{3}$$

The survival function (3) is known as the auto-relevation of $\bar{F}(x)$. Kapodistria and Psarrakos (2012) studied properties and applications of a sequence of random variables with weighted tail distribution functions based on the auto-relevation transform. In this paper we focus our attention on various properties, applications and characterizations of the auto-relevation transform in the context of reliability theory.

The rest of the paper is organized as follows. We provide the concept and basic characteristics of auto-relevation transform in Section 2. Section 3 presents some important characterization results based on reliability and information measures. Various ageing properties and stochastic orders of auto-relevation are presented in Section 4 and Section 5 respectively. Finally, in Section 6, we provide major conclusions of the study.

2. Auto-Relevation Transform (ART)

Let X and Y be two non-negative continuous random variables with survival functions $\bar{F}(x)$ and $\bar{G}(x)$ respectively. Then the survival function of the relevation random variable $Y(X)$ is given in (1). When X and Y are identically distributed, the random variable $X(X)$ is known as the auto-relevation of X. Survival function of $X(X)$ is obtained as

$$\begin{aligned}\bar{T}^*(x) &= \bar{F}(x) - \bar{F}(x)\int_{t=0}^{x} \frac{1}{\bar{F}(x)} d\bar{F}(x) \\ &= \bar{F}(x)(1 - \log(\bar{F}(x)).\end{aligned} \tag{4}$$

The probability density function (p.d.f) of $X(X)$ is obtained as

$$t^*(x) = -f(x)\log(\bar{F}(x)). \tag{5}$$

From (4) and (5), we have, $X(X)$ is the auto-relevation of X if and only if

$$\begin{aligned} h_{X(X)}(x) &= \frac{t^*(x)}{\bar{T}^*(x)} \\ \Leftrightarrow h_{X(X)}(x) &= -\frac{f(x)\log(\bar{F}(x))}{\bar{F}(x)(1-\log(\bar{F}(x)))} \\ \Leftrightarrow h_{X(X)}(x) &= h_X(x)\left(\frac{\log(\bar{F}(x))}{\log(\bar{F}(x))-1}\right) = h_X(x)\left(\frac{\Lambda_X(x)}{1+\Lambda_X(x)}\right), \end{aligned} \tag{6}$$

where $\Lambda_X(x) = -\log(\bar{F}(x))$ is the cumulative hazards function of X.

An important class of distributions used in risk theory and queueing theory is the class $\mathscr{L}$ distribution. A distribution F belongs to the class $\mathscr{L}$ if

$$\lim_{x\to\infty} \frac{\bar{F}(x-y)}{\bar{F}(x)} = 1, \forall\, y \in R. \tag{7}$$

Kluppelberg (1988) showed that,
$F \in \mathscr{L}$ if and only if $\lim_{x\to\infty} h_F(x) = 0$, where $h_F(x)$ is the hazard rate function of $F(x)$.
Proposition 1: if $X \in \mathscr{L}$ then $X(X) \in \mathscr{L}$.

Proof: We have

$$\lim_{x\to\infty} h_{X(X)}(x) = \lim_{x\to\infty} h_X(x) \lim_{x\to\infty}\left(\frac{\log(\bar{F}(x))}{\log(\bar{F}(x))-1}\right). \tag{8}$$

Now by applying L'Hospitals's rule and noting that $\lim_{x\to\infty} h(x) = 0$, we get
$\lim_{x\to\infty} h_{X(X)}(x) = 0$. This completes the proof. □

Let $Q_X(.)$ and $Q_{X(X)}(.)$ be the quantile functions of the random variables X and $X(X)$ with respective distribution functions $F(x)$ and $T^*(x)$. In the following, we establish the relation between the quantile functions of X and $X(X)$.
Proposition 2: Suppose $Q_X(.)$ and $Q_{X(X)}(.)$ are the quantile functions of the random variables X and $X(X)$ respectively. Then

$$Q_X(u) = Q_{X(X)}(u + (1-u)\log(1-u)). \tag{9}$$

Proof: From (4), we have

$$T^*(x) = 1 - \bar{F}(x)(1-\log(\bar{F}(x))). \tag{10}$$

By taking $F(x) = u$ where $u \in (0,1)$, we get $X = Q_X(u)$. Using this in (10), we have

$$\begin{aligned} & T^*(Q_X(u)) = 1-(1-u)(1-\log(1-u)) \\ \Longrightarrow \quad & Q_X(u) = Q_{X(X)}(u+(1-u)\log(1-u)). \end{aligned}$$

□

Remark 1: When the cumulative distribution function of $X(X)$ is non-invertibe, we can effectively employ the identity (9) to simulate random samples of $X(X)$ using the quantile function of X.

3. Characterization Results

Glaser (1980) established a general theorem that facilitates the determination of whether $h_X(x)$ is increasing (IHR), decreasing (DHR), Bath-tub (BT) or upside-down bathtub (UBT). He made use of the function $\psi(x) = -\frac{f'(x)}{f(x)}$, known as the Glaser's function. In the next proposition, we present an interesting identity connecting the Glaser's functions of the random variables X and $X(X)$.

Proposition 3: Let X be a non-negative continuous random variable with survival function $\bar{F}(x)$. Then $X(X)$ is the auto-relevation of X if and only if

$$\psi_{X(X)}(x) = \psi_X(x) - \frac{h_X(x)}{\Lambda_X(x)}, \tag{11}$$

where $\psi_X(x)$ and $\psi_{X(X)}(x)$ are the Glaser's function of X and $X(X)$ respectively.

Proof: If $X(X)$ is the auto-relevation of X then we have

$$\begin{aligned}
\psi_{X(X)}(x) &= -\frac{t^{*\prime}(x)}{t^*(x)} \\
\psi_{X(X)}(x) &= -\frac{f'(x)}{f(x)} + \frac{f(x)}{\bar{F}(x)\log(\bar{F}(x))} \\
\psi_{X(X)}(x) &= \psi_X(x) - \frac{h_X(x)}{\log(\bar{F}(x))}.
\end{aligned} \tag{12}$$

Conversly (11) gives

$$\frac{d}{dx}\left(\log(t^*(x)\right) = \frac{d}{dx}\left(\log(-f(x)\ \log(\bar{F}(x)) + C\right), \tag{13}$$

where C is a constant. Since $t^*(x)$ is a density function, on integration, we get $C = 0$ and (13) reduces to

$$t^*(x) = -f(x)\ \log(\bar{F}(x)). \tag{14}$$

This completes the proof. □ The odds function of a random variable X is defined by

$$\phi_X(x) = \frac{P(X > x)}{P(X \leq x)} = \frac{\bar{F}_X(x)}{F_X(x)}.$$

Note that the odds function is a decreasing function of x. In the coming proposition, we provide an interesting connection between the odds functions of $X(X)$ and X.

Proposition 4: $X(X)$ is the auto-relevated random variable of X if and only if

$$\phi_{X(X)}(x) = \frac{1 + \Lambda(x)}{\phi_X^{-1}(x) - \Lambda(x)}, \tag{15}$$

where $\phi_{X(X)}(x)$ and $\phi_X(x)$ are the odds functions of $X(X)$ and X respectively.

Proof: Assume $X(X)$ is the auto-relevated random variable of X. From (4), We have

$$\begin{aligned} & \phi_{X(X)}(x) = \frac{\bar{T}^*(x)}{1-\bar{T}^*(x)} \\ \Leftrightarrow \quad & \phi_{X(X)}(x) = \frac{\bar{F}(x) - \bar{F}(x)\log(\bar{F}(x))}{F(x) + \bar{F}(x)\log(\bar{F}(x))} \\ \Leftrightarrow \quad & \phi_{X(X)}(x) = \frac{\phi_X(x)(1-\log(\bar{F}(x)))}{1+\phi_X(x)\log(\bar{F}(x))} \\ \Leftrightarrow \quad & \phi_{X(X)}(x) = \frac{1+\Lambda(x)}{\phi_X^{-1}(x) - \Lambda(x)}, \end{aligned} \tag{16}$$

which completes the proof. □

To measure the distance between two probability distributions, Kullback-Leibler divergence (K-L divergence) has been popularly used in modelling of statistical data. The K-L divergence, which is closely related to relative entropy, information divergence, and information for discrimination is a non-symmetric measure of the difference between two probability distributions $f(x)$ and $g(x)$. When $f(x)$ and $g(x)$ are non-negative continuous distributions, then the K-L divergence $I(f,g)$ is defined as

$$I(f,g) = \int_0^\infty f(x)\log\left(\frac{f(x)}{g(x)}\right)dx. \tag{17}$$

Specifically, the K-L divergence of $g(x)$ from $f(x)$, denoted $I(f,g)$, is a measure of the information lost when $g(x)$ is used to approximate $f(x)$. In the following we present a relationship between $I(t^*,f)$ and $I(f,t^*)$ in the context of ART.

Proposition 5: Let $X(X)$ be the ART random variable corresponding to the non-negative random variable X. Then

$$I(X,X(X)) = 1 - I(X(X),X), \tag{18}$$

where $I(X,X(X))$ is the Kullback-Leibler divergence between X and $X(X)$.

Proof: From (17), we have

$$I(X(X),X) = \int_0^\infty t^*(x)\log\left(\frac{t^*(x)}{f(x)}\right)dx. \tag{19}$$

Since $X(X)$ is the ART random variable corresponding to X, using (5) in (19), we get

$$I(X(X),X) = -\int_0^\infty f(x)\log(\bar{F}(x))\log(-\log(\bar{F}(x)))dx. \tag{20}$$

by taking $u = -\log(\bar{F}(x))$, the integral in (20) became

$$I(X(X),X) = \int_0^\infty u\log(u)e^{-u}du.$$

Now by applying integration by parts, we obtain

$$I(X(X),X) = -\underset{x\to\infty}{Lim}\left(\frac{x\log(x)}{e^x}\right) + \underset{x\to 0}{Lim}\left(\frac{x\log(x)}{e^x}\right) + \int_0^\infty (1+\log(x))e^{-x}dx$$
$$= \int_0^\infty (1+\log(x))e^{-x}dx. \tag{21}$$

Again applying integration by parts on (21), we get

$$I(X(X),X) = 1 + \int_0^\infty \log(x)e^{-x}dx = 1-\gamma, \tag{22}$$

where $\gamma = -\int_0^\infty \log(x)e^{-x}dx$ is the Euler–Mascheroni constant ($\gamma \simeq 0.5772$). Now, we have

$$I(X,X(X)) = \int_0^\infty f(x)\log\left(\frac{f(x)}{t^*(x)}\right)$$
$$= -\int_0^\infty f(x)\log(-\log(\bar{F}(x)))dx. \tag{23}$$

Using the transformation $u = -\log(\bar{F}(x))$, (23) becomes

$$I(X,X(X)) = -\int_0^\infty \log(u)e^{-u}du = \gamma. \tag{24}$$

From (22) and (24), the result follows. □

4. Ageing Properties

We describe ageing properties of the relevation random variable $X(X)$ in connection with the ageing behaviour of the baseline random variable X. Various ageing classes and their properties and applications can be seen in Barlow and Proschan (1975), Shaked and Shanthikumar (2007), and Nair *et al.* (2013). From (6), we have

$$h_{X(X)}(x) = h_X(x)\left(\frac{\log(\bar{F}(x))}{\log(\bar{F}(x))-1}\right). \tag{25}$$

Differentiating (25), we obtain

$$h'_{X(X)}(x) = h'_X(x)\left(\frac{\log(\bar{F}(x))}{\log(\bar{F}(x))-1}\right) + \left(\frac{h_X(x)}{(\log(\bar{F}(x)-1)}\right)^2. \tag{26}$$

Note that $\left(\frac{\log(\bar{F}(x))}{\log(\bar{F}(x))-1}\right) > 0$ and $\left(\frac{h_X(x)}{(\log(\bar{F}(x)-1)}\right)^2 > 0$ for all $x > 0$. Thus when X is IHR, we have $h'_X(x) > 0$ for all $x > 0$, which gives $h'_{X(X)}(x) > 0$ for all $x > 0$. Thus $X(X)$ is also IHR. Hence IHR property is preserved under auto-relevation. When X is an exponential random variable with hazard rate $h_X(x) = C$, where $C > 0$, a constant. Then, from (26) we obtain

$$h'_{X(X)}(x) = \left(\frac{C}{(\log(\bar{F}(x)-1)}\right)^2 \geq 0. \tag{27}$$

Thus auto-relevated exponential distribution is always IHR. However, the case when X is DHR gives different options, which is presented in the next proposition.

Proposition 6: Let X be a non-negative continuous random variable with survival function $\bar{F}(x)$. Suppose X is DHR. Then the auto-relevation random variable $X(X)$ is IHR (DHR) if and only if

$$\frac{h'_X(x)}{(h_X(x))^2} \geq (\leq) \frac{-1}{\Lambda_X(x)(\Lambda_X(x)+1)} \text{ for all } x > 0. \tag{28}$$

Proof: We have

$$h'_{X(X)}(x) = h'_X(x)\left(\frac{\log(\bar{F}(x))}{\log(\bar{F}(x))-1}\right) + \left(\frac{h_X(x)}{(\log(\bar{F}(x)-1)}\right)^2. \tag{29}$$

$X(X)$ is IHR(DHR) if and only if $h'_{X(X)}(x) \geq (\leq) 0$. Now, since X is DHR, we have $h'_X(x) < 0$ for all $x > 0$. By using the facts that $\left(\frac{\log(\bar{F}(x))}{\log(\bar{F}(x))-1}\right)$ and $\left(\frac{h_X(x)}{(\log(\bar{F}(x)-1)}\right)^2$ are non-negative, we get $X(X)$ is IHR(DHR) if and only if, for all $x > 0$,

$$\begin{aligned}
&-h'_X(x)\left(\frac{\log(\bar{F}(x))}{\log(\bar{F}(x))-1}\right) \leq (\geq) \left(\frac{h_X(x)}{(\log(\bar{F}(x)-1)}\right)^2 \\
\Leftrightarrow &-\frac{h'_X(x)}{(h_X(x))^2} \leq (\geq) \frac{1}{\log(\bar{F}(x)(\log(\bar{F}(x)-1))} \\
\Leftrightarrow &\frac{h'_X(x)}{(h_X(x))^2} \geq (\leq) \frac{-1}{\Lambda_X(x)(\Lambda_X(x)+1)}, \text{ for all } x > 0.
\end{aligned}$$

□

Remark 2: Note that $X(X)$ accommodates non-monotonic shapes when the equality holds in (28). The change point of the non-monotonic hazard function will be obtained by solving the equality (28).

From Proposition 6, it is clear that the auto-relevation of DHR class of distributions can provide new lifetime models with non-monotonic hazard rate functions. Note that the auto-relevated distribution consists of the same number of parameters as in the parent distribution. Thus we can efficiently use the auto-relevation transformation for developing more flexible lifetime models from the existing ones without introducing additional parameters. To illustrate this, consider the Lomax distribution with survival function

$$\bar{F}(x) = \left(\frac{\alpha}{x+\alpha}\right)^c, \quad \alpha > 0,\, c > 0 \text{ and } 0 < x < \infty, \tag{30}$$

and hazard function

$$h_X(x) = \frac{c}{\alpha + x}. \tag{31}$$

We have $h_X(x)$ is non-increasing for all parameter combinations. Thus X is always DHR. The survival function of the auto-relevated Lomax random variable (ARL) $X(X)$ has the form

$$\bar{T}^*(x) = \left(\frac{\alpha}{\alpha+x}\right)^c \left(1 - \log\left(\left(\frac{\alpha}{\alpha+x}\right)^c\right)\right). \tag{32}$$

The corresponding hazard function is obtained as

$$h_{X(X)}(x) = \frac{c}{\alpha + x}\left(\frac{\log\left(\left(\frac{\alpha}{\alpha+x}\right)^c\right)}{\log\left(\left(\frac{\alpha}{\alpha+x}\right)^c\right) - 1}\right). \tag{33}$$

On differentiating, we get

$$h'_{X(X)}(x) = \frac{c\left(c + \log\left(\left(\frac{\alpha}{\alpha+x}\right)^c\right) - \left(\log\left(\left(\frac{\alpha}{\alpha+x}\right)^c\right)\right)^2\right)}{(\alpha+x)^2\left(\log\left(\left(\frac{\alpha}{\alpha+x}\right)^c\right) - 1\right)^2}. \tag{34}$$

Thus the sign of $h'_{X(X)}(x)$ depends only on the function

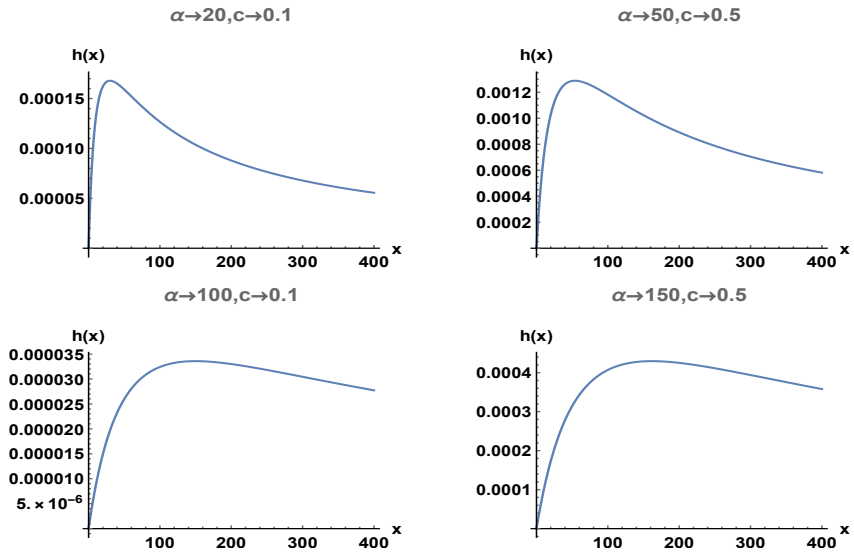


Figure 1: $h_{X(X)}(x)$ **of ARL distribution for different parameter combinations.**

$$\gamma(x) = \left(c + \log\left(\left(\frac{\alpha}{\alpha+x}\right)^c\right) - \left(\log\left(\left(\frac{\alpha}{\alpha+x}\right)^c\right)\right)^2\right).$$

We can write this as

$$\gamma(x) = c + k(x) - (k(x))^2, \tag{35}$$

where $k(x) = \log\left(\left(\frac{\alpha}{\alpha+x}\right)^c\right)$. We can observe that $k(x) < 0$ for all $x > 0$ and strictly decreasing for all $\alpha, c > 0$. Since $k(0) = 0$, it is clear that $\gamma(x)$ takes a positive sign initially and then became negative as x progresses. Correspondingly, the hazard function first increase then decrease in x for all parameter combinations. Thus the hazard function of ARL distribution is always Bathtub shaped. The change point of $h(x)$ will be attained by solving the equation $\gamma(x) = c + k(x) - (k(x))^2 = 0$, which is obtained as

$$x_0 = \alpha\left(e^{\eta} - 1\right), \text{ where } \eta = \frac{1}{2} + \frac{\sqrt{1+4c}}{2}.$$

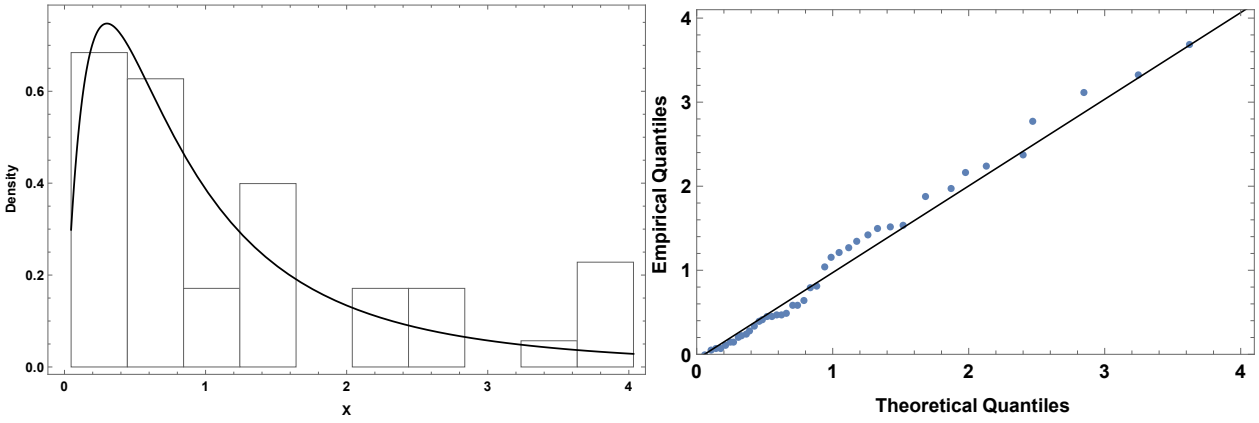
To show the practical importance of the proposed model, we consider a real data reported in Bekker *et al.* (2000), which corresponds to the survival times (in years) of a group of 45 patients given chemotherapy treatment alone. The method of maximum likelihood is employed to estimate the parameters. The estimates obtained are

$$\hat{\alpha} = 0.97003 \quad \text{and} \quad \hat{c} = 2.70067. \tag{36}$$

Recently, Handique and Chakraborty (2016) fitted this data with Beta generalized Kumaraswamy Weibull(BKw-W) distribution and compared with Kumaraswamy Weibull (Kw-W) and Beta generalized Weibull(B-W) distributions. They compared the goodness of fit using the AIC measure. The AIC values of the ARL, BKw-W, Kw-W and B-W models are presented in Table 1.

Table 1: AIC values

Distribution	AIC
ARL	118.831
BKw-W	122.92
Kw-W	123.44
B-W	124.14



(a) Observed histogram and estimated pdf plots **(b) Q-Q plot**

Figure 2

It is evident that the ARL model gives a better fit than the other models concerning the values of AIC. Note that the ARL model contains fewer number of parameters as compared to the competing alternatives. Plot of the fitted density with the histogram of the observed data is given in Figure 2(a). To check the physical closeness of the model, we use the Q-Q plot, which is given in Figure 2(b). We also carry out the Kolmogorov–Smirnov (K–S) goodness of fit test. The K–S test statistic with the associated p-value for the fitted model are 0.093 and 0.80 respectively.

In the context of coherent systems with 'n' identical components, Navarro *et al.* (2013) established that the component survival function $\bar{F}_c(x)$ and the system survival function

$\bar{F}_S(x)$ are connected through the relation

$$\bar{F}_S(x) = q(\bar{F}_c(x)), \tag{37}$$

where $q(u)$ is a distortion function, which is a concave non-decreasing function from $[0,1]$ to $[0,1]$, such that $q(0) = 0$ and $q(1) = 1$.

From (5), the survival function $\bar{T}^*(x)$ satisfies

$$\bar{T}^*(x) = q(\bar{F}(x)), \quad \text{where } q(u) = u\,(1 - \log(u))\ u \in [0,1]. \tag{38}$$

The function $q(u)$ is a concave distortion function. From this, we can infer that $X(X)$ is the distorted random variable obtained from X by the distortion $q(u)$. Distorted random variables have many applications in reliability theory. Navarro *et al.* (2013, 2014) developed various stochastic orders and preservation properties of ageing classes and for the general distorted distributions in the context of coherent systems. For more details on this topic, one could refer to Wang (1996), Sordo and Suarez-Llorens (2011), Sordo *et al.* (2015), and Navarro *et al.* (2016).

Let X and S denotes the lifetimes of the component and system respectively in the context of coherent systems. Then, Navarro *et al.* (2014) showed that If X is NBU (NWU) and $q(u\,v) \leq (\geq)\, q(u)\, q(v)$ for all $0 \leq u,\, v \leq 1$, (submultiplicative (supermultiplicative)) holds then S is NBU (NWU). Similarly, if X is IHRA (DHRA) and $q(u^a) \geq (\leq)\, (q(u))^a$ holds for all $0 \leq u,\, v \leq 1$ and $0 < a < 1$, then S is IHRA (DHRA). Now for the model (4), we have $X(X)$ is the distorted random variable of X, with distortion function $q(u)$ given in (38). We can easily verify that $q(u)$ is submultiplicative and satisfies the condition $q(u^a) \geq (\leq)(q(u))^a$ for all $0 \leq u,\, v \leq 1$ and $0 < a < 1$. Thus, NBU (NWU) and IHRA (DHRA) properties are preserved under auto-relevation transform.

5. Stochastic Orders

There are many situations in practice where we need to compare the characteristics of two distributions. Stochastic orders are used for the comparison of lifetime distributions. In this section, we provide some important stochastic orders between the random variables X and $X(X)$. We shall consider the following stochastic orders. Important properties and interrelations of various stochastic orders can be seen in Shaked and Shanthikumar (2007) and Barlow and Proschan (1975). Suppose $\bar{F}_1(x)$ and $\bar{F}_2(x)$ be the survival functions obtained by distorting $\bar{F}(x)$ using the distortion functions $q_1(u)$ and $q_2(u)$ respectively. Let S_1 and S_2 be the random variables corresponding to $\bar{F}_1(x)$ and $\bar{F}_2(x)$ respectively. Now from Navarro *et al.* (2014) (Theorem 2.5), we have

$$S_1 \leq_{lr} (\geq_{lr})\, S_2 \text{ if and only if } \frac{q_1'(u)}{q_2'(u)} \text{ is increasing (decreasing) in } u \in (0,1), \tag{39}$$

where $q_i'(u)$ is the derivative of $q_i(u)$, $i = 1, 2$. To study different stochastic order relations between X and $X(X)$, we take $S_1 = X(X)$ and $S_2 = X$, with distortion functions $q_1(u) = u\,(1 - \log(u))$ and $q_2(u) = u$ respectively.

Note that,

$$\frac{d}{du}\left(\frac{q_1'(u)}{q_2'(u)}\right) = \frac{d}{du}\left(-\log(u)\right) = -\frac{1}{u} \leq 0.$$

Thus $\frac{q_1'(u)}{q_2'(u)}$ is decreasing in $u \in (0,1)$. Now from (39), we get $X \leq_{lr} X(X)$. Moreover, from Shaked and Shanthikumar (2007), we have the following implications,

$$X \leq_{lr} X(X) \implies X \leq_{hr} X(X) \implies X \leq_{st} X(X).$$

Kochar and Wiens (1987) have defined an IHR order by saying that X is more IHR than Y if $X \underset{c}{\leq} Y$. Further, X is more IHRA (NBU) than Y if $G^{-1}(F(x))$ is star-shaped denoted by $X \leq_* Y$ (super additive denoted by $X \leq_{su} Y$). We have also $X \leq_{DMRL} Y$ if $\frac{m_X(x)}{m_Y(x)}$ is non-decreasing, $X \leq_{NBUE} Y$ if $\frac{m_X(x)}{m_Y(x)} \leq \frac{E(X)}{E(Y)}$, $X \leq_{NBUHR} Y$ if $\frac{h_X(x)}{h_Y(x)} \geq \frac{h_X(0)}{h_Y(0)}$, and $X \leq_{NBUHRA} Y$ if $F_Y^{-1}(F_X(x)) \geq x\left(F_Y^{-1}(F(x)\right)_{x=0}$ (Nair *et al.,* 2013). Among these stochastic orders $X \leq_c Y \implies X \leq_{DMRL} Y \implies X \leq_{NBUE} Y$ and $X \leq_{NBU} Y \implies X \leq_{NBUHRA} Y$. Later Sengupta and Deshpande (1994) proved that $X \underset{c}{\leq} Y$ if and only if $\frac{h_X(x)}{h_Y(x)}$ is non-decreasing in x, provided $h_Y(x) \neq 0$. The following proposition establishes various interrelationships among these orderings.

Proposition 7: Let X be a non-negative random variable and $X(X)$ be the auto-relevation of X with survival function (4). Then $X(X) \underset{c}{\leq} X$.

Proof: From (25), we have

$$\frac{h_{X(X)}(x)}{h_X(x)} = \frac{\log(\bar{F}(x))}{\log(\bar{F}(x)) - 1}.$$

Upon differentiating, we obtain

$$\begin{aligned}\frac{d}{dx}\left(\frac{h_{X(X)}(x)}{h_X(x)}\right) &= \frac{f(x)\,\log(\bar{F}(x))}{\bar{F}(x)(\log(\bar{F}(x)) - 1)^2} - \frac{f(x)\,(\log(\bar{F}(x)) - 1)}{\bar{F}(x)(\log(\bar{F}(x)) - 1)^2} \\ &= \frac{h_X(x)}{\bar{F}(x)(\log(\bar{F}(x)) - 1)^2} \geq 0. \end{aligned} \tag{40}$$

Thus $\frac{h_{X(X)}}{h_X(x)}$ is non-decreasing in x and hence $X(X)$ is more IHR than X. □ The implications, consequence of the Proposition 7, are exhibited in the following diagram;

$$\begin{array}{ccccc} X(X) & \leq_c X & \implies X(X) \leq_* X & \implies X(X) \leq_{su} X \\ \Downarrow & & \Downarrow & & \Downarrow \\ X(X) \leq_{DMRL} X & \implies X(X) \leq_{NBUE} X & \implies X(X) \leq_{NBUHR} X & \implies X(X) \leq_{NBUHRA} X. \end{array}$$

Proposition 8: Let Y_1 and Y_2 be the auto-relevated random variables corresponding to X_1 and X_2 respectively. Then the following results hold;

(i) If $X_1 \leq_{st} X_2$ then $Y_1 \leq_{st} Y_2$.

(ii) If $X_1 \leq_{hr} X_2$ then $Y_1 \leq_{hr} Y_2$.

(iii) If $X_1 \leq_{icx} X_2$ then $Y_1 \leq_{icx} Y_2$.

Proof: The proof for (i) is direct from (4). Now to prove (ii), we have

$$\frac{u\,q'(u)}{q(u)} = \frac{-\log(u)}{1-\log(u)}.$$

Note that $\frac{d}{du}\left(\frac{u\,q'(u)}{q(u)}\right) = -\frac{1}{u(1-\log(u))^2} \leq 0$ for all $u \in (0,1)$. Now from Theorem 2.6 of Navarro *et al.* (2014), we get $Y_1 \leq_{hr} Y_2$. From Theorem 2.6 of Navarro *et al.* (2014), (iii) follows since $q(u)$ is concave in (0,1). □

6. Conclusion

In this paper, we have presented the auto-relevation transform, which is useful in the context of lifetime studies. Various properties and characterizations in terms of reliability measures were presented. Ageing and ordering properties, which will be useful in the reliability context were studied. We also introduced the ARL distribution having non-monotonic hazard function and compared the performance with some existing competing alternatives.

7. Acknowledgement

We thank the referee and the editor for their constructive comments. The first author is thankful to University of Calicut for the financial support.

References

Barlow, R. E. and Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing*. Holt, Rinehart and Winston, New York.

Baxter, L. A. (1982). Reliability applications of the relevation transform. *Naval Research Logistics (NRL)*, **29(2)**, 323–330.

Bekker, A., Roux, J. J. J. and Mosteit, P. J. (2000). A generalization of the compound rayleigh distribution: using a bayesian method on cancer survival times. *Communications in Statistics-Theory and Methods*, **29(7)**:1419–1433.

Chukova, S., Dimitrov, B. and Khalil, Z. (1993). A characterization of probability distributions similar to the exponential. *Canadian Journal of Statistics*, **21(3)**, 269–276.

Glaser, R. E. (1980). Bathtub and related failure rate characterizations. *Journal of the American Statistical Association*, **75(371)**, 667–672.

Grosswald, E., Kotz, S. and Johnson, N. L. (1980). Characterizations of the exponential distribution by relevation-type equations. *Journal of Applied Probability*, **17(3)**, 874–877.

Handique, L. and Chakraborty, S. (2016). Beta generated kumaraswamy-g and other new families of distributions. *ArXiv Preprint*, ArXiv:1603.00634.

Johnson, N. L. and Kotz, S. (1981). Dependent relevations: time-to-failure under dependence. *American Journal of Mathematical and Management Sciences*, **1(2)**, 155–165.

Kapodistria, S. and Psarrakos, G. (2012). Some extensions of the residual lifetime and its connection to the cumulative residual entropy. *Probability in the Engineering and Informational Sciences*, **26(1)**, 129–146.

Klüppelberg, C. (1988). Subexponential distributions and integrated tails. *Journal of Applied Probability*, **25(1)**, 132–141.

Kochar, S. C. and Wiens, D. P. (1987). Partial orderings of life distributions with respect to their aging properties. *Naval Research Logistics*, **34(6)**, 823–829.

Krakowski, M. (1973). The relevation transform and a generalization of the gamma distribution function. *Revue fran¸caise d'automatique, informatique, recherche op´erationnelle. Recherche op´erationnelle*, **7(2)**, 107–120.

Lau, K. S. and Rao, B. P. (1990). Characterization of the exponential distribution by the relevation transform. *Journal of Applied Probability*, **27(3)**, 726–729.

Nair, N. U., and Sankaran, P. G. and Balakrishnan, N. (2013). *Quantile-Based Reliability Analysis*. Springer, Birkhauser, New York.

Navarro, J., del Águila, Y., Sordo, M. A. and Suárez-Llorens, A. (2013). Stochastic ordering properties for systems with dependent identically distributed components. *Applied Stochastic Models in Business and Industry*, **29(3)**, 264–278.

Navarro, J., del Aguila, Y., Sordo, M. A. and Suárez-Llorens, A. (2014). Preservation of reliability classes under the formation of coherent systems. *Applied Stochastic Models in Business and Industry*, **30(4)**, 444–454.

Navarro, J., Del Águila, Y., and Sordo, M. A. and Suárez-Llorens, A. (2016). Preservation of stochastic orders under the formation of generalized distorted distributions. Applications to coherent systems. *Methodology and Computing in Applied Probability*, **18(2)**, 529–545.

Sankaran, P. G. and Dileepkumar, M. (2019). Reliability properties of proportional hazards relevation transform. *Metrika*, **82(4)**, 441–456.

Sengupta, D. and Deshpande, J. V. (1994). Some results on the relative ageing of two life distributions. *Journal of Applied Probability*, **31(4)**, 991–1003.

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic Orders*. Springer Science & Business Media.

Shanthikumar, J. G. and Baxter, L. A. (1985). Closure properties of the relevation transform. *Naval Research Logistics (NRL)*, **32(1)**, 185–189.

Sordo, M. A. and Su´arez-Llorens, A. (2011). Stochastic comparisons of distorted variability measures. *Insurance: Mathematics and Economics*, **49(1)**, 11–17.

Sordo, M. A., Sua´rez-Llorens, A. and Alfonso, J. B. (2015). Comparison of conditional distributions in portfolios of dependent risks. *Insurance: Mathematics and Economics*, **61**, 62–69.

Wang, S. (1996). Premium calculation by transforming the layer premium density. *ASTIN Bulletin: The Journal of the IAA*, **26(1)**, 71–92.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 265–277

Modeling and Analysis of Competing Risks Cure Rate Regression Model with Weibull Distribution

P. G. Sankaran[1] **and P. P. Rejani**[1,2]
[1]*Department of Statistics, Cochin University of Science and Technology, Kerala, India*
[2]*Department of Community Medicine, Govt.Medical College, Kerala, India*

Received: 04 September 2021; Revised: 08 December 2021; Accepted: 30 December 2021

Abstract

Cure rate models have been widely applied in the analysis of lifetime data in the presence of cured fractions. Regression models need more attention when investigators are interested to study the effects of given treatments. The presence of competing risks is an additional challenge for researchers to analyze lifetime data with cured proportion. In this paper, we propose a parametric cure rate regression model incorporating competing risks for the analysis of survival data. The parameters of the model are estimated by the maximum likelihood estimation procedure via EM algorithm. A simulation study is carried out to evaluate the performance of the proposed model. The practical relevance of the model is illustrated by applying the model to a dataset on heart transplantation.

Key words: Cure rate model; Competing risks; Maximum likelihood; EM algorithm; Weibull distribution

AMS Subject Classifications: 62K05, 05B05

1. Introduction

The recent advancements in diagnostic and other drug design experiments resulted increased rate of favorable response of patients to their received treatments and a good proportion of patients have become free from diseases. These disease-free individuals in a set of survival data are said to be immunes and the proportion of immunes that exists in the data is called cured proportion. The presence of immunes in survival data influences the outcome measures in survival studies. While analysing such data, it can be seen that the survival curve does not taper off to zero at the end of the study period. Hence ordinary survival analysis techniques are not suitable to analyse such data and new models have been developed incorporating cured proportions. Such models are said to be cure rate models in survival analysis. Boag (1949) first proposed cure rate model to estimate the cured proportion of breast cancer patients. Cure rate models have been extended its applicability in several areas like financial, criminology, demography, and industrial reliability. Nelson (1982) explained the life expectancy of electric motors with cure rate model. Yamaguchi (1992) applied the cure rate model to describe inter-firm job mobility in Japan. For further reading one can refer to Maller and Zhou (1996), Sy and Tailor (2000), Ortega *et al.* (2014), Shen *et al.* (2019), and Sreedevi and Sankaran (2021).

Corresponding Author: P.P. Rejani
Email: rejstat@gmail.com

Competing risks occur when the study subjects experience more than one events that compete the event of interest. For example, when a researcher observing peritoneal dialysis patients until they develop peritonitis, kidney transplantation can be regarded as a competing cause because the chance of occurrence of peritonitis is very less among patients who have undergone kidney transplantation. The competing risks aspect seeks more attention in the analysis and interpretation of survival data. It is known that age-related mortality is high among older people than others. Also, the probability of death due to the disease is found to be low in clinical trials with the desired effect. In both cases, deaths occur due to other competing causes rather than the event of interest. Hence failure to consider competing risks in the analysis of such data yields reporting of inaccurate and misleading results. The competing risks models are discussed by many authors. Crowder (2001) and Kalbfleisch and Prentice (2011) are prime among them. Wright *et al.* (2020) and Papastefanou *et al.* (2021) are two recent works that draw out the significance of competing risks in the medical field.

In survival studies carry out in the field of medicine and epidemiology, the investigators focus on determining the effect of factors associated with the time to occurrence of the event such as death or disease recurrence. Regression models such as Cox proportional hazards models or parametric models are usually used to study the effect of covariates present in the data. The presence of competing risks, immune proportions and covariates altogether enhance the complexity of data and burden of analysis. All of these prominent scenarios are encountered by formulating competing risks cure rate regression models. Development of such models needs special attention and less available in literature.

In cure rate models, parametric or semiparametric proportional hazards assumptions can be made for lifetime distribution in latency. In recent times, some semiparametric models are proposed for the analysis of competing risks data in the presence of cured proportions. The interested readers can refer to Choi *et al.* (2018) and Rejani and Sankaran (2020). If a particular probability distribution of survival data can be identified and validated, statistical inference based on a parametric regression perceptive will be considered as more efficient and precise than those derived from survival models in the absence of an explicit distributional function (Collett, 2015). Yusuf *et al.* (2016) discussed Weibull distribution as a suitable distribution for the analysis of data in the presence of cured proportion.

In this paper, we introduce a parametric cure rate regression model based on Weibull distribution for the analysis of survival data in competing risks setting. The model and methods focus on the estimation of regression parameters and the probability of cure in the presence of competing risks. The innovative feature of the proposed model is the proficiency to explain the impact of covariates on the survival time of a group of subjects in the presence of immunes and at the same time, the influence of competing causes is also taken into account.

Heart transplantation is the gold standard for the treatment of end-stage heart failure. Rejection and infection are the two major causes of mortality among patients undergoing heart transplantation. Larson and Dince (1985) considered 65 transplant recipient data, there were 29 (45%) rejection deaths, 12 (18%) deaths from other causes, and 24 (37%) censored observations. They analyzed data by mixture model approach without considering the chance of occurrence of cured proportion. A cure rate regression model separates short and long-term survival of patients. It is useful to determine the proportion of cured patients and to identify the associated factors on survival of patients under study. It helps the public

health professionals in decision making. In this context, we use data on heart transplantation in Section 5 for an illustration of our proposed model.

The rest of the paper is structured as follows. We introduce the parametric competing risks cure rate regression model in Section 2. The likelihood function formulation and estimation procedures are explained in Section 3. In Section 4, we report the results of simulation work to explain the bias of estimators on variations in samples size. Section 5 illustrates the application of the proposed model to real data set. Some concluding remarks are given in Section 6.

2. The Model

Suppose that population consists of two groups of subjects say, susceptibles and immunes. Let T be the time to occurrence of the event. Define the indicator variable function to define the status of cure

$$Y = \begin{cases} 1, & \text{if the individual eventually experience the event of interest} \\ 0, & \text{otherwise.} \end{cases}$$

Let p be the probability of occurrence of the event. The survival function of the uncured population at time t is $S(t|Y=1) = P(T > t|Y=1)$. Then survival function of cure rate model is

$$S(t) = (1-p) + pS(t|Y=1) \tag{1}$$

where $t < \infty$. Note that $S(t)$ tends to $(1-p)$ as $t \to \infty$. Let C = cause of death and the probability of uncured subjects $p_j = Pr(Y=1, C=j)$, $j = 1, 2, \ldots, k$. Assume that the time to occurrence of the event T is defined only when $Y=1$ and $C=j$, $j=1,2,\ldots,k$. Let $f_j(t|Y=1)$ be the probability density function and $S_j(t|Y=1)$ be the sub-survival function (Carriere and Kochar (2000)),of the random variable t due to jth cause, $j=1,2,\ldots,k$. For a censored individual, Y is not observed.

In the presence of competing risks, the survival function of cure rate model is

$$S(t) = 1 - \sum_{j=1}^{k} p_j + \sum_{j=1}^{k} p_j S_j(t|Y=1) \tag{2}$$

Let X be a $p+1 \times 1$ vector of covariates at incidence part and Z be a $p \times 1$ covariate vector at latency part of the model that is independent of X. In practical situations, the covariates X and Z can be same or may share common elements between them. Let $b_j = (b_{0j}, b_{1j}, \ldots, b_{pj})'$ be a vector of regression coefficients with $b = (b_1, b_2, \ldots, b_k)'$ for $j = 1, 2, \ldots, k$.

Then, in a competing risks Weibull regression model, the sub-survival function of t due to jth cause of failure with probability density function

$$f_j(t|Y=1,\theta,Z) = \alpha \exp(\beta_j Z)\, t^{\alpha-1} \exp(-t^{\alpha} \exp(\beta_j Z)) \tag{3}$$

is

$$S_j(t|Y=1,\theta,Z) = \exp(-t^{\alpha} \exp(\beta_j Z)) \tag{4}$$

where $\alpha > 0$, $\beta_j = (\beta_{j0}, \beta_{j1}, ...\beta_{jp})'$ is the vector of regression coefficients associated with the covariate Z, $\theta = (\alpha, \beta_j)$ and $\beta = (\beta_1, \beta_2, ...\beta_k)'$ for $j = 1, 2, \ldots, k$.

Under logistic regression model assumption, the probability of occurrence of the event due to jth cause is

$$p_j(b) = Pr(Y = 1, X) = \frac{\exp(b_j' X)}{1 + \sum_{j=1}^{k} \exp(b_j' X)} \tag{5}$$

for $j = 1, 2, \ldots, k$

Let $F_j(t) = \Pr(T \leq t, C = j)$ be the cumulative incidence function due to jth cause which measures the probability of occurrence of the event before time t due to cause j, $j = 1, 2, \ldots, k$.

Now, the cumulative incidence function due to jth cause in the presence of covariates X and Z and in the presence of $Y = 1$ is

$$\begin{aligned} \Pr(T \leq t, C = j | X, Z, Y = 1) &= \Pr(T \leq t | Z, Y = 1, C = j) \Pr(C = j, Y = 1 | X) \\ &= p_j(b)(1 - S_j(t | Y = 1, \theta, Z)) \end{aligned}$$

Now, the survival function of competing risks cure rate regression model is defined as

$$S(t, \Theta) = p_0(b) + \sum_{j=1}^{k} p_j(b) S_j(t | Y = 1, \theta, Z) \tag{6}$$

where $\Theta = (b, \theta)$ denotes the entire set of parameters and $p_0(b) = 1 - \sum_{j=1}^{k} p_j(b)$, the probability of immunes in the model. Suppose that the model parameters are linked to a single covariate Z. (ie., we use the assumption $X = Z$ throughout the paper). We also assume that an independent, non-informative, random censoring model and the censoring variable is statistically independent of Y. Inference procedure of the proposed model is given in the next Section .

3. Inference Procedures

Suppose we have data in the form $(t_{ij}, \delta_{ij}, z_i)$ for $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$ and $i \neq j$ where the notations

t_{ij} = the observed event or censoring time due to jth cause and the n distinct event times be $t_{1j} < t_{2j} < \cdots < t_{nj}$.

$$\delta_{ij} = \begin{cases} 1, & t_{ij} \text{ is uncensored} \\ 0, & \text{otherwise.} \end{cases}$$

and z_i = a vector of covariates.

The likelihood equation under multiple modes of failures is

$$L = \prod_{i=1}^{n} \prod_{j=1}^{k} (f_j(t_i))^{\delta_{ij}} (S(t_i))^{1-\delta_{ij}} \tag{7}$$

Under the model assumptions made, likelihood function of the cure rate regression model is

$$L(\Theta) = \prod_{i=1}^{n} \prod_{j=1}^{k} (p_j(b) f_j(t_i|Y=1,\theta,Z))^{\delta_{ij}} \left(p_0(b) + \sum_{j=1}^{k} p_j(b) S_j(t_i|Y=1,\theta,Z) \right)^{1-\delta_{ij}} \tag{8}$$

Let the complete data be $(t_{ij}, \delta_{ij}, z_i, y_{ij})$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$ which includes the observed data and the unobserved y_{ij}'s, where y_{ij} be the value taken by the random variable Y_i for jth cause. If $\delta_{ij} = 1$, $y_{ij} = 1$ and if $\delta_{ij} = 0$, y_{ij} is unobserved. Then the complete - data full likelihood is

$$\begin{aligned} L_c(\Theta) = \prod_{i=1}^{n} \prod_{j=1}^{k} (p_j(b) f_j(t_i|Y=1,\theta,Z))^{\delta_{ij} y_{ij}} (p_0(b))^{(1-\delta_{ij})(1-\sum_{j=1}^{k} y_{ij})} \\ (p_j(b) S_j(t_i|Y=1,\theta,Z))^{(1-\delta_{ij}) y_{ij}} \end{aligned} \tag{9}$$

By substituting the probability density function and the survival function given in (3) and (4), the above likelihood equation can be expressed as a product of two likelihood functions as

$$L_c(\Theta) = L_1(b) L_2(\beta, \theta) \tag{10}$$

where

$$L_1(b) = \prod_{i=1}^{n} \prod_{j=1}^{k} (p_j(b))^{y_{ij}} (p_0(b))^{(1-\delta_{ij})(1-\sum_{j=1}^{k} y_{ij})}$$

and

$$L_2(\beta, \theta) = \prod_{i=1}^{n} \prod_{j=1}^{k} \left(e^{\beta_j z_i} \alpha {t_i}^{\alpha-1} \right)^{\delta_{ij} y_{ij}} e^{(-{t_i}^{\alpha} \exp(\beta_j z_i) y_{ij})}$$

The likelihood function (10) contains missing observations since partial information of random variable Y is missing. Hence we employ EM Algorithm (Dempster *et al.* (1977)) to estimate the parameters of the model.

3.1. EM algorithm

E-Step : The expectation step (E-step) in the EM algorithm compute the conditional expectation of the complete data log-likelihood function $l(\Theta; y)$ with respect to y_{ij}'s, given the observed data and current estimates of the parameters.

Let the observed data be $\{O = (\text{Observed } y_{ij}\text{'s, } t_{ij}, \delta_{ij}, z_i);\ i = 1, ..., n\}$. Now we have to compute $\pi_j^{(m)} = E(y_{ij}|\Theta^{(m)}, O)$ where $\Theta^{(m)}$ denotes the values of parameters Θ at the mth iteration step. For uncensored i, $E(y_{ij}|\Theta^{(m)}, O) = y_{ij} = 1$. Now for the i'th censored observation, we compute

$$\pi_j^{(m)} = \Pr(Y_i = 1, |T_{ij} > t_{ij}, \delta_{ij} = 0, z_i; \Theta^{(m)})$$

$$= \left[\frac{p_j(b) S_j(t_i|Y=1,\theta,z_i)}{p_0(b) + \sum_{j=1}^{k} p_j(b) S_j(t_i|Y=1,\theta,z_i)} \right]_{|\Theta^{(m)}|}$$

i.e., at the m th iteration, the E-step value of y_{ij} is

$$w_{ij}^{(m)} = \begin{cases} 1, & \text{if the } i^{th} \text{ individual is uncensored} \\ \pi_{ij}{}^{(m)}, & \text{if censored} \end{cases} \tag{11}$$

$$l_c(\Theta; w^{(m)}) = l_1(b; w^{(m)}) + l_2(\theta; w^{(m)}) \tag{12}$$

denote the conditional expectation of the complete data log-likelihood function, where $w^{(m)}$ denote the vector of $w_{ij}^{(m)}$ values.

M-Step : In M-Step, we maximise the conditional expectation of the complete data log-likelihood function $l_c(\Theta; w^{(m)})$ with respect to each parameter in $\Theta = (b, \theta)$ given w_{ij} to obtain an improved estimate $\Theta^{(m+1)}$ at the $(m+1)$th iteration.

The procedures in E-step and M-step are then continued iteratively until we meet the convergence criteria to obtain maximum likelihood estimators of each parameter in the parameter set $\Theta = (b, \theta)$.

3.2. Asymptotic property of estimators

Let $\hat{\Theta} = (\hat{b}, \hat{\theta})$ denote the maximum likelihood estimates of $\Theta = (b, \theta)$, where $\hat{b} = \hat{b}_j$ and $\hat{\theta} = (\hat{\alpha}, \hat{\beta}_j)$, $j = 1, 2, \ldots, k$. Now consider the following regularity conditions.

(a) The first and second order derivatives of the log-likelihood function l with respect to Θ viz., $\frac{\partial l}{\partial \Theta}$ and $\frac{\partial^2 l}{\partial \Theta^2}$ exist and are continuous functions of Θ in a range R (including the true value Θ_0 of the parameter) for almost all t. For every Θ in R

$\left|\frac{\partial l}{\partial \Theta}\right| < H_1(t)$ and $\left|\frac{\partial^2 l}{\partial \Theta^2}\right| < H_2(t)$ where $H_1(t)$ and $H_2(t)$ are integrable functions over $(-\infty, \infty)$.

(b) The third order derivative with respect to Θ, $\frac{\partial^3 l}{\partial \Theta^3}$ exists such that $\left|\frac{\partial^3 l}{\partial \Theta^3}\right| < M(t)$ where $E[M(t)] < Q$, a positive quantity.

(c) For every Θ in R,

$$E\left(-\frac{\partial^2 l}{\partial \Theta^2}\right) = \int_{-\infty}^{\infty} \left(-\frac{\partial^2 l}{\partial \Theta^2}\right) L dt = I(\Theta)$$

is finite and non-zero.

(d) The range of integration is independent of Θ. This assumption is to make differentiation under the integral sign valid.

Under the above mentioned regularity conditions, as $n \to \infty$,

$\sqrt{n}(\Theta - \hat{\Theta}) \to N_8(0, I^{-1}(\Theta))$, where the Fisher information matrix $I(\Theta)$ can be replaced by a consistent estimate $I(\hat{\Theta}) = \left(\frac{-\partial^2 l}{\partial \Theta_i \partial \Theta_j}\right)_{\Theta=\hat{\Theta}}$. The observed information matrix is obtained by applying Louis (1982) method. The variance of the estimates can be determined from diognal elements of $I^{-1}(\hat{\Theta})$. The asymptotic normality property of maximum likelihood estimates is useful to determine the $(1-\alpha) \times 100\%$ confidence interval of each parameter in the parametric set $\Theta = (b, \theta)$. Let $\hat{b}_j$ is the maximum likelihood estimator (MLE) of b_j. Then MLE of cured proportion $1 - p_j$ is $1 - \hat{p}_j = g(\hat{b}_j)$ is also asymptotically normally distributed by the invariance property of maximum likelihood estimators.

4. Simulation Studies

Simulation studies are conducted to evaluate the performance of the proposed model. Let C be the cause of failure and we assumed that there are two causes of failure. We consider a single covariate Z, which is generated from a uniform distribution over the interval (0,1). The censoring variable K is generated from uniform distribution over the interval (0,k) where k chosen in such a way that the lifetimes are mildly or heavily censored. The observations are followed up to a maximum time $\tau = 10$. The data for each observation be (t, δ, Z, C), where $t = \min(T, K, \tau)$ and δ be the event indicator. The data generated from the model with incidence probabilities

$$p_j(b) = \frac{\exp(b_{0j} + b_{1j} Z)}{1 + \sum_{j=1}^{2} \exp(b_{0j} + b_{1j} Z)} \tag{13}$$

for jth cause of failure, $j = 1, 2$. The cause specific survival functions are generated at random using the following sub distribution functions suggested by Dewan and Kulathinal (2007). Let,

$$\begin{aligned} F_1(t) &= P(T \le t, C = 1) = \phi F^a(t) \\ F_2(t) &= P(T \le t, C = 2) = F(t) - \phi F^a(t) \end{aligned} \tag{14}$$

where $1 \leq a \leq 2$, $0 \leq \phi \leq 0.5$ and $F(t)$ is the distribution function at time T. Note that $\phi = P(C = 1)$ and for $a = 1$, T and C are independent. The variables T and C are dependent for other choices of a. The nonnegative condition of cause specific density function of T is maintained by imposing these restrictions on the parameters. We choose the values $\phi = 0.25$ and $a = 1.5$ for simulating data. We fix the initial values $b_{0j} = 2$, $b_{1j} = -1$, $\beta_{0j} = 1.5, j = 1, 2$, $\beta_1 = -0.3$, $\beta_2 = 0.2$ and $\alpha = 0.2$. The initial values of the estimates are chosen using Kaplan-Meier estimate of cured proportion and log-likelihood equation of proposed the model (Balakrishnan and Pal (2012)). We generated random samples of sizes $n = 50$, 100 and 200 and maximum likelihood estimation of the parameters is carried out for the proposed model. The effect of censoring was studied in two situations viz, mild censoring (on an average, 20% of the observations are censored) and heavy censoring (40% of the observations are censored at average level). For the described configuration, 1000 replications are made. The results of absolute bias and MSE of the estimates are reported. The coverage probabilities (CP) of the 95% confidence intervals based on the asymptotic normality of the estimators are also reported. Table 1 shows the average absolute bias and MSE of estimates at different censoring levels. It seems that the proposed model and method work well. The parameters of the model are estimated with lower bias and MSE. There is a slight increase in bias and MSE as the censoring scheme changes from mild to heavy. The coverage probabilities of the asymptotic confidence intervals are also close to the pre-determined levels and it is found to be better for samples of increased size.

Table 1: Absolute Bias and MSE of estimators of parameters

			20% Censored			40% Censored		
Sample size	Parameter	True value	Bias	MSE	CP	Bias	MSE	CP
50	b_{01}	2.0	0.05822	0.013144	95.27273	0.090346	0.016547	95.03546
	b_{11}	-1.0	0.04495	0.01547	95.43568	0.05786	0.003348	95.19231
	b_{02}	2.0	0.07299	0.010114	95.00000	0.07698	0.013841	94.35028
	b_{12}	-1.0	0.05181	0.013698	95.50562	0.05601	0.014436	95.00000
	β_{01}	1.5	0.09006	0.027625	95.23810	0.099391	0.028006	94.83568
	β_{02}	1.5	0.09772	0.211936	95.70896	0.12413	0.242311	95.06173
	β_1	-0.3	0.09006	0.027625	95.23810	0.099391	0.028006	94.83568
	β_2	0.2	0.09772	0.211936	95.70896	0.20013	0.242311	95.06173
	α	0.2	0.00756	0.000681	95.84463	0.00979	0.001025	95.3125
100	b_{01}	2.0	0.02504	0.00976	96.29630	0.05485	0.01495	95.29220
	b_{11}	-1.0	0.04387	0.00360	97.72727	0.04486	0.00360	96.96970
	b_{02}	2.0	0.03678	0.00850	95.13880	0.04312	0.00931	96.23552
	b_{12}	-1.0	0.04452	0.00476	95.74468	0.04532	0.008003	95.55556
	β_{01}	1.5	0.07145	0.027625	95.23810	0.08236	0.028006	94.83568
	β_{02}	-1.5	0.08320	0.211936	95.70896	0.09008	0.242311	95.06173
	β_1	-0.3	0.05586	0.02350	95.58854	0.07920	0.01083	95.45455
	β_2	0.2	0.07491	0.06091	96.31902	0.08921	0.08549	96.31512
	α	0.2	0.00515	0.00049	95.94229	0.00594	0.00034	95.83333
200	b_{01}	2.0	0.02044	0.00112	98.53000	0.02144	0.01180	97.59450
	b_{11}	-1.0	0.01842	0.00335	98.54369	0.02253	0.00356	97.80220
	b_{02}	2.0	0.02052	0.00277	96.90000	0.02385	0.00622	95.45455
	b_{12}	-1.0	0.02765	0.00308	96.50000	0.03839	0.00479	96.35036
	β_{01}	1.5	0.04431	0.027625	95.23810	0.05319	0.028006	94.83568
	β_{02}	1.5	0.04749	0.211936	95.70896	0.05283	0.242311	95.06173
	β_1	-0.3	0.04488	0.01006	97.16981	0.05049	0.01329	95.50000
	β_2	0.2	0.06634	0.03085	97.29730	0.07233	0.05392	96.22302
	α	0.2	0.00254	0.00025	97.73960	0.00437	0.00023	96.51452

5. Data Analysis

To illustrate the applicability of the proposed model, we consider the data set from the Stanford Heart Transplant Program. The data contains the details of 103 patients selected for cardiac transplantation. A detailed description of data is available in Crowley and Hu (1977). We consider a subset of this data set with 63 patients who received the transplant to explain the application of the model. Out of these 63 transplant recipients, there were 27 (43%) deaths that occurred that due to rejection, 12 (19%) deaths from other causes and, the remaining 24 (38%) were censored observations. Survival time was measured in days from the date of transplant surgery. There are nine covariates in the original data set. We select only one covariate, the mismatch score, a key factor that influences survival of patients after heart transplantation (Miller (1976), Opelz, G. and Wujciak, T. (1994), Osorio-Jaramillo *et al.* (2020)) for the analysis of data. The mismatch score measures the degree of dissimilarity between the donor and recipient tissue concerning HLA antigens, and it is therefore related to the phenomenon of rejection of the donor heart by the recipient's immune mechanisms. If the mismatch score is less than one, it is a sign of good match, and if the score is high, greater than one represents a poor match (Miller (1976)). Hence we transform the selected continuous covariate mismatch score into a categorical variable of two categories with cut-off value one as per aforesaid classification criteria of matching and considered for the analysis of data. There are two causes of failure in the data. The cause of death attributable to rejection of the donor heart is labeled as cause 1 and cause of death due to other reasons such as surgical, kidney failure, hepatitis, etc, and not due to rejection of the new heart is labeled as cause 2.

As an initial step of the analysis, Kaplan- Meier plot is drawn for the data and displayed in Figure 1. The plateau in the given survival curve confirms the presence of immunes in the data. Hence the selected data is suitable for the analysis of cure rate models.

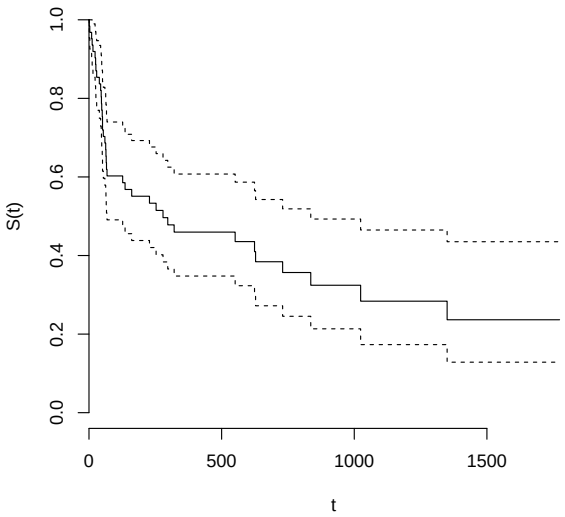


Figure 1: Kaplan-Meier survival curve of heart transplant data

In the present work, we are interested to study the effect of the covariate mismatch score on the survival of patients who have undergone heart transplantation in competing risks setting. The maximum likelihood estimators of regression coefficients are found out using (12) under the given model assumptions. The statistical significance of the regression coefficients is tested by the likelihood ratio test procedure. The estimates of regression coefficients with corresponding standard errors are reported in Table 2. The result shows that the higher mismatch score has a significant effect on rejection-related mortality among patients after heart transplant ($p = 0.013$) but may not affect the survival of patients ($p = 0.137$). The role of mismatch score is negligible on rejection related mortality of patients who died of competing causes.

Figure 2 displays plots of the estimated cumulative incidence rates for mismatch categories. From the Figure, it is obvious that the difference between cause specific failure rates is more in high score (> 1) category of mismatch score compared to low score (< 1) category. This is due to the variations in the influence of mismatch score on mortality of patients due to two causes of failure.

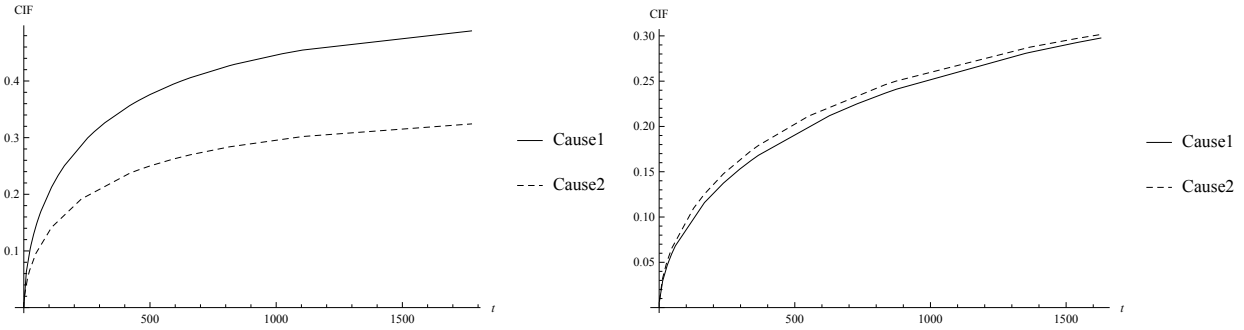


Figure 2: Cumulative incidence curve of mismatch score categories high score (left) and low score (right)

The estimated cured proportion among low score category (17.99%) is greater than that of high score category (12.41%). It bring out the influence of the selected covariate mismatch score on the survival of study subjects. The estimates and 95% Confidence interval of the probability of cure due to rejection and due to other causes obtained from the model are 0.47 (0.34, 0.61) and 0.65 (0.50, 0.81) respectively. The estimated values of cured proportion reveals the presence of cured individuals in the data and confirm the importance of the proposed model. The goodness of fit of latency part of the model is tested using Cox-Snell residuals with the modifications suggested by Peng and Tailor (2017). We consider the Cox-Snell residuals $r_i = -\log S_j(t|Y = 1, \theta, Z)$ using (4). The residuals for each cause of failure estimated with different weights as given in (11) for the censored and uncensored observations. The Kolmogorov - Smirnov test is performed to assess the unit exponentiality of the data and p values obtained as $p = 0.25$ for rejection and $p = 0.12$ for other cause of failure. The values indicate that the model fits well for the given data to explain each cause of failure.

Table 2: Estimates of parameters and Standard Error (SE)

	Rejection ($j=1$)			Other Causes ($j=2$)		
Estimates	Est	SE	P value	Est	SE	P value
b_{0j}	0.861	2.53×10^{-3}	-	0.784	2.59×10^{-3}	-
b_{1j}	0.585	5.14×10^{-3}	0.013	0.251	5.46×10^{-3}	0.654
β_{0j}	-4.135	3.93×10^{-3}	-	-3.950	3.87×10^{-3}	-
β_{1j}	0.730	2.11×10^{-3}	0.137	0.555	2.48×10^{-3}	0.002

6. Conclusion

In this paper, we proposed a regression model with Weibull distribution for the analysis of competing risks data with long term survivors. Maximum likelihood inference via EM algorithm was implemented to estimate the parameters of the model. The goodness of fit of the latency model checked using modified Cox-Snell residuals. The model was illustrated with a real lifetime data on Stanford Heart Transplant Program and distinguished the effect of covariate on short and long term survival of patients after heart transplant in competing risks scenario. This article aimed to evaluate the effect of covariates such as clinico-social variables, different treatment regimens and other prognostic factors on survival of patients suffering from diseases when there is a chance of cure in the presence of competing risks and expected to be useful for investigators in the field of survival analysis. The regression analysis of interval censored data with cured proportion is also challenging in the field of survival analysis. The work in this direction is under progress and it will be communicated in a future paper.

Acknowledgements

The authors are thankful to the referee and editor for the constructive comments and suggestions on earlier version of this manuscript that appreciably improved the article.

References

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B (Methodological)*, **11(1)**, 15-53.

Balakrishnan, N. and Pal, S. (2012). EM algorithm-based likelihood estimation for some cure rate models. *Journal of Statistical Theory and Practice*, **6(4)**, 698-724.

Carriere, K. C. and Kochar, S. C. (2000). Comparing sub-survival functions in a competing risks model. *Lifetime Data Analysis*, **6(1)**, 85-97.

Choi, S., Zhu, L. and Huang, X. (2018). Semiparametric accelerated failure time cure rate mixture models with competing risks. *Statistics in Medicine*, **37(1)**, 48-59.

Collett, D. (2015). *Modelling Survival Data in Medical Research*. CRC Press, London.

Crowder, M. J. (2001). *Classical Competing Risks*. CRC Press, London.

Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, **72(357)**, 27-36.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, **39(1)**, 1-22.

Dewan, I. and Kulathinal, S. (2007). On testing dependence between time to failure and cause of failure when causes of failure are missing, *PLoS One*, **2(12)**, e1255.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*. **360**, John Wiley & Sons, New York.

Larson, M. and Dinse, G. (1985). A mixture model for the regression analysis of competing risks data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **34(3)**, 201-211.

Maller, R. A. and Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. John Wiley& Sons, New York.

Miller, R. G. (1976). Least squares regression with censored data. *Biometrika*, **63(3)**, 449-464.

Nelson, W. B. (2003). *Applied Life Data Analysis*. John Wiley& Sons, New York.

Opelz, G. and Wujciak, T. (1994). The influence of HLA compatibility on graft survival after heart transplantation. *New England Journal of Medicine*, **330(12)**, 816-819.

Ortega, E. M., Cordeiro, G. M. and Hashimoto, E. M. (2011). A log-linear regression model for the Beta-Weibull distribution. *Communications in Statistics: Simulation and Computation*, **40(8)**, 1206-1235.

Osorio-Jaramillo, E., Haasnoot, G. W., Kaider, A., Schaefer, A. K., Haberl, T., Goekler, J.,et al. (2020). Molecular-level HLA mismatch is associated with rejection and worsened graft survival in heart transplant recipients–a retrospective study. *Transplant International*, **33(9)**, 1078-1088.

Papastefanou, I., Nowacka, U., Syngelaki, A., Dragoi, V., Karamanis, G., Wright, D. and Nicolaides, K. H. (2021). Competing-risks model for prediction of small-for-gestational-age neonate from estimated fetal weight at 19–24 weeks' gestation. *Ultrasound in Obstetrics & Gynecology*, **57(6)**, 917-924.

Peng, Y. and Taylor, J. M. (2017). Residual-based model diagnosis methods for mixture cure models. *Biometrics*, **73(2)**, 495-505.

Rejani, P. P. and Sankaran, P. G. (2020). Modeling and Analysis of Proportional Hazards Competing Risks Cure Rate Model. *Journal of the Indian Society for Probability and Statistics*, **21(1)**, 175-185.

Shen, P. S., Chen, H. J., Pan, W. H. and Chen, C. M. (2019). Semiparametric regression analysis for left-truncated and interval-censored data without or with a cure fraction. *Computational Statistics & Data Analysis*, **140**, 74-87.

Sreedevi, E. P. and Sankaran, P. G. (2021). Statistical methods for estimating cure fraction of COVID-19 patients in India. *Model Assisted Statistics and Applications*, **16(1)**, 59-64.

Sy, J. P. and Taylor, J. M. (2000). Estimation in a Cox proportional hazards cure model, *Biometrics*, **56(1)**, 227-236.

Wright, D., Wright, A. and Nicolaides, K. H. (2020). The competing risk approach for prediction of preeclampsia. *American Journal of Obstetrics and Gynecology*, **223(1)**, 12-23.

Yamaguchi, K. (1992). Accelerated failure-time regression models with a regression model of surviving fraction:an application to the analysis of 'permanent employment' in Japan. *Journal of the American Statistical Association*, **87(418)**, 284-292.

Yusuf, M. U. and Bakar, M. R. A. (2016). Cure models based on Weibull distribution with and without covariates using right censored data. *Indian Journal of Science and Technology*, **9(28)**, 1-12.

Algorithm for maximum likelihood estimation of parameters of the model

1. Determine the parameter values b_j, β_j, and α for $j = 1, 2$; (Select the initial values of the parameters and input these values in first stage).
2. For the ith subject, generate the covariate X_i from Uniform(0,1);
3. Find out the probability of incidence $p_j(b)$ for $\forall$ X_i and $j = 1, 2$;
4. Generate censoring variable K_i from Uniform(0,k), where k is set to control the proportion of censored observations;
5. Generate a random variable u_i from Uniform(0,1);
6. Take v_i as the root of $F(t) - u_i = 0$, where $F(t)$ is the distribution function corresponding to the model;
7. Find $t_i = \min(v_i, K_i, \tau)$, τ=10 (assumed). If $t_i < K_i$, set $\delta_i = 1$, otherwise $\delta_i = 0$;
8. Find out survival functions $S_j(t)$ for $j = 1, 2$ and $S(t)$;
9. Find out $\psi_i = 1 - \phi a(1 - S(t_i))^{a-1}$ for $i = 1, 2, \ldots, n$.; (Dewan and Kulathinal (2007))
10. Generate g_i from Uniform(0,1);
11. If $g_i < \psi_i$, set cause = 1, otherwise cause = 2;
12. Now the data set for the ith subject is $(y_{ij}, t_{ij}, \delta_{ij}, X_i)$, $i = 1, 2, \ldots, n$, $j = 1, 2$;
13. Find out the expected value π_j for $\delta_{ij} = 0$, $j = 1, 2$;
14. Assign y_{ij} =1, if $\delta_{ij} = 1$. Otherwise y_{ij} =π_j according to cause j. ($y_{ij} = w_{ij}$);
15. Maximize the complete data log-likelihood function and estimate the parameters;
16. Repeat the procedure of Expectation-Maximization till the convergence criteria is met to get improved estimate (say, $\lambda - \hat{\lambda} < \delta$, a pre defined small quantity for parameter λ)
17. Replicate the required number of data sets.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 279–289

Improvisation of Dataset Efficiency in Visual Question Answering Domain

Sheerin Sitara Noor Mohamed and Kavitha Srinivasan

Department of Computer Science and Engineering,
Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam – 603110, India

Received: 14 August 2021; Revised: 02 December 2021; Accepted: 02 January 2022

Abstract

The technology revolution moves the world towards automation and most of the activities are performed with minimum human intervention. The medical domain is not an exception, few developments in the medical domain helps both the patient and physician to some extent. As a part of this advancement, Visual Question Answering (VQA) in the medical domain is evolved and which helps the physician and partially visually sighted people in clinical decision making and patient education. One of the main disadvantages in achieving this advancement is data limitation problem. In this paper, two methods for handling the data limitation problem are explained and validated using appropriate pre-trained models like VGGNet and ResNet. The methods namely label smoothing and mixup are used to reduce the hard samples and augmentation of the medical data. From the performance analysis, it has been inferred that the highest accuracy and BLEU score are obtained for improved dataset as 0.297 and 0.313 for ResNet with a significant improvement of 7.9% and 5.9% respectively.

Key words: Medical VQA; Data augmentation; Label smoothing; Mixup; VGGNet; ResNet.

1. Introduction

The VQA in medical domain is an emerging field during last few years. But it has many challenges like data limitation, computation time and requires expert radiologist knowledge. Among the challenges, the data limitation issue is chosen based on the suggestions given by few researchers. Saurrouti *et al.* (2021) stated that the new training samples generation reduced the data insufficiency and avoided overfitting for VQA-RAD dataset. He *et al.* (2020) observed that the pathology images are rarely available, and the involvement of well trained pathologists in supporting dataset creation and validation are significantly minimal in real time. Hence, they addressed this issue by extracting pathology images and question-answer pairs from textbooks using semi-automated pipeline. Both the researchers stated the dataset insufficiency problems and adapted different approaches to reduce the issue to some extent. In this section, the existing methods to address the data limitation problem related to medical VQA dataset and its techniques are discussed.

Corresponding Author: Sheerin Sitara Noor Mohamed
E-mail: sheerinsitaran@ssn.edu.in

Nguyen *et al.* (2019) addressed the data limitation problem by combining the denoising auto encoder and meta-learning for large scale unlabelled data, but the compatibility between questions and the visual contents are neglected. Gong *et al.* (2021) proposed the multi-task pre-trained framework, which learns the linguistic compatibility feature set and visual content using classification and segmentation on the external dataset for data limitation problem. Chen *et al.* (2020) suggested few techniques to overcome the data limitation problem such as mixup, label smoothing and adaptive curriculum learning. According to Zhang *et al.* (2020) mixup is a simple but effective augmentation technique based on data centric efficient training. Szegedy *et al.* (2017) concluded that the label smoothing method avoids the model bias on the data by stabilizing the training progress. Bengio *et al.* (2009) stated that the noise in training set is unavoidable and it can be rectified by curriculum learning, which automatically reduces the weights of the samples with higher loss value.

One of the researcher, Chen *et al.* (2020) stated that many solutions are available to overcome data limitation problem but the easiest way is to collect more samples from the available datasets (VQA MED competition) and updating as per the user interest. ImageCLEF is one of the VQA-MED competition forums which has been conducting tasks related to medical image captioning and medical Visual Question Answering since 2018 by providing dataset as open-source. The dataset description of these tasks are given in Table 1 and the importance of each task are as follows: (i). ImageCLEF VQA-MED 2018 task concentrates on VQA dataset related to different organ, plane, modality and abnormality because very few medical VQA dataset was available during that time. (ii). In the ImageCLEF VQA-MED 2019 task, the number of samples for each category is increased to generate a better model but the abnormality type VQA samples in the dataset degrades the overall performance. (iii). ImageCLEF VQA-MED 2020 task concentrates on abnormality type queries for different organ, plane and modality (iv). In the ImageCLEF VQA-MED 2021 task, the number of classes and the equivalent abnormality type samples are increased. The dataset obtained from ImageCLEF VQA MED 2020 and 2021 tasks are augmented and used in this research work for VQA model generation.

As per the literature (Hasan *et al.*, 2018, Abacha *et al.*, 2019 and Abacha *et al.*, 2020), different techniques are used in medical VQA for visual and text feature extraction process. The visual feature extractions techniques are Convolutional Neural Network (CNN) or pre-trained models like VGGNet, ResNet or Inception – ResNet and, the text feature extraction techniques are Long Short Term Memory (LSTM), Bidirectional Long Short Term Memory (Bi-LSTM) or Bidirectional Encoder Representations from Transformers (BERT). Then the extracted features are encoded using Stacked Attention Networks (SAN), Bilinear Attention Network (BAN) or Multi-modal Factorized High-order pooling (MFH) for attention based feature fusion for training the model. One of the researcher, Aisha *et al.* (2020) proposed a VQA model for ImageCLEF VQA-MED 2020 dataset using VGG16, ResNet or DensNet where the last layer is equivalent to the number of classes of the dataset.

As of now, the data limitation problem are addressed in a few ways: (i). Extracting dataset from the textbook using semi-automated pipeline for pathology medical dataset (ii). Combining the denoising autoencoder and meta-learning approach for unlabelled large dataset (iii). By learning the visual content and linguistic compatibility in the feature set for the VQA-RAD dataset. The limitations of existing methods includes: (i). The generated dataset has limited set of queries (ii). The compatibility between the question and the visual

content are neglected (iii). The cross model self-attention approach captures the long-range dependency between the question and visual content but the resulted number of samples are not increased. To overcome this specific issue, the number of samples are increased explicitly by mixup and label smoothing methods in the proposed work. Also, the modified dataset (improved) can be used directly in model creation, which in turn will reduce the overall computation time. For validating the model, different pre-defined techniques are available among which VGGNet and ResNet are significantly better than other techniques. Also, the datasets of ImageCLEF VQA-MED 2020 and 2021 are more suitable to analyse abnormality type questions as given in Table 1.

Table 1: ImageCLEF VQA – MED dataset description

ImageCLEF VQA-MED Dataset	Training Set		Validation Set		Test Set		Categories
	Images	QApairs	Images	QApairs	Images	QApairs	
Hasan *et al.* (2018)	2278	5413	324	500	264	500	Organ, plane, modality and abnormality
Abacha *et al.* (2019)	3200	12792	500	2000	500	500	Organ, plane, modality and abnormality
Abacha *et al.* (2020)	4000	4000	500	500	500	500	Abnormality
Abacha *et al.* (2021)	4500	4500	500	500	500	500	Abnormality

Source: ImageCLEF VQA-MED 2018 to 2021 tasks

The remaining sections of this paper are organised as follows. In Section 2, the dataset description and design of the proposed VQA model with dataset improvisation methods are briefly explained. In Section 3, the experimental setup and the results obtained are analysed with suitable quantitative metrics. Finally, summarized with conclusion and future work by stating the importance of dataset improvisation in medical domain.

2. Proposed Methodology

In the proposed system, the dataset is improvised by removing the hard samples and augmenting the data and, the pre-trained models are used to perform Visual Question Answering. For dataset improvisation, the methods label smoothing and mixup are used sequentially and vice-versa to improve the efficiency of the model. This can be further validated by the pre-trained models (VGGNet and ResNet) with the number of nodes in the last layer equivalent to the number of classes for the original and improved datasets. The overall system design of the proposed system is shown in Figure 1.

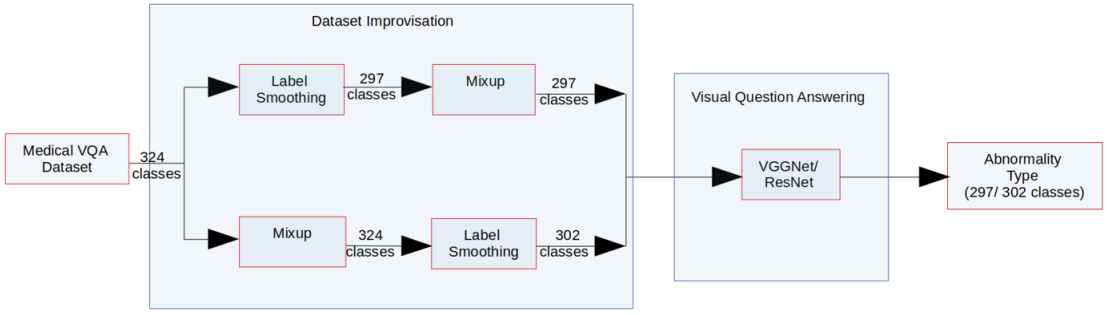


Figure 1: System design

2.1. Dataset description

The dataset used in this work is collected from ImageCLEF VQA-MED 2020 and 2021 for model generation and validation is mentioned in Table 2. The overall training set images are 5000 samples of radiology images with equivalent question-answer pairs. Similarly the validation set and test set comprises of 500 radiology images with its respective 500 question-answer pairs.

Table 2: Dataset description

ImageCLEF VQA-MED	Training Set		Validation Set		Test Set	
Dataset (Year)	Images	QApairs	Images	QApairs	Images	QApairs
2020	500	500	-	-	-	-
2021	4500	4500	500	500	500	500
Total	5000	5000	500	500	500	500

2.2. Label smoothing

The label smoothing methodology removes hard samples by adjusting the probability of target label is referred from (Gong *et al.*, 2021) as given in Equation (1),

$$p_a = f(x) = \begin{cases} 1-\varepsilon, & \text{if } a = b. \\ \frac{\varepsilon}{M-1}, & \text{otherwise.} \end{cases} \tag{1}$$

where ε is the small constant, M is the number of classes, and p_a denotes the probability of category a. This method is more suitable, when there are an elevated number of hard samples in the dataset and which affects the accuracy considerably. The hard samples are removed by grouping the representation of the samples from the same class into a tight cluster to improve the generalization ability. In general, the role of label smoothing in medical VQA dataset is to reduce the hard samples by adjusting the probability of target samples.

2.3. Mixup

The mixup methodology alleviates the data limitation problem by augmenting the dataset. Given the two samples (x_a and y_a) and (x_b and y_b), the new image $\hat{x}$ and $\hat{y}$ are created by linear interpolation by the equation is referred from (Gong *et al.*, 2021) as given in Equations (2) and (3),

$$\hat{x} = \lambda x_a + (1-\lambda)\, x_b \tag{2}$$

$$\hat{y} = \lambda y_a + (1-\lambda)\, y_b \tag{3}$$

$$p_a = f(x) = \begin{cases} Random(\beta(\alpha,\alpha)), & \text{if } \alpha > 0. \\ 1, & \text{otherwise.} \end{cases} \tag{4}$$

$$\beta(\alpha, \alpha) = \frac{\gamma(\alpha + \alpha))}{(\gamma(\alpha) * \gamma(\alpha))} \tag{5}$$

$$\gamma(\alpha) = (\alpha - 1)! \tag{6}$$

where $\alpha\epsilon\Big[0, 1\Big]$ is the shape parameter, γ is the factorial function to capture the continuous change, $\beta\epsilon\Big[0, 1\Big]$ is the target probability distribution value modifier and $\lambda\epsilon\Big[0, 1\Big]$ is a random value used to create new samples during the training process are given in Equations (6), (5) and (4), respectively.

The parameter α modifies the shape of the probability distribution and γ function is used to compute the range of probability distribution values. The α parameter in the beta distribution function controls the interpolation between feature-target pair using γ value. The beta distribution is chosen for two reasons, such as (i) to compute the probability distribution value from the range of alpha values (ii) the probability distribution function of beta distribution is approximately normal if $\gamma\Big(\alpha + \alpha\Big)$ is large. The role of mixup in medical VQA tasks is to augment the dataset by generating the new images from the existing images by linear interpolation.

2.4. VGGNet and ResNet

The medical VQA dataset improvised by label smoothing and mixup is given as input to the pre-trained models. In the pre-trained models, the last layer (fully connected layer) is frozen and the resultant model is used in the training process. The last layer is frozen because it is trained for the ImageNet dataset but the output dimension needs to be equivalent to the modified number of abnormality classes of the dataset to be validated. For this reason, the fully connected layer is frozen to predict the abnormality class types. The architecture of the pre-trained models, such as VGGNet and ResNet are referred from Simonyan *et al.* (2015) and He *et al.* (2016) for the implementation of proposed system.

3. Experiments and Results

In this section, the implementation requirement and experimental setup are discussed for the proposed system. Then the significance of label smoothing and mixup methods are analysed from the results of the proposed model along with hyper parameters for two datasets (original and improved).

The implementation platform (hardware) for the proposed system are: (i). Intel x64 Processor (ii). 16 GB RAM (iii). 1TB Memory (including 50 GB disk space) (iv). SSD drive to support high speed Input/Output (v). Graphics Processing Unit. The software requirements includes: (i). Ubuntu 16.04 (ii). Python 3.6 (iii). Tensorflow library. The following paragraphs explains the significance of proposed system developed with this environmental setup.

In label smoothing, the number of hard samples are removed by adjusting the probability of target label using the parameters ϵ and M. The determination of appropriate

class label based on target label probability is shown in Figure 2. In this, y-axis denotes the probability of the particular class and hence it ranges from 0.0 to 1.0 and x-axis denotes the comparison between target and predicted value and it achieves its peek value when both values are comparatively equal and it varies for each samples.

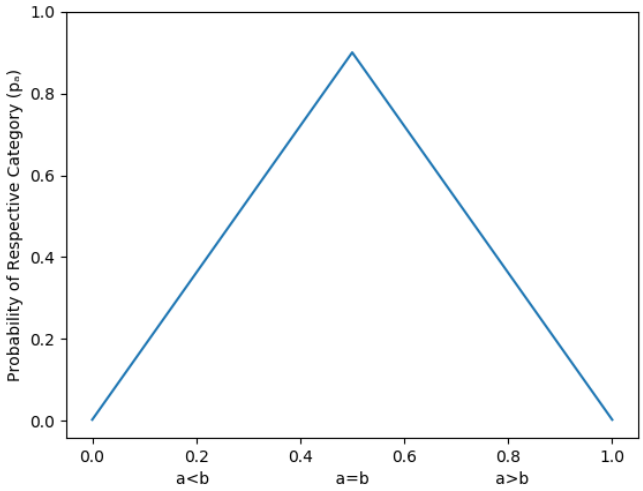


Figure 2: Appropriate class label with respect to target label probability

The hyper parameters and its values of label smoothing method is shown in Table 3 whereas common parameters are given in Table 5. Among the hyper parameters, Multi-StepLR is used to modify the learning rate based on Stochastic Gradient Descent, which is updated whenever number of epochs reaches one of the two milestone (initially it starts with 0.1).

Table 3: Label smoothing - Hyper parameters for improved dataset

Hyper parameters	**# value(s)**
Learning rate	Starts with 0.1, update the value at 30^{th}, 60^{th} and 90^{th} epoches
Epoch	120
Pooling	1X1 (Adaptive Average Pooling)

In Mixup, the new image is generated using two images with appropriate parameters such as α, γ, β and λ. The choice of alpha value plays a significant role in linear interpolation of new image because it acts as a basic element for all required computation. The variation of beta value distribution with respect to α value is shown in Figure 3. The alpha value can be represented as $0 < \alpha \leq 1$. The value of α never be zero, because at this point beta distribution is undefined and hence the scale is 0.1 to 1.0 with an interval of 0.1 in x-axis. The resulted beta distribution ranges from 0.20 to 0.50 with an interval of 0.05 is the scale of y-axis.

The hyper parameters specific to mixup method is given in Table 4 and, the hyper-paramters and its value common to both mixup and label smoothing for improved dataset is given in Table 5. In Table 3, the learning rate is decreased by 10% after 100^{th} epoch then

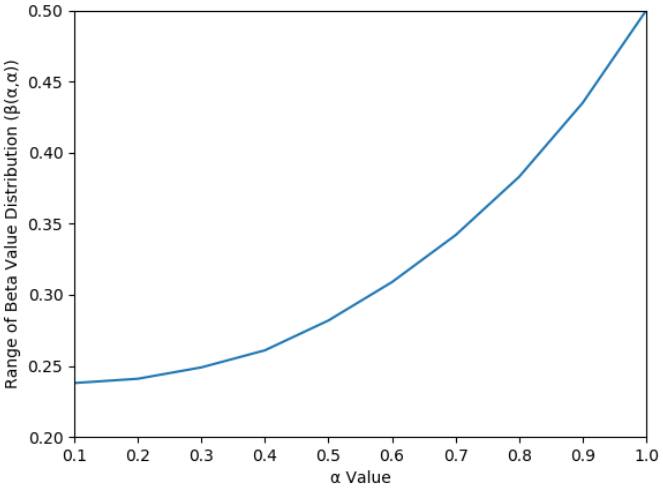


Figure 3: Range of beta value distribution with respect to α value

again 10% after 150^{th} epoch because smaller learning rate allows the model to learn more optimal set of weights but takes significantly longer time to train the model.

Table 4: Mixup - Hyper parameters for improved dataset

Hyper parameters	**# value(s)**
Learning rate	0.1, 0.01 and 0.001
Epoch	200
$WeightDecay$	0.0001
Pooling	4X4 (Average Pooling)

As a result of these two methods, the dataset comprises of 5000 VQA-MED samples with 324 classes is updated. The modifications in the number of samples and classes for label smoothing followed by mixup and vice versa are given in Table 6. In mixup followed by label smoothing method, the number of samples are augmented and then removed and hence few of the least contributing samples with higher loss values are also augmented.

The importance of improved dataset generated from label smoothing and mixup are validated using pre-trained models. The common hyperparameters used for VGGNet and ResNet for validating the model is shown in Table 7.

The results are analysed using the quantitative metrics namely accuracy and BLEU score for three cases such as, without dataset improvisation, with dataset improvisation (Label smoothing followed by mixup, mixup followed by label smoothing) as mentioned in Table 8.

From the overall results given in Tables 6 and 8 , some of the interesting inferences are: (i) The label smoothing followed by Mixup gives comparatively better results even though the number of samples are reduced (ii) Improvised dataset gives better results for both cases

Table 5: Hyper parameters common to label smoothing and Mixup for improved dataset

Hyper parameters	# value
WeightDecay	0.0001
Momentum	0.9
Normalization	64 (Batch Normalization)
Kernel size	3
Stride	1
Padding	1
Batch size	64
Type of Optimizer	Stochastic Gradient Descent
Type of Activation function	Rectified Linear Unit

Table 6: Improved dataset description

Label Smoothing followed by Mixup			**Mixup followed by Label Smoothing**		
Execution Sequence	**Number of Samples**	**Number of Classes**	**Execution Sequence**	**Number of Samples**	**Number of Classes**
Label Smoothing	4294	297	*Mixup*	5134	324
Mixup	4513	297	*Label Smoothing*	4700	302

(VGGNet and ResNet) (iii) For improved dataset, the overall accuracy is increased by 3.8% and 7.9% for VGGNet and ResNet respectively and (iv) For augmented dataset the accuracy and BLEU score are increased by 7.9% and 5.9% respectively for ResNet. In addition, the results of two metrics is graphically represented in Figure 4, for Datasets Vs ResNet only.

Table 7: VGGNet and ResNet - Hyper parameters

Hyper parameters	# value
Batchsize	128
Epoch	100
Dropout	0.2
Learning rate	0.001
Type of Optimizer	RMSPROP

In Figure 4, x-axis denotes the dataset and y-axis denotes the performance value achieved for three datasets which ranges between 0.20 and 0.32 with an interval of 0.02. The WoDI , LS_MU and MU_LS used in the graph represents Without Dataset Improvisation, Label Smoothing followed by Mixup and Mixup followed by Label Smoothing respectively. From Figure 4, it is clear that Label smoothing followed by mixup data improvisation achieved better accuracy and BLEU score using ResNet.

Table 8: Performance analysis

Performance Metrics	Label Smoothing followed by Mixup		Mixup followed by Label Smoothing		Without improvisation of dataset	
	VGGNet	ResNet	VGGNet	ResNet	VGGNet	ResNet
Accuracy	0.234	0.297	0.230	0.284	0.196	0.218
BLEU Score	0.257	0.313	0.248	0.299	0.227	0.254

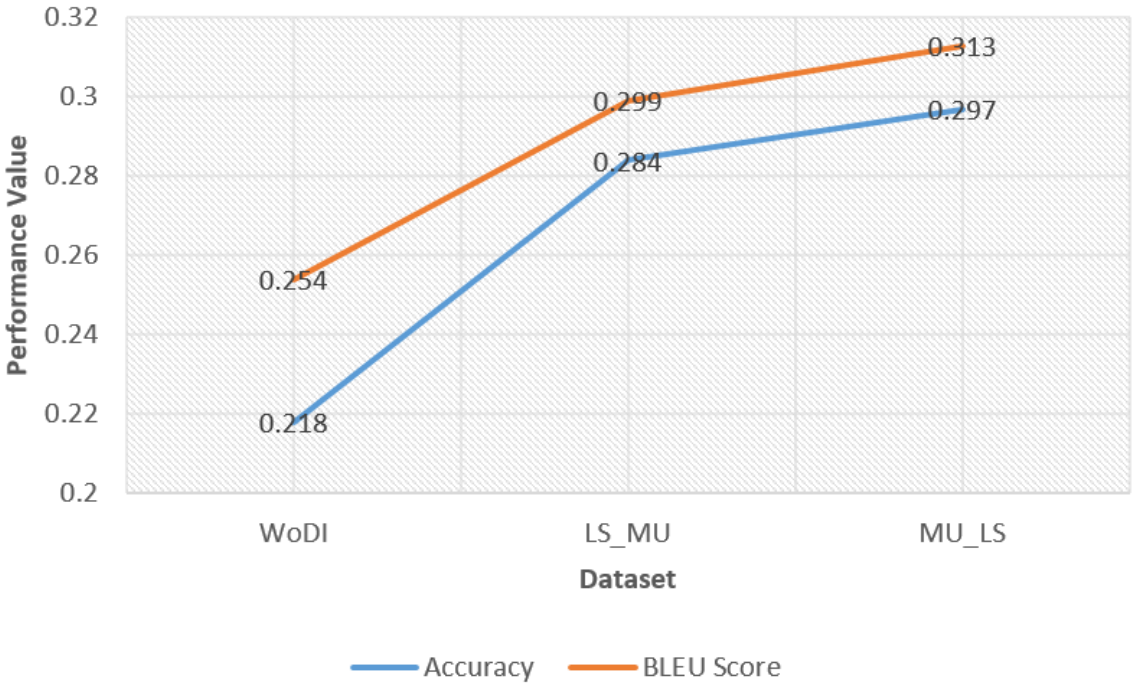


Figure 4: Dataset Vs performance analysis for ResNet

WoDI: Without Dataset Improvisation; LS_MU: Label Smoothing followed by Mixup; MU_LS: Mixup followed by Label Smoothing

4. Conclusion and Future Work

In this research, to strengthen the dataset of medical VQA, two methods namely label smoothing and mixup are chosen and its parameters are analysed and modified to improve the dataset. In label smoothing method, the hard samples are removed by adjusting the probability of target samples and mixup method augmented the new samples from the existing samples by linear interpolation. These methods improvises the efficiency of the dataset and overcomes the data limitation problem in the medical domain to some extent. The importance of dataset improvisation is validated using the pre-trained models (VGGNet and ResNet) with appropriate hyperparameters. The accuracy and BLEU score is improved by 3.8% and 3.0% for VGGNet, and 7.9% and 5.9% for ResNet respectively using the improved dataset. From the results, it has been inferred that the removal of hard samples and data augmentation improved the performance of the model significantly.

The important future direction is the creation of larger and varied dataset by increasing the number of samples in each category of medical domain with enhanced quality. Using this dataset, an improved VQA system can be developed to answer all medical queries. The VQA system development can be enhanced by selecting suitable hyper parameters to increase the efficiency and reliability of the system.

Acknowledgements

Our profound gratitude to Sri Sivasubramaniya Nadar College of Engineering, Department of CSE, for allowing us to utilize the High Performance Computing Laboratory and GPU Server for the execution of this research work successfully.

References

Abacha, A. B., Datla, V. V., Sadid A. Hasan, S. A., Demner-Fushman, D. and Muller, H. (2020). Overview of the VQA-Med Task at ImageCLEF 2020: Visual Question Answering and Generation in the Medical Domain. *CLEF 2020 Working Notes, CEUR Workshop Proceedings, Greece*, 1–9.

Abacha, A. B., Hasan, S. A., Datla, V. V., Liu, J., Demner-Fushman, D. and Miller, H. (2019). VQAMed: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019. *CLEF 2019 Working Notes, CEUR Workshop Proceedings, Switzerland*, 1–11.

Abacha, A. B., Sarrouti, M., Demner-Fushman, D., Hasan, S.A. and Muller, H. (2021). Overview of the VQA-Med Task at ImageCLEF 2021: Visual Question Answering and Generation in the Medical Domain. *CLEF 2021 Working Notes, CEUR Workshop Proceedings, Romania*, 1–8.

Bengio, Y., Louradour, J., Collobert, R. and Weston, J. (2009). Curriculum Learning. *International Conference on Machine Learning, Canada*, 41–48.

Chen, G., Gong, H. and Li, G. (2020). HCP-MIC at VQA-Med 2020: Effective Visual Representation for Medical Visual Question Answering. *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Greece*, 1–11.

Gong, H., Chen, G., Liu, S., Yu, Y. and Li, G. (2021). Cross-Modal Self-Attention with Multi-task Pretraining for Medical Visual Question Answering. *ACM International Conference on Multimedia Retrieval (ICMR), Taiwan*, 456–460.

Gong, H., Huang, R., Chen, G. and Li, G. (2021). SYSU-HCP at VQA-Med 2021: A Data-Centric Model with Efficient Training Methodology for Medical Visual Question Answering. *CLEF 2021 Working Notes, CEUR Workshop Proceedings, Romania*, 1–11.

Hasan, S. A., Ling, Y., Farri, O., Liu, J., Miller, H. and Lungren, M. (2018). Overview of ImageCLEF 2018 Medical Domain Visual Question Answering Task. *CLEF 2018 Working Notes, CEUR Workshop Proceedings, Switzerland*, 1–8.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 770 –778.

He, X., Zhang, Y., Mou, L., Xing, E. and Xie, P. (2020). PathVQA: 3000+ Questions for medical question answering. *arXiv preprint arXiv:2003.10286*, 1–12.

Muller, R., Kornblith, S. and Hinton, G. (2020). What does label smoothing help?. *Advances in Neural Information Processing System*, **32**, 1–10.

Nguyen, B. D., Do, T. T., Nguyen, B. X., Do, T., Tjiputra, E. and Tran, Q. D. (2019). Overcoming Data Limitation in Medical Visual Question Answering. *International Conference on Medical Image Computing and Computer-Assisted Intervention, Turkey*, 522–530.

Sarrouti, M., Abacha, A. B. and Demner-Fushman, D. (2021). Goal-driven visual question generation from radiology images. *Information*, **12(8)**, 1–16.

Sheerin, S. N. M. and Kavitha, S. (2020). ImageCLEF 2020: An approach for Visual Question Answering using VGG-LSTM for different datasets. *CLEF 2020 Working Notes, CEUR Workshop Proceedings, Greece*, 1–10.

Sheerin, S. N. M. and Kavitha, S. (2021). SSN MLRG at VQA-MED 2021: An Approach for VQA to Solve Abnormality Related Queries using Improved Datasets. *CLEF 2021 Working Notes, CEUR Workshop Proceedings, Romania*, 1–10.

Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-scale Image Recognition. *International Conference on Learning Expectation, Canada*, 1–14.

Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A. A. (2017). Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning. *AAAI Conference on Artificial Intelligence, California*, 4278–4284.

Zhang, H., Cisse, M., Dauphin, Y. N. and Lopez-Paz, D. (2018). Mixup: Beyond Empirical Risk Minimization. *International Conference on Learning Representations, Canada*, 1–13

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M. and Smola, A. J. (2020). Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*,1–12.

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
2022

Printed by : Galaxy Studio & Graphics
Mob: +91 9818 35 2203, +91 9582 94 1203