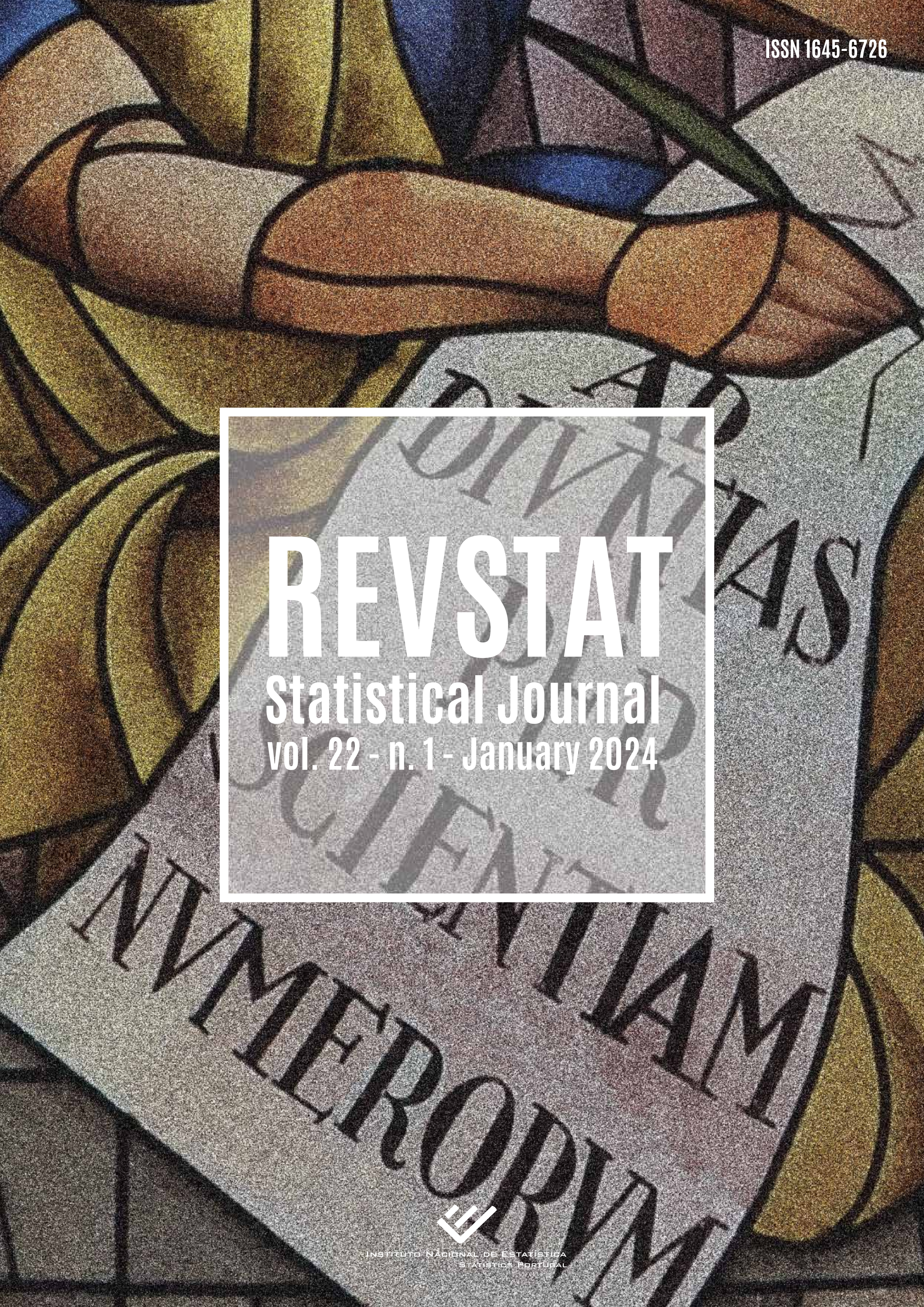
ISSN 1645-6726



REVSTAT

Statistical Journal

vol. 22 - n. 1 - January 2024

Instituto Nacional de Estatística
Statistics Portugal

REVSTAT — Statistical Journal, vol.22, n. 1 (January 2024)

vol.1, 2003- . - Lisbon : Statistics Portugal, 2003- .

Continues: Revista de Estatística = ISSN 0873-4275.

ISSN 1645-6726 ; e-ISSN 2183-0371

Publisher – *Statistics Portugal*

Layout-Graphic Design – *Carlos Perpétuo* | **Cover Design*** – *Helena Nogueira*

Edition - 130 copies | **Legal Deposit Registration** - 191915/03 | **Price** [VAT included] - € 9,00

**image*: stain glass window by Abel Manta (1888-1982)

EDITORIAL

Letter from the Editor-in-Chief

Dear REVSTAT community,

I am very honored to have been appointed as the new Editor-in-Chief of *REVSTAT — Statistical Journal* effective January 1st, 2024. First and foremost, I am indebted to my predecessor, Isabel Fraga Alves, for her leadership and development of the journal over the past five years. Isabel greatly contributed to the high standards of the journal and to refresh journal's website, which currently includes an effective automatic editorial platform. With unfailing courtesy and enthusiasm, Isabel also provided invaluable advice and assistance during the transition, as did Olga Bessa Mendes (Statistics Portugal) the Executive Editor of the journal.

As Editor-in-Chief, assisted by the co-Editor Cláudia Nunes Philippart, it is my responsibility to ensure that *REVSTAT — Statistical Journal* continues to grow its impact and influence within the statisticians' community. To this end, the team of Associate Editors (AEs) will play a key role in overseeing and managing the peer-review process, supporting the Editor's decision-making. REVSTAT - Statistical Journal is delighted to announce 12 new AEs who have joined the Editorial Board with a 2-year appointment (2024-2025). We welcome **Abdelhakim Aknouche** (Qassim University), **Andrés Alonso** (Carlos III University of Madrid), **Wagner Barreto-Souza** (University College Dublin), **Francisco Blasques** (VU Amsterdam), **Rui Castro** (Eindhoven University of Technology), **Sónia Gouveia** (University of Aveiro), **Raquel Menezes** (University of Minho), **Cláudia Neves** (King's College London), **Carlos Oliveira** (Norwegian University of Science and

Technology), **Rosário Oliveira** (University of Lisbon), **Jacobo de Uña-Álvarez** (University of Vigo), and **Christian Weiß** (Helmut Schmidt University). Furthermore, **Barry Arnold** (University of California), **Narayanaswamy Balakrishnan** (McMaster University), **Paula Brito** (University of Porto), **Valérie Chavez-Demoulin** (University of Lausanne), **David Conesa** (University of Valencia), **Charmaine Dean** (University of Waterloo), **Fernanda Figueiredo** (University of Porto), **Jorge Milhazes Freitas** (University of Porto), **Stéphane Girard** (Inria Grenoble Rhône-Alpes), **Victor Leiva** (Pontificia Universidad Católica de Valparaíso), **Artur Lemonte** (Federal University of Rio Grande do Norte), **Shuangzhe Liu** (University of Canberra), **Fernando Moura** (Federal University of Rio de Janeiro), **John Nolan** (American University), **Paulo Eduardo Oliveira** (University of Coimbra), **Pedro Oliveira** (University of Porto), **Gilbert Saporta** (Conservatoire National des Arts et Métiers), **Alexandra M. Schmidt** (McGill University), and **Lisete Sousa** (University of Lisbon) have agreed to extend their initial appointment (2019-2023) to 2024-2025. Finally, **Marília Antunes** (University of Lisbon), **Jan Beirlant** (Katholieke Universiteit Leuven), **Graciela Boente** (University of Buenos Aires), **Alan Gelfand** (Duke University), **Marie Kratz** (ESSEC Business School), **Maria Nazaré Mendes-Lopes** (University of Coimbra), **Arthur Pewsey** (University of Extremadura), **Milan Stehlík** (University of Valparaíso), and **María Dolores Ugarte** (Public University of Navarre) step down as AEs. To all of you, thank you for your dedication, commitment, and support to the success of the journal.

A final word to the authors who choose *REVSTAT – Statistical Journal* as the outlet for their work. Despite of the overwhelming number of submissions, the Editorial Board is very committed in ensuring that manuscripts move through the revision system as efficiently as possible and making recommendations about acceptance or rejection of a manuscript timely.

We are looking forward to receiving your submissions!

Sincerely,

Manuel G. Scotto

INDEX

Abstracted/indexed in: *Journal Citation Reports™ – JCR (Clarivate); DOAJ – Directory of Open Access Journals; Current Index to Statistics; Google Scholar; Mathematical Reviews® (MathSciNet®); Zentralblatt für Mathematic; SCImago Journal & Country Rank; Scopus®.*

REVSTAT – Statistical Journal
Volume 22, Number 1, January 2024, 1–24
https://doi.org/10.57805/revstat.v22i1.436

A Study on Zografos-Balakrishnan Log-Normal Distribution: Properties and Application to Cancer Dataset

Authors: D.S. Shibu
– Department of Statistics, University College,
Thiruvananthapuram, India – 695034
statshibu@gmail.com

S.L. Nitin ✉
– Department of Statistics, University College,
Thiruvananthapuram, India – 695034
nitinstat24@gmail.com

M.R. Irshad
– Department of Statistics, CUSAT,
Cochin, India – 682022
irshadm24@gmail.com

Received: October 2020 Revised: March 2022 Accepted: March 2022

Abstract:

- In this article, we studied a generalization of the log-normal distribution called Zografos-Balakrishnan log-normal distribution, and investigate its various important properties and functions including moments, quantile function, various reliability measures, Rényi entropy, and some inequality measures. The estimation of unknown parameters is discussed by the methods of maximum likelihood, and the Bayesian technique and their simulation studies are also carried out. The applicability of the distribution is illustrated utilizing a real dataset. A likelihood ratio test is utilized for testing the efficiency of the third parameter. The effectiveness of this model for the dataset is also established using the parametric bootstrap approach.

Keywords:

- *Zografos-Balakrishnan-G family; reliability measures; maximum likelihood estimation; Bayesian estimation; bootstrap confidence interval; likelihood ratio test.*

AMS Subject Classification:

- 60E05, 62F10, 62F15.

✉ Corresponding author.

1. INTRODUCTION

Recently, an increasing interest can be observed for the art of adding parameters to some well-known existing distributions for getting different shapes of hazard rate or failure rate functions for applying it in various real-life situations and also for analyzing data with a high degree of skewness and kurtosis. In principle, the log-normal distribution is defined as the continuous probability distribution of a random variable whose logarithm is normally distributed. It is one of the most widely used distributions for asymmetric datasets. Thus, it has been widely applied in many different aspects of life sciences, including biology, geology, ecology, and meteorology as well as in economics, finance, and risk analysis (see [15]), and also attracts attention quite often in environmental sciences, physics, astrophysics, and cosmology (see [3], [4], [22]).

On many occasions, the significance of the LN distribution in biological science has been acknowledged. Bentley (1954) ([9]) provides numerous generic resources for statistical data generated from biological and agricultural sources. A study on the complexities of the biochemical mechanisms associated with gene expression has created an emergent LN distribution of expression levels, according to [2]. Carvalho (2018) ([5]) found that a form of the LN distribution fits the postpartum blood loss data from numerous geographical areas quite well, suggesting that the LN distribution may fit postpartum blood loss generally. Hence, in this article, we utilize a cancer dataset as an application that is related to biological science.

The probability density function (pdf) for a log-normal random variable W is given by

$$q(w) = \frac{1}{\sqrt{2\pi}\sigma w}\ \exp\left[-\frac{(\log w - \mu)^2}{2\sigma^2}\right], \quad w > 0,\ \mu \in \mathbb{R},\ \sigma > 0.$$

Zografos and Balakrishnan (2009) (see [26]) proposed a novel family of univariate distributions generated by gamma random variables. Further Nadarajah *et al.* (2015) (see [20]) provides a comprehensive treatment of the general mathematical properties of this family and denote it with the prefix "Zografos-Balakrishnan-G" or "ZB-G" distributions. They discuss the estimation of parameters by maximum likelihood and provide an application to a real dataset and also propose a bivariate generalization. For any baseline cumulative distribution function (cdf) $G(x)$, and $x \in \mathbb{R}$, Zografos and Balakrishnan (2009) ([26]) defined a distribution with pdf $f(x)$ and cumulative distribution function (cdf) $F(x)$ given by

$$f(x) = \frac{1}{\Gamma(\alpha)}\{-\log[1 - G(x)]\}^{\alpha-1} g(x), \tag{1.1}$$

and

$$F(x) = \frac{\gamma(\alpha, -\log[1 - G(x)])}{\Gamma(\alpha)} = \frac{1}{\Gamma(\alpha)} \int_0^{-\log[1-G(x)]} t^{\alpha-1} \exp(-t) dt, \tag{1.2}$$

respectively for $\alpha > 0$, where $g(x) = dG(x)/dx$, $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \exp(-t) dt$ denotes the gamma function, and $\gamma(\alpha, z) = \int_0^\infty t^{\alpha-1} \exp(-t) dt$ denotes the incomplete gamma function. The corresponding hazard rate function (hrf) is

$$h(x) = \frac{\{-\log[1 - G(x)]\}^{\alpha-1} g(x)}{\Gamma(\alpha, -\log[1 - G(x)])},$$

where $\Gamma(\alpha, z) = \int_0^z t^{\alpha-1} \exp(-t) dt$ denotes the complementary incomplete gamma function.

Moreover, using the generalization in (1.1) and considering the immense applicability of the log-normal distribution, Nadarajah *et al.* (2015) ([20]) also suggests the generalization of log-normal distribution called Zografos-Balakrishnan log-normal (ZBLN) distribution. However, little is known in terms of general mathematical properties and in terms of application for this generalization.

The aim of this article is to derive some mathematical properties of Zografos--Balakrishnan log-normal distribution in the most simple, explicit and general forms and apply it to biological sciences and other reliability analyses. The main motivation for considering this lifetime model is to study the flexibility of the distribution that can be used to model lifetime data in a wider class of biological data and reliability problems.

The rest of the paper is organized as follows. In Section 2, we present the definition of the ZBLN distribution and obtain the weighted form of the same. The moments of the distribution are obtained in Section 3. The quantile function and some of its associated measures are obtained in Section 4. The various functions and the moments related to the reliability measures are discussed in Section 5. Section 6 deals with the derivation of the Rényi entropy, and Section 7 deals with the discussion of some inequality measures associated with the ZBLN distribution. The distributions of order statistics are derived in Section 8. In order to estimate the unknown parameters of the ZBLN model, the method of maximum likelihood estimation, and the Bayesian estimation procedure are employed, and also a parametric bootstrap method of simulation is presented in Section 9. To analyze the longstanding performances of maximum likelihood estimators, and the Bayesian estimators of the parameters, a simulation study has been conducted in Section 10. To illustrate the potentiality of the ZBLN distribution over competing distributions, one real dataset is analyzed in Section 11. The final concluding remarks are given in Section 12.

2. DEFINITION OF THE DISTRIBUTION

In this section, we present the definition and some important features of the ZBLN distribution.

Definition 2.1. Let X be a random variable which follows ZBLN distribution (see [20]) with parameters α, μ and σ, then its pdf is given by

$$f(x) = \frac{1}{\sigma\; x\; \Gamma(\alpha)} \left\{ -\log\left[1 - \Phi\left(\frac{\log x - \mu}{\sigma}\right)\right]\right\}^{\alpha-1} \phi\left(\frac{\log x - \mu}{\sigma}\right), \tag{2.1}$$

and the cdf is given by

$$\begin{aligned} F(x) &= \frac{\gamma\left(\alpha, -\log\left[1 - \Phi\left(\frac{\log x - \mu}{\sigma}\right)\right]\right)}{\Gamma(\alpha)} \\ &= \frac{1}{\Gamma(\alpha)} \int_0^{-\log\left[1-\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right]} t^{\alpha-1} \exp(-t) dt, \end{aligned} \tag{2.2}$$

where $x > 0$, $\mu \in \mathbb{R}$ and $\alpha, \sigma > 0$. Also, $\Phi(.)$ and $\phi(.)$ are respectively the cdf and pdf of the standard normal distribution.

Note that, ZBLN distribution reduces to the two-parameter log-normal if $\alpha = 1$. The plot in Figure 1 portrays the pdf of ZBLN distribution, and we observe that the pdf may be decreasing and unimodal with a certain flexibility in the mode and tails. It is, however, mainly right-skewed or almost symmetrical.

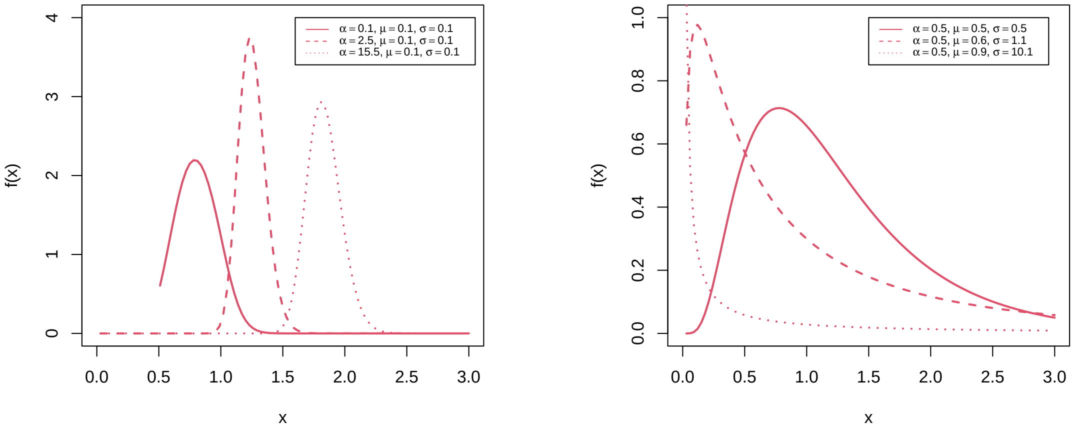


Figure 1: Plots of pdf of the ZBLN distribution.

2.1. Expansions for pdf and cdf

Nadarajah *et al.* (2015) ([20]) derived some useful expansions for (1.1) and (1.2) using the concept of exponentiated distributions. For an arbitrary baseline cdf $G(x)$, a random variable is said to have the exponentiated-G distribution with parameter $\alpha > 0$, say $X \sim$ exp-$G(\alpha)$, if its pdf and cdf are respectively given by

$$f_\alpha^*(x) = \alpha G^{\alpha-1}(x) g(x), \quad \text{and} \quad F_\alpha^*(x) = G^\alpha(x).$$

The important properties of exponentiated distributions have been studied by several authors; for examples, see [18] for exponentiated Weibull, [10] for exponentiated Pareto, [12] for exponentiated exponential, [19] for exponentiated Gumbel and [21] for exponentiated gamma distributions.

Note that, for any real parameter $\alpha > 0$, the following formula holds.

$$\{-\log[1-G(x)]\}^{\alpha-1} = (\alpha-1) \sum_{k=0}^{\infty} \binom{k+1-\alpha}{k} \sum_{j=0}^{k} \binom{k}{j} \frac{(-1)^{j+k}\; p_{j,k}}{(\alpha-1-j)} \{G(x)\}^{\alpha+k-1},$$

where the constants $p_{j,k}$ can be calculated recursively through the relation,

$$p_{j,k} = \frac{1}{k} \sum_{m=1}^{k} [k - m(j+1)] c_m \; p_{j,k-m},$$

for $k = 1, 2, ...$ with $p_{j,0} = 1$ and $c_k = (-1)^{k+1}(k+1)^{-1}$. Thus, Nadarajah *et al.* (2015) ([20]) demonstrated that (1.1), and the corresponding (1.2) can be expressed as

$$f(x) = \sum_{k=0}^{\infty} b_k \ f^*_{\alpha+k}(x), \quad \text{and} \quad F(x) = \sum_{k=0}^{\infty} b_k \ F^*_{\alpha+k}(x),$$

where $f^*_{\alpha+k}(x)$ and $F^*_{\alpha+k}(x)$ respectively denotes the corresponding pdf and cdf of the exp-$G(\alpha+k)$ distribution and for any real parameter $\alpha > 0$, and

$$b_k = \frac{\binom{k+1-\alpha}{k}}{(\alpha+k)\Gamma(\alpha-1)} \sum_{j=0}^{k} \binom{k}{j} \frac{(-1)^{j+k} \ p_{j,k}}{(\alpha-1-j)}. \tag{2.3}$$

Thus, the cdf and pdf of the ZBLN distribution respectively obtained as

$$F(x) = \sum_{k=0}^{\infty} b_k \left[\Phi\left(\frac{\log x - \mu}{\sigma}\right)\right]^{\alpha+k}, \tag{2.4}$$

and

$$f(x) = \sum_{k=0}^{\infty} b_k \frac{(\alpha+k)}{\sigma x} \phi\left(\frac{\log x - \mu}{\sigma}\right) \left[\Phi\left(\frac{\log x - \mu}{\sigma}\right)\right]^{\alpha+k-1}. \tag{2.5}$$

Thus, ZBLN distribution can be expressed as the infinite weighted sum of Exponentiated log-normal distributions indexed by power parameter $\alpha + k$.

3. MOMENTS

In this section, we derive the expression for the r^{th} raw moment of ZBLN distribution. From Equation (2.5), the moments of the ZBLN distribution can be written as the weighted sum of probability-weighted moments of the log-normal distribution. Thus, the r^{th} raw moment of the distribution is given by

$$\mu'_r = E(X^r) = \sum_{k=0}^{\infty} (\alpha+k) b_k \ \mu'_{r,\alpha+k},$$

where

$$\mu'_{r,\alpha+k} = E\left\{X^r \left[\Phi\left(\frac{\log x - \mu}{\sigma}\right)\right]^{\alpha+k-1}\right\}$$

$$\Rightarrow \qquad \mu'_{r,\alpha+k} = \int_0^{\infty} \frac{x^r}{\sigma x} \ \phi\left(\frac{\log x - \mu}{\sigma}\right) \left[\Phi\left(\frac{\log x - \mu}{\sigma}\right)\right]^{\alpha+k-1} dx,$$

is the probability weighted moments of the log-normal distribution.

4. QUANTILE FUNCTION AND ASSOCIATED MEASURES

Generally, a probability distribution can be specified either in terms of the distribution function or by the quantile function. Quantile functions have several interesting properties that are not shared by distributions, which makes them more convenient and flexible for analysis. Moreover, the random numbers from any distribution can be generated using appropriate quantile functions. So, in this section, we derive an explicit expression for the quantile function of ZBLN distribution and some of its associated measures.

Theorem 4.1. *If X follows ZBLN distribution as given in (2.2), then the p^{th} quantile, $Q_p = F^{-1}(p)$ of the distribution is given by*

$$Q_p = \exp\big\{\mu + \sigma\ \Phi^{-1}\big[1 - \exp\big(-Q^{-1}(\alpha, 1-p)\big)\big]\big\},$$

where $\Phi^{-1}(.)$ is the quantile function of standard normal variate.

Proof: For the ZBLN distribution, Q_p is the solution of the equation

$$Q\left(\alpha, -\log\left[1 - \Phi\left(\frac{\log(Q_p)-\mu}{\sigma}\right)\right]\right) = 1-u, \quad p \in (0,1)$$

$$\Rightarrow -\log\left[1 - \Phi\left(\frac{\log(Q_p)-\mu}{\sigma}\right)\right] = Q^{-1}(\alpha, 1-p) \tag{4.1}$$

On simplifications, (4.1) reduces to

$$\Phi\left(\frac{\log(Q_p)-\mu}{\sigma}\right) = 1 - \exp\big(-Q^{-1}(\alpha, 1-p)\big)$$

$$\Rightarrow \frac{\log(Q_p)-\mu}{\sigma} = \Phi^{-1}\big[1 - \exp\big(-Q^{-1}(\alpha, 1-p)\big)\big]$$

$$\Rightarrow Q_p = \exp\big\{\mu + \sigma\ \Phi^{-1}\big[1 - \exp\big(-Q^{-1}(\alpha, 1-p)\big)\big]\big\}. \tag{4.2}$$

□

Remark 4.1. Since $\Phi^{-1}(.)$ is the quantile function of standard normal variate, Q_p in Equation (4.2) also written in the form

$$Q_p = \exp\left\{\mu + \sigma\sqrt{2}\ \text{erf}^{-1}\left[1 - 2\exp\big(-Q^{-1}(\alpha, 1-p)\big)\right]\right\}, \tag{4.3}$$

where $\text{erf}^{-1}(.)$ is the inverse error function.

Now, by putting $p = 0.5$, in Equation (4.3), we get the median (M) of ZBLN distribution and is given by

$$\text{M} = Q_{0.5} = \exp\left\{\mu + \sigma\sqrt{2}\ \text{erf}^{-1}\left[1 - 2\exp\big(-Q^{-1}(\alpha, 1/2)\big)\right]\right\}.$$

For $p = 1/4$ and $p = 3/4$, Equation (4.3) respectively gives first and third quartiles of the ZBLN distribution.

5. RELIABILITY MEASURES

Many domains of practical studies, such as physics, engineering, psychology, and others, rely heavily on reliability measures. As a reason, providing expressions for various reliability measures is critical. Due to these facts, in this section, we derive expressions for various measures of reliability.

5.1. Hazard rate function

The hazard rate provides the instantaneous risk that the event of interest happens, within a very narrow time frame. As a function of age x, the hazard rate function is also referred to as the failure rate function, instantaneous death rate, force of mortality, and intensity function in other areas of study like survival analysis, actuarial science, biosciences, demography, and extreme value theory. Thus, it also plays a substantial role in lifetime data analysis, mainly in survival and reliability studies. Indeed, the mathematical characterization of a lifetime distribution for a certain life phenomenon can be made on the basis of its failure rate pattern. Most commonly, the hazard function can be increasing, decreasing, upside-down bathtub or bathtub shaped.

By definition, the hazard function $h(x)$ can be defined as $h(x) = f(x)/S(x)$, where $S(x) = 1 - F(x)$ is the survival function. Obviously, the survival function of ZBLN distribution is given as

$$S(x) = 1 - \frac{\gamma\Big(\alpha, -\log\Big[1 - \Phi\Big(\frac{\log x - \mu}{\sigma}\Big)\Big]\Big)}{\Gamma(\alpha)}.$$

Thus, the hazard function of ZBLN distribution is obtained as

$$h(x) = \frac{\phi\Big(\frac{\log x-\mu}{\sigma}\Big)\Big\{-\log\Big[1 - \Phi\Big(\frac{\log x-\mu}{\sigma}\Big)\Big]\Big\}^{\alpha-1}}{\sigma x\;\Gamma\Big(\alpha, -\log\Big[1 - \Phi\Big(\frac{\log x-\mu}{\sigma}\Big)\Big]\Big)},$$

where $\Gamma(\alpha, z) = \int_z^{\infty} t^{\alpha-1} \exp(-t)dt$ denotes the complementary incomplete gamma function. Also, plots in Figure 3 refers the hazard rate function and observed that ZBLN distribution possess increasing, decreasing, bathtub, and upside-down bathtub shapes. In this scenario, the capability of our model to construct a bathtub-shaped failure rate function with a significantly longer flat region is one of its unique advantages. Nonetheless, this region is crucial in real-world applications, underscoring the importance of effective flat region modeling (see [14]). Again, from Figure 2, it can be seen in further detail that the hazard rate function graph for the shape bathtub happens when $\alpha = 0.0001$, $\mu = 1.5$, $0.2 \le \sigma \le 0.31$. When $\alpha \ge 0.1$, $\mu = 0.01$, and $\sigma = 1.1$, the shapes also change from decreasing to increasing via an upside-down bathtub.

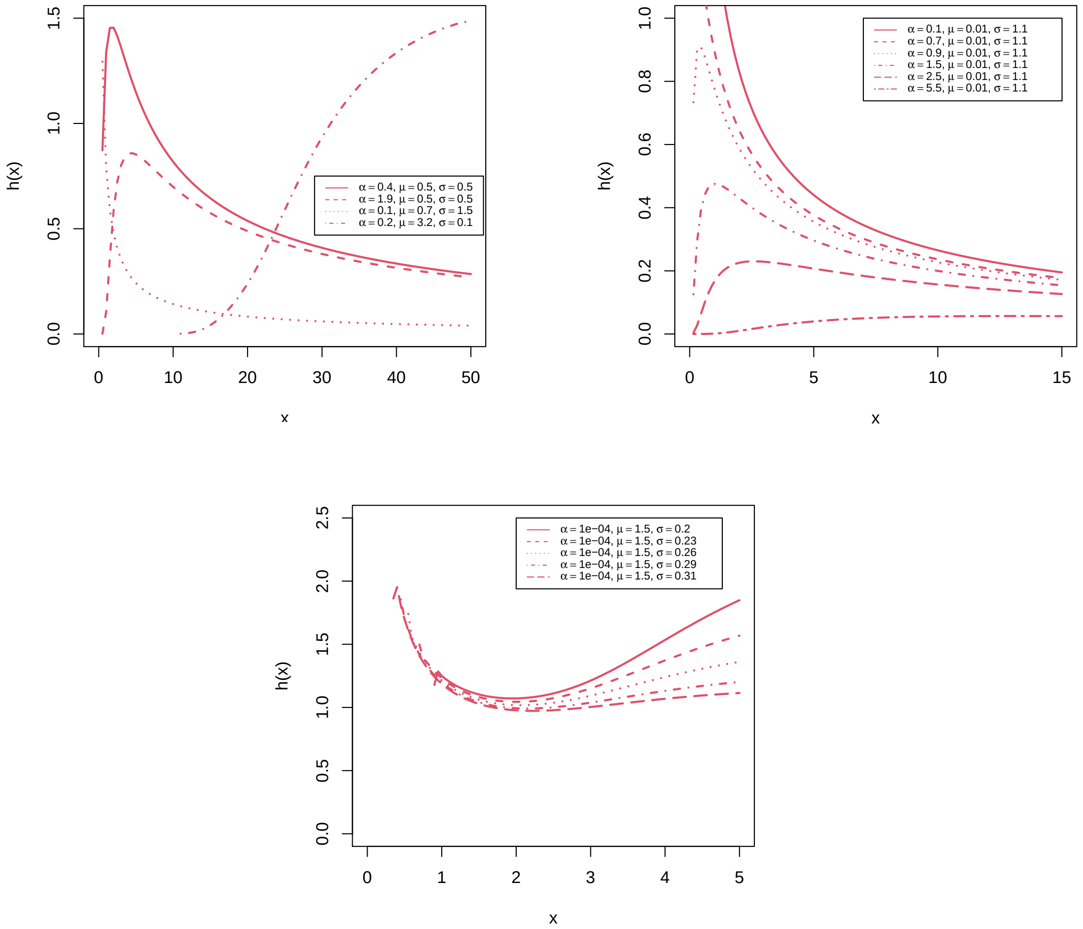


Figure 2: Plots of the hazard rate function of the ZBLN distribution.

5.2. Cumulative hazard rate function

The cumulative hazard rate function, also known as the integrated hazard function, is the overall number of failures or deaths over a period of time. Like the hazard function, the cumulative hazard function $H(x)$ is not a probability, but still a measure of risk. The greater the value of $H(x)$, the greater the risk of failure by time x.

By definition, $H(x) = -\log\{S(x)\}$. Thus, the cumulative hazard rate function of ZBLN distribution is given by

$$H(x) = -\log\left\{1 - \frac{\gamma\left(\alpha, -\log\left[1-\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right]\right)}{\Gamma(\alpha)}\right\}. \tag{5.1}$$

Note that, $\log(1-x) = -\sum_{n=1}^{\infty}\frac{x^n}{n}$, and also from Equation (2.4), $H(x)$ in Equation (5.1) can be simplified as

$$H(x) = \sum_{n=1}^{\infty}\frac{1}{n}\left\{\sum_{k=0}^{\infty} b_k\left[\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right]^{\alpha+k}\right\}^{n}.$$

5.3. Reversed hazard rate function

Reversed hazard rate (RHR) function is an important measure as a tool in the analysis of the reliability of both natural and man-made systems. Recently, the properties of the RHR have attracted considerable interest from researchers (see for examples [6] and [11]). The RHR function is defined as $r(x) = f(x)/F(x)$. Thus, the RHR function of ZBLN distribution is given by

$$r(x) = \frac{\phi\left(\frac{\log x-\mu}{\sigma}\right)\left\{-\log\left[1-\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right]\right\}^{\alpha-1}}{\sigma x\ \gamma\left(\alpha, -\log\left[1-\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right]\right)}.$$

5.4. Conditional moments

For lifetime distributions, it is of greater interest to know the conditional moments which are important in prediction. The conditional moments of any distribution is defined as

$$E(X^r|X>t) = \frac{1}{S(t)}\int_t^\infty x^r f(x)dx.$$

Thus, the conditional moments of ZBLN distribution is given by

$$E(X^r|X>t) = \frac{1}{S(t)}\sum_{k=0}^{\infty}\left(\frac{\alpha+k}{\sigma}\right)\ b_k\ I_1(r,k), \tag{5.2}$$

where $S(.)$ is the survival function, b_k is given in Equation (2.3) and $I_1(r,k)$ is given as

$$I_1(r,k) = \int_t^\infty x^{r-1}\ \phi\left(\frac{\log x-\mu}{\sigma}\right)\left[\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right]^{\alpha+k-1} dx. \tag{5.3}$$

5.5. Vitality function

In modeling lifetime data, the vitality function is a very valuable tool. This function plays important role in reliability engineering, biomedical science, and survival analysis. It is worth mentioning that the rapid aging of a component needs to low vitality relatively, whereas high vitality implies relatively slow aging during the given time period. For more details on the vitality function see [16].

For $r = 1$, in Equation (5.2), gives the vitality function of ZBLN distribution, and is given by

$$V(t) = E(X|X>t) = \frac{1}{S(t)}\int_t^\infty xf(x)dx = \frac{1}{S(t)}\sum_{k=0}^{\infty}\left(\frac{\alpha+k}{\sigma}\right)b_k\ I_1(1,k), \tag{5.4}$$

where $I_1(1,k)$ is obtained by putting $r = 1$ in Equation (5.3), and is given by

$$I_1(1,k) = \int_t^\infty \phi\left(\frac{\log x-\mu}{\sigma}\right)\left[\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right]^{\alpha+k-1} dx.$$

5.6. Geometric vitality function

The concept of geometric vitality function is based on the geometric mean of the residual lifetime. If X be a random variable that represents the lifetime of a component, then $\log G(t) = E(\log X|X > t)$ represents the geometric mean of lifetimes of components that have survived up to time t. For a non-negative random variable X follows an absolutely continuous distribution function, with $E(\log X) < 1$, the geometric vitality function is defined as

$$\log G(t) = E(\log X|X > t) = \frac{1}{S(t)} \int_t^\infty \log x \ f(x)dx.$$

Now, the geometric vitality function of the ZBLN distribution is given by

$$\log G(t) = \frac{1}{S(t)} \sum_{k=0}^\infty (\alpha + k) \ b_k \ I_2(k),$$

where $I_2(k)$ can be expressed as

$$I_2(k) = \int_t^\infty \left(\frac{\log x}{\sigma x}\right) \phi\left(\frac{\log x - \mu}{\sigma}\right) \left[\Phi\left(\frac{\log x - \mu}{\sigma}\right)\right]^{\alpha+k-1} dx.$$

5.7. Moments of residual life

In reliability theory, the concept of residual life is very noteworthy. It represents the life remaining in a unit after it has attained age t.

The r^{th} order moment of the residual life of the ZBLN distribution is given as

$$\begin{aligned}\mu_r(t) = E[(X-t)^r|X > t] &= \frac{1}{S(t)} \int_t^\infty (x-t)^r \ f(x)dx \\ &= \frac{1}{S(t)} \sum_{i=0}^r \binom{r}{i} (-1)^{r-i} \ t^{r-i} \int_t^\infty x^i \ f(x)dx,\end{aligned}$$

which can be simplified as

$$\mu_r(t) = \frac{1}{S(t)} \sum_{i=0}^r \sum_{k=0}^\infty \binom{r}{i} (-1)^{r-i} \ t^{r-i} \left(\frac{\alpha + k}{\sigma}\right) \ b_k \ I_1(i, k),$$

where $I_1(r, k)$ is given in Equation (5.3). Now, for $r = 1$ and using Equation (2.5), we get the expression for mean residual life (MRL) function, and is given by

$$\begin{aligned}\mu_1(t) = E(X - t|X > t) &= \frac{1}{S(t)} \int_t^\infty (x-t) \ f(x)dx \\ &= \frac{1}{S(t)} \sum_{k=0}^\infty \left(\frac{\alpha + k}{\sigma}\right) \ b_k \ I_3(k),\end{aligned}$$

where

$$I_3(k) = \int_t^\infty \frac{(x-t)}{x} \phi\left(\frac{\log x - \mu}{\sigma}\right) \left[\Phi\left(\frac{\log x - \mu}{\sigma}\right)\right]^{\alpha+k-1} dx.$$

Hence, $\mu_1(t)$ also gets the form, $\mu_1(t) = V(t) - t$, where $V(t)$ is given in Equation (5.4). Similarly, the second moment of the residual lifetime of the ZBLN distribution is given by

$$\mu_2(t) = \frac{1}{S(t)} \sum_{k=0}^{\infty} \left(\frac{\alpha+k}{\sigma}\right) b_k \; I_1(2,k) - \frac{2t\;V(t)}{S(t)} + t^2,$$

where $I_1(2,k)$ is given as

$$I_1(2,k) = \int_t^{\infty} x\;\phi\left(\frac{\log x - \mu}{\sigma}\right)\left[\Phi\left(\frac{\log x - \mu}{\sigma}\right)\right]^{\alpha+k-1} dx.$$

Thus, the variance of the residual life function of the ZBLN distribution can be obtained using $\mu_1(t)$ and $\mu_2(t)$.

5.8. Moments of reversed residual life

The r^{th} order moment of the reversed residual life of the ZBLN distribution is given by

$$\begin{aligned} m_r(t) = E[(t-X)^r | X \leq t] &= \frac{1}{F(t)} \int_0^t (t-x)^r \; f(x) dx \\ &= \frac{1}{F(t)} \sum_{i=0}^{r} \binom{r}{i} (-1)^i \; t^{r-i} \int_0^t x^i \; f(x) dx. \end{aligned}$$

On simplification, $m_r(t)$ gets the form

$$m_r(t) = \frac{1}{F(t)} \sum_{i=0}^{r} \sum_{k=0}^{\infty} \binom{r}{i} (-1)^{r-i} \; t^{r-i} \left(\frac{\alpha+k}{\sigma}\right) \; b_k \; I_4(,k), \tag{5.5}$$

where $I_4(i,k)$ is given as

$$I_4(i,k) = \int_0^t x^{i-1} \;\phi\left(\frac{\log x - \mu}{\sigma}\right)\left[\Phi\left(\frac{\log x - \mu}{\sigma}\right)\right]^{\alpha+k-1} dx.$$

Now, the mean $(m_1(t))$ and second moment $(m_2(t))$ of the reversed residual life of the ZBLN distribution can be obtained by setting $r = 1, 2$; respectively in Equation (5.5). Again, using $m_1(t)$ and $m_2(t)$, one can obtain the variance of the reversed residual life function of the distribution.

6. RÉNYI ENTROPY

Entropy is considered to be the measure of uncertainty of a system and it is typically used in physical sciences. The study of entropy has gained momentum in the theoretical perspective as well as in terms of its applications in the field of applied research. Among the number of entropies available in the literature, one of the most popular entropy measures is Rényi entropy (see [23]).

By definition, for any random variable Y with pdf $g(y)$, the Rényi entropy is defined as

$$H_\gamma(y) = \frac{1}{1-\gamma} \log \int_{\mathbb{R}} g^\gamma(y)dy; \quad \text{for} \quad \gamma > 0 \quad \text{and} \quad \gamma \neq 1.$$

Let $f(x)$ be the density function of the ZBLN distribution, then standard calculations show that the Rényi entropy of the distribution can be written as

$$H_\gamma(x) = \frac{1}{1-\gamma} \log \int_0^\infty f^\gamma(x)dx$$

in which, by using (2.1),

$$\int_0^\infty f^\gamma(x)dx = \left(\frac{1}{\sigma\,\Gamma(\alpha)}\right)^\gamma \int_0^\infty \tau^\gamma(x)dx,$$

where

$$\tau^\gamma(x) = \left\{\frac{1}{x}\,\phi\left(\frac{\log x-\mu}{\sigma}\right)\left\{-\log\left[1-\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right]\right\}^{\alpha-1}\right\}^\gamma.$$

On simplification, the Rényi entropy of ZBLN distribution gets the expression

$$H_\gamma(x) = (1-\gamma)^{-1}\log\int_0^\infty \tau^\gamma(x)dx - \gamma(1-\gamma)^{-1}\log(\sigma) - \gamma(1-\gamma)^{-1}\log(\Gamma(\alpha)).$$

7. INEQUALITY MEASURES

Lorenz and Bonferroni curves are income inequality measures that are widely useful and applicable to some other areas including reliability, demography, medicine, and insurance. Also, the Zenga curve introduced by Zenga (2007) (see [25]) is another widely used inequality measure. The Lorenz, Bonferroni, and Zenga curves for the ZBLN distribution will be derived in this section. The Lorenz curve is defined by

$$L_F(x) = \frac{1}{E(X)}\int_0^x t\ f(t)dt.$$

Simple algebra provides the Lorenz curve for ZBLN distribution, and is given by

$$L_F(x) = \frac{\sum\limits_{k_1=0}^{\infty} (\alpha+k_1)\ b_{k_1}\ I_4(k_1)}{\sum\limits_{k_2=0}^{\infty} (\alpha+k_2)\ b_{k_2}\ I_5(k_2)},$$

where b_k is given in equation (2.3),

$$I_4(k_1) = \int_0^x \phi\left(\frac{\log t-\mu}{\sigma}\right)\left[\Phi\left(\frac{\log t-\mu}{\sigma}\right)\right]^{\alpha+k_1-1}, \quad \text{and}$$

$$I_5(k_2) = \int_0^\infty \phi\left(\frac{\log x-\mu}{\sigma}\right)\left[\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right]^{\alpha+k_2-1}.$$

Also, the Bonferroni curve is defined by

$$B_F(x) = \frac{1}{E(X)F(x)} \int_0^x t\ f(t)dt.$$

Thus, the Bonferroni curve of ZBLN distribution gets expression given by

$$B_F(x) = \frac{\sum\limits_{k_1=0}^{\infty} (\alpha + k_1)\ b_{k_1}\ I_4(k_1)}{\left\{\sum\limits_{k_2=0}^{\infty} (\alpha + k_2)\ b_{k_2}\ I_5(k_2)\right\}\left\{\sum_{k=0}^{\infty} b_k \left[\Phi\left(\frac{\log x - \mu}{\sigma}\right)\right]^{\alpha+k}\right\}}.$$

Now, the Zenga curve is defined as

$$A_F(x) = 1 - \frac{\mu^-(x)}{\mu^+(x)}, \tag{7.1}$$

where

$$\mu^-(x) = \frac{1}{F(x)} \int_0^x t\ f(t)dt, \quad \text{and} \quad \mu^+(x) = \frac{1}{S(x)} \int_x^\infty t\ f(t)dt.$$

Therefore, $\mu^-(x)$ and $\mu^+(x)$ of ZBLN distribution are respectively given by

$$\mu^-(x) = \frac{\sum\limits_{k_1=0}^{\infty} \left(\frac{\alpha+k_1}{\sigma}\right) b_{k_1}\ I_4(k_1)}{\sum\limits_{k=0}^{\infty} b_k \left[\Phi\left(\frac{\log x - \mu}{\sigma}\right)\right]^{\alpha+k}}, \quad \text{and} \quad \mu^+(x) = V(x),$$

where $V(x)$ is the vitality function of ZBLN distribution in x, such that the expression for vitality function of the distribution is given in (5.4). Substituting the values of $\mu^-(x)$ and $\mu^+(x)$ in (7.1), gets the expression of $A_F(x)$ for the ZBLN distribution.

8. ORDER STATISTICS

Let $X_1, X_2, ..., X_n$ be a random sample from the ZBLN distribution and its order statistics is $X_{1:n}, X_{2:n}, ..., X_{n:n}$. Let $F_{i:n}(x)$ and $f_{i:n}(x)$ denote the cdf and pdf of the i^{th} order statistic $X_{i:n}$, respectively. Hence, using the standard expressions of order statistics, $F_{i:n}(x)$ and $f_{i:n}(x)$ of ZBLN distribution is respectively given by

$$F_{i:n}(x) = \sum_{j=i}^{n} \binom{n}{j} F^j(x)\ [1 - F(x)]^{n-j},$$

and

$$\begin{aligned}
f_{i:n}(x) &= \frac{n!}{(i-1)!\;(n-i)!}\;[F(x)]^{i-1}\;[1-F(x)]^{n-i}\;f(x)\\
&= \frac{1}{\mathcal{B}(i,n-i+1)}\sum_{k_3=0}^{n-i}(-1)^{k_3}\binom{n-i}{k_3}[F(x)]^{k_3+i-1}f(x)\\
&= \frac{1}{\mathcal{B}(i,n-i+1)}\sum_{k_3=0}^{n-i}(-1)^{k_3}\binom{n-i}{k_3}\\
&\quad\times\left\{\sum_{k=0}^{\infty}b_k\left[\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right]^{\alpha+k}\right\}^{k_3+i-1}\\
&\quad\times\sum_{k=0}^{\infty}b_k\frac{(\alpha+k)}{\sigma x}\phi\left(\frac{\log x-\mu}{\sigma}\right)\left[\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right]^{\alpha+k-1},
\end{aligned}$$

where $\mathcal{B}(a,b)=\Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the Beta function. Now, for $i=1$ and n, one can get the pdf of $X_{(1)}=min\{X_1,X_2,...,X_n\}$ and $X_{(n)}=max\{X_1,X_2,...,X_n\}$ for ZBLN distribution, respectively.

9. ESTIMATION OF PARAMETERS

In this section, we'll look at how to estimate the parameters of the ZBLN distribution using two widely used methods: maximum likelihood (ML) and Bayesian methods.

9.1. Maximum likelihood estimation

This subsection considers the maximum likelihood estimation for the ZBLN model parameters α,μ, and σ. Let $X_1,X_2,...,X_n$ be a random sample taken from the ZBLN distribution, and $x_1,x_2,...,x_n$ are the corresponding observed values. Then the log-likelihood function can be expressed as

$$\begin{aligned}
\mathcal{L}_n = &-n\log(\sigma)-n\log(\Gamma(\alpha))-\sum_{i=1}^{n}\log(x_i)+\sum_{i=1}^{n}\log\left[\phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)\right]\\
&+\;(\alpha-1)\sum_{i=1}^{n}\log\left\{-\log\left[1-\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right]\right\}.
\end{aligned}$$

The score function associated with the log-likelihood function is

$$\mathbf{U}=\left(\frac{\partial\mathcal{L}_n}{\partial\alpha},\frac{\partial\mathcal{L}_n}{\partial\mu},\frac{\partial\mathcal{L}_n}{\partial\sigma}\right)^T.$$

Now, by solving $\frac{\partial\mathcal{L}_n}{\partial\alpha}=0$, $\frac{\partial\mathcal{L}_n}{\partial\mu}=0$ and $\frac{\partial\mathcal{L}_n}{\partial\sigma}=0$, we get the associated nonlinear log-likelihood equations and are respectively given by

$$\sum_{i=1}^{n}\log\left\{-\log\left[1-\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right]\right\}-n\;\psi(\alpha)=0, \tag{9.1}$$

$$\sum_{i=1}^{n}\frac{\log(x_i)-\mu}{\sigma^2}+\left(\frac{\alpha-1}{\sigma}\right)\sum_{i=1}^{n}\frac{\phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)}{\left[1-\Phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)\right]\log\left[1-\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right]}=0, \tag{9.2}$$

$$\frac{-n}{\sigma}+\sum_{i=1}^{n}\frac{(\log(x_i)-\mu)^2}{\sigma^3}+\sum_{i=1}^{n}\frac{(\alpha-1)\left(\frac{\log(x_i)-\mu}{\sigma}\right)\phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)}{\left[1-\Phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)\right]\log\left[1-\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right]}=0, \tag{9.3}$$

where $\psi(\alpha) = d\{\log\Gamma(\alpha)\}/d\alpha$ is the digamma function. Now, by solving the equations (9.1), (9.2) and (9.3) simultaneously, we obtain the maximum likelihood estimators (MLEs) $(\hat{\alpha}, \hat{\mu}, \hat{\sigma})$ of the model parameters (α, μ, σ).

Now, we construct the asymptotic confidence intervals for parameters α, μ and σ. On taking the second partial derivatives of equations (9.1), (9.2) and (9.3), the Hessian matrix of ZBLN distribution can be obtained, and denoted as $H(\Theta)$, where $\Theta = \{\alpha, \mu, \sigma\}$. Now, the observed Fisher's information matrix $J(\Theta)$ can be obtained by taking the negative of Hessian matrix. That is, $J(\Theta) = -H(\Theta)$. Hence, the inverse of observed Fisher's information matrix will provide the variance-covariance matrix of the MLEs, which is given by

$$\Sigma = J^{-1}(\Theta) = \{\Sigma_{ij},\ i,j = 1,2,3\},$$

and $\Sigma_{ij} = \Sigma_{ji}$ for $i \neq j = 1,2,3$. Again, it is well established that the MLEs are asymptotically normally distributed. That is, $\sqrt{n}(\Theta - \hat{\Theta}) \sim N_3(0,\Sigma)$, where n is the sample size and $\hat{\Theta}$ is the MLEs of Θ.

Thus, we obtain $100 \times (1-\delta)\%$ asymptotic confidence intervals of the parameters using the following formulae:

$$\alpha \in \left\{\hat{\alpha} \mp Z_{\delta/2}\sqrt{\Sigma_{11}}\right\},\ \mu \in \left\{\hat{\mu} \mp Z_{\delta/2}\sqrt{\Sigma_{22}}\right\},\ \text{and } \sigma \in \left\{\hat{\sigma} \mp Z_{\delta/2}\sqrt{\Sigma_{33}}\right\},$$

where Z_δ is the upper δ^{th} percentile of the standard normal distribution.

9.2. Bayesian estimation

The Bayesian analysis for the ZBLN model parameters is performed in this subsection. Each parameter should have a prior density in order to do so. For this, we utilize two types of priors: half-Cauchy (HC) and normal (N) priors. The pdf of the HC distribution with scale parameter a is defined as

$$f_{HC}(x_*) = \frac{2a}{\pi(x_*^2 + a^2)}, \quad x_* > 0,\ a > 0.$$

The HC distribution has no mean or variance. Meanwhile, its mode is equal to 0. Since the pdf of the HC is virtually flat but not totally flat at scale value equals 25, which verges on acquiring adequate information for the numerical approximation algorithm to continue looking at the target posterior pdf, the HC distribution with $a = 25$ is recommended as a noninformative prior. Gelman and Hill (2006) ([8]) suggested that the uniform distribution, or whether more information is required, is a superior alternative to the HC distribution. As a result, for the parameters α and σ, the HC distribution with $a = 25$ is chosen as a noninformative prior distribution in this article. Thus, we set the prior distributions of the parameters to be $\mu \sim N(0, 1000)$, and $\alpha, \sigma \sim HC(25)$. Thus, we obtain the joint posterior pdf as given by

$$\pi(\mu, \alpha, \sigma|x) \propto L_n \times \pi(\mu) \times \pi(\alpha) \times \pi(\sigma), \tag{9.4}$$

where L_n is the likelihood funtion for ZBLN distribution. From Equation (9.4), it is obvious that there is no analytical solution to find out the Bayesian estimates. Thus, we use a remarkable method of simulation, namely the Metropolis-Hastings algorithm of the Markov Chain Monte Carlo (MCMC) method.

9.3. Bootstrap confidence intervals

In this subsection, we use the parametric bootstrap method to approximate the distribution of the maximum likelihood estimators of the ZBLN parameters. Then, we can use the bootstrap distribution to estimate the confidence intervals on each parameter of the fitted ZBLN distribution. Let $\hat{\Theta} = \Theta(X)$ be a ML estimator of the set of parameters of interest $\Theta = \{\alpha, \mu, \sigma\}$ using a given dataset $X = \{x_1, x_2, ..., x_n\}$. The bootstrap is a method to estimate the distribution of the statistic $\hat{\Theta}$ by getting a random sample $\Theta_1^*, \Theta_2^*, ..., \Theta_B^*$ for Θ based on B random samples that are drawn with replacement from the original data $X = \{x_1, x_2, ..., x_n\}$ (see [24]). The bootstrap sample $\Theta_1^*, \Theta_2^*, ..., \Theta_B^*$ can be used to construct bootstrap confidence intervals for the parametric set $\Theta = \{\alpha, \mu, \sigma\}$ of ZBLN distribution.

Thus, we obtain $100 \times (1 - \delta)\%$ bootstrap confidence intervals of the parameters using the following formulae:

$$\alpha \in \left\{\hat{\alpha} \mp z_{\delta/2}\ \hat{se}_{\alpha,boot}\right\},\ \mu \in \left\{\hat{\mu} \mp z_{\delta/2}\ \hat{se}_{\mu,boot}\right\},\ \text{and}\ \sigma \in \left\{\hat{\sigma} \mp z_{\delta/2}\ \hat{se}_{\sigma,boot}\right\},$$

where z_δ denotes the δ^{th} percentile of the bootstrap sample and for $\Theta = \{\alpha, \mu, \sigma\}$

$$\hat{se}_{\Theta,boot} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\Theta_b^* - \frac{1}{B}\sum_{b=1}^{B}\Theta_b^*\right)^2}.$$

10. PERFORMANCE OF THE ESTIMATORS USING SIMULATION STUDY

In this section, we conduct simulation experiments to assess the long-run performances of MLEs and Bayesian estimates of the ZBLN parameters for some finite sample sizes. We have simulated datasets of sizes $n = 50, 100$, and 250 from the ZBLN distribution for the parameter values $\alpha = 0.2,\ \mu = 3.5,\ \sigma = 0.5$ and iterated each sample for 500 times. Then, we compute the average biases and MSEs for the MLEs to all replications in the relevant sample sizes.

That is, the analysis computes the values by the given formulae. The equation for average bias of the simulated estimates equals $\frac{1}{500}\sum_{i=1}^{500}(\hat{\Theta}_i - \Theta)$, and the equation for average

MSE of the simulated estimates equals $\frac{1}{500}\sum_{i=1}^{500}(\hat{\Theta}_i - \Theta)^2$, where $\hat{\Theta} = (\hat{\alpha}, \hat{\mu}, \hat{\sigma})$ are estimates of the parameter vector $\Theta = (\alpha, \mu, \sigma)$. The results to the simulation for MLEs are reported in Table 1. It can be concluded that the M.S.E of all the estimators decreases with increasing sample size. This shows the consistency of the estimators.

Table 1: Estimates, Average bias and MSE values of MLEs from simulation of the ZBLN distribution.

Parameters	**Sample Size**	**Estimates**	**Bias**	**M.S.E**
α	50	0.6892	0.4892	1.9199
	100	0.4514	0.2514	0.7520
	250	0.2597	0.0597	0.0384
μ	50	3.0965	−0.4035	1.1623
	100	3.2030	−0.2970	0.7376
	250	3.3702	−0.1298	0.2029
σ	50	0.5418	0.0418	0.0385
	100	0.5345	0.0345	0.0297
	250	0.5215	0.0215	0.0112

Now, in the case of Bayesian simulation, we consider the prior distributions for the ZBLN parameters as given in Subsection 9.2. For the respective sample sizes, the posterior summary results such as mean, standard deviation (SD), Monte Carlo error (MCE), 95% confidence interval (CI), and median are presented in Table 2. It is observed that the SD and MCE decrease as the sample size increases, which predicts the consistency of Bayesian estimates of the ZBLN distribution parameters.

Table 2: Posterior summary results for Bayesian simulation.

Parameters	n	**Mean**	**SD**	**MCE**	95% **CI**	**Median**
α	50	1.6267	1.5925	1.0016	(0.4615, 5.0629)	1.4301
	100	0.5894	0.2646	0.1755	(0.1748, 0.8089)	0.7889
	250	0.2115	0.0919	0.0458	(0.1748, 0.4371)	0.1931
μ	50	2.5282	0.7701	0.5653	(1.1535, 3.4592)	2.3215
	100	2.9731	0.3274	0.2064	(2.7282, 3.5331)	2.8252
	250	3.4679	0.1632	0.0813	(3.0676, 3.5331)	3.4131
σ	50	0.6999	0.1494	0.0667	(0.4728, 0.8246)	0.7105
	100	0.6444	0.1065	0.0591	(0.4704, 0.7237)	0.6218
	250	0.4959	0.0640	0.0319	(0.4704, 0.6530)	0.5715

11. APPLICATION AND EMPIRICAL STUDY

To demonstrate the applicability of the ZBLN distribution, we consider a real dataset based on a cancer survival study, and the parameters are estimated by using maximum likelihood, and the Bayesian estimation methods to compare the data modeling ability of the ZBLN distribution over some competitive distributions. The dataset is taken from Lee & Wang (2003) (see [17]), which corresponds to the remission times (in months) of a random sample of 128 bladder cancer patients. The summary statistics of the dataset is given in Table 3.

Table 3: Summary statistics of real dataset.

n	M	Md	SD	Sk	Ku	min	max
128	9.2094	6.28	10.4026	3.3987	16.3942	0.08	79.05

Now, we study the empirical hazard function of the datasets using the concept of total time on test (TTT) plot. The TTT plot is a graph that mainly serves to discriminate between different types of aging represented in hazard rate shapes. For details, the readers are referred to [1]. The TTT plot is drawn by plotting

$$T\left(\frac{i}{n}\right) = \frac{\sum\limits_{r=1}^{i} x_{r:n} + (n-i)x_{i:n}}{\sum\limits_{r=1}^{n} x_{r:n}}$$

against i/n, where $i = 1, 2, ..., n$ and $x_{r:n}$, $r = 1, 2, ..., n$ are the order statistics of the sample. Figure 3 indicates that the above-given dataset has an upside-down bathtub shape for the empirical hazard function. Therefore, the ZBLN distribution can be a credible pick for the given dataset, since its hazard function satisfies the upside-down bathtub shape.

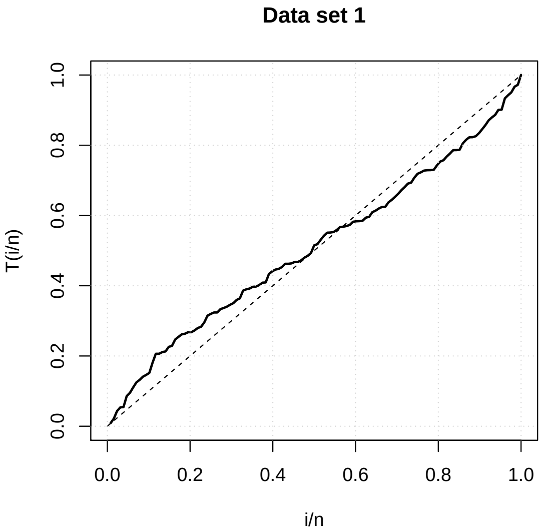


Figure 3: The TTT plot of real dataset.

11.1. Maximum likelihood estimation

To illustrate the potentiality of the ZBLN distribution, the following distributions are considered for comparison.

- The two-parameter Log-normal (LN) distribution with pdf
$$f(x) = \frac{1}{\sqrt{2\pi}\ \sigma x}\ \exp\left[-\frac{(\log x - \mu)^2}{2\sigma^2}\right], \quad x > 0,\ \mu \in \mathbb{R},\ \sigma > 0.$$
- The Exponentiated Log-normal (ELN) distribution with pdf
$$f(x) = \frac{\alpha}{x\sigma}\phi\left(\frac{\log x - \mu}{\sigma}\right)\left[\Phi\left(\frac{\log x - \mu}{\sigma}\right)\right]^{\alpha-1}, \ x > 0,\ \mu \in \mathbb{R},\ \alpha, \sigma > 0.$$
- The Weibull distribution with pdf
$$f(x) = \frac{\alpha}{\sigma}\left(\frac{x}{\sigma}\right)^{\alpha-1} e^{-(x/\sigma)^{\alpha}}, \quad x > 0,\ \alpha, \sigma > 0.$$
- The New Generalized Lindley distribution (NGLD) (see [7]) with pdf
$$f(x) = \frac{e^{-\mu x}}{1+\mu}\left(\frac{\mu^{\alpha+1}x^{\alpha-1}}{\Gamma(\alpha)} + \frac{\mu^{\sigma}x^{\sigma-1}}{\Gamma(\sigma)}\right), \quad x > 0,\ \alpha, \mu, \sigma > 0.$$
- The Zografos-Balakrishnan Lindley distribution (ZBLD) (see [13]) with pdf
$$f(x) = \frac{1}{\Gamma(\alpha)}\left[\log\left(\frac{1+\theta}{1+\theta+\theta x}e^{\theta x}\right)\right]^{\alpha-1}\frac{\theta^2}{\theta+1}(1+x)e^{-\theta x}, \quad x > 0,\ \alpha, \sigma > 0.$$

We apply the following statistical tools in order to find out the goodness-of-fit of distributions to the real dataset; log-likelihood (LL), Kolmogorov-Smirnov (KS), Cramér-von Misses (W^*), Anderson-Darling (A^*) statistics, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) values, and are presented in Table 4. We use the RStudio software for numerical evaluations.

Moreover, Table 4 shows the MLEs and goodness-of-fit statistics of the distributions for the corresponding dataset. It can be seen that the KS, W^*, A^*, AIC, and BIC values of the ZBLN distribution are smaller than that of other distributions. We also present other important graphs which consist of empirical density plot, empirical cdf plot, Q-Q, and P-P plots for the real dataset in Figure 4. It again gives some superimposed curves of those fitted and empirical functions. Thus, we conclude that the ZBLN is the most suitable distribution for the given dataset while comparing other distributions.

We also utilized the likelihood ratio (LR) test for comparing ZBLN distribution having additional parameter α with LN distribution. That is, we test H_0 : LN against H_A : ZBLN and obtain critical values for the LR test statistics for the cancer dataset. Thus we get the LR test statistic value as 6.663 and the corresponding p-value as 0.0098 for the given dataset. Given the value of the test statistics and the associated p-value, we reject the null hypotheses for the dataset and conclude that the ZBLN model provides a significantly better representation for the dataset than the LN distribution.

Table 4: Maximum-likelihood estimates, goodness-of-fit statistics, AIC and BIC values based on the bladder cancer dataset.

Estimates	**LN**	**ELN**	**Weibull**	**NGLD**	**ZBLD**	**ZBLN**
$\hat{\alpha}$	—	0.1516	1.0514	1.1852	0.7353	0.2425
$\hat{\mu}$	1.7422	3.0494	—	0.1287	—	2.9666
$\hat{\sigma}$	1.0646	0.5404	9.4172	1.1850	0.1569	0.6730
LL	-412.6565	-410.0441	-411.8925	-411.0846	-413.5513	-409.3414
KS	0.0644	0.0562	0.0721	0.0760	0.0901	0.0542
W^*	0.1313	0.0846	0.1666	0.1416	0.2230	0.0736
A^*	0.8708	0.5589	1.0488	0.8235	1.2465	0.4828
AIC	829.3131	826.0883	827.7849	828.1691	831.1025	824.6828
BIC	835.0171	834.6444	833.4890	836.7252	836.8066	833.2389

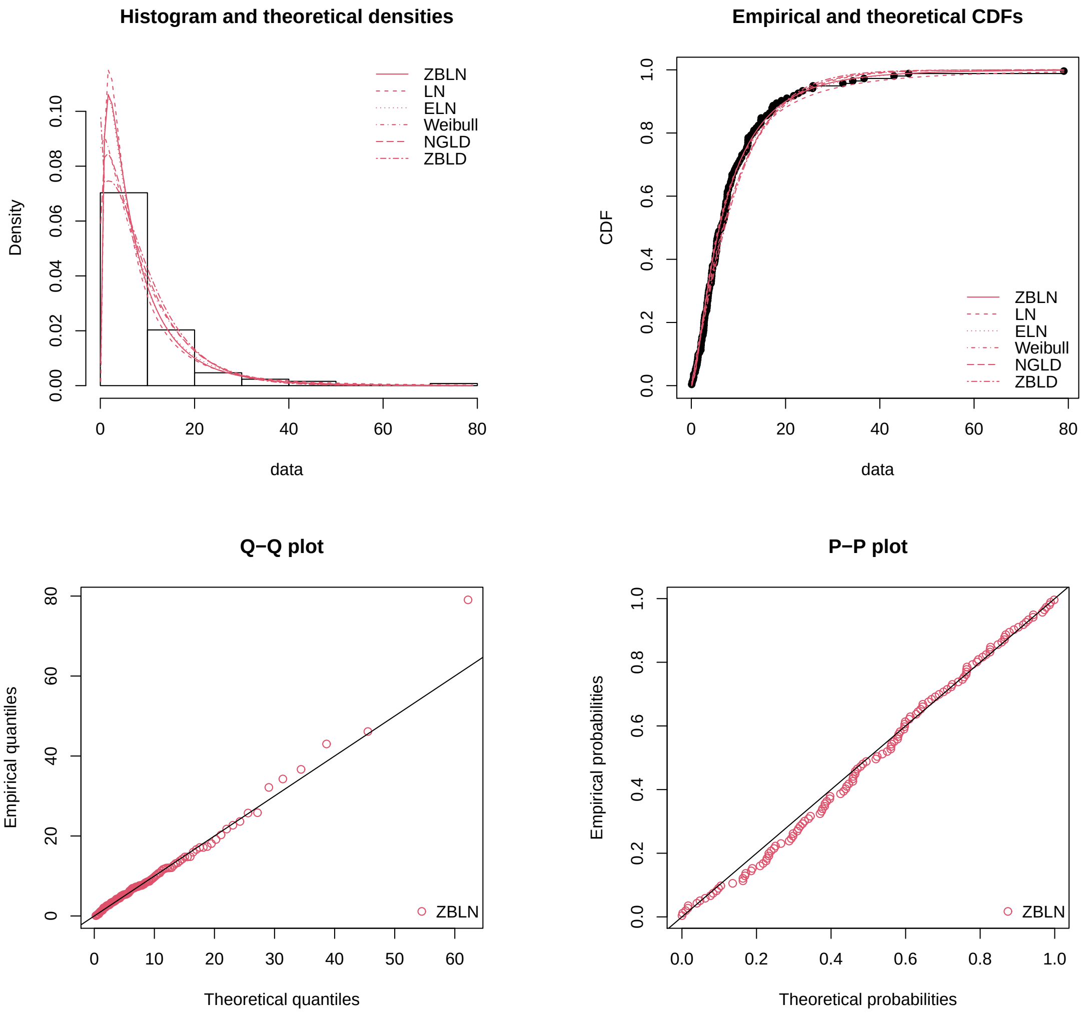


Figure 4: Various empirical plots of bladder cancer dataset.

Now, the Hessian matrix corresponding to real dataset is obtained as

$$H(\Theta) = \begin{pmatrix} 2330.9119 & 794.1654 & -1451.9253 \\ 794.1654 & 131391.5635 & -191.6978 \\ -1451.9253 & -191.6978 & 132403.6280 \end{pmatrix}.$$

Hence, the asymptotic variance-covariance matrix for real dataset is obtained as

$$\Sigma = \begin{pmatrix} 0.0793 & -0.0065 & 0.0118 \\ -0.0065 & 0.0189 & -0.0005 \\ 0.0118 & -0.0005 & 0.0189 \end{pmatrix}.$$

Again, the 95% asymptotic confidence intervals of the ZBLN parameters are given in Table 5.

Table 5: The 95% asymptotic confidence intervals of the ZBLN parameters based on bladder cancer dataset.

Parameter	**Lower**	**Upper**
α	0.2017	0.2833
μ	2.9612	2.9720
σ	0.6676	0.6784

Now, we use the obtained MLEs to derive the 95% bootstrap confidence intervals for the parameters α, μ, and σ. We simulate 1001 samples of size as in the real dataset we studied, from ZBLN distribution with true values of the parameters taken as MLEs of the parameters. For each obtained sample, we have estimated the MLEs $\hat{\alpha}_b^*$, $\hat{\mu}_b^*$, and $\hat{\sigma}_b^*$, for $b \in \{1, 2, ..., 1001\}$. The median and 95% bootstrap confidence interval for parameters α, μ, and σ of the given dataset is presented in Table 6. It is also interesting to look at the joint distribution of the bootstrapped values in a matrix of scatter plots in order to understand the potential structural correlation between parameters. The plots in Figure 5 consist of matrix scatterplots of the bootstrapped values of ZBLN parameters providing a representation of the joint uncertainty distribution of the fitted parameters.

Table 6: The median and 95% bootstrap confidence interval for ZBLN parameters of the bladder cancer dataset.

Parameter	**Median**	**Bootstrap CI**
α	0.2294	(0.1203, 3.9249)
μ	2.9653	(−0.4722, 3.3680)
σ	0.6549	(0.4867, 1.2079)

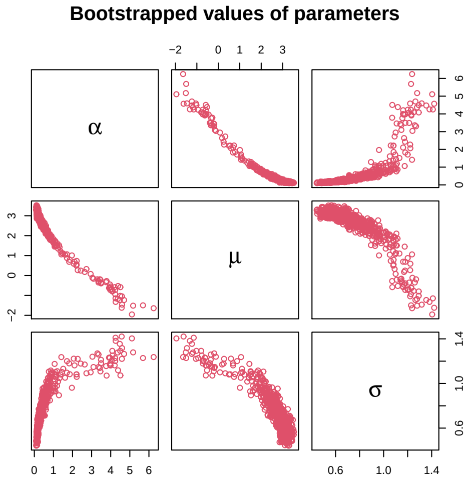


Figure 5: Matrix scatter plots of bootstrappped values of ZBLN parameters due to the bladder cancer dataset.

11.2. Bayesian estimation

Here, we focus on estimating the parameters of the ZBLN distribution using the Bayesian procedure based on the same univariate bladder cancer survival dataset which we discussed in the above subsection. In the context of Bayesian estimation, the analysis was performed using the Metropolis-Hastings algorithm of the MCMC method with 1001 iterations. For comparing Bayes estimates with the MLEs, both the estimates of the ZBLN parameters with corresponding standard error (SE) and Monte Carlo standard error (MCSE) for the real dataset are given in Table 7. The numerical computations on Bayesian estimation are also done using RStudio software.

Table 7: MLEs and Bayesian estimates of the ZBLN parameters on bladder cancer dataset.

Parameter	MLE (SE)	Bayes (MCSE)
α	0.2425 (0.0208)	0.2471 (0.0402)
μ	2.9666 (0.0028)	3.0206 (0.10003)
σ	0.6730 (0.0028)	0.7127 (0.0343)

12. CONCLUDING REMARKS

In this paper, we studied a distribution that generalizes the log-normal distribution. We refer to the model as the Zografos-Balakrishnan log-normal (ZBLN) distribution and study its mathematical and statistical properties. We provide explicit expressions for the moments, quantile function, various reliability measures, Rényi entropy, and some inequality measures associated with the ZBLN distribution. It is worth noting that the hazard rate function supports all of the standard shapes, including increasing, decreasing, bathtub, and upside-down bathtub. The model parameters are estimated by using the Bayesian technique, and the method of maximum likelihood, and also, the observed information matrix is presented. Further, we adopt the parametric bootstrap technique to obtain confidence intervals for the model parameters. Moreover, the simulation studies based on the defined estimation methods are also done to confirm the parameter consistencies. The usefulness of the new model is illustrated by an application to the real dataset based on a cancer survival study using goodness-of-fit tests. The model provides a consistently better fit than other models available in the literature. We hope the model may attract wider applications for modeling positive real datasets in many areas such as physics, engineering, medicine, survival analysis, hydrology, economics, and so on.

ACKNOWLEDGMENTS

The authors would like to thank the Editor and the unknown reviewers for the constructive comments, which greatly improved the present version of this article.

REFERENCES

[1] Aarset, M.V. (1987). How to identify a bathtub hazard rate, *IEEE Transactions on Reliability*, **R-36**(1), 106–108.

[2] Beal, J. (2017). Biochemical complexity drives log-normal variation in genetic expression, *Engineering Biology*, **1**(1), 55–60.

[3] Bernardeau, F. and Kofman, L. (1995). Properties of the cosmological density distribution function, *The Astrophysical Journal*, **443**, 479.

[4] Blasi, P.; Burles, S.M. and Olinto, A.V. (1999). Cosmological magnetic field limits in an inhomogeneous universe, *The Astrophysical Journal*, **514**(2), L79–L82.

[5] Carvalho, J.; Piaggio, G.; Wojdyla, D.; Widmer, M. and Gülmezoglu, A. (2018). Distribution of postpartum blood loss: modeling, estimation and application to clinical trials, *Reproductive Health*, **15**(1), 199.

[6] Chandra, N. and Roy, D. (2001). Some results on reversed hazard rate, *Probability in the Engineering and Informational Sciences*, **15**(1), 95–102.

[7] Elbatal, I.; Merovci, F. and Elgarhy, M. (2013). A new generalized Lindley distribution, *Mathematical Theory and Modeling*, **3**, 30–47.

[8] Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models (Analytical Methods for Social Research)*, Cambridge University Press.

[9] Glass, B. (1954). Statistics for medical and other biological students. By L. Bernstein and M. Weatherall, *The Quarterly Review of Biology*, **29**(3), p. 303.

[10] Gupta, R.; Gupta, P. and Gupta, R. (1998). Modeling failure time data by Lehmann alternatives, *Communications in Statistics: Theory and Methods*, **27**(4), 887–904.

[11] Gupta, R. and Nanda, A. (2001). Some results on reversed hazard rate ordering, *Communications in Statistics: Theory and Methods*, **30**(11), 2447–2457.

[12] Gupta, R.D. and Kundu, D. (1999). Theory & Methods: generalized exponential distributions, *Australian & New Zealand Journal of Statistics*, **41**, 173–188.

[13] Irshad, M.R.; D'Cruz, V. and Maya, R. (2021). The Zografos-Balakrishnan Lindley distribution: properties and applications, *Statistica*, **81**(1), 45–64.

[14] Irshad, M.R.; Maya, R. and Krishna, A. (2021). Exponentiated power Muth distribution and associated inference, *Journal of the Indian Society for Probability and Statistics*, **22**, 265–302.

[15] Jobe, J.; Crow, E.L. and Shimizu, K. (1989). Lognormal distributions: theory and applications, *Technometrics*, **31**, 392.

[16] Kupka, J. and Loo, S. (1989). The hazard and vitality measures of ageing, *Journal of Applied Probability*, **26**(3), 532–542.

[17] Lee, E.T. and Wang, J.W. (2013). *Statistical Methods for Survival Data Analysis*, 4th ed., Wiley Publishing.

[18] Mudholkar, G. and Srivastava, D. (1993). Exponentiated Weibull family for analyzing bathtub failure-rate data, *Reliability, IEEE Transactions*, **42**(2), 299–302.

[19] Nadarajah, S. (2006). The exponentiated Gumbel distribution with climate application, *Environmetrics*, **17**, 13–23.

[20] Nadarajah, S.; Cordeiro, G. and Ortega, E. (2015). The Zografos-Balakrishnan-G family of distributions: mathematical properties and applications, *Communications in Statistics: Theory and Methods*, **44**(1), 186–215.

[21] Nadarajah, S. and Gupta, A. (2007). The exponentiated Gamma distribution with application to drought data, *Calcutta Statistical Association Bulletin*, **59**, 29–54.

[22] Parravano, A.; Sanchez, N. and Alfaro, E. (2012). The dependence of prestellar core mass distributions on the structure of the parental cloud, *Astrophysical Journal*, **754**, 150.

[23] Rényi, A. (1961). *On measures of entropy and information.* In: "Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability", Volume 1: Contributions to the Theory of Statistics, Berkeley, Calif. University of California Press, pp. 547–561.

[24] Wasserman, L. (2006). *All of Nonparametric Statistics*, Springer Texts in Statistics, Springer New York.

[25] Zenga, M. (2007). Inequality curve and inequality index based on the ratios between lower and upper arithmetic means, *Statistica & Applicazioni*, **4**, 3–27.

[26] Zografos, K. and Balakrishnan, N. (2009). On families of beta- and generalized gamma-generated distributions and associated inference, *Statistical Methodology*, **6**(4), 344–362.

REVSTAT – Statistical Journal
Volume 22, Number 1, January 2024, 25–43
https://doi.org/10.57805/revstat.v22i1.437

Bounds on Negative Binomial Approximation to Call Function

Author: Amit N. Kumar
– Department of Mathematical Sciences, Indian Institute of Technology (BHU),
Varanasi (Uttar Pradesh) – 221005, India
amit.kumar2703@gmail.com

Received: September 2021 Revised: March 2022 Accepted: March 2022

Abstract:

- In this paper, we develop Stein's method for negative binomial distribution using call function defined by $f_z(k) = (k - z)^+ = \max\{k - z, 0\}$, for $k \geq 0$ and $z \geq 0$. We obtain error bounds between $\mathbb{E}[f_z(\mathrm{N}_{r,p})]$ and $\mathbb{E}[f_z(V)]$, where $\mathrm{N}_{r,p}$ follows negative binomial distribution and V is the sum of locally dependent random variables, using certain conditions on moments. We demonstrate our results through an interesting application, namely, collateralized debt obligation (CDO), and compare the bounds with the existing bounds.

Keywords:

- *negative binomial distribution; call function; error bounds; Stein's method; CDO.*

AMS Subject Classification:

- Primary: 62E17, 62E20; Secondary: 60F05, 60E05.

1. INTRODUCTION

The call function is a non-negative real-valued function of the form

$$f_z(k) = (k-z)^+ = \max\{k-z, 0\}, \quad \text{for } k \geq 0 \text{ and } z \geq 0. \tag{1.1}$$

It has been used in several areas of probability and statistics, for example, finance, risk theory, and derivative pricing, among many others. In particular, it has been successfully applied to the collateralized debt obligation (CDO). For more details, see Karoui and Jiao [5], Karoui *et al.* [6], Hull and White [7], Neammanee and Yonghint [12], Yonghint *et al.* [17], and references therein.

For a random variable (rv) W, the study of $\mathbb{E}[f_z(W)]$ plays an important role in many real-life applications. For example, if W is the sum of Bernoulli random variables (rvs) then $\mathbb{E}[f_z(W)]$ is used to compute the mean value of total percentage loss for each tranche in CDO (see Neammanee and Yonghint [12], and Yonghint *et al.* [17] for details). Also, if W has a complicated structure, for example, W is the sum of locally dependent or independent (but non-identical) rvs, then $\mathbb{E}[f_z(W)]$ becomes difficult to compute in practice. In such cases, an approximation to standard and easy-to-use distribution is of interest. Approximation to call function has been studied by several authors in the literature, for example, Poisson approximation has been studied by Neammanee and Yonghint [12], and Yonghint *et al.* [17], and Normal approximation has been studied by Karoui and Jiao [5], and Karoui *et al.* [6].

In this paper, we study negative binomial (NB) approximation to call function using certain conditions on moments. The main advantage of NB distribution over Poisson distribution is the extra flexibility parameter that builds our bounds more shaper compare to the existing bounds for Poisson approximation. Throughout this paper, let $\mathrm{N}_{r,p}$ follow NB distribution with probability mass function

$$\mathbb{P}(\mathrm{N}_{r,p} = k) = \binom{r+k-1}{k} p^r q^k, \quad k \in \mathbb{Z}_+, \tag{1.2}$$

where $r > 1$, $q = 1-p \in (0,1)$ and $\mathbb{Z}_+ = \{0, 1, 2, ...\}$, the set of non-negative integers. From Neammanee and Yonghint [12], and Yonghint *et al.* [17], We observe that the call function can be studied under a locally dependent or independent setup. Therefore, we consider the following locally dependent structure that can be used for both cases.

Let J be a finite subset of $\mathbb{N} = \{1, 2, ...\}$, the set of all positive integers, and $\{\zeta_i\}_{i\in J}$ be a collection of non-negative rvs. For each i, let $i \in A_i \subseteq B_i \subset J$ be such that ζ_i is independent of $\zeta_{A_i^c}$ and ζ_{A_i} is independent of $\zeta_{B_i^c}$, where ζ_A is the collection of rvs $\{\zeta_i\}_{i\in A}$ and A^c denotes the complement of the set A. See Section 3 of Röllin [13] and Section 2 of Kumar [9] for a similar type of locally dependent structure. Define

$$V = \sum_{i\in J} \zeta_i, \tag{1.3}$$

the sum of locally dependent rvs. Note that if $A_i = B_i = \{i\}$ then V is the sum of independent rvs. Throughout this paper, we let $\zeta_A = \sum_{i\in A} \zeta_i$, for a set $A \subset J$, and $\mathcal{D}(W) := 2d_{TV}(W, W+1)$, for a rv W, where $d_{TV}(X,Y)$ denotes the total variation distance between X and Y.

In this paper, our aim is to study the proximity between $\mathbb{E}[(V-z)^+]$ and $\mathbb{E}[(\mathrm{N}_{r,p}-z)^+]$. That is, our interest is to obtain the upper bound for

$$\big|\mathbb{E}[(V-z)^+]-\mathbb{E}[(\mathrm{N}_{r,p}-z)^+]\big|. \tag{1.4}$$

We use Stein's method to obtain the bound for the above expression discussed in Section 2.

This paper is organized as follows. In Section 2, we develop Stein's method for NB distribution using the call function. In Section 3, we obtain uniform and non-uniform bounds for the expression given in (1.4) and compare our results with the existing results. In Section 3, we give an application of our results to CDO and give some numerical comparisons. Finally, in Appendix A, we give some inequalities and their proofs that are useful to develop Stein's method for NB distribution.

2. STEIN'S METHOD

Stein's method (Stein [14]) is a tool for obtaining error bounds between two probability distributions. This method is mainly based on obtaining the solution of the Stein equation given by

$$\mathcal{A}g(k)=f(k)-\mathbb{E}f(X), \quad \text{for } k\in\mathbb{Z}_+, \tag{2.1}$$

where $\mathcal{A}$ is a Stein operator for a rv X such that $\mathbb{E}[\mathcal{A}g(X)]=0$, f and g are real-valued bounded functions on $\mathbb{Z}_+$. Stein's method has been developed for NB distribution by Brown and Phillips [4] and Barbour *et al.* [1] for total variation distance and Wasserstein distance, respectively. In this section, we develop Stein's method for NB distribution when f is a call function, defined in (1.1), which is used to obtain upper bounds for the expression given in (1.4). The NB approximation via Stein's method has been studied by several authors such as Barbour *et al.* [1], Brown and Phillips [4], Vellaisamy *et al.* [15], Wang and Xia [16], Kumar and Upadhye [10], among many others.

Next, let $X=\mathrm{N}_{r,p}$ and $f=f_z$, defined in (1.1), then the Stein equation (2.1) leads to

$$\mathcal{A}g(k)=(k-z)^+-\mathbb{E}[(\mathrm{N}_{r,p}-z)^+], \quad \text{for } k\in\mathbb{Z}_+ \text{ and } z\geq 0. \tag{2.2}$$

Also, let $g=g_z$ be the solution of the above equation. Now, replacing k by V in (2.2) and taking expectation, we get

$$|\mathbb{E}[\mathcal{A}g_z(V)]|=\big|\mathbb{E}[(V-z)^+]-\mathbb{E}[(\mathrm{N}_{r,p}-z)^+]\big|. \tag{2.3}$$

Therefore, to obtain the upper bound for the expression given in (1.4), it is enough to obtain the upper bound for $|\mathbb{E}[\mathcal{A}g_z(V)]|$.

Next, the Stein operator of $\mathrm{N}_{r,p}$ is given by

$$\mathcal{A}g(k)=q(r+k)g(k+1)-kg(k), \quad \text{for } k\in\mathbb{Z}_+. \tag{2.4}$$

See Lemma 1 of Brown and Phillips [4] for details. Substituting (2.4) in (2.2), we get

$$q(r+k)g(k+1)-kg(k)=(k-z)^+-\mathbb{E}[(\mathrm{N}_{r,p}-z)^+]. \tag{2.5}$$

It can be easily verified that the solution of (2.5) is

$$g_z(k) = \begin{cases} 0 & \text{if } k = 0; \\ -\displaystyle\sum_{j=k}^{\infty} \frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k-1)} \frac{(k-1)!}{j!} q^{j-k} & \\ \quad \times[(j-z)^+ - \mathbb{E}[(\mathrm{N}_{r,p} - z)^+]] & \text{if } k \geq 1. \end{cases} \tag{2.6}$$

For more details, see Section 2 of Kumar *et al.* [11, p. 4] with appropriate changes. Now, we move to obtain uniform and non-uniform upper bound for $|g_z(\cdot)|$ and $|\Delta g_z(\cdot)|$, where $\Delta g(k) = g(k+1) - g(k)$ denotes the first forward difference operator. Some of the proofs of the following results are similar to the proofs given by Neammanee and Yonghint [12].

Lemma 2.1. *For $k \geq 0$ and $z \geq 0$, g_z defined in (2.6) satisfies the following:*

(i) $|g_z(k)| \leq p^{-(r+1)}$.

(ii) $|\Delta g_z(k)| \leq 2p^{-(r+1)} - p^{-1}$.

Proof:

(i) As $g_z(0) = 0$, it is enough to prove the result for $k \geq 1$. Consider

$$\begin{aligned} 0 &< \sum_{j=k}^{\infty} \frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k-1)} \frac{(k-1)!}{j!} q^{j-k}(j-z)^+ \\ &\leq 1 + \sum_{j=k+1}^{\infty} \frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k-1)} \frac{(k-1)!}{(j-1)!} q^{j-k} \\ &= 1 + \sum_{j=k+1}^{\infty} \frac{(r+k)\cdots(r+j-1)}{k(k+1)\cdots(j-1)} q^{j-k} \\ &= 1 + \sum_{j=1}^{\infty} \frac{(r+k)\cdots(r+j+k-1)}{k(k+1)\cdots(j+k-1)} q^{j} \\ &\leq p^{-(r+1)}, \quad \text{(using Lemma A.2(i))}. \end{aligned} \tag{2.7}$$

Next, consider

$$\begin{aligned} 0 &< \sum_{j=k}^{\infty} \frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k-1)} \frac{(k-1)!}{j!} q^{j-k} \\ &\leq 1 + \sum_{j=k+1}^{\infty} \frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k-1)} \frac{(k-1)!}{j!} q^{j-k} \\ &= 1 + \sum_{j=k+1}^{\infty} \frac{(r+k)\cdots(r+j-1)}{k(k+1)\cdots j} q^{j-k} \\ &= 1 + \sum_{j=1}^{\infty} \frac{(r+k)\cdots(r+j+k-1)}{k(k+1)\cdots(j+k)} q^{j} \\ &\leq 1 + \sum_{j=1}^{\infty} \frac{(r+k)\cdots(r+j+k-1)}{(k+1)\cdots(j+k)} q^{j} \\ &\leq \frac{p^{-r} - 1}{rq}, \quad \text{(using Lemma A.2(ii))}. \end{aligned} \tag{2.8}$$

Therefore, from Lemma A.1(i), we have

$$0 < \sum_{j=k}^{\infty} \frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k-1)} \frac{(k-1)!}{j!} q^{j-k} \mathbb{E}[(\mathrm{N}_{r,p}-z)^+] \\ \leq p^{-(r+1)} - p^{-1}. \tag{2.9}$$

Hence, from (2.6), (2.7), and (2.9), we get

$$|g_z(k)| = \left| \sum_{j=k}^{\infty} \frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k-1)} \frac{(k-1)!}{j!} q^{j-k} [(j-z)^+ - \mathbb{E}[(\mathrm{N}_{r,p}-z)^+]] \right| \\ \leq p^{-(r+1)}.$$

This proves (i).

(ii) Note that, for $k=0$,

$$|\Delta g_z(0)| = |g_z(1)| \leq p^{-(r+1)} \leq 2p^{-(r+1)} - p^{-1}.$$

Now, we prove the result for $k \geq 1$. Let

$$A_1(k) = \sum_{j=k}^{\infty} \frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k-1)} \frac{(k-1)!}{j!} q^{j-k} (j-z)^+ \\ - \sum_{j=k+1}^{\infty} \frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k)} \frac{k!}{j!} q^{j-k-1} (j-z)^+$$

and

$$A_2(k) = \sum_{j=k+1}^{\infty} \frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k)} \frac{k!}{j!} q^{j-k-1} \mathbb{E}[(\mathrm{N}_{r,p}-z)^+] \\ - \sum_{j=k}^{\infty} \frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k-1)} \frac{(k-1)!}{j!} q^{j-k} \mathbb{E}[(\mathrm{N}_{r,p}-z)^+].$$

Then

$$\Delta g_z(k) = g_z(k+1) - g_z(k) = A_1(k) + A_2(k),$$

Hence, using (2.7) and (2.9), we have

$$|\Delta g_z(k)| \leq |A_1(k)| + |A_2(k)| \leq 2p^{-(r+1)} - p^{-1}.$$

This proves (ii).

□

Lemma 2.2. *For $k \geq 1$ and $z > 1$, g_z defined in (2.6) satisfies the following:*

$$|\Delta g_z(k)| \leq \begin{cases} \frac{1}{z}\left(2p^{-(r+1)} - p^{-1}\right) & \text{if } k \geq z; \\ \frac{1}{z}\left((1+q^{-1})p^{-(r+2)} - p^{-2}\right) & \text{if } 2 \leq k < z; \\ \frac{(r+1)}{z}\left(2p^{-(r+2)} - p^{-2}\right) & \text{if } k = 1. \end{cases}$$

Proof: Let $k \geq z$. First, consider

$$\begin{aligned}
A_1(k) &= \sum_{j=k}^{\infty} \frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k)} \frac{(k-1)!}{(j+1)!} q^{j-k} \, [(r+k)(j+1)(j-z)^+ - k(r+j)(j+1-z)^+] \\
&= \sum_{j=k}^{\infty} \frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k)} \frac{(k-1)!}{(j+1)!} q^{j-k} \, [(r+k)(j+1)(j-z) - k(r+j)(j+1-z)].
\end{aligned}$$

Observe that

(2.10)

$$\begin{aligned}
|(r+k)(j+1)(j-z) - k(r+j)(j+1-z)| &= |r(j+1)(j-k) - r(j-k)z - (k+r)z| \\
&\leq |r(j+1)(j-k) - r(j-k)z| + (k+r)z \\
&= r(j+1)(j-k) - (r(j-k) - k - r)z \\
&\leq \begin{cases} (k+r)z & \text{if } j = k; \\ r(j+1)(j-k) & \text{if } j > k. \end{cases}
\end{aligned}$$

Therefore,

$$\begin{aligned}
|A_1(k)| &\leq \frac{z}{k(k+1)} + r \sum_{j=k+1}^{\infty} \frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k)} \frac{(k-1)!(j-k)}{j!} q^{j-k} \\
&\leq \frac{z}{k(k+1)} + \frac{r}{k} \sum_{j=k+1}^{\infty} \frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k)} \frac{k!}{(j-1)!} q^{j-k} \\
&= \frac{z}{k(k+1)} + \frac{rq}{k} + \frac{r}{k} \sum_{j=k+2}^{\infty} \frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k)} \frac{k!}{(j-1)!} q^{j-k} \\
&= \frac{z}{k(k+1)} + \frac{rq}{k} + \frac{r}{k} \sum_{j=k+2}^{\infty} \frac{(r+k+1)\cdots(r+j-1)}{(k+1)\cdots(j-1)} q^{j-k} \\
&= \frac{1}{z} \left(1 + rq + r \sum_{j=2}^{\infty} \frac{(r+k+1)\cdots(r+j+k-1)}{(k+1)\cdots(j+k-1)} q^j \right) \\
&\leq \frac{p^{-(r+1)}}{z}, \quad \text{(using Lemma A.2(iii)).}
\end{aligned} \tag{2.11}$$

Now, consider

$$\begin{aligned}
\sum_{j=k}^{\infty} \frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k-1)} \frac{(k-1)!}{j!} q^{j-k} &= \frac{1}{k} \sum_{j=k}^{\infty} \frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k-1)} \frac{k!}{j!} q^{j-k} \\
&= \frac{1}{z} \left(1 + \sum_{j=k+1}^{\infty} \frac{(r+k)\cdots(r+j-1)}{(k+1)\cdots j} q^{j-k} \right) \\
&= \frac{1}{z} \left(1 + \sum_{j=1}^{\infty} \frac{(r+k)\cdots(r+j+k-1)}{(k+1)\cdots(j+k)} q^{j} \right) \\
&\leq \frac{p^{-r} - 1}{rqz}, \quad \text{(using Lemma A.2(ii)).}
\end{aligned}$$

Therefore, from Lemma A.1(i), we have

$$\sum_{j=k}^{\infty}\frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k-1)}\frac{(k-1)!}{j!}q^{j-k}\mathbb{E}[(\mathrm{N}_{r,p}-z)^{+}]\le\frac{p^{-(r+1)}-p^{-1}}{z}. \tag{2.12}$$

Hence, for $k\ge z$, from (2.11) and (2.12), we have

$$|\Delta g(k)|\le|A_1(k)|+|A_2(k)|\le\frac{1}{z}\Big(2p^{-(r+1)}-p^{-1}\Big).$$

Next, let $k<z$ and consider

$$\begin{aligned}
|A_1(k)|&\le\sum_{j=\lceil z\rceil-1}^{\infty}\frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k)}\frac{(k-1)!}{(j+1)!}q^{j-k}\\
&\qquad\times|(r+k)(j+1)(j-z)^{+}-k(r+j)(j+1-z)^{+}|\\
&\le\frac{r(r+1)\cdots(r+\lceil z\rceil-1)}{r(r+1)\cdots(r+k)}\frac{k!}{(\lceil z\rceil)!}(\lceil z\rceil-z)q^{\lceil z\rceil-1-k}\\
&\qquad+\frac{r}{\lceil z\rceil}\sum_{j=\lceil z\rceil}^{\infty}\frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k)}\frac{(k-1)!}{(j-2)!}q^{j-k}\quad\text{(using (2.10))}\\
&\le\frac{r(r+1)\cdots(r+\lceil z\rceil-1)}{r(r+1)\cdots(r+k)}\frac{k!}{(\lceil z\rceil)!}(\lceil z\rceil-z)q^{\lceil z\rceil-1-k}\\
&\qquad+\frac{r}{z}\frac{r(r+1)\cdots(r+\lceil z\rceil-1)}{r(r+1)\cdots(r+k)}\frac{(k-1)!}{(\lceil z\rceil-2)!}q^{\lceil z\rceil-k}\\
&\qquad+\frac{r}{z}\sum_{j=\lceil z\rceil+1}^{\infty}\frac{(r+k+1)\cdots(r+j-1)}{k(k+1)\cdots(j-2)}q^{j-k},
\end{aligned} \tag{2.13}$$

where $\lceil z\rceil$ is the smallest integer greater than or equal to z. At $k=\lceil z\rceil-1$, we have

$$|A_1(k)|\le\frac{1}{z}\left(1+rq+r\sum_{j=\lceil z\rceil+1}^{\infty}\frac{(r+\lceil z\rceil)\cdots(r+j-1)}{(\lceil z\rceil-1)(\lceil z\rceil)\cdots(j-2)}q^{j-\lceil z\rceil+1}\right) \tag{2.14}$$

$$\begin{aligned}
&=\frac{1}{z}\left(1+rq+r\sum_{j=\lceil z\rceil+1-k}^{\infty}\frac{(r+\lceil z\rceil)\cdots(r+j+k-1)}{(\lceil z\rceil-1)(\lceil z\rceil)\cdots(j+k-2)}q^{j+k-\lceil z\rceil+1}\right)\\
&=\frac{1}{z}\left(1+rq+r\sum_{j=2}^{\infty}\frac{(r+k+1)\cdots(r+j+k-1)}{k(k+1)\cdots(j+k-2)}q^{j}\right)
\end{aligned}$$

$$=\frac{1}{z}\left(1+rq+rq\sum_{j=1}^{\infty}\frac{(r+k+1)\cdots(r+j+k)}{k(k+1)\cdots(j+k-1)}q^{j}\right) \tag{2.15}$$

$$\le\frac{p^{-(r+2)}}{z}\quad\text{(using Lemma A.2(iv)).}$$

Now, let $k<\lceil z\rceil-1$. From (2.13), we have

$$\begin{aligned}
|A_1(k)|&\le\frac{r(r+1)\cdots(r+\lceil z\rceil-1)}{r(r+1)\cdots(r+k)}\frac{k!}{(\lceil z\rceil)!}(\lceil z\rceil-z)q^{\lceil z\rceil-1-k}\\
&\qquad+\frac{r}{\lceil z\rceil}\sum_{j=\lceil z\rceil}^{\infty}\frac{(r+k+1)\cdots(r+j-1)}{k(k+1)\cdots(j-2)}q^{j-k}
\end{aligned} \tag{2.16}$$

$$
\begin{aligned}
&\leq \frac{1}{z}\frac{(r+k+1)\cdots(r+\lceil z\rceil-1)}{k(k+1)\cdots(\lceil z\rceil-2)}(\lceil z\rceil-z)q^{\lceil z\rceil-1-k}\\
&\quad+\frac{r+1}{z}\sum_{j=\lceil z\rceil}^{\infty}\frac{(r+k+1)\cdots(r+j)}{k(k+1)\cdots(j-1)}q^{j-k}\\
&\leq \frac{r+1}{z}\sum_{j=\lceil z\rceil-1}^{\infty}\frac{(r+k+1)\cdots(r+j)}{k(k+1)\cdots(j-1)}q^{j-k}\\
&\leq \frac{r+1}{z}\sum_{j=k+1}^{\infty}\frac{(r+k+1)\cdots(r+j)}{k(k+1)\cdots(j-1)}q^{j-k}\\
&\leq \frac{r+1}{z}\sum_{j=1}^{\infty}\frac{(r+k+1)\cdots(r+j+k)}{k(k+1)\cdots(j+k-1)}q^{j} \qquad (2.17)\\
&\leq \frac{p^{-(r+2)}}{qz}, \quad \text{(using Lemma A.2(iv)).}
\end{aligned}
$$

Next, for $k\geq 2$, consider

$$
\begin{aligned}
\sum_{j=k}^{\infty}\frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k-1)}\frac{(k-1)!}{j!}q^{j-k} &= \frac{1}{k}+\sum_{j=k+1}^{\infty}\frac{(r+k)\cdots(r+j-1)}{k(k+1)\cdots j}q^{j-k}\\
&\leq \frac{1}{2}+\sum_{j=1}^{\infty}\frac{(r+k)\cdots(r+j+k-1)}{k(k+1)\cdots(j+k)}q^{j}\\
&\leq \frac{p^{-r}-1}{r(r+1)q^2}, \quad \text{(using Lemma A.2(v)).}
\end{aligned}
\tag{2.18}
$$

Therefore, from Lemma A.1(ii) and (2.18), we get

$$
\sum_{j=k}^{\infty}\frac{r(r+1)\cdots(r+j-1)}{r(r+1)\cdots(r+k-1)}\frac{(k-1)!}{j!}q^{j-k}\mathbb{E}[(\mathrm{N}_{r,p}-z)^{+}]\leq\frac{p^{-(r+2)}-p^{-2}}{z}. \tag{2.19}
$$

Hence, for $k<z$, from (2.14), (2.16), and (2.19), we have

$$
|\Delta g(k)|\leq|A_1(k)|+|A_2(k)|\leq\frac{1}{z}\Big((1+q^{-1})p^{-(r+2)}-p^{-2}\Big).
$$

Next, at $k=1$, from (2.15), we have

$$
\begin{aligned}
|A_1(1)| &\leq \frac{1}{z}\left(1+rq+rq\sum_{j=1}^{\infty}\binom{r+j+1}{j}q^j\right)\\
&\leq \frac{1}{z}\Big(1+rqp^{-(r+2)}\Big)\leq\frac{(r+1)p^{-(r+2)}}{z}
\end{aligned}
\tag{2.20}
$$

and, at $k=1$, from (2.17), we have

$$
|A_1(1)|\leq\frac{r+1}{z}\sum_{j=1}^{\infty}\binom{r+j+1}{j}q^j=\frac{(r+1)\big(p^{-(r+2)}-1\big)}{z}\leq\frac{(r+1)p^{-(r+2)}}{z}. \tag{2.21}
$$

Also, using Lemma A.1(ii), it can be easily verified that

(2.22)

$$
|A_2(1)| \leq \frac{1}{r}\sum_{j=1}^{\infty}\frac{r(r+1)\cdots(r+j-1)}{j!}q^{j-1}\mathbb{E}[(\mathrm{N}_{r,p}-z)^{+}] \leq \frac{(r+1)\big(p^{-(r+2)}-p^{-2}\big)}{z}.
$$

Hence, at $k = 1$, from (2.20), (2.21), and (2.22), we have

$$|\Delta g(1)| \le |A_1(1)| + |A_2(1)| \le \frac{(r+1)}{z}\Big(2p^{-(r+2)} - p^{-2}\Big).$$

This proves the result. □

Remark 2.1. From Lemma 2.2, a rather crude uniform bound is given by

$$\|\Delta g_z\| \le \vartheta_{r,p,z} := \frac{r+1}{z}\Big((1+q^{-1})p^{-(r+2)} - p^{-2}\Big), \text{ for } k \ge 1 \text{ and } z > 1. \tag{2.23}$$

3. BOUNDS FOR NB APPROXIMATION

In this section, we obtain error bounds between $\mathbb{E}[(\mathrm{N}_{r,p} - z)^+]$ and $\mathbb{E}[(V - z)^+]$ such that $\mathrm{N}_{r,p}$ follows NB distribution and $V = \sum_{i \in J} \zeta_i$, where $\{\zeta_i\}_{i \in J}$ is a collection of $\mathbb{Z}_+$-valued rvs. Throughout this section, let μ_X and σ_X denote the mean and variance for the rv X. The following theorem gives the bound for the locally dependent setup.

Theorem 3.1. *Let $\mathbb{E}(\zeta_i^3) < \infty$ and V be the sum of locally dependent rvs as defined in (1.3). Then*

1. *(uniform bound)* $\sup_{z \ge 0} |\mathbb{E}[\mathcal{A}g_z(V)]| \le \big(2p^{-(r+1)} - p^{-1}\big)U_J$
2. *(non-uniform bound)* $|\mathbb{E}[\mathcal{A}g_z(V)]| \le \vartheta_{r,p,z}U_J$, *for all $z > 1$,*

where

$$U_J = \begin{cases} \displaystyle\sum_{i \in J}[p\mathbb{E}(\zeta_i)\mathbb{E}(\zeta_{A_i}) + q\mathbb{E}(\zeta_i\zeta_{A_i}) + \mathbb{E}(\zeta_i(\zeta_{A_i} - 1))] & \text{if } \mu_{\mathrm{N}_{r,p}} = \mu_V; \\ \displaystyle p\sum_{i \in J}\mathbb{E}(\zeta_i)\mathbb{E}[\zeta_{A_i}(2\zeta_{B_i} - \zeta_{A_i} - 1)\mathcal{D}(V|\zeta_{A_i}, \zeta_{B_i})] & \text{if } \mu_{\mathrm{N}_{r,p}} = \mu_V \text{ and} \\ \displaystyle +q\sum_{i \in J}\mathbb{E}[\zeta_i\zeta_{A_i}(2\zeta_{B_i} - \zeta_{A_i} - 1)\mathcal{D}(V|\zeta_i, \zeta_{A_i}, \zeta_{B_i})] & \quad \sigma_{\mathrm{N}_{r,p}} = \sigma_V. \\ \displaystyle +\sum_{i \in J}\mathbb{E}[\zeta_i(\zeta_{A_i} - 1)(2\zeta_{B_i} - \zeta_{A_i} - 2)\mathcal{D}(V|\zeta_i, \zeta_{A_i}, \zeta_{B_i})] & \\ \displaystyle +\sum_{i \in J}|p\mathbb{E}(\zeta_i)\mathbb{E}(\zeta_{A_i}) + q\mathbb{E}(\zeta_i\zeta_{A_i}) - \mathbb{E}(\zeta_i(\zeta_{A_i} - 1))| & \\ \qquad \times\mathbb{E}[\zeta_{B_i}\mathcal{D}(V|\zeta_{B_i})] & \end{cases}$$

and $\vartheta_{r,p,z}$ is defined in (2.23).

Proof: Consider the Stein operator given in (2.4) and taking expectation with respect to V, we get

$$\begin{aligned} \mathbb{E}[\mathcal{A}g_z(V)] &= rq\mathbb{E}[g_z(V+1)] + q\mathbb{E}[Vg_z(V+1)] - \mathbb{E}[Vg_z(V)] \\ &= p\sum_{i \in J}\mathbb{E}(\zeta_i)\mathbb{E}[g_z(V+1)] + q\sum_{i \in J}\mathbb{E}[\zeta_i g_z(V+1)] - \sum_{i \in J}\mathbb{E}[\zeta_i g_z(V)], \end{aligned} \tag{3.1}$$

where the last expression is obtained by using $\mu_{\mathrm{N}_{r,p}} = \mu_V$. Now, let $V_i = V - \zeta_{A_i}$ then ζ_i and V_i are independent rvs. Also, note that

$$p\sum_{i\in J}\mathbb{E}(\zeta_i)\mathbb{E}[g_z(V_i+1)] + q\sum_{i\in J}\mathbb{E}[\zeta_i g_z(V_i+1)] - \sum_{i\in J}\mathbb{E}[\zeta_i g_z(V_i+1)] = 0. \tag{3.2}$$

Using (3.2) in (3.1), we get

$$\begin{aligned}\mathbb{E}[\mathcal{A}g_z(V)] &= p\sum_{i\in J}\mathbb{E}(\zeta_i)\mathbb{E}[g_z(V+1) - g_z(V_i+1)] \\ &\quad + q\sum_{i\in J}\mathbb{E}[\zeta_i(g_z(V+1) - g_z(V_i+1))] \\ &\quad - \sum_{i\in J}\mathbb{E}[\zeta_i(g_z(V) - g_z(V_i+1))] \\ &= p\sum_{i\in J}\mathbb{E}(\zeta_i)\mathbb{E}\left[\sum_{j=1}^{\zeta_{A_i}}\Delta g_z(V_i+j)\right] + q\sum_{i\in J}\mathbb{E}\left[\zeta_i\sum_{j=1}^{\zeta_{A_i}}\Delta g_z(V_i+j)\right] \\ &\quad - \sum_{i\in J}\mathbb{E}\left[\zeta_i\sum_{j=1}^{\zeta_{A_i}-1}\Delta g_z(V_i+j)\right].\end{aligned} \tag{3.3}$$

Therefore,

$$|\mathbb{E}[\mathcal{A}g_z(V)]| \le \|\Delta g_z\|\sum_{i\in J}[p\mathbb{E}(\zeta_i)\mathbb{E}(\zeta_{A_i}) + q\mathbb{E}(\zeta_i\zeta_{A_i}) + \mathbb{E}(\zeta_i(\zeta_{A_i}-1))].$$

Hence, using Lemma 2.1(ii) and (2.23), the result follows when $\mu_{\mathrm{N}_{r,p}} = \mu_V$. Next, using $\mu_{\mathrm{N}_{r,p}} = \mu_V$ and $\sigma_{\mathrm{N}_{r,p}} = \sigma_V$, it can be easily verified that

$$\left[p\sum_{i\in J}\mathbb{E}(\zeta_i)\mathbb{E}[\zeta_{A_i}] + q\sum_{i\in J}\mathbb{E}[\zeta_i\zeta_{A_i}] - \sum_{i\in J}\mathbb{E}[\zeta_i(\zeta_{A_i}-1)]\right]\mathbb{E}[g_z(V+1)] = 0. \tag{3.4}$$

Let $V_i^* = V - \zeta_{B_i}$ then ζ_i and ζ_{A_i} are independent of V_i^*. Now, using (3.4) in (3.3), we get

$$\begin{aligned}\mathbb{E}[\mathcal{A}g_z(V)] &= p\sum_{i\in J}\mathbb{E}(\zeta_i)\mathbb{E}\left[\sum_{j=1}^{\zeta_{A_i}}(\Delta g_z(V_i+j) - \Delta g_z(V_i^*+1))\right] \\ &\quad + q\sum_{i\in J}\mathbb{E}\left[\zeta_i\sum_{j=1}^{\zeta_{A_i}}(\Delta g_z(V_i+j) - \Delta g_z(V_i^*+1))\right] \\ &\quad - \sum_{i\in J}\mathbb{E}\left[\zeta_i\sum_{j=1}^{\zeta_{A_i}-1}(\Delta g_z(V_i+j) - \Delta g_z(V_i^*+1))\right] \\ &\quad - \sum_{i\in J}[p\mathbb{E}(\zeta_i)\mathbb{E}(\zeta_{A_i}) + q\mathbb{E}(\zeta_i\zeta_{A_i}) - \mathbb{E}(\zeta_i(\zeta_{A_i}-1))] \\ &\qquad \times\mathbb{E}[g_z(V+1) - g_z(V_i^*+1)] \\ &= p\sum_{i\in J}\mathbb{E}(\zeta_i)\mathbb{E}\left[\sum_{j=1}^{\zeta_{A_i}}\sum_{\ell=1}^{\zeta_{B_i\setminus A_i}+j-1}\Delta^2 g_z(V_i+\ell)\right] \\ &\quad + q\sum_{i\in J}\mathbb{E}\left[\zeta_i\sum_{j=1}^{\zeta_{A_i}}\sum_{\ell=1}^{\zeta_{B_i\setminus A_i}+j-1}\Delta^2 g_z(V_i+\ell)\right]\end{aligned} \tag{3.5}$$

$$
\begin{aligned}
&-\sum_{i\in J}\mathbb{E}\left[\zeta_i\sum_{j=1}^{\zeta_{A_i}-1}\sum_{\ell=1}^{\zeta_{B_i\setminus A_i}+j-1}\Delta^2 g_z(V_i+\ell)\right]\\
&-\sum_{i\in J}[p\mathbb{E}(\zeta_i)\mathbb{E}(\zeta_{A_i})+q\mathbb{E}(\zeta_i\zeta_{A_i})-\mathbb{E}(\zeta_i(\zeta_{A_i}-1))]\\
&\qquad\times\mathbb{E}\left[\sum_{\ell=1}^{\zeta_{B_i}}\Delta^2 g_z(V_i+\ell)\right]\\
&=p\sum_{i\in J}\mathbb{E}(\zeta_i)\mathbb{E}\left[\sum_{j=1}^{\zeta_{A_i}}\sum_{\ell=1}^{\zeta_{B_i\setminus A_i}+j-1}\mathbb{E}[\Delta^2 g_z(V_i+\ell)|\zeta_{A_i},\zeta_{B_i}]\right]\\
&+q\sum_{i\in J}\mathbb{E}\left[\zeta_i\sum_{j=1}^{\zeta_{A_i}}\sum_{\ell=1}^{\zeta_{B_i\setminus A_i}+j-1}\mathbb{E}[\Delta^2 g_z(V_i+\ell)|\zeta_i,\zeta_{A_i},\zeta_{B_i}]\right]\\
&-\sum_{i\in J}\mathbb{E}\left[\zeta_i\sum_{j=1}^{\zeta_{A_i}-1}\sum_{\ell=1}^{\zeta_{B_i\setminus A_i}+j-1}\mathbb{E}[\Delta^2 g_z(V_i+\ell)|\zeta_i,\zeta_{A_i},\zeta_{B_i}]\right]\\
&-\sum_{i\in J}[p\mathbb{E}(\zeta_i)\mathbb{E}(\zeta_{A_i})+q\mathbb{E}(\zeta_i\zeta_{A_i})-\mathbb{E}(\zeta_i(\zeta_{A_i}-1))]\\
&\qquad\times\mathbb{E}\left[\sum_{\ell=1}^{\zeta_{B_i}}\mathbb{E}[\Delta^2 g_z(V_i+\ell)|\zeta_{B_i}]\right].
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
|\mathbb{E}[\mathcal{A}g_z(V)]|\le\|\Delta g_z\|\Bigg\{&p\sum_{i\in J}\mathbb{E}(\zeta_i)\mathbb{E}[\zeta_{A_i}(2\zeta_{B_i}-\zeta_{A_i}-1)\mathfrak{D}(V|\zeta_{A_i},\zeta_{B_i})]\\
&+q\sum_{i\in J}\mathbb{E}[\zeta_i\zeta_{A_i}(2\zeta_{B_i}-\zeta_{A_i}-1)\mathfrak{D}(V|\zeta_i,\zeta_{A_i},\zeta_{B_i})]\\
&+\sum_{i\in J}|p\mathbb{E}(\zeta_i)\mathbb{E}(\zeta_{A_i})+q\mathbb{E}(\zeta_i\zeta_{A_i})-\mathbb{E}(\zeta_i(\zeta_{A_i}-1))|\mathbb{E}[\zeta_{B_i}\mathfrak{D}(V|\zeta_{B_i})]\\
&+\sum_{i\in J}\mathbb{E}[\zeta_i(\zeta_{A_i}-1)(2\zeta_{B_i}-\zeta_{A_i}-2)\mathfrak{D}(V|\zeta_i,\zeta_{A_i},\zeta_{B_i})]\Bigg\}.
\end{aligned}
$$

Hence, using Lemma 2.1(ii) and (2.23), the result follows when $\mu_{\mathrm{N}_{r,p}}=\mu_V$ and $\sigma_{\mathrm{N}_{r,p}}=\sigma_V$. □

Corollary 3.1. *Let $V_1=\sum_{i\in J}\zeta_i$ with $p_i=\mathbb{P}(\zeta_i=1)$ and $p_{i,j}=\mathbb{P}(\zeta_i=1,\zeta_j=1)$. Then, for $\mu_{N_{r,p}}=\mu_{V_1}$, we have*

$$\sup_{z\ge0}|\mathbb{E}[\mathcal{A}g_z(V_1)]|\le\left(2p^{-(r+1)}-p^{-1}\right)\sum_{i\in J}\left[(1+q)\sum_{j\in A_i}p_{i,j}+p_i\left(p\sum_{j\in A_i}p_j-1\right)\right]. \tag{3.6}$$

Remark 3.1.

(i) In Theorem 3.1, note that we have the flexibility to choose one parameter (either r or p) of our choice when $\mu_{\mathrm{N}_{r,p}}=\mu_V$. Also, the bound is valid only if $\mathbb{E}(V)<\mathrm{Var}(V)$ when $\mu_{\mathrm{N}_{r,p}}=\mu_V$ and $\sigma_{\mathrm{N}_{r,p}}=\sigma_V$.

(ii) Observe that V can be expressed as a conditional sum of independent rvs and hence, Subsections 5.3 and 5.4 of Röllin [13] can be used to obtain the bound of $\mathcal{D}(V|\cdot)$. For more details, see Remark 3.1(ii) of Kumar *et al.* [11].

Next, the following theorem gives the bound for independent setup.

Theorem 3.2. *Let $\mathbb{E}(\zeta_i^3) < \infty$ and V be the sum of independent rvs. Then*

1. *(uniform bound)* $\sup_{z\geq 0} |\mathbb{E}[\mathcal{A}g_z(V)]| \leq \big(2p^{-(r+1)} - p^{-1}\big)U_J^*$
2. *(non-uniform bound)* $|\mathbb{E}[\mathcal{A}g_z(V)]| \leq \vartheta_{r,p,z}U_J^*$, *for all $z > 1$,*

where

$$U_J^* = \begin{cases} \displaystyle\sum_{i\in J}\sum_{k=1}^{\infty} k|(p\mathbb{E}(\zeta_i)+qk)\gamma_{i,k} - (k+1)\gamma_{i,k+1}| & \text{if } \mu_{N_{r,p}} = \mu_V; \\ \displaystyle\sqrt{\frac{2}{\pi}}\left(\frac{1}{4}+\sum_{j\in J}\delta_j - \delta^*\right)^{-\frac{1}{2}} & \text{if } \mu_{N_{r,p}} = \mu_V \\ \displaystyle\Bigg\{\sum_{i\in J}\mathbb{E}(\zeta_i)|p\mathbb{E}(\zeta_i)^2 + q\mathbb{E}(\zeta_i^2) - \mathbb{E}(\zeta_i(\zeta_i-1))| & \text{and } \sigma_{N_{r,p}} = \sigma_V \\ \displaystyle\quad + \sum_{i\in J}\sum_{k=2}^{\infty}\frac{k(k-1)}{2}|(p\mathbb{E}(\zeta_i)+qk)\gamma_{i,k} - (k+1)\gamma_{i,k+1}|\Bigg\}, & \end{cases}$$

$\vartheta_{r,p,z}$ is defined in (2.23), $\gamma_{i,k} = \mathbb{P}(\zeta_i = k)$, $\delta_j = \min\{\frac{1}{2}, 1 - d_{TV}(\zeta_j, \zeta_j+1)\}$, and $\delta^* = \max_{j\in J}\delta_j$.

Proof: Substituting $A_i = \{i\}$ in (3.3), we get

$$\begin{aligned} \mathbb{E}[\mathcal{A}g_z(V)] &= p\sum_{i\in J}\mathbb{E}(\zeta_i)\mathbb{E}\left[\sum_{j=1}^{\zeta_i}\Delta g_z(V_i+j)\right] + q\sum_{i\in J}\mathbb{E}\left[\zeta_i\sum_{j=1}^{\zeta_i}\Delta g_z(V_i+j)\right] \\ &\quad - \sum_{i\in J}\mathbb{E}\left[\zeta_i\sum_{j=1}^{\zeta_i-1}\Delta g_z(V_i+j)\right] \\ &= p\sum_{i\in J}\sum_{k=1}^{\infty}\sum_{j=1}^{k}\mathbb{E}(\zeta_i)\mathbb{E}[\Delta g_z(V_i+j)]\gamma_{i,k} \\ &\quad + q\sum_{i\in J}\sum_{k=1}^{\infty}\sum_{j=1}^{k}k\mathbb{E}[\Delta g_z(V_i+j)]\gamma_{i,k} \\ &\quad - \sum_{i\in J}\sum_{k=2}^{\infty}\sum_{j=1}^{k-1}k\mathbb{E}[\Delta g_z(V_i+j)]\gamma_{i,k} \\ &= \sum_{i\in J}\sum_{k=1}^{\infty}[(p\mathbb{E}(\zeta_i)+qk)\gamma_{i,k} - (k+1)\gamma_{i,k+1}]\sum_{j=1}^{k}\mathbb{E}[\Delta g_z(V_i+j)]. \end{aligned}$$

Therefore,

$$|\mathbb{E}[\mathcal{A}g_z(V)]| \leq \|\Delta g_z\| \sum_{i\in J}\sum_{k=1}^{\infty} k|(p\mathbb{E}(\zeta_i)+qk)\gamma_{i,k}-(k+1)\gamma_{i,k+1}|.$$

Hence, using Lemma 2.1(ii) and (2.23), the result follows when $\mu_{\mathrm{N}_{r,p}}=\mu_V$.
Next, substituting $A_i=B_i=\{i\}$ in (3.5), we get

$$\begin{aligned}
\mathbb{E}[\mathcal{A}g_z(V)] = p\sum_{i\in J}\mathbb{E}(\zeta_i)\mathbb{E}&\left[\sum_{j=1}^{\zeta_i}\sum_{\ell=1}^{j-1}\mathbb{E}[\Delta^2 g_z(V_i+\ell)|\zeta_i]\right]\\
&+q\sum_{i\in J}\mathbb{E}\left[\zeta_i\sum_{j=1}^{\zeta_i}\sum_{\ell=1}^{j-1}\mathbb{E}[\Delta^2 g_z(V_i+\ell)|\zeta_i]\right]\\
&-\sum_{i\in J}[p\mathbb{E}(\zeta_i)^2+q\mathbb{E}(\zeta_i^2)-\mathbb{E}(\zeta_i(\zeta_i-1))]\mathbb{E}\left[\sum_{\ell=1}^{\zeta_i}\mathbb{E}[\Delta^2 g_z(V_i+\ell)|\zeta_i]\right]\\
&-\sum_{i\in J}\mathbb{E}\left[\zeta_i\sum_{j=1}^{\zeta_i-1}\sum_{\ell=1}^{j-1}\mathbb{E}[\Delta^2 g_z(V_i+\ell)|\zeta_i]\right]\\
=p&\sum_{i\in J}\sum_{k=1}^{\infty}\sum_{j=1}^{k}\sum_{\ell=1}^{j-1}\mathbb{E}(\zeta_i)\mathbb{E}[\Delta^2 g_z(V_i+\ell)]\gamma_{i,k}\\
&+q\sum_{i\in J}\sum_{k=1}^{\infty}\sum_{j=1}^{k}\sum_{\ell=1}^{j-1}k\mathbb{E}[\Delta^2 g_z(V_i+\ell)]\gamma_{i,k}\\
&-\sum_{i\in J}\sum_{k=1}^{\infty}\sum_{\ell=1}^{k}[p\mathbb{E}(\zeta_i)^2+q\mathbb{E}(\zeta_i^2)-\mathbb{E}(\zeta_i(\zeta_i-1))]\mathbb{E}[\Delta^2 g_z(V_i+\ell)]\gamma_{i,k}\\
&-\sum_{i\in J}\sum_{k=2}^{\infty}\sum_{j=1}^{k-1}\sum_{\ell=1}^{j-1}k\mathbb{E}[\Delta^2 g_z(V_i+\ell)]\gamma_{i,k}\\
=&\sum_{i\in J}\sum_{k=1}^{\infty}\sum_{j=1}^{k}\sum_{\ell=1}^{j-1}[(p\mathbb{E}(\zeta_i)+qk)\gamma_{i,k}-(k+1)\gamma_{i,k+1}]\mathbb{E}[\Delta^2 g_z(V_i+\ell)]\\
&-\sum_{i\in J}\sum_{k=1}^{\infty}\sum_{\ell=1}^{k}[p\mathbb{E}(\zeta_i)^2+q\mathbb{E}(\zeta_i^2)-\mathbb{E}(\zeta_i(\zeta_i-1))]\mathbb{E}[\Delta^2 g_z(V_i+\ell)]\gamma_{i,k}.
\end{aligned}$$

Note that $|\mathbb{E}(\Delta^2 g_z(V_i+\cdot))|\leq \delta\|\Delta g_z\|$, where $\delta=2\max_{i\in J} d_{TV}(V_i,V_i+1)$ (see Barbour and Xia [3], and Barbour and Čekanavičius [2, p. 517])). Also, from Corollary 1.6 of Brown and Phillips [4] (see also Remark 4.1 of Vellaisamy *et al.* [15]), we have $\delta\leq\sqrt{\frac{2}{\pi}}\left(\frac{1}{4}+\sum_{j\in J}\delta_j-\delta^*\right)^{-1/2}$ with $\delta_j=\min\{\frac{1}{2},1-d_{TV}(\zeta_j,\zeta_j+1)\}$ and $\delta^*=\max_{j\in J}\delta_j$. Therefore,

$$\begin{aligned}
(3.7)\qquad |\mathbb{E}[\mathcal{A}g_z(V)]| \leq \|\Delta g_z\|\sqrt{\frac{2}{\pi}}&\left(\frac{1}{4}+\sum_{j\in J}\delta_j-\delta^*\right)^{-\frac{1}{2}}\\
&\times\left\{\sum_{i\in J}\mathbb{E}(\zeta_i)|p\mathbb{E}(\zeta_i)^2+q\mathbb{E}(\zeta_i^2)-\mathbb{E}(\zeta_i(\zeta_i-1))|\right.\\
&\left.+\sum_{i\in J}\sum_{k=2}^{\infty}\frac{k(k-1)}{2}|(p\mathbb{E}(\zeta_i)+qk)\gamma_{i,k}-(k+1)\gamma_{i,k+1}|\right\}.
\end{aligned}$$

Hence, using Lemma 2.1(ii) and (2.23), the result follows when $\mu_{\mathrm{N}_{r,p}} = \mu_V$ and $\sigma_{\mathrm{N}_{r,p}} = \sigma_V$. □

Next, for $J = \{1, 2, ..., n\}$, we present and compare our results for the sum of Bernoulli and geometric rvs as special cases.

Remark 3.2.

(i) Note that the expression U_J^* in Theorem 3.2 is similar to the expression given in Theorems 3.1 and 4.1 of Vellaisamy *et al.* [15]. Also, for total variation distance ($\|\Delta g\| \leq 1/rq$), the bound given in (3.7) is an improvement over Theorem 4.1 of Kadu [8].

(ii) For $J = \{1, 2, ..., n\}$, the bounds given in Theorem 3.2 are of $O(np^{-n})$ when $\mu_{N_{r,p}} = \mu_V$ and $O(\sqrt{n}p^{-n})$ when $\mu_{N_{r,p}} = \mu_V$ and $\sigma_{N_{r,p}} = \sigma_V$. These bounds improved the existing bounds given by Neammanee and Yonghint [12] which is of $O(ne^n)$. Moreover, our bounds are more suitable for sufficiently large values of p.

(iii) Let $V_2 = \sum_{i=1}^{n} \zeta_i$ be the sum of independent Bernoulli rvs. Then, from Theorem 3.2, we have

$$\sup_{z\geq 0} |\mathbb{E}[\mathcal{A}g_z(V_2)]| \leq \left(2p^{-(r+1)} - p^{-1}\right) \sum_{i=1}^{n} p_i(1 - pq_i), \tag{3.8}$$

where $p_i = 1 - q_i = \mathbb{P}(\zeta_i = 1)$ and $r(1-p) = p\sum_{i=1}^{n} p_i$. Note that we can not obtain the bound by matching mean and variance as $\mathbb{E}(V_2) > \mathrm{Var}(V_2)$. From Corollary 1 of Neammanee and Yonghint [12], we have

$$\sup_{z\geq 0} |\mathbb{E}[\mathcal{A}g_z(V_2)]| \leq (2e^{\lambda} - 1) \sum_{i=1}^{n} p_i^2, \tag{3.9}$$

where $\lambda = \sum_{i=1}^{n} p_i$. Observe that the bound given in (3.8) is either comparable to or an improvement over the bound given in (3.9), for example, some numerical comparisons are given in Table 2.

(iv) Let $V_3 = \sum_{i=1}^{n} \zeta_i$ be the sum of independent geometric rvs with $\mathbb{P}(\zeta_i = k) = q_i^k p_i$, for $k \in \mathbb{Z}_+$, and $q_i \leq 1/2$. Then, from Theorem 3.2, we have

$$\sup_{z\geq 0} |\mathbb{E}[\mathcal{A}g_z(V_2)]| \leq \begin{cases} \left(2p^{-(r+1)} - p^{-1}\right) \displaystyle\sum_{i=1}^{n} \frac{|p - p_i|q_i}{p_i^2} & \text{if } \mu_{\mathrm{N}_{r,p}} = \mu_{V_3}; \\ 3\left(2p^{-(r+1)} - p^{-1}\right) & \text{if } \mu_{\mathrm{N}_{r,p}} = \mu_{V_3} \\ \times\sqrt{\frac{2}{\pi}}\left(\sum_{j=1}^{n} q_j - \frac{1}{4}\right)^{-1/2} & \text{and } \sigma_{\mathrm{N}_{r,p}} = \sigma_{V_3}, \\ \times\displaystyle\sum_{i=1}^{n} \frac{|p - p_i|q_i^2}{p_i^3} & \end{cases} \tag{3.10}$$

where $\sum_{i=1}^{n} q_i > 1/4$ when $\mu_{\mathrm{N}_{r,p}} = \mu_{V_3}$ and $\sigma_{\mathrm{N}_{r,p}} = \sigma_{V_3}$. Note that if $p_i = p$, for all $1 \leq i \leq n$, then $\sup_{z\geq 0} |\mathbb{E}[\mathcal{A}g_z(V_2)]| = 0$, as expected. From Theorem 1 and Corollary 2 of Neammanee and Yonghint [12], we have

$$\sup_{z\geq 0} |\mathbb{E}[\mathcal{A}g_z(V_2)]| \leq (2e^{\lambda} - 1) \sum_{i=1}^{n} \frac{(8 - 7p_i)q_i^2}{p_i^3}, \tag{3.11}$$

where $\lambda = \sum_{i=1}^{n}(q_i/p_i)$. The above bound is better than the bound given by Jiao and Karoui [5] (shown in Remark 1(1) by Neammanee and Yonghint [12]). Note that our bound is better than the bound given in (3.11). For instance, let $n = 75$ and q_i, $1 \leq i \leq 75$, be defined as follows:

Table 1: The values of q_i.

i	q_i	i	q_i	i	q_i	i	q_i	i	q_i	i	q_i
0–10	0.05	11–20	0.10	21–30	0.15	31–40	0.20	41–50	0.25	51–75	0.30

Then, choose $r = n$ if $\mu_{\mathrm{N}_{r,p}} = \mu_V$, the following table gives a comparison between our bounds and the existing bounds under Bernoulli and geometric setup.

Table 2: Comparison of bounds.

n	For Bernoulli setup		For geometric setup		
	From (3.9)	From (3.8)	From (3.11)	From (3.10) $(\mu_{\mathrm{N}_{r,p}} = \mu_V)$	From (3.10) $(\mu_{\mathrm{N}_{r,p}} = \mu_V$ and $\sigma_{\mathrm{N}_{r,p}} = \sigma_V)$
10	21.5280	22.2920	0.09390	9.47×10^{-17}	9.07×10^{-17}
20	158.986	161.239	2.53041	0.41416	0.06280
30	1438.02	1348.40	60.4516	7.17325	1.23534
40	22467.0	17633.1	2117.84	195.211	27.7360
50	745974	423881	142995	7902.23	1079.63

For large values of n, note that our bounds are an improvement over the existing bounds for various values of q_i. Moreover, for the geometric setup, the bounds are much sharper than the existing bounds as NB and the sum of geometric rvs consists of similar properties. Also, observe that the bounds computed by matching mean and variance are better than the bounds computed by matching mean only, as expected.

4. AN APPLICATION TO CDO

The CDO is a financial tool that transfers a pool of assets such as auto loans, credit card debt, mortgages, and corporate debt, among many others, into a product and sold to investors. The assets are divided into several tranches, that is, the set of repayment. Each tranche has various credit quality and risk levels. The primary tranches in CDOs are senior, mezzanine, and equity. The investors can opt for multiple tranches to invest as per their interest. For more details, see Neammanee and Yonghint [12], Yonghint *et al.* [17], Kumar [9], and reference therein.

It is known that the CDO occurs in both, locally dependent and independent setup (see Yonghint *et al.* [17] and Neammanee and Yonghint [12] for more details), and therefore, the results obtained in this paper are useful in applications. Consider the similar type of CDO discussed by Yonghint *et al.* [17]. Suppose there are N assets that have a constant recovery rate R then the percentage cumulative loss in CDO up to time T is

$$L(T) = \frac{1-R}{N}\sum_{i=1}^{N}\xi_i, \tag{4.1}$$

where $\xi_i = \mathbf{1}_{\{\tau_i \leq T\}}$, τ_i is the default time of the i^{th} asset, and $\mathbf{1}_A$ denotes the indicator function of A. The expression in (4.1) can be rewritten as

$$\mathbb{E}[(L(T)-z^*)^+] = \frac{1-R}{N}\mathbb{E}[(V_4 - z^*)^+], \tag{4.2}$$

where $z^* = (1-R)z/N > 0$ is the attachment or the detachment point of the tranche and $V_4 = \sum_{i=1}^{N}\xi_i$. Therefore, the problem is reduced to obtain error bounds for $\mathbb{E}[(V_4 - z^*)^+]$, and hence, Corollary 3.1 and Remark 3.2(iii) are useful in applications. For more details, we refer the reader to Yonghint *et al.* [17], Kumar [9], and reference therein.

Next, we compare our results with the existing results under the locally dependent and independent setup. For the independent setup, Neammanee and Yonghint [12] gives the bound discussed in (3.9) and, for the locally dependent setup, from Theorem 2 of Yonghint *et al.* [17], we have

$$\sup_{z\geq 0}|\mathbb{E}[\mathcal{A}g_z(V^*)]| \leq \left(2e^{\lambda}-1\right)\sum_{i=1}^{n}\left(\sum_{j\in A_i\setminus\{i\}} p^*_{i,j} + \sum_{j\in A_i} p^*_i p^*_j\right), \tag{4.3}$$

where $\lambda = \sum_{i=1}^{n} p^*_i$, $p^*_i = \mathbb{P}(\xi_i = 1)$, and $p^*_{i,j} = \mathbb{P}(\xi_i = 1, \xi_j = 1)$. Note that our bound given in (3.6) is better than the bound given in (4.3). For instance, let $r = n$, $p^*_{i,j} = p^*$, $1 \leq i,j \leq n$, $A_i = \{i-1, i, i+1\}$, and q_i as defined in Table 1, $1 \leq i \leq 75$, then, the following table gives a comparison between the upper bounds given in (3.8), (3.9), (3.6), and (4.3) for different values of p^* and q_i.

Table 3: Comparison for the locally dependent and independent setup.

n	For independent setup		For locally dependent setup		
	From (3.9)	From (3.8)	p^*	From (4.3)	From (3.6)
15	64.0726	65.9832	0.4	247.274	206.347
35	5747.39	4964.44		22958.3	15690.1
55	6.79×10^6	3.00×10^6	0.7	3.37×10^7	1.39×10^7
75	4.49×10^{10}	6.22×10^9		2.31×10^{11}	3.00×10^{10}

For large values of n, note that our bounds are better than the existing bounds for various values of p^* and q_i.

A. APPENDIX: SOME USEFUL INEQUALITIES

Here we give some inequalities and their proofs that have used in Lemmas 2.1 and 2.2. Recall that f_z is a call function, defined in (1.1), and $\mathrm{N}_{r,p}$ follows the negative binomial distribution, defined in (1.2). The following lemma gives uniform and non-uniform upper bounds for $\mathbb{E}[f_z(\mathrm{N}_{r,p})] = \mathbb{E}[(\mathrm{N}_{r,p} - z)^+]$.

Lemma A.1. *The following inequalities hold:*

(i) $\mathbb{E}[(N_{r,p} - z)^+] \leq \frac{rq}{p}$, *for* $z \geq 0$.

(ii) $\mathbb{E}[(N_{r,p} - z)^+] \leq \frac{r(r+1)q^2}{zp^2}$, *for* $z > 1$.

Proof:

(i) For $z \geq 0$, we have

$$\begin{aligned}\mathbb{E}[(\mathrm{N}_{r,p} - z)^+] &= \sum_{k=1}^{\infty} (k-z)^+ \binom{r+k-1}{k} p^r q^k \\ &\leq rp^r \sum_{k=1}^{\infty} \binom{r+k-1}{k-1} q^k = \frac{rq}{p}.\end{aligned}$$

This proves (i).

(ii) For $z > 1$, we have

$$\begin{aligned}\mathbb{E}[(\mathrm{N}_{r,p} - z)^+] &= \sum_{k=\lceil z \rceil}^{\infty} (k-z) \binom{r+k-1}{k} p^r q^k \\ &\leq \frac{p^r}{\lceil z \rceil} \sum_{k=\lceil z \rceil}^{\infty} r(r+1)\cdots(r+k-1) \frac{(k-z)}{(k-1)!} q^k \\ &\leq \frac{p^r}{z} \sum_{k=\lceil z \rceil}^{\infty} \frac{r(r+1)\cdots(r+k-1)}{(k-2)!} q^k \\ &\leq \frac{r(r+1)p^r}{z} \sum_{k=2}^{\infty} \binom{r+k-1}{k-2} q^k = \frac{r(r+1)q^2}{zp^2}.\end{aligned}$$

This proves (ii).

□

Next, the following lemma gives some inequalities related to the parameters r and p of $\mathrm{N}_{r,p}$.

Lemma A.2. *The following inequalities hold:*

(i) $$\sum_{j=1}^{\infty}\frac{(r+k)\cdots(r+j+k-1)}{k(k+1)\cdots(j+k-1)}q^j\le p^{-(r+1)}-1,\ \textit{for } k\ge 1.$$

(ii) $$\sum_{j=1}^{\infty}\frac{(r+k)\cdots(r+j+k-1)}{(k+1)\cdots(j+k)}q^j\le \frac{p^{-r}-1}{rq}-1,\ \textit{for all } k\ge 1.$$

(iii) $$\sum_{j=2}^{\infty}\frac{(r+k+1)\cdots(r+j+k-1)}{(k+1)\cdots(j+k-1)}q^j\le \frac{p^{-(r+1)}-1}{r}-q,\ \textit{for } k\ge 1.$$

(iv) $$\sum_{j=1}^{\infty}\frac{(r+k+1)\cdots(r+j+k)}{k(k+1)\cdots(j+k-1)}q^j\le \frac{p^{-(r+2)}-1}{(r+1)q}-1,\ \textit{for } k\ge 2.$$

(v) $$\sum_{j=1}^{\infty}\frac{(r+k)\cdots(r+j+k-1)}{k(k+1)\cdots(j+k)}q^j\le \frac{p^{-r}-1}{r(r+1)q^2}-\frac{1}{2},\ \textit{for } k\ge 2.$$

Proof: Note that, for $k=1$,

$$\sum_{j=1}^{\infty}\frac{(r+1)\cdots(r+j)}{1\cdot 2\cdots j}q^j=\sum_{j=1}^{\infty}\binom{r+j}{j}q^j=p^{-(r+1)}-1.$$

Therefore, the inequality (i) holds for $k=1$. Now, suppose it holds for $k=m$, that is,

$$\sum_{j=1}^{\infty}\frac{(r+m)\cdots(r+j+m-1)}{m(m+1)\cdots(j+m-1)}q^j\le p^{-(r+1)}-1. \tag{A.1}$$

Observe that

$$\begin{aligned}&\sum_{j=1}^{\infty}\frac{(r+m+1)\cdots(r+j+m)}{(m+1)\cdots(j+m)}q^j\\ &=\sum_{j=1}^{\infty}\frac{m(r+j+m)}{(r+m)(m+j)}\,\frac{(r+m)\cdots(r+m+j-1)}{m(m+1)\cdots(j+m-1)}q^j\\ &\le\sum_{j=1}^{\infty}\frac{(r+m)\cdots(r+m+j-1)}{m(m+1)\cdots(j+m-1)}q^j\\ &\le p^{-(r+1)}-1\quad(\text{using (A.1)}).\end{aligned}$$

This implies that the inequality (i) holds for $k=m+1$, and hence it holds for all $k\ge 1$. Following similar steps, the inequalities (ii)-(v) can be easily proved. □

ACKNOWLEDGMENTS

The author is grateful to the associate editor and reviewers for valuable suggestions that improved the presentation of the paper.

REFERENCES

[1] Barbour, A.D.; Gan, H.L. and Xia, A. (2015). Stein factors for negative binomial approximation in Wasserstein distance, *Bernoulli*, **21**(2), 1002–1013.

[2] Barbour, A.D. and Čekanavičius, V. (2002). Total variation asymptotes for sum of independent integer random variables, *Ann. Prob.*, **30**(2), 509–545.

[3] Barbour, A.D. and Xia, A. (1999). Poisson perturbation, *ESAIM Probab. Statist.*, **3**, 131–150.

[4] Brown, T.C. and Phillips, M.J. (1999). Negative binomial approximation with Stein's method, *Methodol. Comput. Appl. Probab.*, **1**(4), 407–421.

[5] El Karoui, N. and Jiao, Y. (2009). Stein's method and zero bias transformation for CDO tranche pricing, *Finance Stoch.*, **13**, 151–180.

[6] El Karoui, N.; Jiao, Y. and Kurtz, D. (2008). Gaussian and Poisson approximation: applications to CDOs tranche pricing, *J. Comput. Finance*, **12**(2), 31–58.

[7] Hull, J.C. and White, A.D. (2004). Valuation of a CDO and an n^{th} to default CDS without Monte Carlo simulation, *J. Deriv.*, **12**(2), 8–23.

[8] Kadu, P.E. (2022). Approximation results for sums of independent random variables, *REVSTAT – Statistical Journal*, **20**(3), 373–385.

[9] Kumar, A.N. (2021). Approximations to weighted sums of random variables, *Bull. Malays. Math. Sci. Soc.*, **44**(4), 2447–2464.

[10] Kumar, A.N. and Upadhye, N.S. (2017). On perturbations of stein operator, *Comm. Statist. Theory Methods*, **46**(18), 9284–9302.

[11] Kumar, A.N.; Upadhye, N.S. and Vellaisamy, P. (2022). Approximations related to the sums of m-dependent random variables, *Braz. J. Probab. Stat.*, **36**, 349–368.

[12] Neammanee, K. and Yonghint, N. (2020). Poisson approximation for call function via Stein-Chen method, *Bull. Malays. Math. Sci. Soc.*, **43**, 1135–1152.

[13] Röllin, A. (2008). Symmetric and centered binomial approximation of sums of locally dependent random variables, *Electron. J. Probab.*, **13**, 756–776.

[14] Stein, C. (1972). *A bound for the error in the normal approximation to the distribution of a sum of dependent random variables.* In: "Proc. Sixth Berkeley Symp. Math. Statist. Probab. II. Probability Theory", Univ. California Press, Berkeley, Calif., 583–602.

[15] Vellaisamy, P.; Upadhye, N.S. and Čekanavičius, V. (2013). On negative binomial approximation, *Theory Probab. Appl.*, **57**(1), 97–109.

[16] Wang, X. and Xia, A. (2008). On negative approximation to k-runs, *J. Appl. Probab.*, **45**(2), 456–471.

[17] Yonghint, N.; Neammanee, K. and Chaidee, N. (2022). Poisson approximation for locally dependent CDO, *Comm. Statist. Theory Methods*, **51**, 2073–2081.

REVSTAT – Statistical Journal
Volume 22, Number 1, January 2024, 45–60
https://doi.org/10.57805/revstat.v22i1.454

Tales of the Wakeby Tail and Alternatives when Modelling Extreme Floods

Author: Jesper Rydén
– Department of Energy and Technology, Swedish University of Agricultural Sciences, Uppsala, Sweden
jesper.ryden@slu.se

Received: March 2020 Revised: January 2022 Accepted: January 2022

Abstract:

- Estimation of return levels, based on extreme-value distributions, is of importance in the earth and environmental sciences. The selection of an appropriate probability distribution is crucial. The Wakeby distribution has shown to be an interesting alternative. By simulation studies, we investigate by various means of minimum distance to distinguish between common distributions when modelling extreme events in hydrology. Estimation of parameters is performed by L-moments. Moreover, time series of annual maximum floods from major unregulated rivers in Northern Sweden were analysed with respect to fitting an appropriate distribution. The results of the simulation study shows that the Wakeby distribution has the best fit of the tail for a wide range of sample sizes. For the analysis of extreme floods, the Wakeby distribution is in the majority of cases the best fit by means of minimum distance. However, when considering estimation of return levels by competing distributions, results can vary considerably for longer return periods.

Keywords:

- *Wakeby distribution; GEV distribution; L-moments; distance measures; floods.*

AMS Subject Classification:

- 62G32, 62P12.

1. INTRODUCTION

In the earth sciences, statistical modelling of extreme events is of importance; in fields like hydrology and oceanography there is a need to estimate return levels, for instance for the sake of engineering design. For this purpose, quantiles of probability distributions are of key interest, and hence choice of distribution is crucial. When using statistical methodology based on likelihood functions, criteria like Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) can be employed for model choice [2], [25]. Often in the applied literature, goodness-of-fit tests are employed as measures of deviation between the empirical distribution and the potential distribution family.

The tail behaviour of the probability distribution is a key factor in extreme-value analysis, e.g. when estimating return levels. When a block-maximum approach is chosen for studying annual maxima of some quantity (e.g. daily maximum rainfall), extreme-value theory tells that certain distributions serve as limiting distributions, that can be summarised in the Generalised Extreme Value (GEV) distribution. However, the results are valid asymptotically and not seldom only small samples are available. In hydrology, one occasionally also considers the lognormal distribution and other alternatives, like the five-parameter Wakeby distribution, first presented in 1978 by Houghton [16] and to be investigated more closely in the sequel of this paper. Griffiths [10] claims that "the distribution has a secure theoretical basis and is hydrologically more realistic". A list of applications of this distribution is given in a recent paper by Busababodhin *et al.* along with proposed estimation techniques [7].

In recent years, generalisations of conventional distributions have been introduced, with the intention of being more flexible. Exponentiated distributions have for instance been proposed: exponentiated exponential, exponentiated Gumbel etc. For an investigation of the exponentiated Gumbel applied to series of significant wave height, see Persson and Rydén [23]. Another generalisation of the Gumbel distribution, the so-called Beta Gumbel distribution, was studied by Jonsson and Rydén [17], where this distribution was compared to the Gumbel and GEV distributions in a case study of extreme precipitation. However, for that study the difference between Beta Gumbel and GEV was minor, with respect to infomation criteria as well as estimated return levels and their uncertainties.

This paper serves two purposes. First, to check the intended flexibility of the Wakeby distribution through simulation studies for various sample sizes. Estimation of parameters will be made conveniently by L-moments [13], and hence the likelihood-based AIC and BIC are not options for model choice. Examinations of the differences between simulated samples and candidate distributions are based on various measures of minimum distance. As the tail behaviour is of particular interest for typical applications, upper quantiles are also compared. The second purpose is to study data of unregulated extreme floods in northern Sweden. Several distributions are considered, in particular the effects on estimated return levels due to various distribution assumptions on the tails. Moreover, the influence of record length is of interest.

The paper is outlined as follows. Section 2 serves as a background, introducing first of all the Wakeby distribution. Further, a review of the methodology for estimation by L-moments is given as well as presentation of the approaches for discerning distributions.

The simulation study is outlined and its main findings given in Section 3, and in Section 4 the case study of extreme floods in Sweden is presented, including estimated return levels for various situations.

2. BACKGROUND

2.1. The Wakeby distribution with applications

The Wakeby distribution was presented by Houghton [16], along with results of goodness-of-fit tests for observations of extreme floods. We here give the parametrisation by Hosking and Wallis found in [15], a five-parameter distribution:

$$x(F) = \xi + \frac{\alpha}{\beta}\big[1-(1-F)^{\beta}\big] - \frac{\gamma}{\delta}\big[1-(1-F)^{-\delta}\big], \tag{2.1}$$

where $F \in [0,1]$. The following parameter restrictions are valid: either $\beta+\delta > 0$ or $\beta = \gamma = \delta = 0$; if $\alpha = 0$ then $\beta = 0$; if $\gamma = 0$ then $\delta = 0$. The generalised Pareto distribution follows with the formulation in equation (2.1) as the special $\alpha = 0$ or $\gamma = 0$. Note that the definition is stated in terms of the quantile function, which faciliates estimation of return levels. In addition, simulation of random numbers can be performed by the inverse method.

This distribution has been applied successfully for various quantities in the earth sciences. A list of applications is given in [7]. In his landmark paper [16], Houghton examined the fit of observations of floods from stations in the United States, and Griffiths [10] investigated flood data from New Zealand.

2.2. Estimation of parameters

In this study, we employ estimation by L-moments, which is convenient for the five-parameter Wakeby distribution. For instance, Busababodhin *et al.* point out that maximum-likelihood estimates are not easily obtained [7]. Moreover, the methodology is in widespread use in many countries; see [6] for a list of studies performed by L-moments. Hosking [13] claims that estimation of parameters by L-moments is occasionally more accurate in small samples. Furthermore, quantile functions can be expressed in terms of L-moments, a clear advantage in hydrological sciences when estimating return levels. For the computational work in this paper, the implementations in the R packages `lmom` and `lmomco` were used ([14], [5]), following the parameterisation in equation (2.1).

2.2.1. Introduction to L-moments

The methodology with L-moments was introduced by Hosking [13]. The L-moments are the quantities λ_r as follows, and are linear functions of order statistics:

$$\lambda_r = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k\binom{r-1}{k}\mathsf{E}[X_{r-k:r}], \quad r = 1, 2, ...,$$

where $X_{1:n} \leq \cdots \leq X_{n:n}$ are the order statistics of a random sample of size n drawn from the distribution of a random variable X. In applied studies, moments could be standardised, becoming independent of the units of measurement. These so-called L-moment ratios are the quantities

$$\tau_r = \lambda_r/\lambda_2, \quad r = 3, 4, \dots .$$

The measures τ_3 and τ_4 can be regarded as measures of skewness and kurtosis. For instance, for a symmetrical distribution, $\tau_3 = 0$. Further details on L-moments are found in Appendix.

2.2.2. Remarks on estimation methodologies

The notion of L-moments has been extended, for instance trimmed L-moments (TL moments), [9]. TL moments with the smallest value trimmed with an application to the generalised Pareto distribution were considered in [1]. An estimation method using higher-order L-moments, so-called LH-moments, was presented in [7].

Recently, versions of L-moments as well as maximum-likelihood methods for estimation of high quantiles of the generalised Pareto and generalised extreme-value distribution have been compared [27]. The authors concluded that “there are small differences when estimating high quantiles of the GPD or GEV distributions. It was revealed that L-moment and maximum likelihood methods outperform LQ- and TL-moment methods: the L-moment method is preferred for heavy-tailed distributions, while the maximum likelihood method is recommended for light-tailed distributions.” Thus, from these findings, we are motivated in the choice of estimation by L-moments in this paper.

2.3. Evaluating candidate distributions

Already in the paper by Houghton [16], goodness-of-fit tests were considered in the analysis, and many papers in e.g. hydrology apply various versions of such tests. However, as pointed out by Wilks [28]: “Of substantially more interest is the closeness of fit on the right tail, since it is here that extrapolations of relevance to engineering design and other applications will be made.” In the literature, there seems to be no concensus on a specific procedure (in the forms of visualisations, goodness-of-fit measures, computer-intensive methods) to apply. Uses of criteria like AIC and BIC for model choice is often a convenient strategy, not the least to compare distributions (or models) with varying number of parameters.

However, in this paper we use L-moments for estimation, not maximum-likelihood estimation, and hence other approaches have to be taken.

In the sequel of this paper, we will perform simulations from a particular distribution and compare to candidate distributions (to be described in detail in Section 3). The so resulting samples will be compared by firstly, two general distance measures, secondly, comparison of high quantiles. Moreover, when analysing observed river-flow data from stations in Section 4, the so-called L-moment diagram will assist in interpretations.

2.3.1. Distance measures

In the literature, there is a substantial number of distance, or similarity, measures in various scientific fields. A review is given by Cha, where measures also are categorised [8]. In the presentation below, we assume that two probability densities P and Q, each in a discretised "histogram" form of B values, are to be compared.

Some measures are said to belong to Shannon's entropy family. In this paper, we chose the Kullback–Leibler distance [18]:

$$d_{\mathrm{KL}} = \sum_{j=1}^{B} p_i \ln \frac{p_i}{q_i}.$$

In another category, the measures are based on geometric means: the fidelity or squared-chord family. The simplest version was chosen in this paper, the Fidelity similarity measure:

$$s_{\mathrm{F}} = \sum_{j=1}^{B} \sqrt{p_j q_j}.$$

Other alternatives in this category are Bhattacharyya and Hellinger distances. Both d_{KL} and s_{F} are interpreted that the smaller the value, the two objects (here, distributions) are closer and have a higher degree of similarity.

2.3.2. Comparison of quantiles

For a simulated sample from a specified random variable X, the upper quantile $x_{0.99}$ for which $\mathsf{P}(X > x_{0.99}) = 0.01$, is estimated, resulting in $x^*_{0.99}$, say. Based on the simulated sample, candidate distributions are fitted with L moments, and the related upper quantiles are estimated. The absolute differences between these estimates and $x^*_{0.99}$ are finally calculated. Further details on the simulation procedure are given in Section 3.1.

3. SIMULATION STUDY: DIFFERENCES AMONG DISTRIBUTIONS

In this section, we investigate how distances between distributions differ, given simulated observations from a parent distribution. We will use the approaches presented in Section 2.3. In addition to the Wakeby distribution, we will consider two other distributions often encountered in the earth sciences or hydrology: the generalised extreme value (GEV) distribution and the three-parameter lognormal (LN3). Though familiar and well known in the research domains mentioned, we present them below in order to present their parameters.

The GEV distribution has three parameters (location μ, scale σ and shape ξ), and is commonly stated by its distribution function:

$$F(x;\mu,\sigma,\xi) = \begin{cases} \exp\Big\{-\big[1+\xi\frac{x-\mu}{\sigma}\big]^{-1/\xi}\Big\}, & \xi \neq 0, \\ \exp\big\{-\exp\big[-\frac{x-\mu}{\sigma}\big]\big\}, & \xi = 0, \end{cases}$$

where $\mu \in \mathbb{R}$, $\sigma > 0$ and $\xi \in \mathbb{R}$. The shape parameter ξ affects the support of this distribution: when $\xi = 0$, the GEV distribution is the Gumbel distribution (with support $\mathbb{R}$). When $\xi > 0$, the distribution corresponds to the Fréchet distribution with support $x \geq \mu + \sigma/\xi$, and when $\xi < 0$ it corresponds to the reversed Weibull distribution with support $x \leq \mu - \sigma/\xi$.

Consider the LN3 distribution with distribution function

$$F(x) = \Phi(y), \quad x > 0,$$

where $y = (\ln(x-\zeta) - \mu)/\sigma$ and $\Phi(y)$ is the distribution function of the standard normal distribution. In other words, the density function of X is given as

$$f(x;\mu,\sigma,\zeta) = \frac{1}{(x-\zeta)\sigma\sqrt{2\pi}} \exp\bigg\{-\frac{(\ln(x-\zeta)-\mu)^2}{2\sigma^2}\bigg\}, \quad x > \zeta \geq 0,$$

where $\mu \in \mathbb{R}$, $\sigma > 0$. If X is distributed as above, $Y = \ln(X - \zeta)$ has a normal distribution with mean μ and variance σ^2.

3.1. Algorithm of simulation study

In this study, we simulate from a given parent distribution: GEV or LN3. The sample size n was chosen in the range from 25 to 200, in steps of 25 at the lower sample sizes.

1. For each sample size, simulate $N = 5000$ samples from a parent distribution.
2. For each sample, compute the L-moments by the R package `lmomco` and then the probability-density functions for the candidate distributions LN3, Wakeby and GEV, evaluated at the sample points.
3. Estimate the probability-density function for the sample (by the R routine `density`), and compute measures s_{F} and d_{KL} for comparison with the densities obtained in step 2. In addition, compute the upper 0.99 quantiles for the sample and the candidate distributions.
4. Register which distribution alternative had the smallest deviation from the simulated sample, in terms of s_{F}, d_{KL} and upper quantile, respectively. Over the N samples the overall proportions of "winners" (in terms of smallest distance) from the three distribution alternatives can be collected, resulting in a triple with the three components summing up to one. For instance, with d_{KL} considered, GEV, LN3 and Wakeby could result in the triple (0.25, 0.15, 0.60), i.e. Wakeby here resulted in the smallest distance in the majority of cases.

3.2. L-moment ratio diagram

Before turning to the simulations outlined above, let us illustrate the notion of the so-called L-moment ratio diagram with simulations from a GEV distribution. In such a diagram is found τ_3 on the abscissa and τ_4 on the ordinate. Probability distributions can be illustrated as curves and in some cases as points. For instance, the uniform distribution has $(\tau_3, \tau_4) = (0, 0)$, for the Gumbel distribution $(\tau_3, \tau_4) \doteq (0.17, 0.15)$, and the GEV distribution forms a curve in the τ_3-τ_4 plane [13].

In Figure 1, the curve for the GEV distribution is drawn along with dots corresponding to L-moments from 500 simulated samples from a GEV distribution. The left panel shows the result for sample size $n = 25$, the right panel shows the case $n = 100$. We note for the smaller sample size a considerable spreading in the (τ_3, τ_4) plane, relatively the larger sample size. This feature could be kept in mind, when facing real data in Section 4.

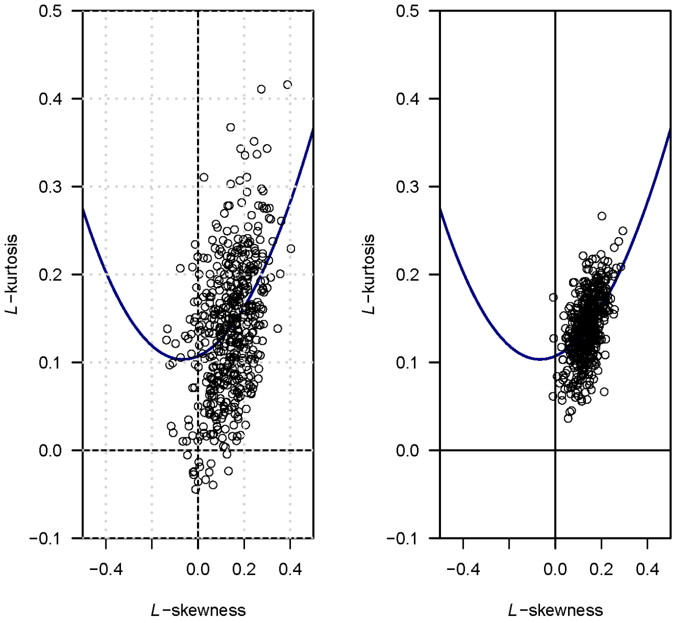


Figure 1: Simulation from a GEV distribution with $\tau_3 = \tau_4 = 0.14$. The solid curve represents a GEV distribution in the (τ_3, τ_4) space. Left panel: sample size 25 (500 samples). Right panel: sample size 100 (500 samples).

An illustration how this type of plot can assist in distinguishing between distributions is found in [13], Section 3.5. This visualisation tecnique has shown itself useful in hydrology [22], [20].

3.3. Case 1. Simulation from GEV distribution

We first study the case of the parent distribution being the standard Gumbel distribution:

$$F(x) = \exp(-\mathrm{e}^{-x}), \quad x \in \mathbb{R}.$$

In Figure 2, the proportions of smallest distance are shown for the three potential distributions as function of sample size. Thus, for each sample size, the proportions obviously sum to one. The left panel shows results based on distance in upper quantile, middle panel d_{KL} and right panel s_{F}.

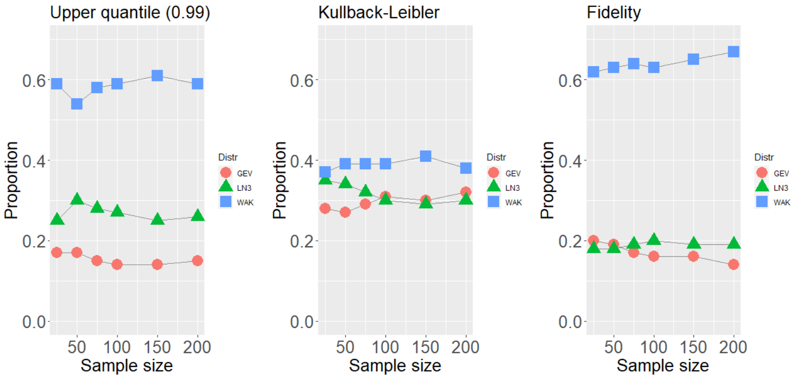


Figure 2: Simulation from a Gumbel distribution ($\tau_3 = 0.17$, $\tau_4 = 0.15$). For each sample size, the proportions of smallest distance for each distribution family are displayed.

From the plots in Figure 2, the Wakeby distribution is for all the measures considered, and regardless of the sample size, the choice which for a majority of cases is the closest to the simulated sample.

In the next example, a GEV distribution with $\tau_3 = 0.07$, $\tau_4 = 0.25$ is the parent distribution, and results are shown in Figure 3. The conclusions are similar to the preceding case: the Wakeby gives in the majority of cases the best fit of the simulated data.

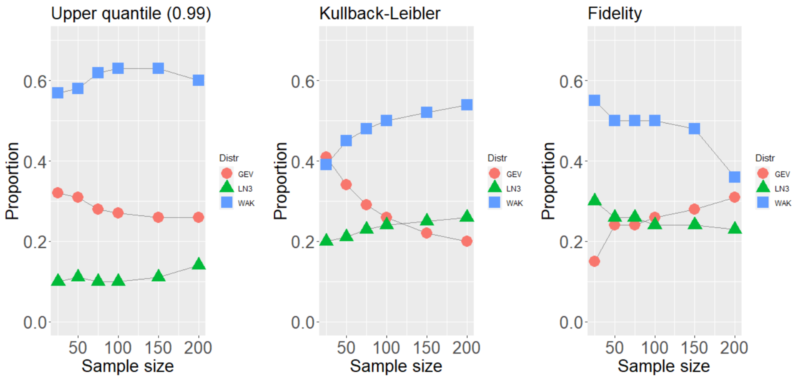


Figure 3: Simulation from a GEV distribution ($\tau_3 = 0.07$, $\tau_4 = 0.25$). For each sample size, the proportions of smallest distance for each distribution family are displayed.

3.4. Case 2. Simulation from log-normal distribution (LN3)

We here simulate from the LN3 distribution with $\tau_3 = 0.07$, $\tau_4 = 0.25$. Results for the three measures are shown in Figure 4. Again, the Wakeby distribution is the best option, regardless of sample size or measure.

Remark. Note that the GEV and LN3 simulations had the same choices of τ_3 and τ_4, respectively. Actually, in order to have realistic values, Station 11 in the next section was used here: fitting each distribution in case by L-moments and rendering parameters in the relevant distribution for the actual simulation study. Obviously, the estimates of τ_3 and τ_4 remain the same, since the same original sample is considered.

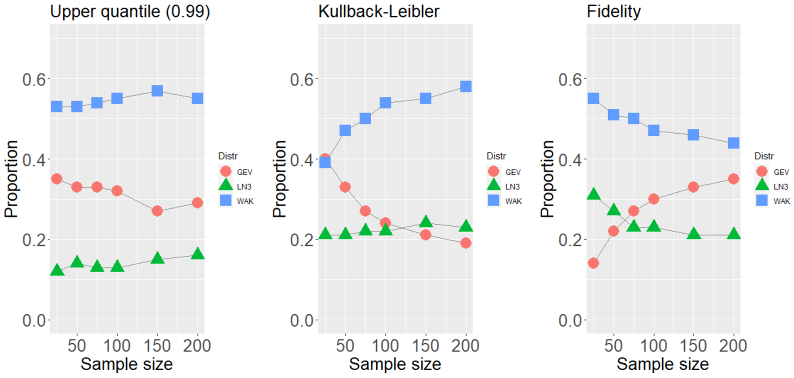


Figure 4: Simulation from an LN3 distribution ($\tau_3 = 0.07$, $\tau_4 = 0.25$). For each sample size, the proportions of smallest distance for each distribution family are displayed.

4. CASE STUDY: FLOOD FLOWS IN SWEDEN

In this section we focus on flood flows from northern Sweden. The aim is to fit annual maximum flows by the three distributions considered earlier in the paper (GEV, LN3, Wakeby). We do not consider possible non-stationary effects due to climate change; for further discussion, see [4], where no significant trends were discerned in annual maximum daily flow in Sweden over the past 100 years.

Data are available from Swedish Meteorological and Hydrological Institute (SMHI), online address: `http://vattenwebb.smhi.se/station/#`. Unregulated rivers in northern Sweden were considered. As long series as possible were chosen, for possible GEV asymptotics to work. In all, eleven stations were selected and descriptions are given in Table 1. For the rest of the paper, they will for simplicity be referred to as Station 1, ..., Station 11.

Table 1: Information on selected stations.

Nr.	Station	Name	River ID	River	Area (km^2)	Start	End
1	4	Junosuando	1000	Torne	4348.0	1968	2019
2	957	Övre Abiskojokk	1000	Torne	566.3	1986	2019
3	2012	Pajala pumphus	1000	Torne	11038.1	1970	2019
4	2357	Abisko	1000	Torneträsk	3345.5	1985	2019
5	2395	Kallio 2	1000	Muonio älv	14477.1	1988	2019
6	16722	Kukkolankoski övre	1000	Torne	33929.6	1911	2019
7	11	Männikkö	4000	Tärendö	5856.2	1976	2019
8	17	Räktfors	4000	Kalix	23102.9	1937	2019
9	1456	Kaalasjärvi	4000	Kalix	1472.5	1975	2019
10	2159	Killingi	4000	Kalix	2345.5	1976	2019
11	2358	Tärendö 2	4000	Kalix	13000.0	1985	2019

For these stations, the L-moments were estimated for each time series, and the results are shown in an L-moment diagram (Figure 5). One could note that stations 7, 9, 10 and 11 tend to form a group in the plane. Indeed, these stations belong to the same river system (Kalix, River ID 4000). Moreover, Stations 2, 4 and 6 are located close to the curve for GEV distribution.

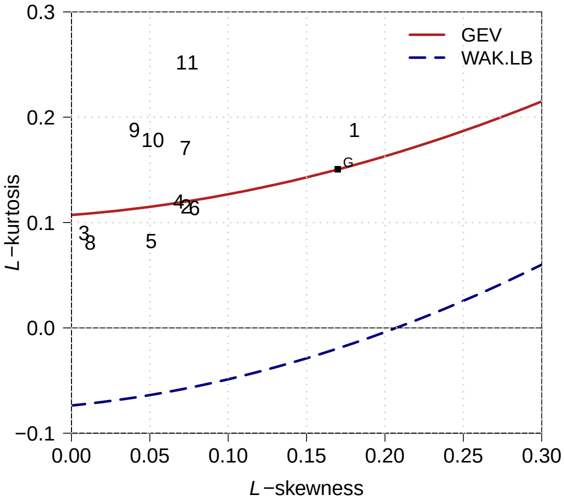


Figure 5: L-moment ratio diagram. Solid line: GEV distribution. Dashed line: Wakeby distribution, lower bound. Numbers: Stations 1–11.

4.1. Estimated quantities

The measures d_{KL} and s_{F} were computed, comparing the original sample and the three candidate distributions with parameters fitted by L-moments. These measures are presented in Table 2 along with estimates of τ_3, τ_4 and the shape parameter ξ in the GEV distribution.

Table 2: Stations 1–11: Estimates of L-moment ratios τ_3 and τ_4; estimate of shape parameter ξ in GEV; distance measures d_{KL} and s_{F} respectively, between sample and fitted candidat distribution.

Station	τ_3	τ_4	Shape ξ	$d_{\mathrm{KL}}^{\mathrm{GEV}}$	$d_{\mathrm{KL}}^{\mathrm{LN3}}$	$d_{\mathrm{KL}}^{\mathrm{WAK}}$	$s_{\mathrm{F}}^{\mathrm{GEV}}$	$s_{\mathrm{F}}^{\mathrm{LN3}}$	$s_{\mathrm{F}}^{\mathrm{WAK}}$
1	0.18	0.19	−0.016	7.304	7.289	7.596	1.765	1.766	1.731
2	0.07	0.12	0.16	13.239	13.262	13.232	2.217	2.215	2.210
3	0.008	0.09	0.27	6.175	6.209	6.130	1.918	1.913	1.926
4	0.07	0.12	0.16	11.667	11.696	11.627	2.269	2.267	2.271
5	0.05	0.08	0.19	12.054	12.077	12.049	2.403	2.397	2.425
6	0.08	0.11	0.15	2.123	2.135	2.091	1.264	1.264	1.264
7	0.07	0.17	0.16	8.862	8.863	9.131	1.949	1.948	1.934
8	0.01	0.08	0.26	2.704	2.690	2.649	1.449	1.446	1.442
9	0.04	0.19	0.21	9.230	9.224	9.039	1.844	1.843	1.825
10	0.05	0.18	0.19	14.759	14.778	15.108	2.062	2.061	2.052
11	0.07	0.25	0.16	11.415	11.417	11.817	1.766	1.765	1.720

From this table, we may reflect upon the following:

- Absolute differences between the measures are generally quite small; the distributions are, in this meaning, close for description of data.
- For each distance measure, d_{KL} and s_{F} respectively, Wakeby gives the closest fit in a majority of cases (7 out of 11 for each measure). For 5 out of 11 samples, *both* measures d_{KL} and s_{F} gave preference for Wakeby.
- Station 6 has the longest period of observations, 109 years. Here one can note that for the measure s_{F}, all three distribution options yield results equal up to the third decimal.
- The GEV distribution is, interestingly, seldom the distribution with minimum distance to the sample. The asympotics of the maximum distribution seems not to have been attained for these samples. From the L-moment ratio diagram in Figure 5, Stations 2, 4 and 6 are close to the GEV curve, but for the measures considered, there are only minor differences between the distribution options.

4.2. Return levels

A T-year return level x_T is often defined as the high quantile for which the probability that the annual maximum exceeds this quantile is $1/T$, hence $F(x_T) = 1 - 1/T$ where $F(.)$ is the distribution function for the series of maxima. We consider T in the range from 10 to 1000 years and estimate return levels based on quantiles for the three distribution families considered above: GEV, LN3, Wakeby.

In Figure 6, we note minor differences between distributions for low T values, but for most stations a considerable spreading for $T = 1000$. In particular, for Stations 2, 9 and 11, the Wakeby 1000-year return level is remarkably higher than the alternatives. Station 6, with the largest observation period, also has a notable difference between distribution choices at $T = 1000$, with the LN3 alternative resulting in the highest levels.

5. CONCLUDING REMARKS

In this paper, we have investigated the use of the Wakeby distribution. Through simulation studies, we found that based on several distance measures, the Wakeby distribution has a good fit to the tail, regardless of the distribution of origin (lognormal or generalised extreme-value distribution) and sample size. One could remark that although a certain distribution was registred as the "winner" (in terms of smallest distance), for a particular sample and choice of distance measure, often in practice the differences between distributions are quite small; cf. the detailed numerical outcomes for the flood data (Table 2).

In addition, we examined annual extreme floods with respect to fit of distribution and estimation of return levels. Uncertainties of return levels, e.g. in the form of confidence intervals, were not provided in the study, but although intervals in this context typically tend to be wide, the selection of the very distribution shows itself to be of interest for high quantiles ($T = 1000$).

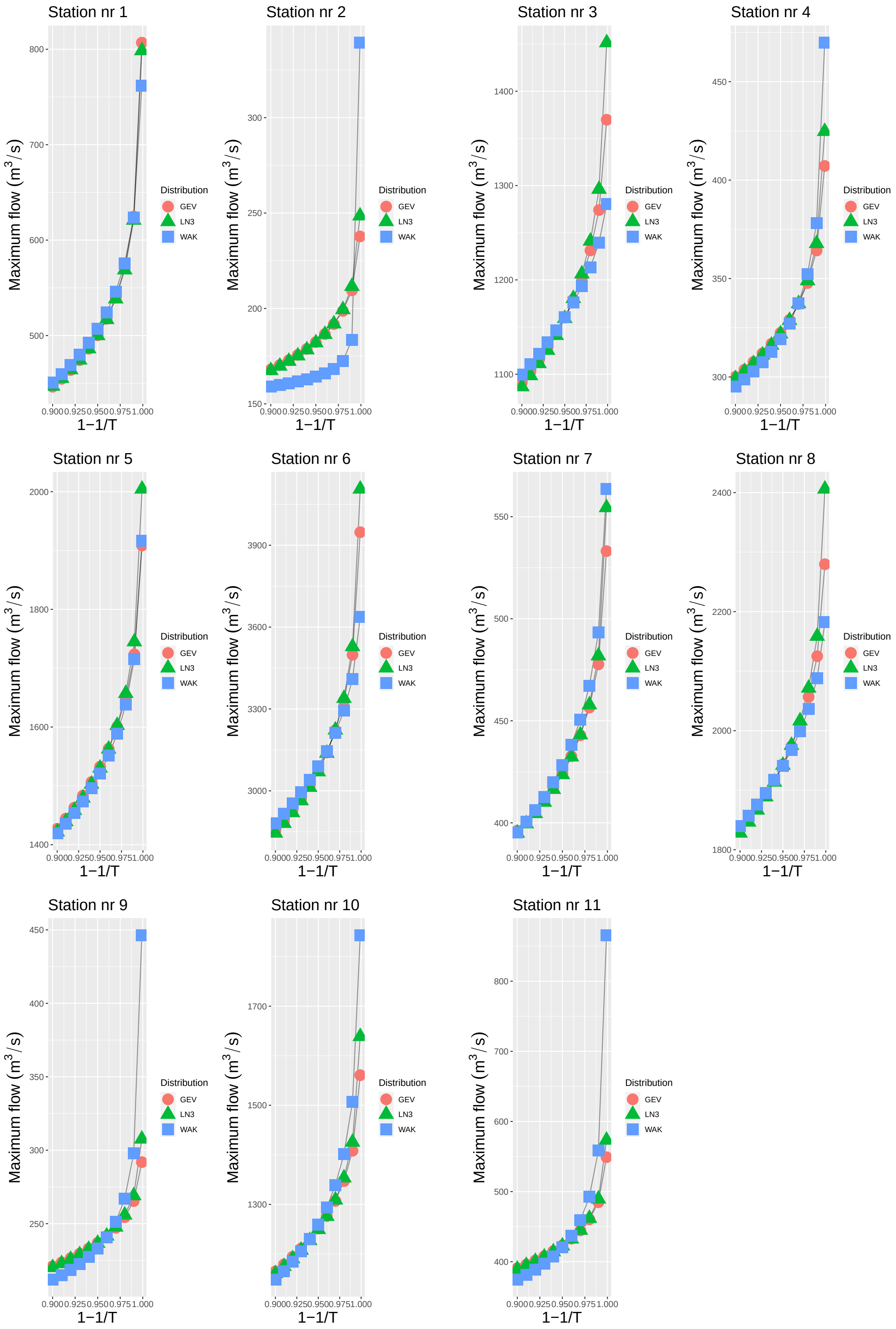


Figure 6: Return levels, based on station data. Abscissa: $1 - 1/T$, where T is the return period, starting from $T = 10$, final value $T = 1000$. The ordinate shows the related estimated return level.

We studied here the quantity of extreme floods, like in the vintage paper on Wakeby distribution [16]. In the literature, the Wakeby distribution has also been employed to model rainfall, [21] (although these authors did not motivate the choice of Wakeby distribution, compared to other alternatives). Further studies could be performed to investigate extreme rainfall. Moreover, concerning extreme daily rainfall, Papalexiou and Koutsoyiannis found by fitting GEV distributions to records worldwide that the record length strongly affects the estimate of the GEV shape [20]. Furthermore, the influence on the shape parameter was analysed. Further studies in this direction, employing a Bayesian approach are found in [24]. To conclude, the longer the observation series, the more likely the GEV distribution might be attained. Further studies on extreme floods with different observation lengths would be interesting.

Several options for analysis of minimum distance are available; for a review, see [8]. A version of the Anderson–Darling test statistic for analysis of tail deviation at the upper tail was suggested in [26]. The author experimented with that measure, but overall conclusions in the simulation studies were as for the measures presented in this paper: the Wakeby distribution gives the better fit.

To end this paragraph, and indeed the paper, we cite Haktanir and Horlacher [11]:

> “Because of the ample availability of computers nowadays, a single-site flood frequency analysis should be done with the inclusion of many standard probability distributions, and a final decision should be made combining experience with engineering judgement.”

Even more today, some decades later, computers and related software are important tools. In a strategy for estimation for a certain region, one could still agree that several potential distributions are possible. Methodology for selection of candidates is of interest to further analyse.

APPENDIX

Let X be a real-valued random variable with distribution function $F(x)$ and quantile function $x(F)$. Moreover, denote by $X_{1:n} \leq X_{2:n} \leq X_{n:n}$ the order statistics for a random sample of size n. The L-moments are defined in [13] as the quantities

$$\lambda_r = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k \binom{r-1}{k} \mathsf{E}[X_{r-k:r}], \qquad r = 1, 2, \dots \, .$$

The first four L-moments can be shown to be

$$\begin{aligned}
\lambda_1 = \mathsf{E}[X] &= \int_0^1 x(F)\,\mathrm{d}F, \\
\lambda_2 = \mathsf{E}[X_{2:2} - X_{1:2}] &= \int_0^1 x(F)(2F-1)\,\mathrm{d}F, \\
\lambda_3 = \frac{1}{3}\mathsf{E}[X_{3:3} - 2X_{2:3} + X_{1:3}] &= \int_0^1 x(F)(6F^2 - 6F + 1)\,\mathrm{d}F, \\
\lambda_4 = \frac{1}{4}\mathsf{E}[X_{4:4} - 3X_{3:4} + 3X_{2:4} - x_{1:4}] &= \int_0^1 x(F)(20F^3 - 30F^2 + 12F - 1)\,\mathrm{d}F.
\end{aligned}$$

In practice, L-moments must be estimated from samples. The rth sample L-moment ℓ_r, estimated as a U-statistic, can be computed by

$$\ell_r = \frac{1}{n}\sum_{k=0}^{r-1}\sum_{i=1}^{n}(-1)^{r-1-k}\binom{r-1}{k}\binom{r-1+k}{k}\frac{(i-1)(i-2)\cdots(i-k)}{(n-1)(n-2)\cdots(n-k)}x_{i:n}.$$

Note, for instance, that $\ell_1 = \bar{x} = n^{-1}\sum_i x_i$.

L-moment ratios are L-moments that are standardized:

$$\tau_r = \frac{\lambda_r}{\lambda_2}, \qquad r = 3, 4, \dots \, .$$

Values of τ_3 and τ_4 are often plotted against each other, resulting in an L-moment diagram.

Hosking presents in [13], Table 1, the L-moment ratios for some common distributions. For instance, with relevance for this article, the Gumbel distribution has

$$\begin{aligned}
\tau_3 &= \ln(9/8)/\ln 2 \doteq 0.17, \\
\tau_4 &= (16\ln 2 - 10\ln 3)/\ln 2 \doteq 0.15.
\end{aligned}$$

For a GEV distribution with shape parameter ξ,

$$\begin{aligned}
\tau_3 &= 2(1-3)^{-\xi}/(1-2^{-\xi}) - 3, \\
\tau_4 &= \big[5(1-4^{-\xi}) - 10(1-3^{-\xi}) + 6(1-2^{-\xi})\big]/(1-2^{-\xi})
\end{aligned}$$

(the formula for τ_4 is here given following the report [12]; there seems to be a misprint in [13]).

ACKNOWLEDGMENTS

I am grateful to Staffan Betnér for discussion about visualisations and related coding. Thanks also to an anonymous referee for valuable comments.

REFERENCES

[1] Ahmad, U.N.; Shabri, A. and Zakaria, Z.A. (2011). Trimmed L-moments (1,0) for the generalized Pareto distribution, *Hydrological Sciences Journal*, **56**(6), 1053–1060.

[2] Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, 716–723.

[3] Anderson, T.W. and Darling, D.A. (1952). Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes, *Annals of Mathematical Statistics*, **23**, 93–212.

[4] Arheimer, B. and Lindström, G. (2015). Climate impact on floods: changes in high flows in Sweden in the past and the future (1911–2100), *Hydrology and Earth System Sciences*, **19**, 771–784.

[5] Asquith, W. (2021). *lmomco – L-moments, censored L-moments, trimmed L-moments, L-comoments, and many distributions*, R package version 2.3.7.

[6] Aydoğan, D.; Kankal, M. and Önsoy, H. (2016). Regional flood frequency analysis for Çoruh Basin of Turkey with L-moments approach, *Journal of Flood Risk Management*, **9**(1), 69–86.

[7] Busababodhin, P.; Seo, Y.A.; Park, J.-S. and Kumphon, B. (2016). LH-moment estimation of Wakeby distribution with hydrological applications, *Stochastic Environmental Research and Risk Assessment*, **30**(6), 1757–1767.

[8] Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions, *International Journal of Mathematical Models and Methods in Applied Sciences*, **4**(1), 300–307.

[9] Elamir, E.A.H. and Seheult, A.H. (2003). Trimmed L-moments, *Computational Statistics & Data Analysis*, **43**(3), 299–314.

[10] Griffiths, G.A. (1987). A theoretically based Wakeby distribution for annual flood series, *Hydrological Sciences Journal*, **34**(3), 231–248.

[11] Haktanir, T. and Horlacher, H.B. (1993). Evaluation of various distributions for flood frequency analysis, *Hydrological Sciences Journal*, **35**, 15–32.

[12] Hosking, J.R.M. (1986). *The theory of probability weighted moments*, "Research Report RC12210", IBM Research Division, Yorktown Heights, New York.

[13] Hosking, J.R.M. (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics, *Journal of the Royal Statistical Society, Series B*, **52**(1), 105–124.

[14] Hosking, J.R.M. (2019). *L-Moments*, R package, version 2.8.
`https://CRAN.R-project.org/package=lmom`

[15] Hosking, J.R.M. and Wallis, J.R. (1997). *Regional Frequency Analysis: An Approach Based on L-moments*, Cambridge University Press, Cambridge, United Kingdom.

[16] Houghton, J.C. (1978). Birth of a parent: the Wakeby distribution for modeling flood flows, *Water Resources Research*, **14**, 1105–1110.

[17] Jonsson, F. and Rydén, J. (2017). Statistical studies of the Beta Gumbel distribution: estimation of extreme levels of precipitation, *Statistica Applicata*, **29**, 5–27.

[18] Kullback, S. and Leibler, R.A. (1951). On information and sufficiency, *Annals of Mathematical Statistics*, **22**, 79–86.

[19] Önöz, B. and Bayazit, M. (1995). Best-fit distributions of largest available flood samples, *Journal of Hydrology*, **167**, 195–208.

[20] Papalexiou, S.M. and Koutsoyiannis, D. (2013). Battle of extreme value distributions: a global survey on extreme daily rainfall, *Water Resources Research*, **49**, 187–201.

[21] Park, J.-S.; Jung, H.-S.; Kim, R.-S. and Oh, J.-H. (2001). Modelling summer extreme rainfall over the Korean peninsula using Wakeby distribution, *International Journal of Climatology*, **21**, 1371–1384.

[22] Peel, M.C.; Wang, Q.J.; Vogel, R.M. and McMahon, T.A. (2001). The utility of L-moment ratio diagrams for selecting a regional probability distribution, *Hydrological Sciences – Journal des Sciences Hydrologiques*, **46**, 147–155.

[23] Persson, K. and Rydén, J. (2010). Exponentiated Gumbel distribution for estimation of return levels of significant wave height, *Journal of Environmental Statistics*, **1**(3), 1–12.

[24] Ragulina, G. and Reitan, T. (2017). Generalized extreme value shape parameter and its nature for extreme precipitation using long time series and the Bayesian approach, *Hydrological Sciences Journal*, **62**(6), 863–879.

[25] Schwarz, G.E. (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**, 461–464.

[26] Shin, H.; Jung, Y.; Jeong, C. and Heo, J.-H. (2012). Assessment of modified Anderson-Darling test statistics for the generalized extreme value and generalized logistic distributions, *Stochastic Environmental Research and Risk Assessment*, **26**, 105–114.

[27] Šimková, T. and Picek, J. (2017). A comparison of L-, LQ-, TL-moment and maximum likelihood high quantile estimates of the GPD and GEV distribution, *Communications in Statistics – Simulation and Computation*, **46**(8), 5991–6010.

[28] Wilks, D.S. (1993). Comparison of three-parameter probability distributions for representing annual extreme and partial duration precipitation series, *Water Resources Research*, **29**, 3543–3549.

REVSTAT – Statistical Journal
Volume 22, Number 1, January 2024, 61–86
https://doi.org/10.57805/revstat.v22i1.456

On q-Generalized Extreme Values under Power Normalization with Properties, Estimation Methods and Applications to COVID-19 Data

Authors: Mohamed S. Eliwa
– Department of Statistics and Operation Research, College of Science, Qassim University, Buraydah 51482, Saudi Arabia
– Department of Statistics and Computer Science, Faculty of Science, Mansoura University, Mansoura 35516, Egypt
mseliwa@mans.edu.eg

E.O. Abo Zaid
– Department of Mathematics and Computer Science, Faculty of Science, Suez University, Egypt
Esraa.AboZaid@sci.suezuni.edu.eg

Mahmoud El-Morshedy ✉
– Department of Mathematics, College of Science and Humanities in Al-Kharj, Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
– Department of Mathematics, Faculty of Science, Mansoura University, Mansoura 35516, Egypt
m.elmorshedy@psau.edu.sa

Received: July 2021 Revised: March 2022 Accepted: April 2022

Abstract:

- This paper introduces the q-analogues of the generalized extreme value distribution and its discrete counterpart under power normalization. The inclusion of the parameter q enhances modeling flexibility. The continuous extended model can produce various types of hazard rate functions, with supports that can be finite, infinite, or bounded above or below. Additionally, these new models can effectively handle skewed data, particularly those with highly extreme observations. Statistical properties of the proposed continuous distribution are presented, and the model parameters are estimated using various approaches. A simulation study evaluates the performance of the estimators across different sample sizes. Finally, three distinct real datasets are analyzed to demonstrate the versatility of the proposed model.

Keywords:

- *extreme value theory; survival discretization approach; entropy; estimation methods; simulation; COVID-19.*

AMS Subject Classification:

- 60E05, 62E10, 62E15, 62F10, 62N05.

✉ Corresponding author.

1. INTRODUCTION

In the last two decades, there has been an increasing interest in building statistical models for estimating the probability of rare and extreme events. These models involving extreme value theory (EVT) are of a great interest in environmental sciences, engineering, finance, insurance, and many other disciplines. Especially in finance, extreme price movement of a financial asset or a market index can be defined as the lowest and highest costs in an observed period (see Gilli, 2006 [19]). EVT shows that the asymptotic minimum and maximum returns have a definite shape that is independent of the return process itself. The EVT deals with the probabilistic description of the extremes of a stochastic sequence. The fundamental results of Fisher and Tippett (1928) [17] constitute the backbone of the classical EVT. The fundamental theorem states that maxima of independent and identically distributed random variables have one of the three extreme value distributions: Fréchet distribution, with infinite upper and heavy tail, Gumbel distribution, whose upper tail is also infinite, but lighter than the Fréchet distribution, and inverse Weibull distribution with finite upper tail. The three previous models can be gathered in the following family

$$(1.1)\qquad G_\xi(x;\mu,\sigma,\xi)=\begin{cases}\exp\{-(1+\xi(\frac{x-\mu}{\sigma}))^{\frac{-1}{\xi}}\}; & \xi\neq 0,\\ \exp\{-\exp(-\frac{x-\mu}{\sigma})\}; & \xi\to 0,\end{cases}$$

and

$$(1.2)\qquad g_\xi(x;\mu,\sigma,\xi)=\begin{cases}\frac{1}{\sigma}\exp\{-(1+\xi(\frac{x-\mu}{\sigma}))^{\frac{-1}{\xi}}\}(1+\xi(\frac{x-\mu}{\sigma}))^{\frac{-1}{\xi}-1}; & \xi\neq 0,\\ \frac{1}{\sigma}\exp\{-\exp(-(\frac{x-\mu}{\sigma}))\}\exp(-(\frac{x-\mu}{\sigma})); & \xi\to 0,\end{cases}$$

where μ is a location parameter, σ is a positive scale parameter, and ξ is the shape parameter, for more detail (see De Haan and Ferreira, 2007 [14]). The cumulative distribution function (CDF) and probability density function (PDF) in Equations (1.1) and (1.2), respectively, are known as the generalized extreme value distribution under linear normalization (GEVL). Another reason for using the power normalization in EVT is concerning the possibility of getting a better rate of convergence in EVT (see Barakat *et al.*, 2010 [2]). The CDF F is said to belong to the max stable model under power normalization or simply p-max domain of attraction of a non-degenerate CDF H, denote by $F\in D_p(H)$, if for some norming constants $\alpha_n>0$ and $\beta_n>0$, we have

$$(1.3)\qquad P(|\frac{X_{n:n}}{\alpha_n}|^{1/\beta_n}\operatorname{sign}(X_{n:n})\leq x)=F^n(\alpha_n|x|^{\beta_n}\operatorname{sign}(x))\xrightarrow[n]{w}H(x),$$

where $\operatorname{sign}(x)=-1$, or 0, or 1, according as $x<0$, or $x=0$, or $x>0$. Pantcheva (1985) [22] proved that $H(x)$ belongs to one p-type of the following six classes of extreme value distributions

$$\text{Type-I}:\qquad H_{1,\beta}(x)=\begin{cases}0; & x\leq 1,\\ \exp\{-(\log x)^{-\beta}\}; & x>1,\beta>0,\end{cases}$$

$$\text{Type-II}:\qquad H_{2,\beta}(x)=\begin{cases}0; & x\leq 0,\\ \exp\{-(-\log x)^{\beta}\}; & -1\leq x\leq 1,\\ 1; & x>1,\end{cases}$$

$$\text{Type-III}: \quad H_{3,\beta}(x) = \begin{cases} 0; & x \leq -1, \\ \exp\{-(-\log(-x)^{-\beta}\}; & -1 \leq x \leq 0, \\ 1; & x > 0, \end{cases}$$

$$\text{Type-IV}: \quad H_{4,\beta}(x) = \begin{cases} \exp\{-(\log(-x))^{\beta}\}; & x \leq -1, \\ 1; & x > -1, \end{cases}$$

$$\text{Type-V}: \quad H_{5}(x) = \begin{cases} 0; & x \leq 0, \\ \exp\{-x^{-1}\}; & x > 0, \end{cases}$$

and

$$\text{Type-VI}: \quad H_{6}(x) = \begin{cases} \exp\{x\}; & x \leq 0, \\ 1; & x > 0. \end{cases} \tag{1.4}$$

Nasri-Roudsari (1999) [28] demonstrated that the six p-max stable laws in can be represented as two families. We call them log-GEVL distribution in positive support, and negative log-GEVL distribution in negative support, i.e.:

(1) For $x^0 > 0$, $x > 0$ and $1 + \frac{\xi}{\sigma}\log(e^{-\mu}x) > 0$

$$H_{\xi,1}(x;\mu,\sigma) = \begin{cases} \exp\{-(1 + \frac{\xi}{\sigma}\log(e^{-\mu}x))^{\frac{-1}{\xi}}\}; & \xi \neq 0, \\ \exp\{-(xe^{-\mu})^{\frac{1}{\sigma}}\}; & \xi \to 0. \end{cases} \tag{1.5}$$

(2) For $x^0 \leq 0$, $x \leq 0$ and $1 - \frac{\xi}{\sigma}\log(-e^{-\mu}x) > 0$

$$H_{\xi,2}(x;\mu,\sigma) = \begin{cases} \exp\{-(1 - \frac{\xi}{\sigma}\log(-e^{-\mu}x))^{\frac{-1}{\xi}}\}; & \xi \neq 0, \\ \exp\{-(-xe^{-\mu})^{\frac{-1}{\sigma}}\}; & \xi \to 0. \end{cases} \tag{1.6}$$

The corresponding density function to Equation (1.5) can be formulated as

$$h_{\xi}(x;\mu,\sigma,\xi) = \begin{cases} \frac{1}{\sigma x}\exp\{-(1 + \frac{\xi}{\sigma}\log(xe^{-\mu})\,\mathrm{sign}(x))^{\frac{-1}{\xi}}\}((1 + \frac{\xi}{\sigma}\log(xe^{-\mu})\,\mathrm{sign}(x)))^{\frac{-1}{\xi}-1}; & \xi \neq 0, \\ \exp\{-(xe^{-\mu}\,\mathrm{sign}(x))^{\frac{-1}{\sigma}}\}\,\frac{e^{-\mu}}{\sigma}(xe^{-\mu}\,\mathrm{sign}(x))^{\frac{-1}{\sigma}-1}; & \xi \to 0. \end{cases} \tag{1.7}$$

The results of Gnedenko *et al.* (1943) [21] and De Haan (1971) [13] concerning linear normalization were extended to p-max stable laws. They showed that every CDF attracted to linear max stable law is necessarily attracted to some p-max stable, and that p-max stable laws, in fact attract more. For more information about the extreme under power normalization and its applications, see Galambos (1987) [18], Nasri-Roudsari (1999) [28], Barakat *et al.* (2010 [2], 2013 [3], 2014a [4], 2014b [5], 2015 [6], 2019 [7]), among others.

In mathematical physics and probability, the q-distribution is more general than classical distribution. It was introduced by Diaz and Pariguan (2009) [12] and Diaz *et al.* (2010) [11] in the continuous case, and by Charalambides (2010) [9] in the discrete version. The construction of a q-distribution is the construction of a q-analogue of ordinary distribution. Mathai and Provost (2006) [27] introduced the q-analogue of the gamma distribution with respect to Lebesgue measure. Recently, several q-type super statistical distributions such as the

q-exponential, q-Weibull, and q-logistic were developed in the context of statistical mechanics, information theory and reliability modelling, as discussed for instance in Chung *et al.* (1994) [10], Picoli *et al.* (2003) [24], Gauchman (2004) [20], De Sole and Kac (2003) [31], Mathai (2005) [26], Srivastava and Choi (2012) [30], among others. Provost *et al.* (2018) [23] introduced the CDF and PDF of q-generalized extreme value under linear normalization (q-GEVL) and q-Gumbel distributions as

$$F(x;\mu,\sigma,\xi,q)=\begin{cases}[1+q(\xi(sx-m)+1)^{-\frac{1}{\xi}}]^{-\frac{1}{q}}; & \xi\neq 0,\quad q\neq 0,\\ (1+qe^{-(sx-m)})^{-\frac{1}{q}}; & \xi\to 0,\quad q\neq 0,\end{cases}\tag{1.8}$$

and

$$f(x;\mu,\sigma,\xi,q)=\begin{cases}s(1+\xi(sx-m))^{\frac{-1}{\xi}-1}[1+q(\xi(sx-m)+1)^{-\frac{1}{\xi}}]^{-\frac{1}{q}-1}; & \xi\neq 0,\quad q\neq 0,\\ (1+qe^{-(sx-m)})^{-\frac{1}{q}-1}se^{-(sx-m)}; & \xi\to 0,\quad q\neq 0,\end{cases}\tag{1.9}$$

where $s=\frac{1}{\sigma}$ and $m=\frac{\mu}{\sigma}$. In this paper, we propose the q-analogues of the generalized extreme value under power normalization (q-GEVP) to construct heavy-tailed distributions for modeling real data; to propose various types of the hazard rate function; and to generate flexible distributions with left-skewed and right-skewed shape, which can be utilized effectively in modeling extreme observations.

The paper is organized as follows. In Section 2, the q-GEVP model is reported. Some mathematical properties such as quantile function, moments, moment generating function and Shanon entropy are derived in Section 3. Section 4, explains how to determine the maximum likelihood, Cramer-von Mises minimum distance, ordinary and weighted least-square estimators of the model parameters. A Monte Carlo simulation study is carried out in Section 5, to compare the behavior of the different estimation techniques which used in the estimation of the unknown parameters of the model. In Section 6, we fit some models to COVID-19 in three countries, Japan, Saudi Arabia and Romania. Also, some statistics are employed in order to assess goodness of fit. Finally, some concluding remarks are introduced in the last section.

2. ON q-GENERALIZED EXTREME DISTRIBUTION UNDER POWER NORMALIZATION

The CDF and PDF of the q-GEVP model and q-distribution "$\xi\to 0$" are, respectively, given by:

(1) For $x^0>0$, $x>0$ and $1+\frac{\xi}{\sigma}\log(e^{-\mu}x)>0$

$$H_{q,\xi,1}(x;\mu,\sigma)=\begin{cases}(1+q(1+\frac{\xi}{\sigma}\log(e^{-\mu}x))^{\frac{-1}{\xi}})^{\frac{-1}{q}}; & \xi\neq 0,\quad q\neq 0,\\ (1+q(xe^{-\mu})^{\frac{-1}{\sigma}})^{\frac{-1}{q}}; & \xi\to 0,\quad q\neq 0,\end{cases}\tag{2.1}$$

and

(2.2)

$$h_{q,\xi,1}(x;\mu,\sigma,\xi)=\begin{cases}\left(1+q\left(1+\frac{\xi}{\sigma}\log(e^{-\mu}x)\right)^{\frac{-1}{\xi}}\right)^{\frac{-1}{q}-1}\frac{1}{\sigma x}\left(1+\frac{\xi}{\sigma}\log(e^{-\mu}x)\right)^{\frac{-1}{\xi}-1}; & \xi\neq 0,\ q\neq 0,\\ \left(1+q(xe^{-\mu})^{\frac{-1}{\sigma}}\right)^{\frac{-1}{q}-1}\frac{e^{-\mu}}{\sigma}(xe^{-\mu})^{\frac{-1}{\sigma}-1}; & \xi\to 0,\ q\neq 0.\end{cases}$$

(**2**) For $x^0\leq 0$, $x\leq 0$ and $1-\frac{\xi}{\sigma}\log(-e^{-\mu}x)>0$

$$H_{q,\xi,2}(x;\mu,\sigma)=\begin{cases}\left(1+q\left(1-\frac{\xi}{\sigma}\log(-xe^{-\mu})\right)^{\frac{-1}{\xi}}\right)^{\frac{-1}{q}}; & \xi\neq 0,\ \ q\neq 0,\\ \left(1+q(-xe^{-\mu})^{\frac{1}{\sigma}}\right)^{\frac{-1}{q}}; & \xi\to 0,\ \ q\neq 0,\end{cases}\tag{2.3}$$

and

(2.4)

$$h_{q,\xi,2}(x;\mu,\sigma,\xi)=\begin{cases}\left(1+q\left(1-\frac{\xi}{\sigma}\log(-xe^{-\mu})\right)^{\frac{-1}{\xi}}\right)^{\frac{-1}{q}-1}\frac{1}{\sigma x}\left(1-\frac{\xi}{\sigma}\log(-xe^{-\mu})\right)^{\frac{-1}{\xi}-1}; & \xi\neq 0,\ q\neq 0,\\ \left(1+q(-xe^{-\mu})^{\frac{1}{\sigma}}\right)^{\frac{-1}{q}-1}\frac{e^{-\mu}}{\sigma}(-xe^{-\mu})^{\frac{1}{\sigma}-1}; & \xi\to 0,\ q\neq 0,\end{cases}$$

where $x^0=\sup\{x:F(x)<1\}$. Figures 1 and 2 show the PDF of the q-GEVP model in case of $\xi\neq 0$ and $\xi\to 0$, respectively, for various values of the parameters.

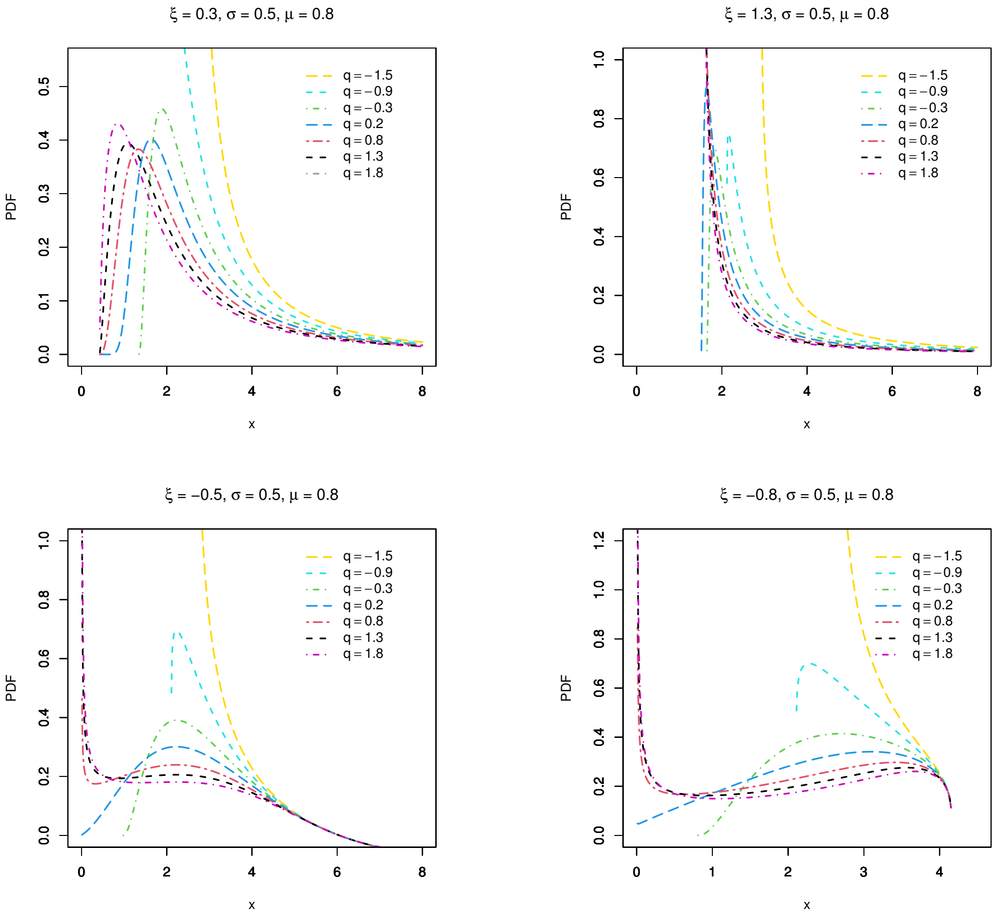


Figure 1: The PDF plots of the q-GEVP distribution in case of $\xi\neq 0$.

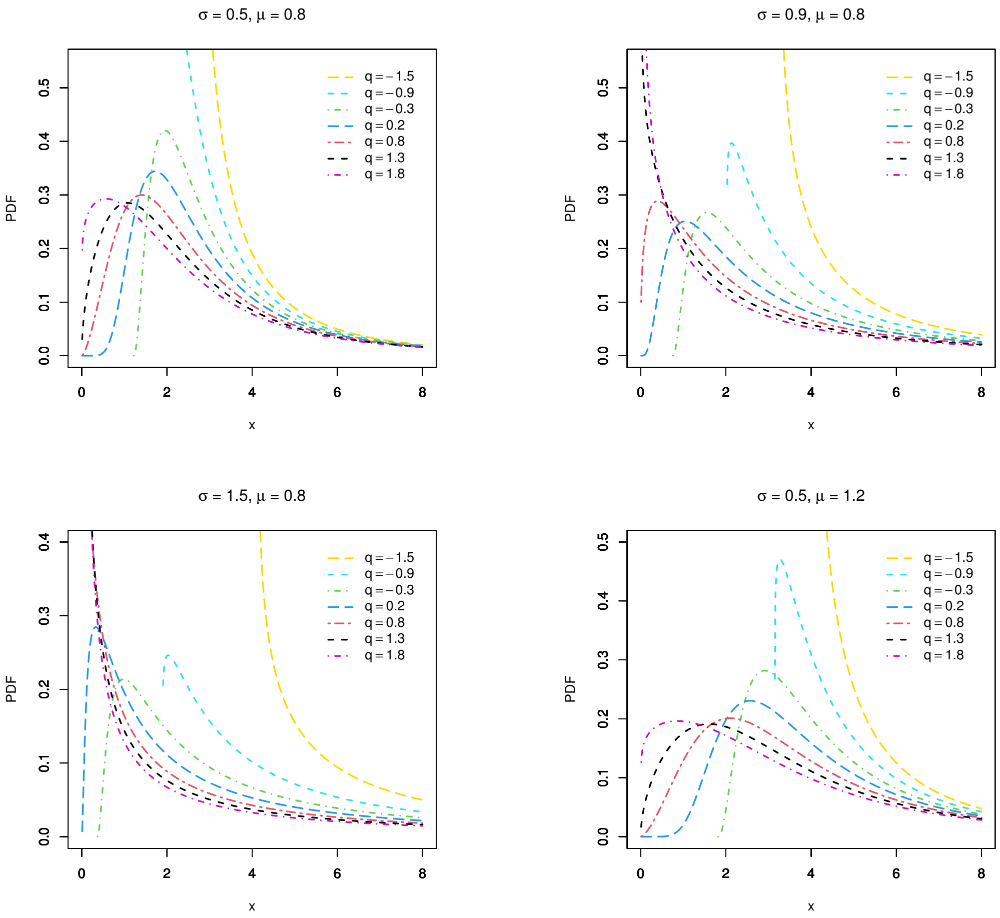


Figure 2: The PDF plots of the q-GEVP distribution in case of $\xi \to 0$.

According to Figures 1 and 2, it is noted that the proposed distribution can be used to model left and right skewed data. Moreover, the shape of the PDF can be unimodal and bimodal, which makes the proposed model can be utilized for modeling various data in different fields.

The hazard function (also called the force of mortality, instantaneous failure rate, instantaneous death rate, or age-specific failure rate) is a way to model data distribution in survival analysis. The most common use of the function is to model a participant's chance of death as a function of their age. However, it can be used to model any other time-dependent event of interest. The hazard function (HF) is defined as $\frac{h(x)}{1-H(x)}$. Figures 3 and 4 display the HF for the proposed model for $\xi \neq 0$ and $\xi \to 0$, respectively, and it is noted that the HF has various shapes including increasing, decreasing, unimodal, or bathtub.

In several cases, lifetimes need to be recorded on a discrete scale rather than on a continuous analogue. Due to the previous reason, discretizing continuous distributions has received much attention in the statistical literature. See for example, Bebbington *et al.* (2012) [8], Nekoukhou and Bidram (2015) [29], El-Morshedy *et al.* (2020) [15], Eliwa *et al.* (2020) [16], Altun *et al.* (2020) [1], and references cited therein. Based on discretization survival approach, the CDF and probability mass function (PMF) of the discrete q-GEVP

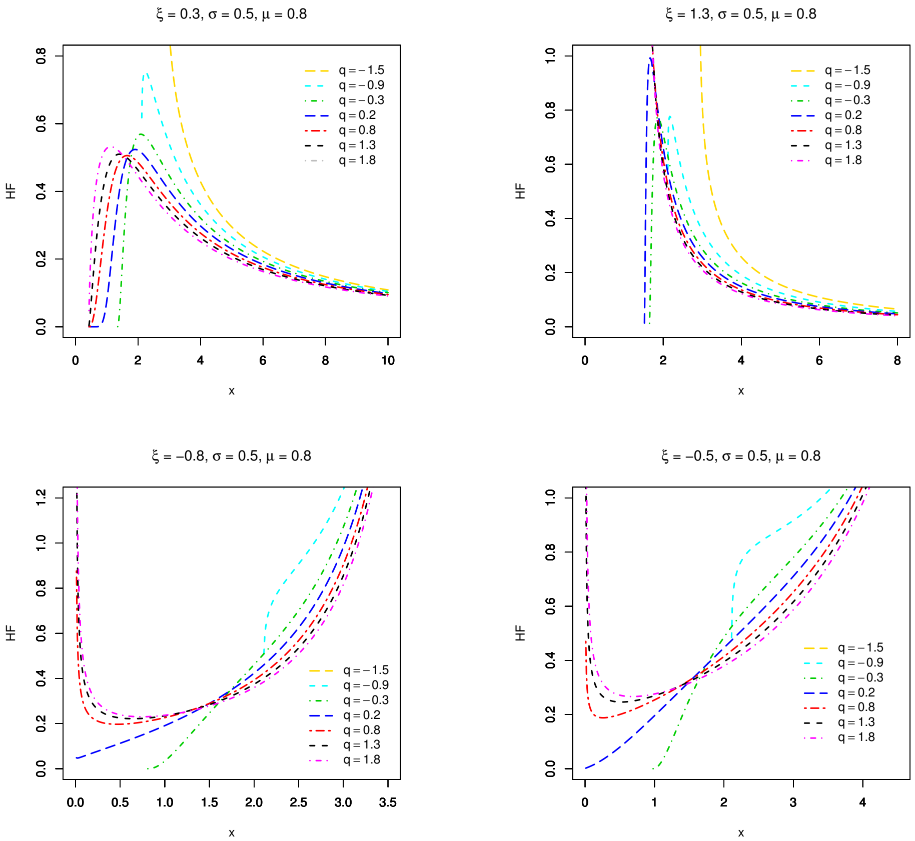


Figure 3: The HRF plots of the q-GEVP distribution in case of $\xi \neq 0$.

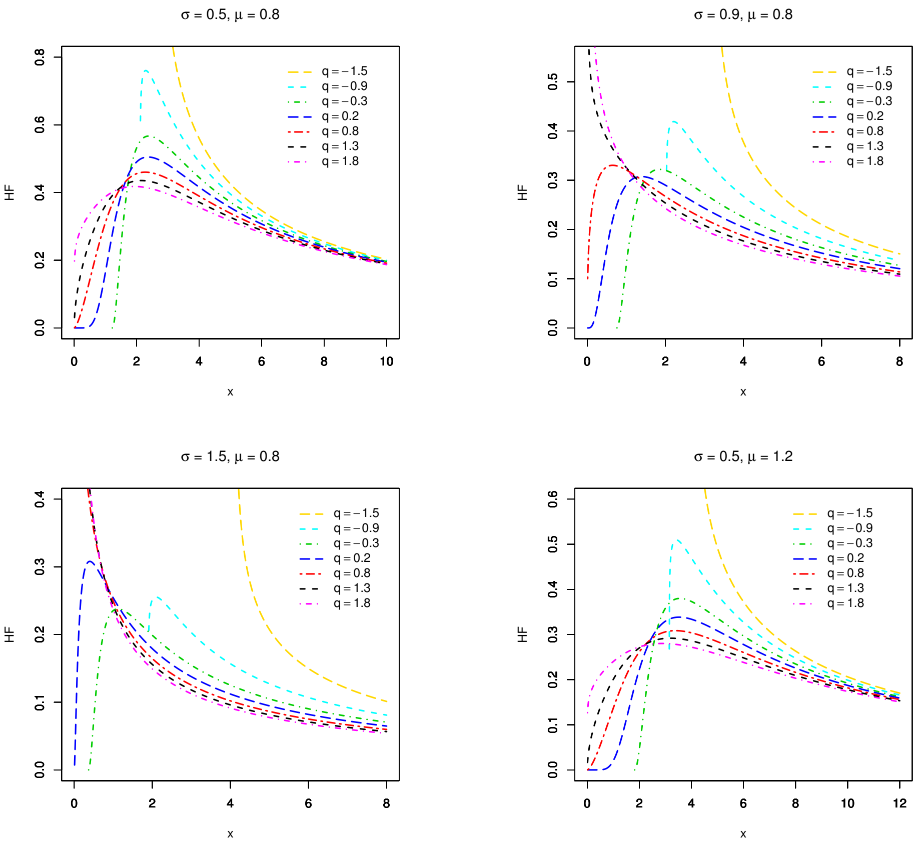


Figure 4: The HRF plots of the q-GEVP distribution in case of $\xi \to 0$.

(Dq-GEVP) model can be formulated as

$$H_{q,\xi,1}(x;\mu,\sigma)=\begin{cases}(1+q(1+\frac{\xi}{\sigma}\log(e^{-\mu}(x+1)))^{\frac{-1}{\xi}})^{\frac{-1}{q}}; & \xi\neq 0,\ \ q\neq 0,\\ (1+q((x+1)e^{-\mu})^{\frac{-1}{\sigma}})^{\frac{-1}{q}}; & \xi\to 0,\ \ q\neq 0,\end{cases}\tag{2.5}$$

and

(2.6)

$$f(x;\mu,\sigma,\xi,q)=\begin{cases}(1+q(1+\frac{\xi}{\sigma}\log(e^{-\mu}(x+1)))^{\frac{-1}{\xi}})^{\frac{-1}{q}}-(1+q(1+\frac{\xi}{\sigma}\log(e^{-\mu}x))^{\frac{-1}{\xi}})^{\frac{-1}{q}};\\ \hfill \xi\neq 0,\quad q\neq 0,\\ (1+qe^{-(sx-m)})^{-\frac{1}{q}-1}se^{-(sx-m)}-(1+q(xe^{-\mu})^{\frac{-1}{\sigma}})^{\frac{-1}{q}};\\ \hfill \xi\to 0,\quad q\neq 0,\end{cases}$$

respectively. Figure 18 shows the PMF and HRF of the Dq-GEVP models for various values of the model parameters, and it is found that the PMF can be used to model asymmetric data which have extreme observations. Further, the HRF can be utilized to model data with have decreasing failure shape.

3. STATISTICAL PROPERTIES

3.1. Quantile function and moments

The quantile function (QF) is frequently utilized for determining confidence intervals or eliciting certain properties of a distribution. In order to obtain the QF of a random variable (RV) X, that is, one has to solve the equation $F(x)=p$ with respect to x for some fixed $p\in(0,1)$, where $F(x)$ denotes the CDF of X. The QFs of the q-GEVP ($\xi\neq 0$) and q-distribution ($\xi\to 0$) can be listed as

$$x_p=H^{-1}(q,\xi,1)=\begin{cases}e^{\frac{\sigma}{\xi}(q^{\xi}(p^{-q}-1)^{-\xi}-1)+\mu}; & \xi\neq 0,\ \ q\neq 0,\\ q^{\sigma}(p^{-q}-1)^{-\sigma}e^{\mu}; & \xi\to 0,\ \ q\neq 0,\end{cases}\tag{3.1}$$

and

$$x_p=H^{-1}(q,\xi,2)=\begin{cases}-e^{\frac{\sigma}{\xi}(1-q^{\xi}(p^{-q}-1)^{-\xi})+\mu}; & \xi\neq 0,\ \ q\neq 0,\\ -q^{-\sigma}(p^{-q}-1)^{\sigma}e^{\mu}; & \xi\to 0,\ \ q\neq 0,\end{cases}\tag{3.2}$$

respectively. Assume non-negative RV have a q-GEVP model, then the n-th moment, and moment generating function of X, are given, respectively, as follows:

$$\begin{aligned}\mathbf{E}(X^n)&=\int_0^{\infty}x^n h(x;\mu,\sigma,\xi)\ dx\\ &=\Upsilon_{(n,\mu)}^{(\sigma,\xi,q)}\sum_{j=0}^{\infty}\left(\frac{n\sigma q^{\xi}}{\xi}\right)^j\frac{\Gamma(1-\xi j)\Gamma\left(\frac{1}{q}+\xi j\right)}{j!\Gamma\left(\frac{1}{q}+1\right)}\end{aligned}$$

and

$$\mathbf{M}_X(t) = \int_0^\infty \exp(tx)h(x;\mu,\sigma,\xi)\ dx$$
$$= \Theta^{(\sigma,\xi)}_{(\mu)} \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} t^j \left(\frac{\sigma q^\xi j}{\xi}\right)^k \frac{\Gamma(1-\xi k)\Gamma\left(\frac{1}{q}+\xi k\right)}{j!k!\Gamma\left(\frac{1}{q}+1\right)},$$

where $\Theta^{(\sigma,\xi)}_{(\mu)} = \frac{1}{2}\exp\Big(t\exp\Big\{\mu - \frac{\sigma}{\xi}\Big\}\Big)$, $\Upsilon^{(\sigma,\xi,q)}_{(n,\mu)} = \frac{1}{q}\exp\Big\{n\Big(\mu - \frac{\sigma}{\xi}\Big)\Big\}$, and the terms $(1-\xi j)$, $\Big(\frac{1}{q}+\xi j\Big)$, $\Big(\frac{1}{q}+1\Big)$, $(1-\xi k)$, $\Big(\frac{1}{q}+\xi k\Big)$ and $\Big(\frac{1}{q}+1\Big)$ should be greater than 0. Figure 5 shows the skewness and kurtosis under different values of the model parameters "$\xi = -0.5$ and $\sigma = 0.2$" in the left panel, and "$\xi = -1.5$ and $\sigma = 1.2$" in the right panel, respectively.

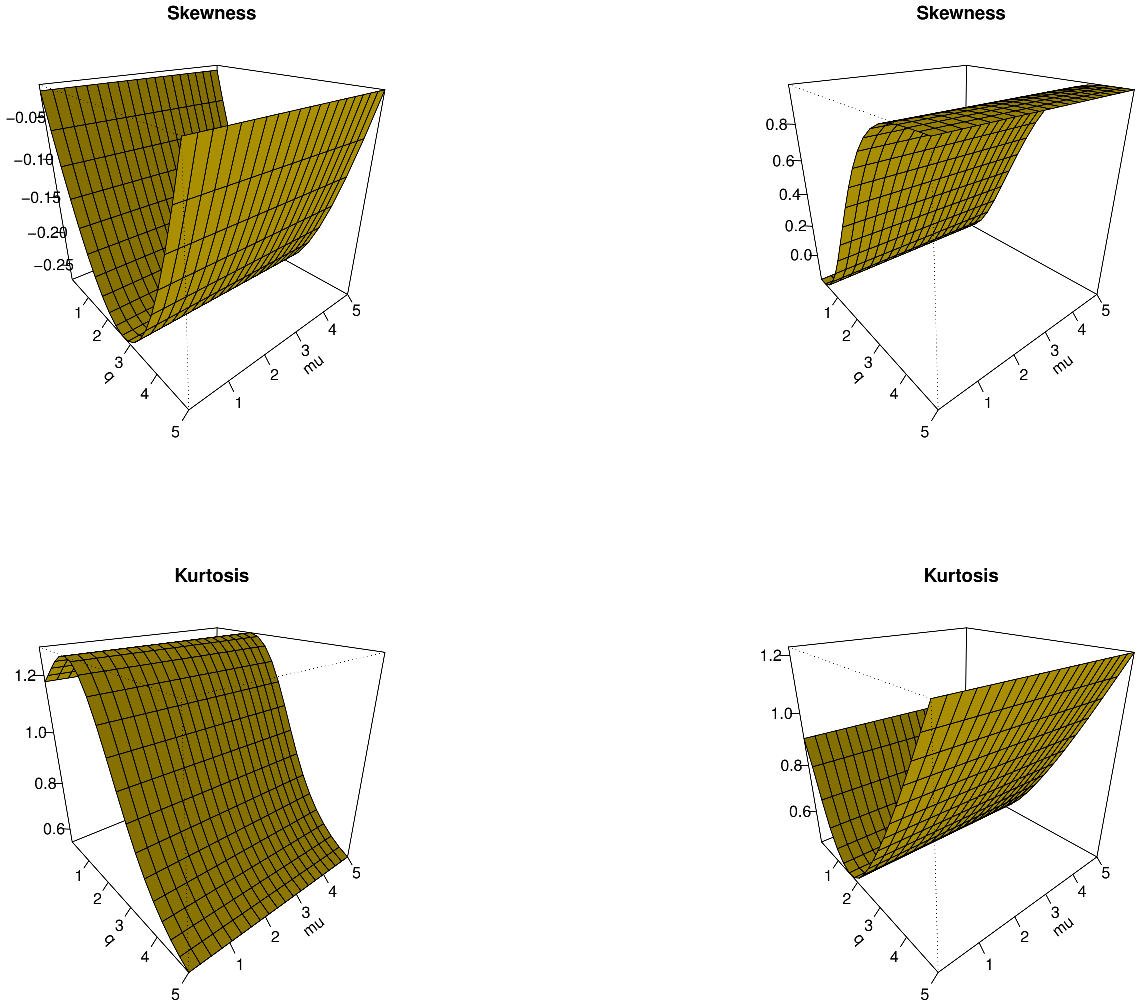
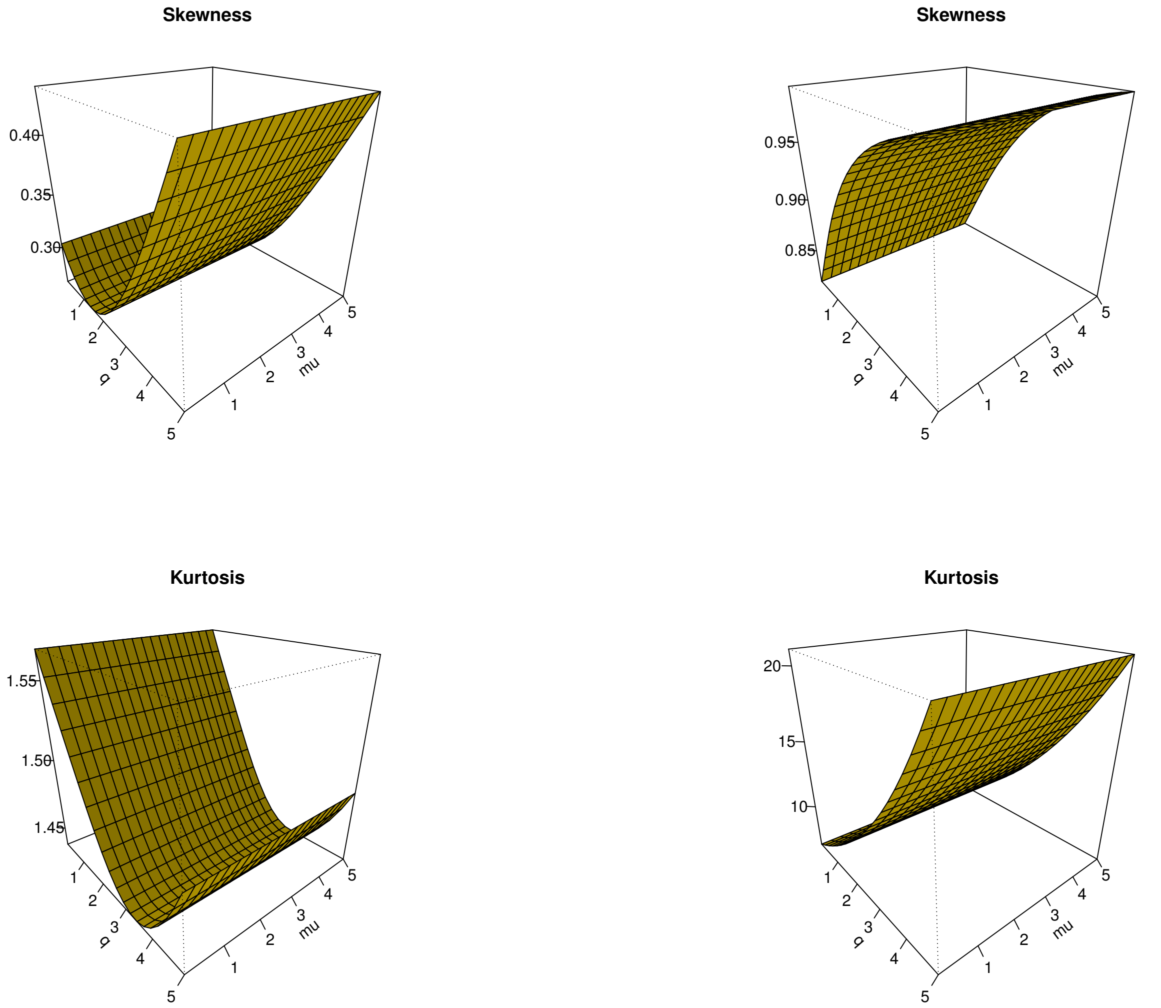


Figure 5: The skewness and kurtosis of the q-GEVP distribution in case of $\xi \neq 0$.

Figure 6 shows the skewness and kurtosis in case of $\xi \to 0$ with $\sigma = 0.5$ "left panel" and $\sigma = 2.5$ "right panel", respectively, which support our results.


Skewness

Skewness

Kurtosis

Kurtosis

Figure 6: The skewness and kurtosis of the q-GEVP distribution in case of $\xi \to 0$.

3.2. Entropy

An entropy of RV X is a measure of variation of the uncertainty. Shannon entropy (SnEy) is defined by

$$H(X) = -\int_A f(x)\log f(x)dx, \tag{3.3}$$

where $A = x : f(x) > 0$. The SnEy of the GEVL family can be expressed as

$$H(X) = \log\hat{\sigma} + (\hat{\xi} + 1)\gamma + 1. \tag{3.4}$$

The SnEy of six classes of extreme value distributions which was mentioned in Section 1, is evaluated by Ravi and Saeb (2012) [25]. Herein, the SnEy of GEVP, q-GEVL and q-GEVP families are listed in the following theorems.

Theorem 3.1. *If X is a RV with CDF GEVP for $\xi<0$, then the SnEy of X is given by*

$$H(X)=\mu+\log\hat{\sigma}+(\hat{\xi}+1)\gamma+\frac{\widehat{\sigma}}{\widehat{\xi}}[\Gamma(1-\widehat{\xi})-1]+1. \tag{3.5}$$

Proof: From Equations (1.7) and (3.3), we have

$$H(X)=-\mathbf{E}(\log h_{\mu,\xi,\sigma,q})=\log\sigma+\mathbf{E}(\log|X|)+\Delta_{(\mu,\xi,\sigma)}+\Delta^{(\mu,\xi,\sigma)},$$

where $\Delta_{(\mu,\xi,\sigma)}=\mathbf{E}(1+\frac{\xi}{\sigma}\log|X|e^{-\mu})^{\frac{-1}{\xi}}$ and $\Delta^{(\mu,\xi,\sigma)}=(1+\frac{1}{\xi})\mathbf{E}(\log(1+\frac{\xi}{\sigma}\log|X|e^{-\mu}))$. Let $Y=(1+\frac{\xi}{\sigma}\log|x|e^{-\mu}))^{\frac{-1}{\xi}}$, which have the standard exponential (StEx) distribution, then

$$\mathbf{E}(\log|X|)=\frac{\sigma}{\xi}\mathbf{E}(Y^{-\xi}-1)+\mu=\frac{\sigma}{\xi}[\Gamma(1-\xi)-1]+\mu, \tag{3.6}$$

$$\Delta_{(\mu,\xi,\sigma)}=\mathbf{E}(Y)=1 \tag{3.7}$$

and

$$\Delta^{(\mu,\xi,\sigma)}=(1+\frac{1}{\xi})\mathbf{E}(-\xi\log Y)=-(1+\xi)\mathbf{E}(\log Y)=(1+\xi)\gamma, \tag{3.8}$$

where $\gamma=-\int_0^\infty\log ye^{-y}dy$. From Equations (3.6)–(3.8), Equation (3.5) can be derived. □

Theorem 3.2. *If X is a RV with CDF q-GEVL for $\xi<0$, then the SnEy of X is given by*

$$H(X)=\log\hat{\sigma}+(\hat{\xi}+1)\gamma+(1+q)\left[1-\sum_{n=2}^{\infty}(-1)^{n+1}q^{n-1}\Gamma(n-1)\right]. \tag{3.9}$$

Proof: Since the PDF of the q-GEVL model can be listed as

$$f_X(x)=\frac{1}{\sigma}(1+\frac{\xi}{\sigma}(x-\mu))^{-\frac{1}{\xi}-1}[1+q(1+\frac{\xi}{\sigma}(x-\mu))^{-\frac{1}{\xi}}]^{-\frac{1}{q}-1}.$$

Then,

$$H(X)=-\mathbf{E}(\log f_X(X))=\log\sigma+\Delta^*_{(\mu,\xi,\sigma)}+\Delta_*^{(\mu,\xi,\sigma)},$$

where $\Delta^*_{(\mu,\xi,\sigma)}=(1+\frac{1}{\xi})\mathbf{E}(\log(1+\frac{\xi}{\sigma}(X-\mu)))$ and $\Delta_*^{(\mu,\xi,\sigma)}=(1+\frac{1}{q})\mathbf{E}(\log(1+q(1+\frac{\xi}{\sigma}(X-\mu))^{\frac{-1}{\xi}})$. Assume $Y=(1+\frac{\xi}{\sigma}(x-\mu))^{\frac{-1}{\xi}}$, which have the StEx distribution, then Equation (3.9) can be derived. □

Theorem 3.3. *If X is a RV with CDF q-GEVP for $\xi<0$, then the SnEy of X is given by*

$$H(X)=\mu+\log\hat{\sigma}+(\hat{\xi}+1)\gamma+\frac{\xi}{\sigma}\mathbf{E}\{\mathbf{sign}(X)[\Gamma(1-\xi)-1]\}+(1+q)\left[1-\sum_{n=2}^{\infty}(-1)^{n+1}q^{n-1}\Gamma(n-1)\right]. \tag{3.10}$$

Proof: Since the RV have the q-GEVP distribution, then

$$H(X) = \log\sigma + \mathbf{E}(\log|X|) + \mathbf{E}\left(1+\frac{\xi}{\sigma}\log|X|e^{-\mu}\right)^{\frac{-1}{\xi}} + \left(1+\frac{1}{\xi}\right)\Omega_{(\mu,\xi,\sigma)} + \left(1+\frac{1}{q}\right)\Omega^{(\mu,\xi,\sigma)},$$

where $\Omega_{(\mu,\xi,\sigma)} = \mathbf{E}\Big(\log(1+\frac{\xi}{\sigma}\log|X|e^{-\mu})\Big)$ and $\Omega^{(\mu,\xi,\sigma)} = \mathbf{E}(\log(1+q(1+\frac{\xi}{\sigma}\log|X|e^{-\mu})^{\frac{-1}{\xi}})$.

Let $Y = \left(1+\frac{\xi}{\sigma}\log|x|e^{-\mu}\right)^{\frac{-1}{\xi}}$ which have the StEx distribution, then

$$\mathbf{E}(\log|X|) = \frac{\sigma}{\xi}\mathbf{E}(Y^{-\xi}-1)+\mu = \frac{\sigma}{\xi}[\Gamma(1-\xi)-1]+\mu, \tag{3.11}$$

$$(1+\frac{1}{\xi})\Omega_{(\mu,\xi,\sigma)} = (1+\frac{1}{\xi})\mathbf{E}(-\xi\log Y) = (1+\xi)\gamma \tag{3.12}$$

and

$$\begin{aligned}\left(1+\frac{1}{q}\right)\Omega^{(\mu,\xi,\sigma)} &= (1+\frac{1}{q})\int_0^\infty \log(1+qy))e^{-y}dy\\ &= (1+\frac{1}{q})\left[1-\sum_{n=2}^{\infty}(-1)^{n+1}q^{n-1}\Gamma(n-1)\right].\end{aligned} \tag{3.13}$$

From Equations (3.11)–(3.13), Equation (3.10) can be derived. □

Hint: If $q \to 0$ in Equations (3.9) and (3.10), we get Equations (3.4) and (3.5).

4. VARIOUS ESTIMATION APPROACHES

4.1. Maximum likelihood estimation

In order to estimate the parameters of the q-GEVP model and q-distribution whose density functions are in (2.2), one has to maximize their respective log-likelihood functions with respect to the model parameters. Given the observations $x_i, i = 1, ..., n$, the log-likelihood functions of the q-GEVP model and q-distribution are, respectively, given by

$$\ell(\mu,\ \sigma,\ \xi,\ q) = -n\log\sigma - \sum_{i=1}^{n}\log x_i - (1+\frac{1}{q})\sum_{i=1}^{n}\log[1+q\ A_i^{\frac{-1}{\xi}}] - (1+\frac{1}{\xi})\sum_{i=1}^{n}\log A_i \tag{4.1}$$

and

$$\ell^*(\mu,\ \sigma,\ q) = -n\log\sigma + \frac{n\mu}{\sigma} - (1+\frac{1}{\sigma})\sum_{i=1}^{n}\log x_i - (1+\frac{1}{q})\sum_{i=1}^{n}\log[1+q\ B_i^{\frac{-1}{\sigma}}], \tag{4.2}$$

where $A_i = 1+\frac{\xi}{\sigma}\log B_i$ and $B_i = x_i e^{-\mu}$. The associated log-likelihood system of equations are, respectively,

$$\frac{\partial \ell}{\partial \mu} = (\frac{\xi+1}{\sigma})\sum_{i=1}^{n} A_i^{\frac{1}{\xi}} - (\frac{q+1}{\sigma})\sum_{i=1}^{n} \frac{A_i^{\frac{-1}{\xi}-1}}{1+qA_i^{\frac{-1}{\xi}}},$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + (\frac{\xi+1}{\sigma^2})\sum_{i=1}^{n} \frac{\log B_i}{A_i} - (\frac{q+1}{\sigma^2})\sum_{i=1}^{n} \frac{A_i^{\frac{-1}{\xi}-1}\log B_i}{1+qA_i^{\frac{-1}{\xi}}},$$

$$\frac{\partial \ell}{\partial \xi} = \frac{1}{\xi^2}\sum_{i=1}^{n} \log A_i - (\frac{\xi+1}{\xi\sigma})\sum_{i=1}^{n} \frac{\log B_i}{A_i} + (\frac{q+1}{\xi\sigma})\sum_{i=1}^{n} \frac{A_i^{\frac{-1}{\xi}-1}\log B_i}{1+qA_i^{\frac{-1}{\xi}}} - (\frac{q+1}{\xi^2})\sum_{i=1}^{n} \frac{A_i^{\frac{-1}{\xi}}\log A_i}{1+qA_i^{\frac{-1}{\xi}}},$$

$$\frac{\partial \ell}{\partial q} = \frac{1}{q^2}\sum_{i=1}^{n} \log[1+qA_i^{\frac{-1}{\xi}}] - (\frac{1}{q}+1)\sum_{i=1}^{n} \frac{A_i^{\frac{-1}{\xi}}}{1+qA_i^{\frac{-1}{\xi}}} \tag{4.3}$$

and

$$\frac{\partial \ell^*}{\partial \mu} = \frac{n}{\sigma} - (\frac{1+q}{\sigma})\sum_{i=1}^{n} \frac{B_i^{\frac{-1}{\sigma}}}{1+qB_i^{\frac{-1}{\sigma}}},$$

$$\frac{\partial \ell^*}{\partial \sigma} = -\frac{n}{\sigma} - \frac{n\mu}{\sigma^2} + \frac{1}{\sigma^2}\sum_{i=1}^{n} \log x_i - (\frac{q+1}{\sigma^2})\sum_{i=1}^{n} \frac{B_i^{\frac{-1}{\sigma}}\log B_i}{1+qB_i^{\frac{-1}{\sigma}}},$$

$$\frac{\partial \ell^*}{\partial q} = \frac{1}{q^2}\sum_{i=1}^{n} \log[1+qB_i^{\frac{-1}{\sigma}}] - (\frac{1}{q}+1)\sum_{i=1}^{n} \frac{B_i^{\frac{-1}{\sigma}}}{1+qB_i^{\frac{-1}{\sigma}}}. \tag{4.4}$$

Solving the nonlinear systems specified by the sets of equations yields the maximum likelihood estimates (MLE's) of the parameters of the q-GEVP model and q-distribution. Since these equations cannot be solved analytically; iterative method such as the Newton–Raphson technique is required.

4.2. Ordinary and weighted least-square estimators

Let $x_{(1)}, x_{(2)}, ..., x_{(r)}$ be the order statistics (OS) of the random sample of size r from $F(x; q, \xi, \sigma, \mu)$. The least square estimators (LSEs) of the q-GEVP parameters, say, $\widehat{q}_{LS}$, $\widehat{\xi}_{LS}$, $\widehat{\sigma}_{LS}$ and $\widehat{\mu}_{LS}$ can be obtained by solving the non-linear equations

$$\sum_{d=1}^{r}\left[F\big(x_{(d)}|\, q,\xi,\sigma,\mu\big) - \frac{d}{r+1}\right]\Delta_{\varrho}\big(x_{(d)}|\, q,\xi,\sigma,\mu\big) = 0, \quad \varrho = 1,2,3,4,$$

where

$$\begin{cases} \Delta_1\big(x_{(d)}|\, q,\xi,\sigma,\mu\big) = \dfrac{\partial}{\partial q}F\big(x_{(d)}|\, q,\xi,\sigma,\mu\big), \;\; \Delta_2\big(x_{(d)}|\, q,\xi,\sigma,\mu\big) = \dfrac{\partial}{\partial \xi}F\big(x_{(d)}|\, q,\xi,\sigma,\mu\big), \\ \Delta_3\big(x_{(d)}|\, q,\xi,\sigma,\mu\big) = \dfrac{\partial}{\partial \sigma}F\big(x_{(d)}|\, q,\xi,\sigma,\mu\big), \; \Delta_4\big(x_{(d)}|\, q,\xi,\sigma,\mu\big) = \dfrac{\partial}{\partial \mu}F\big(x_{(d)}|\, q,\xi,\sigma,\mu\big). \end{cases} \tag{4.5}$$

Whereas the weighted least squares estimators (WLSEs), say, $\widehat{q}_{WLS}$, $\widehat{\xi}_{WLS}$, $\widehat{\sigma}_{WLS}$ and $\widehat{\mu}_{WLS}$ can be reported by solving the non-linear equations

$$\sum_{d=1}^{r}\frac{(r+1)^2(r+2)}{d(r-d+1)}\left[F\big(x_{(d)}\,|\,q,\xi,\sigma,\mu\big)-\frac{d}{r+1}\right]\Delta_{\varrho}\big(x_{(d)}\,|\,q,\xi,\sigma,\mu\big)=0,\quad \varrho=1,2,3,4,$$

where $\Delta_1(\cdot\,|\,q,\xi,\sigma,\mu)$, $\Delta_2(\cdot\,|\,q,\xi,\sigma,\mu)$, $\Delta_3(\cdot\,|\,q,\xi,\sigma,\mu)$ and $\Delta_4(\cdot\,|\,q,\xi,\sigma,\mu)$ are provided in Equation (4.5).

4.3. Cramer-von Mises minimum distance estimators

The CVMEs of the q-GEVP parameters are derived by solving the non-linear equations

$$\sum_{d=1}^{r}\left[F\big(x_{(d)}\,|\,q,\xi,\sigma,\mu\big)-\frac{2d-1}{2r}\right]\Delta_{\varrho}\big(x_{(d)}\,|\,q,\xi,\sigma,\mu\big)=0,\qquad \varrho=1,2,3,$$

where $\Delta_1(\cdot\,|\,q,\xi,\sigma,\mu)$, $\Delta_2(\cdot\,|\,q,\xi,\sigma,\mu)$, $\Delta_3(\cdot\,|\,q,\xi,\sigma,\mu)$ and $\Delta_4(\cdot\,|\,q,\xi,\sigma,\mu)$ are defined in Equation (4.5).

5. THE MONTE CARLO SIMULATION STUDY

Here, we have conducted a Monte Carlo simulation study to compare the behavior of the different estimation techniques (MLEs, LSEs, WLSEs, and CVMEs) used in the estimation of the unknown parameters of the q-GEVP model in case of $\xi \neq 0$, and $\xi \to 0$. We have drawn 1000 samples of size $n = 20, 50, 100, 150, 200, 250, 300, 500$ from q-GEVP$(0.5, 0.5, 0.8, 0.5)$ and q-GEVP$(0.8, \xi \to 0, 0.5, 0.3)$, respectively, through the **R** software. We have calculated the MLEs, LSEs, WLSEs, and CVMEs for each of the 1000 samples, say, $\widehat{q}_k, \widehat{\xi}_k, \widehat{\sigma}_k$ and $\widehat{\mu}_k$ for $k = 1, 2, ..., 1000$. We have calculated the biases and mean-squared errors (MSEs) for $\Upsilon = q, \xi, \sigma$, and μ through the following formulas

$$\text{Bias}=\frac{1}{1000}\sum_{k=1}^{1000}\Big(\widehat{\Upsilon}_k-\Upsilon\Big)\text{ and MSE}=\frac{1}{1000}\sum_{k=1}^{1000}\Big(\widehat{\Upsilon}_k-\Upsilon\Big)^2.$$

The empirical results are given in Figures 7 and 8.

From Figures 7 and 8 the following observations can be made:

1. As the value of n increases, the magnitude of the bias decreases towards zero.
2. The MSEs of all the estimators decrease when we increase the value of the sample size n. This finding supports the first-order asymptotic theory.
3. In view of MSEs, clearly, MLE, LSE, WLSE, and CVME techniques perform satisfactorily in the estimation of q-GEVP parameters in case of $\xi \neq 0$, and $\xi \to 0$.

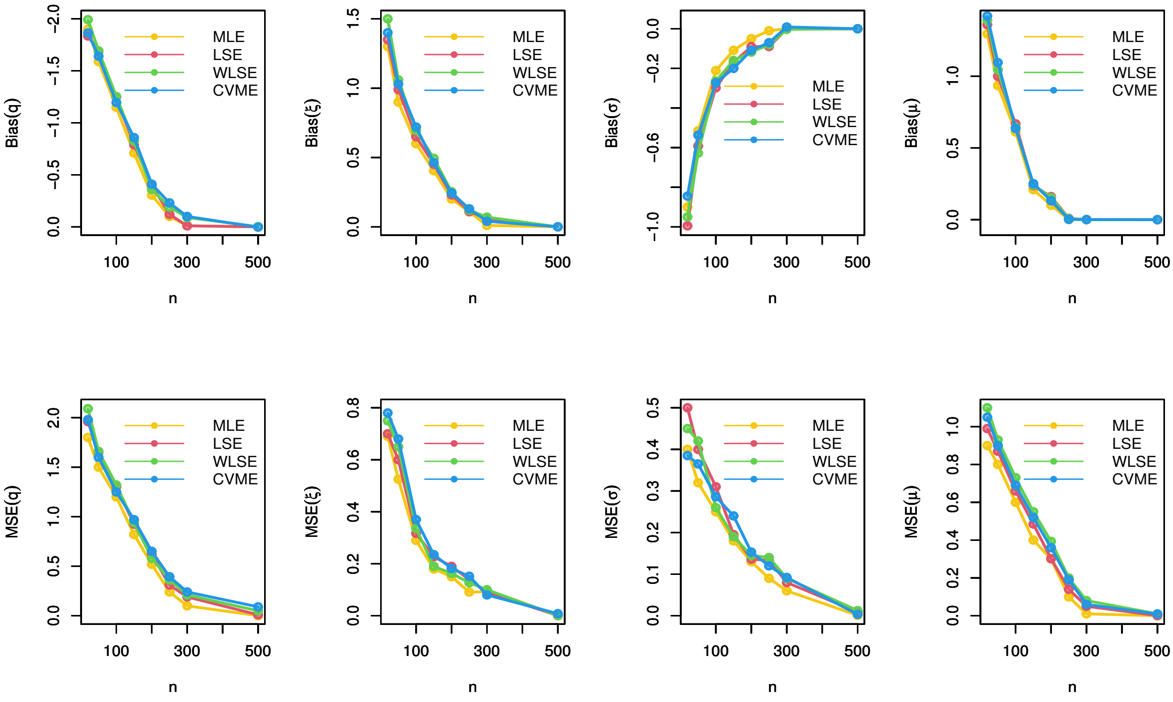


Figure 7: The bias of $\widehat{q}, \widehat{\xi}, \widehat{\sigma}$ and $\widehat{\mu}$ versus n for the q-GEVP$(0.5, 0.5, 0.8, 0.5)$ model.

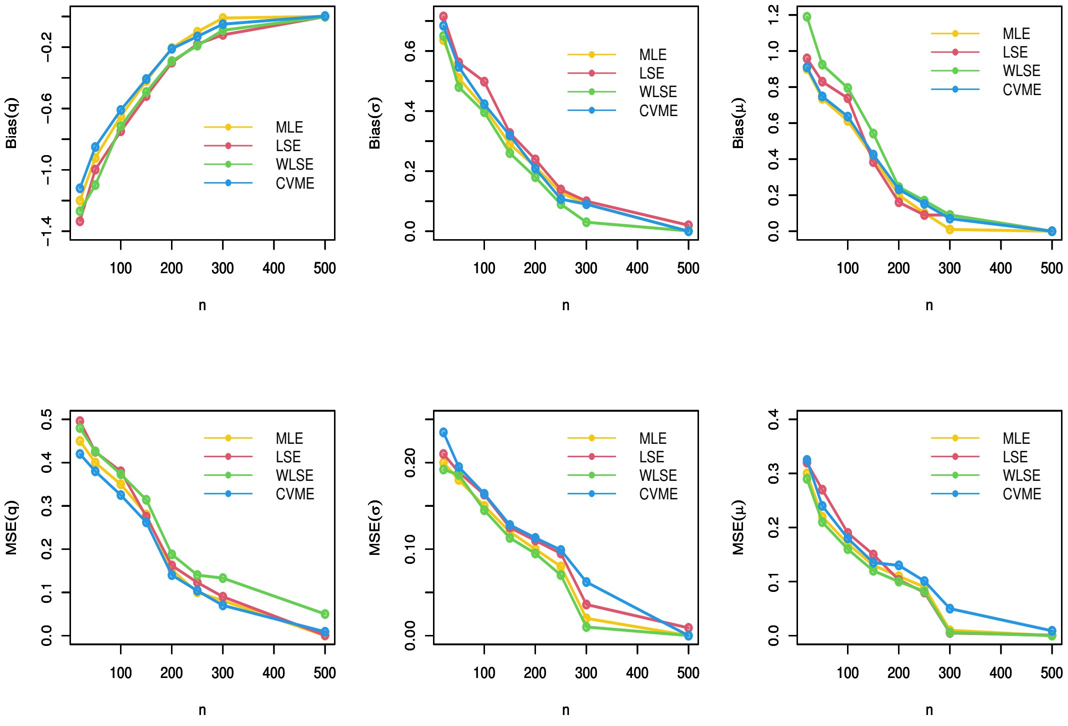


Figure 8: The bias of $\widehat{q}, \widehat{\xi}, \widehat{\sigma}$ and $\widehat{\mu}$ versus n for the q-GEVP$(0.8, \xi \to 0, 0.5, 0.3)$ model.

6. DATA ANALYSIS

In this section, we discuss the empirical importance of the q-GEVP model in case of $\xi \neq 0$, and $\xi \to 0$ for a positive random variable by using three applications to COVID-19 data. The fitted distributions are compared utilizing some criteria namely, Cramér-von Mises (CM), Anderson-Darling (AD) statistics, and Kolmogorov-Smirnov (KS) statistic with their p-values. Moreover, Akaike information criterion (AIC) with its corrected value (CAIC) beside Bayesian information criterion (BIC) and Hannan-Quinn information criterion (HQIC) have been used as a part from these criteria. We shall compare the fits of the q-GEVP distribution with some competitive models like GEVP-type I (GEVP-I), inverse Weibull (IW), Gumbel (Gu), Weibull (W), generalized inverse Weibull (GIW), Gumbel inverse Weibull (GuIW), and type I generalized exponential inverse Weibull (T1GEIW) in case of $\xi \neq 0$ "see data sets I and II", and $\xi \to 0$ "see data set III".

6.1. Data set I: COVID-19 in Japan

This data is listed in (`https://www.worldometers.info/coronavirus/country/japan/`) which represents the maximum value of the new deaths per a week due to COVID-19 in Japan from 7 Mar 2020 up to 20 Feb 2021. Initial density shape is explored using the nonparametric "Kernel density estimation (KDE)" approach in Figure 9, and it is noted that the density is asymmetric and multimodal functions. The "normality" condition is checked via the "quantile-quantile (Q-Q) plot" in Figure 9. The extremes are spotted from the "box plot" in Figure 9, and it is showed that some extreme observations were founded.

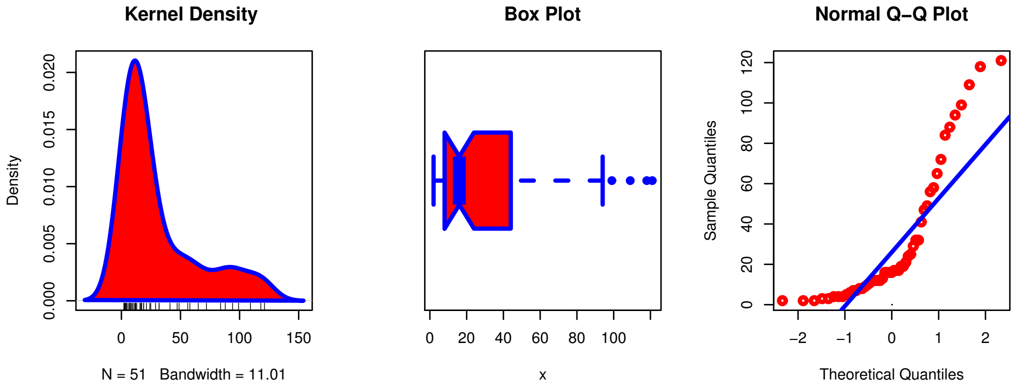


Figure 9: The KDE, Q-Q, and box plots for data set I.

Table 1 lists the MLEs with its standard errors (SE) in parentheses, whereas the goodness-of-fit (GOF) measures have been reported in Table 2 for data sets I.

From Table 2, it is noted that the q-GEVP model provides the best fit among all competitive distributions because it has the smallest value of CM, AD, KS, AIC, CAIC, BIC, and HQIC as well as it has the highest p-value. The empirical PDF, CDF, SF and P-P plots for data set I are displayed in Figure 10, which indicates that the data set plausibly came from q-GEVP model.

Table 1: The MLEs with its SE in parentheses for data set I.

Model	**MLEs(SE)**			
q-**GEVP**(q,ξ,σ,μ)	$-0.8659(0.5492)$	$-0.9892(0.0236)$	3.5633(0.1379)	1.1944(0.2913)
GEVP-I(α,β)	$-0.6454(0.0632)$	1.0229(0.1503)	–	–
IW(α,β)	9.6646(1.5479)	0.9269(0.0971)	–	–
Gu(μ,σ)	17.0872(2.8921)	19.9133(2.4818)	–	–
W(α,β)	30.1519(4.6867)	0.9545(0.1020)	–	–
GIW(α,β,γ)	1.9247(407.6033)	0.9269(0.0971)	4.4629(876.0921)	–
GuIW$(\gamma,\delta,\alpha,\beta)$	3.3919(0.5459)	8.9611(0.9268)	2.3058(0.0273)	7.8184(0.0301)
T1GEIW$(\gamma,\delta,\alpha,\beta)$	$388.4329(2.4\times10^3)$	0.1385(1.5242)	0.1299(1.5230)	0.9254(0.0974)

Table 2: The GOF measures for data set I.

GOF	**Model**							
	q-**GEVP**	**GEVP-I**	**IW**	**Gu**	**W**	**GIW**	**GuIW**	**T1GEIW**
KS **p-value**	0.0981 0.7109	0.3458 ≤ 0.001	0.1079 0.5929	0.2047 0.0279	0.1333 0.3252	0.1079 0.5929	0.1141 0.5206	0.1166 0.4916
$\mathbf{A}^*$ **p-value**	0.6679 0.5855	3.1846 0.1510	0.7298 0.5337	3.0261 0.0267	0.8439 0.4498	0.7298 0.5337	0.8678 0.4340	1.5855 0.1574
$\mathbf{W}^*$ **p-value**	0.0842 0.6701	0.6189 0.1172	0.0982 0.5958	0.4801 0.0444	0.1421 0.4157	0.0982 0.5958	0.1160 0.5134	0.1846 0.2999
-L **AIC** **CAIC** **BIC** **HQIC**	217.4310 442.8620 443.7316 450.5893 445.8148	252.9968 509.9936 510.2436 513.8573 511.4700	226.6616 457.3232 457.5732 461.1868 458.7996	238.7913 481.5826 481.8326 485.4462 483.059	225.7776 455.5552 455.8052 459.4188 457.0316	226.6616 459.3232 459.8338 465.1186 461.5378	225.3522 458.7045 459.5741 466.4318 461.6573	229.6592 467.3185 468.1881 475.0458 470.2713

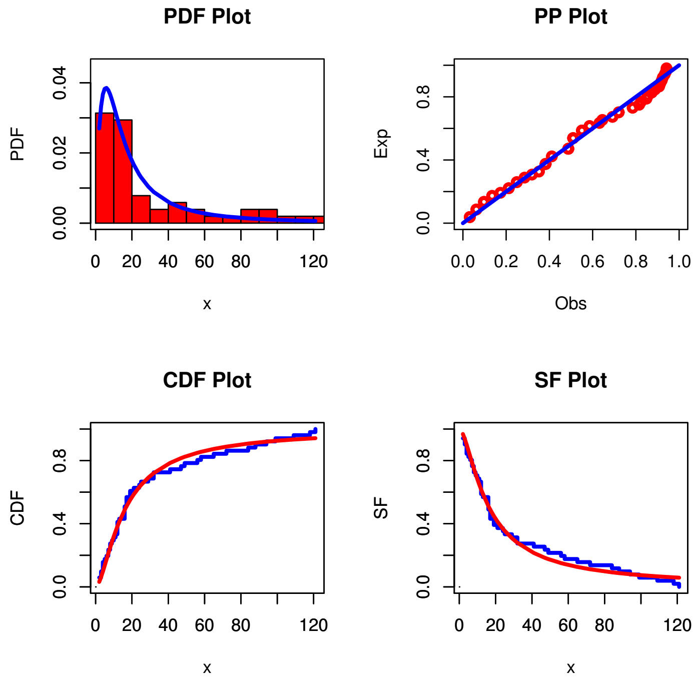


Figure 10: The fitted PDF, P-P, estimated CDF, and empirical SF plots for data set I.

Table 3 lists the estimates of the unknown parameters using three estimation methods for data set I.

Table 3: Various estimators of the q-GEVP model for data set I.

Parameters and GOF	**Methods**			
	MLE	**LSE**	**WLSE**	**CVME**
q	-0.8659	0.5785	0.2283	0.5634
ξ	-0.9892	-0.0333	-0.3006	-0.0338
σ	3.5633	0.8705	1.1325	0.8480
μ	1.1944	2.6773	2.5464	2.6744
KS	0.0981	0.0744	0.0808	0.0693
p-value	0.7109	0.9405	0.8929	0.9672
$\mathbf{A}^*$	0.6679	0.3633	0.4353	0.3454
p-value	0.5855	0.8837	0.8124	0.9002
$\mathbf{W}^*$	0.0842	0.0395	0.0554	0.0381
p-value	0.6701	0.9375	0.8454	0.9444

Table 3 illustrates that all estimation methods work quite well beside the MLE method, but the CVME approach is the best for data set I. Figure 11 shows the fitted PDFs, estimated CDFs, and empirical SF plots for data set I utilizing the estimators in Table 3, which support our results.

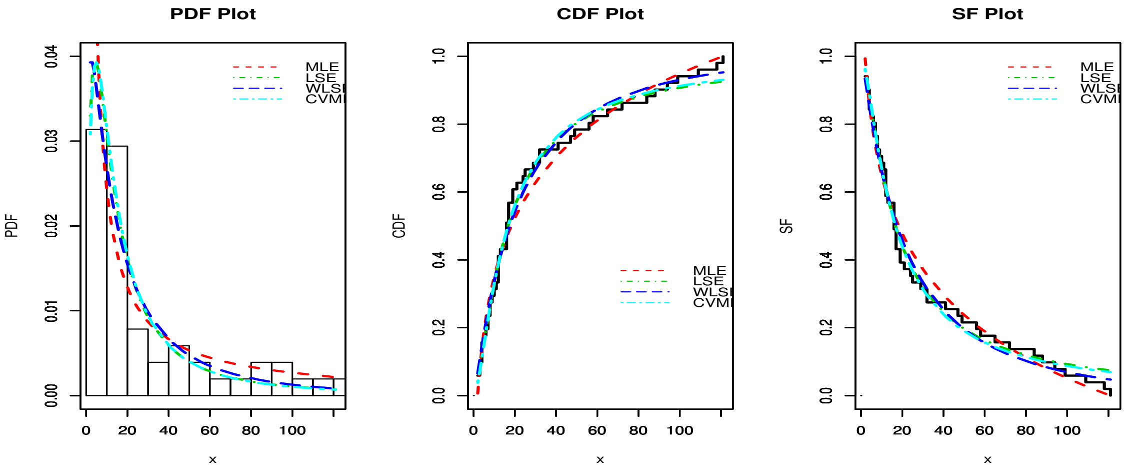


Figure 11: The fitted PDF, estimated CDF, and empirical SF plots based on various estimators for data set I.

6.2. Data set II: COVID-19 in Saudi Arabia

This data is reported in (`https://www.worldometers.info/coronavirus/country/saudi-arabia/`) which represents the maximum value of the new deaths per a week due to COVID-19 in Saudi Arabia from 28 Mar 2020 up to 20 Feb 2021. Initial density shape is explored utilizing the KDE approach in Figure 12, and it is noted that the density is asymmetric and bimodal functions. Further, the Q-Q and box plots are displayed in the same Figure.

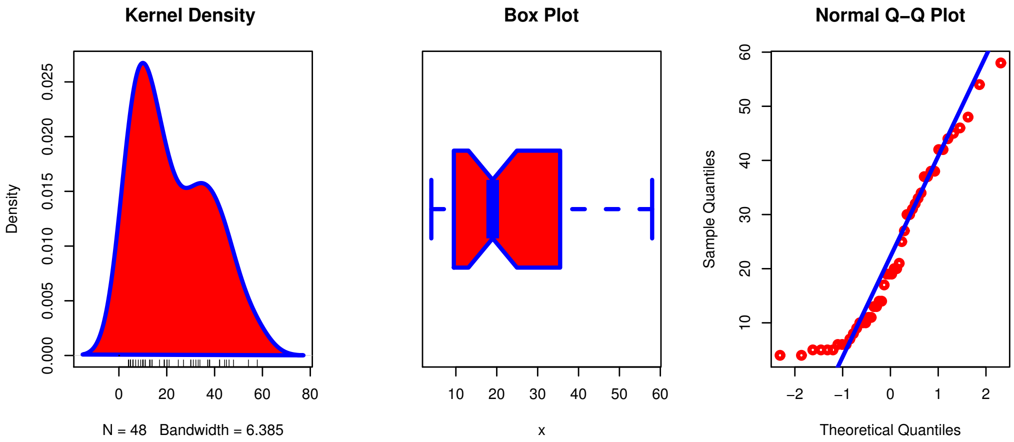


Figure 12: The KDE, Q-Q, and box plots for data set II.

Tables 4 and 5 list the MLEs, SE, and GOF measures for data sets II.

Table 4: The MLEs with its SE in parentheses for data set II.

Model	**MLEs(SE)**			
q-**GEVP**(q,ξ,σ,μ)	$-0.7945(0.0231)$	$-0.9354(0.1254)$	2.0494(0.2415)	1.8730(0.1478)
GEVP-I(α,β)	$-1.4687(0.1341)$	1.1769(0.2136)	–	–
IW(α,β)	24.0124(7.9202)	1.3077(0.1431)	–	–
Gu(μ,σ)	15.3915(1.8026)	11.8673(1.4129)	–	–
W(α,β)	1.5059(0.1742)	25.0505(2.5344)	–	–
GIW(α,β,γ)	3.5549(910.6226)	1.3077(0.1431)	4.5719(1531.5631)	–
GuIW$(\gamma,\delta,\alpha,\beta)$	3.0714(0.4487)	9.4339(0.9552)	4.4974(NaN)	11.7043(NaN)
T1GEIW$(\gamma,\delta,\alpha,\beta)$	213.3478(1411.1180)	0.2929(6.3869)	0.4758(8.0940)	1.3044(0.1445)

Table 5: The GOF measures for data set II.

GOF	**Model**							
	q-**GEVP**	**GEVP-I**	**IW**	**Gu**	**W**	**GIW**	**GuIW**	**T1GEIW**
KS **p-value**	0.1049 0.6659	0.3303 ≤ 0.001	0.1417 0.2898	0.1218 0.4751	0.1057 0.6568	0.1417 0.2898	0.1398 0.3053	0.1420 0.2875
$\mathbf{A}^*$ **p-value**	0.5464 0.6992	8.6403 ≤ 0.001	1.383 0.2070	1.1593 0.2833	0.8067 0.4756	1.3830 0.2070	1.3421 0.2191	1.3869 0.2059
$\mathbf{W}^*$ **p-value**	0.0724 0.7394	1.7523 ≤ 0.001	0.21031 0.2487	0.1802 0.3101	0.1243 0.4798	0.2103 0.2487	0.2049 0.2585	0.2111 0.2473
-L **AIC** **CAIC** **BIC** **HQIC**	184.2950 376.5900 377.5202 384.0748 379.4185	215.4770 434.9540 435.2207 438.6964 436.3683	196.8051 397.6103 397.8769 401.3527 399.0245	195.6623 395.3245 395.5912 399.0669 396.7388	192.2741 388.5482 388.8149 392.2906 389.9625	196.8051 399.6103 400.1557 405.2239 401.7316	195.3561 398.7121 399.6423 406.1969 401.5406	196.8159 401.6319 402.5621 409.1167 404.4604

From Table 5, it is noted that the q-GEVP distribution provides the best fit among all competitive models. The empirical PDF, CDF, SF and P-P plots for data set II are displayed in Figure 13.

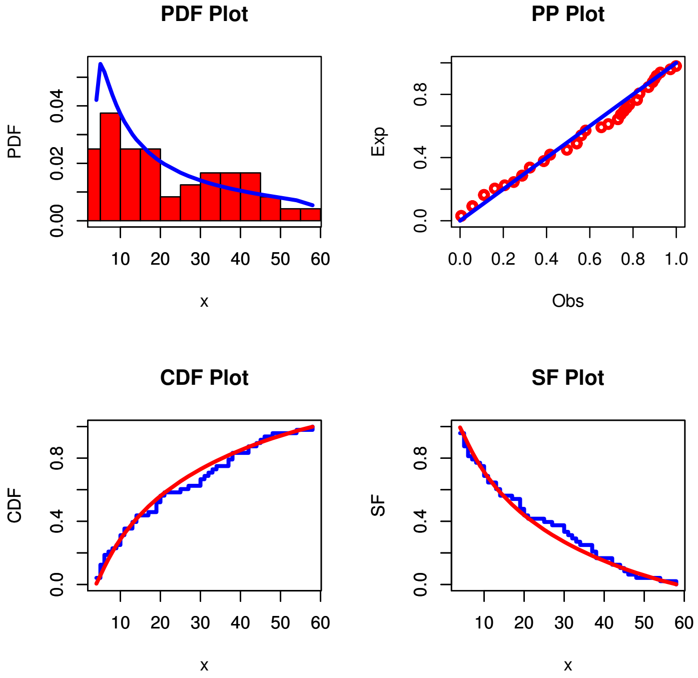


Figure 13: The fitted PDF, P-P, estimated CDF, and empirical SF plots for data set II.

Table 6 lists the estimates of the unknown parameters using various estimation methods for data set II.

Table 6: Various estimators of the q-GEVP model for data set II.

Parameters and GOF	**Methods**			
	MLE	**LSE**	**WLSE**	**CVME**
q	−0.7945	0.1733	0.1982	0.1614
ξ	−0.9354	−0.5311	−0.6364	−0.5547
σ	2.0494	0.9121	0.9321	0.9113
μ	1.8730	2.6956	2.7251	2.6978
KS	0.1049	0.0949	0.1097	0.0939
p-value	0.6659	0.7793	0.6107	0.7911
A*	0.5464	0.5914	0.5793	0.5514
p-value	0.6992	0.6552	0.6669	0.6942
W*	0.0724	0.0582	0.0616	0.0569
p-value	0.7394	0.8279	0.8060	0.8353

From Table 6, it is clear that all estimation techniques work quite well beside the MLE method, but the CVME approach is the best for data set II. Figure 14 shows the fitted PDFs, estimated CDFs, and empirical SF plots for data set II by using the estimators in Table 6, which support our results.

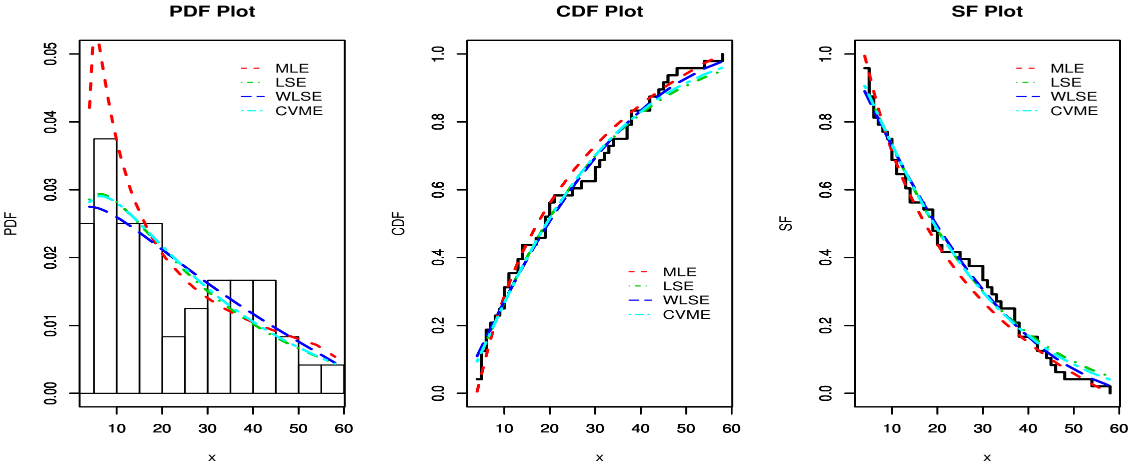


Figure 14: The fitted PDF, estimated CDF, and empirical SF plots based on various estimators for data set II.

6.3. Data set III: COVID-19 in Romania

This data is reported in (`https://www.worldometers.info/coronavirus/country/romania/`) which represents the maximum value of the new deaths per a week due to COVID-19 in Romania from 7 Mar 2020 up to 20 Feb 2021. Initial density shape is explored using the KDE method in Figure 15, and it is clear that the density is asymmetric and bimodal functions. Moreover, the Q-Q and box plots are displayed in the same Figure.

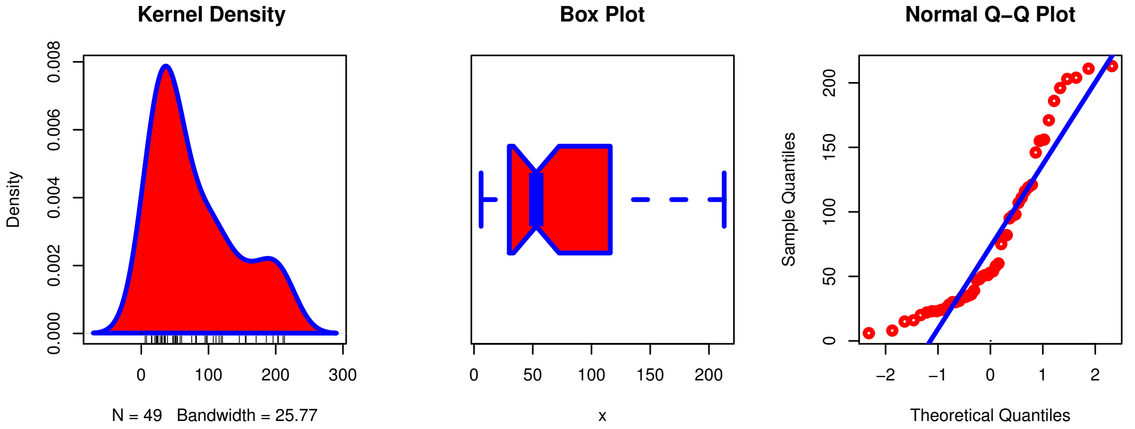


Figure 15: The KDE, Q-Q, and box plots for data set III.

Tables 7 and 8 report the MLEs, SE, and the GOF measures for data sets III.

From Table 8, it is noted that the q-GEVP model provides the best fit among all competitive distributions. The empirical PDF, CDF, SF and P-P plots for data set III are displayed in Figure 16.

Table 7: The MLEs with its SE in parentheses for data set III.

Model	**MLEs(SE)**			
q-**GEVP**(q,σ,μ)	1.1182(0.7912)	0.4969(0.1408)	4.0881(0.2569)	–
GEVP-I(α,β)	−1.5821(0.0265)	0.7867(0.1254)	–	–
IW(α,β)	52.0870(19.6569)	1.1022(0.1117)	–	–
Gu(μ,σ)	51.7563(6.5298)	43.6582(5.2948)	–	–
W(α,β)	1.3167(0.1469)	86.5845(9.9331)	–	–
GIW(α,β,γ)	5.6304(2654.1739)	1.1022(0.1117)	7.7531(4028.5284)	–
GuIW$(\gamma,\delta,\alpha,\beta)$	9.7866(2.2779)	6.4026(0.6714)	4.5341(0.1529)	7.0421(0.1946)
T1GEIW$(\gamma,\delta,\alpha,\beta)$	1115.2175(5733.0614)	0.1679(2.2473)	0.3115(3.9096)	1.1012(0.1116)

Table 8: The GOF measures for data set III.

GOF	**Model**							
	q-**GEVP**	**GEVP-I**	**IW**	**Gu**	**W**	**GIW**	**GuIW**	**T1GEIW**
KS	0.0900	0.4171	0.1318	0.1438	0.1151	0.1318	0.1322	0.1319
p-value	0.8221	≤ 0.001	0.3621	0.2627	0.5351	0.3621	0.3584	0.3611
A*	0.5499	12.1030	0.9625	1.4830	0.7140	0.9625	0.9693	0.9652
p-value	0.6957	≤ 0.001	0.3771	0.1806	0.5464	0.3771	0.3733	0.3756
W*	0.0821	2.5283	0.1097	0.2313	0.1172	0.1097	0.1106	0.1101
p-value	0.6817	≤ 0.001	0.5411	0.2145	0.5084	0.5411	0.5372	0.5392
-L	262.1081	311.5210	266.2566	265.1471	260.7146	266.2566	266.2533	266.2694
AIC	530.2162	627.0420	536.5132	534.2943	525.4292	538.5132	540.5066	540.5388
CAIC	530.7495	627.3029	536.7741	534.5551	525.6901	539.0465	541.4157	541.4479
BIC	535.8916	630.8256	540.2968	538.0779	529.2128	544.1886	548.0739	548.1061
HQIC	532.3694	628.4775	537.9487	535.7298	526.8647	540.6664	543.3777	543.4098

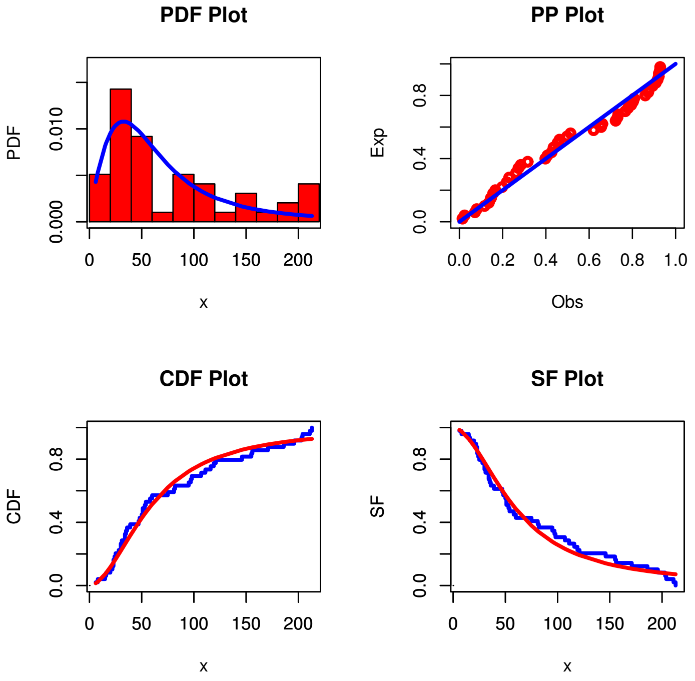


Figure 16: The fitted PDF, P-P, estimated CDF, and empirical SF plots for data set III.

Table 9 reports the estimates of the unknown parameters using various estimation approaches for data set III.

Table 9: Various estimators of the q-GEVP model for data set III.

Parameters and GOF	**Methods**			
	MLE	**LSE**	**WLSE**	**CVME**
q σ μ	1.1182 0.4969 4.0881	0.6959 0.6599 3.9588	2.2430 0.4102 4.3003	0.6642 0.6448 3.9499
KS **p-value**	0.0900 0.8221	0.1084 0.6128	0.0859 0.8620	0.1024 0.6827
$\mathbf{A}^*$ **p-value**	0.5499 0.6957	0.5272 0.7184	0.6501 0.6011	0.4884 0.7578
$\mathbf{W}^*$ **p-value**	0.0821 0.6817	0.0550 0.8476	0.0793 0.6979	0.0535 0.8571

Table 9 illustrates that all estimation methods work quite well besides the MLE method. Figure 17 shows the fitted PDFs, estimated CDFs empirical SF plots for data set III using the estimators in Table 9, which support our results.

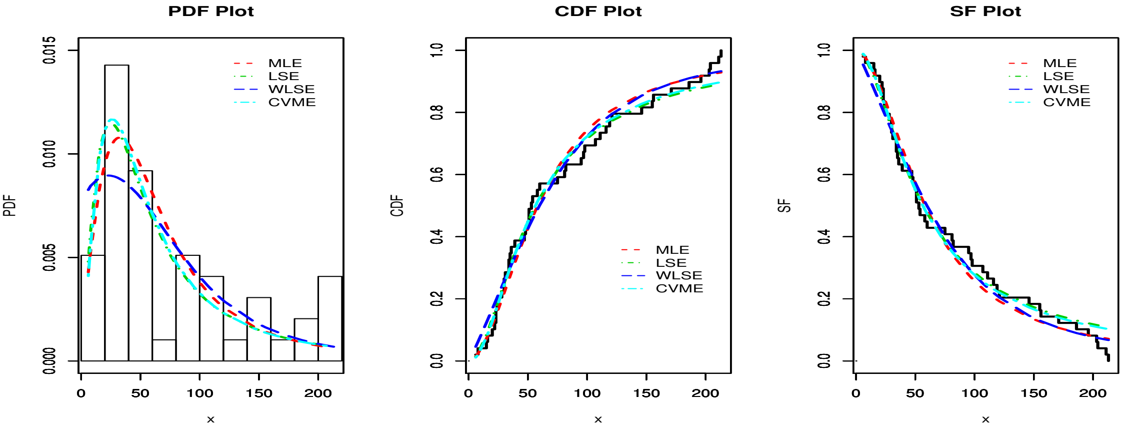


Figure 17: The fitted PDF, estimated CDF, and empirical SF plots based on various estimators for data set III.

7. CONCLUSIONS

In this paper, we proposed q-generalized extreme values model and its discrete version under power normalization technique. Its various statistical features have been derived in detail. It was found that the proposed models are a proper for modelling skewed data sets, especially which have very extreme observations. Moreover, the new model provides a wide variation in the shape of the HRF, including decreasing, increasing, unimodal, and bathtub shapes, and consequently the proposed distribution can be utilized in modelling several different kinds of data. The model parameters have been estimated using four different estimation approaches, namely, MLE, LSE, WLSE, and CVME. A simulation has been performed based on different sample sizes, and it was found that the four methods work quit effectively in estimating the model parameters. Three distinctive data sets "COVID-19" have been analyzed to illustrate and prove the flexibility of the proposed model. Finally, the q-generalized extreme values model under power normalization technique would be a better alternative to other lifetime models available in existing literature, especially, in extreme values field.

APPENDIX

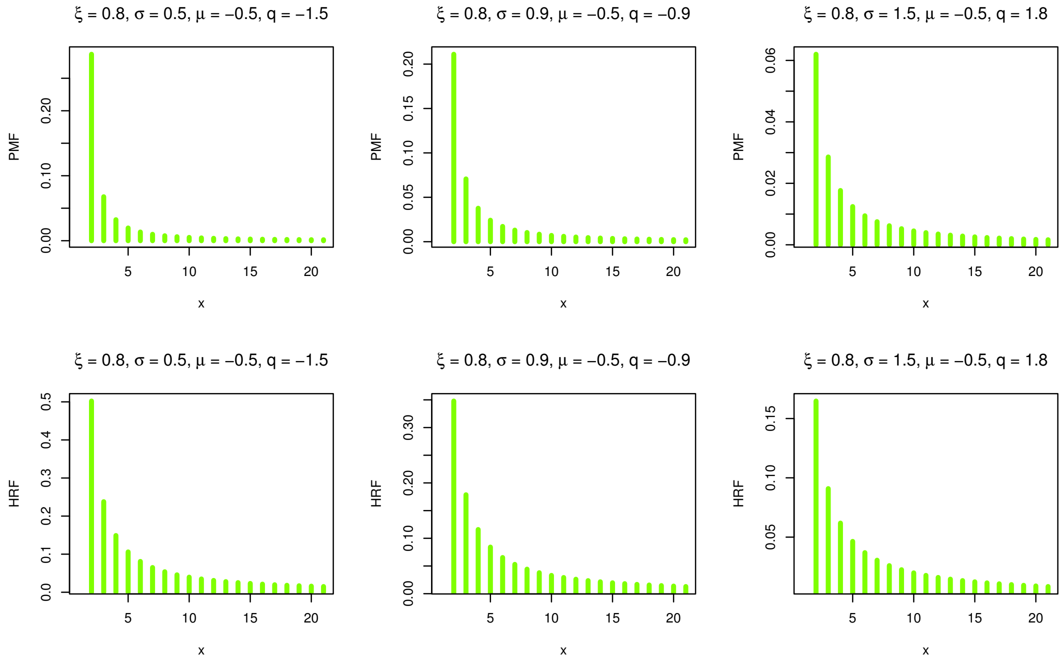


Figure 18: The PMF and HRF plots of the Dq-GEVP model for some parameter values.

ACKNOWLEDGMENTS

This project was supported by the deanship of scientific research at Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia. This study is supported via funding from Prince Sattam bin Abdulaziz University, project number (PSAU/2023/R/1444).

REFERENCES

[1] Altun, E.; El-Morshedy, M. and Eliwa, M.S. (2020). A study on discrete Bilal distribution with properties and applications on integer-valued autoregressive process, *REVSTAT – Statistical Journal*, **18**, 70–99.

[2] Barakat, H.M.; Nigm, E.M. and El-Adll, M.E. (2010). Comparison between the rates of convergence of extremes under linear and under power normalization, *Statistical Papers*, **51**(1), 149–164.

[3] Barakat, H.M.; Nigm, E.M. and Khaled, O.M. (2013). Extreme value modeling under power normalization, *Applied Mathematical Modelling*, **37**(24), 10162–10169.

[4] Barakat, H.M.; Nigm, E.M. and Khaled, O.M. (2014a). Statistical modeling of extremes under linear and power normalizations with applications to air pollution, *Kuwait Journal of Science*, **41**(1).

[5] Barakat, H.M.; Nigm, E.M.; Khaled, O.M. and Alaswed, H.A. (2014b). The counterparts of Hill estimators under power normalization, *Journal of Applied Statistical Science*, **22**(1/2), 87.

[6] Barakat, H.M.; Nigm, E.M.; Khaled, O.M. and Momenkhan, F.A. (2015). Bootstrap method for order statistics and modeling study of the air pollution, *Communications in Statistics – Simulation and Computation*, **44**(6), 1477–1491.

[7] Barakat, H.M.; Nigm, E.S.M. and Khaled, O.M. (2019). *Statistical techniques for modelling extreme value data and related applications*, Cambridge Scholars Publishing.

[8] Bebbington, M.; Lai, C.D.; Wellington, M. and Zitikis, R. (2012). The discrete additive Weibull distribution: a bathtub-shaped hazard for discontinuous failure data, *Reliability Engineering & System Safety*, **106**, 37–44.

[9] Charalambides, C.A. (2010). Discrete q-distributions on Bernoulli trials with a geometrically varying success probability, *Journal of Statistical Planning and Inference*, **140**(9), 2355–2383.

[10] Chung, K.S.; Chung, W.S.; Nam, S.T. and Kang, H.J. (1994). New q-derivative and q-logarithm, *International Journal of Theoretical Physics*, **33**(10), 2019–2029.

[11] Diaz, R.; Ortiz, C. and Pariguan, E. (2010). On the k-gamma q-distribution, *Open Mathematics*, **8**(3), 448–458.

[12] Diaz, R. and Pariguan, E. (2009). On the Gaussian q-distribution, *Journal of Mathematical Analysis and Applications*, **358**(1), 1–9.

[13] De Haan, L. (1971). A form of regular variation and its application to the domain of attraction of the double exponential distribution, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **17**(3), 241–258.

[14] De Haan, L. and Ferreira, A. (2007). *Extreme Value Theory: An Introduction*, Springer Science & Business Media.

[15] El-Morshedy, M.; Eliwa, M.S. and Altun, E. (2020). Discrete Burr-Hatke distribution with properties, estimation methods and regression model, *IEEE Access*, **8**, 74359–74370.

[16] Eliwa, M.S.; Altun, E.; El-Dawoody, M. and El-Morshedy, M. (2020). A new three-parameter discrete distribution with associated INAR (1) process and applications, *IEEE Access*, **8**, 91150–91162.

[17] Fisher, R.A. and Tippett, L.H.C. (1928). *Limiting forms of the frequency distribution of the largest or smallest member of a sample.* In "Mathematical Proceedings of the Cambridge Philosophical Society", **24**(2), pp. 180–190, Cambridge University Press.

[18] Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*, 2nd ed., Wiley, New York, NY, USA.

[19] Gilli, M. (2006). An application of extreme value theory for measuring financial risk, *Computational Economics*, **27**(2), 207–228.

[20] Gauchman, H. (2004). Integral inequalities in q-calculus, *Computers & Mathematics with Applications*, **47**(2-3), 281–300.

[21] Gnedenko, B.V.; Kotz, S. and Johnson, N.L. (1943). Sur la distribution limité du terme d'une série aléatoire, *Annals of Mathematics*, **44**, 423–453; *Breakthroughs in Statistics*, 195–225.

[22] Pantcheva, E. (1985). *Limit theorems for extreme order statistics under nonlinear normalization.* In "Stability Problems for Stochastic Models", pp. 284–309, Springer, Berlin, Heidelberg.

[23] Provost, S.B.; Saboor, A.; Cordeiro, G.M. and Mansoor, M. (2018). On the q-generalized extreme value distribution, *REVSTAT – Statistical Journal*, **16**(1), 45–70.

[24] Picoli Jr, S.; Mendes, R.S. and Malacarne, L.C. (2003). q-exponential, Weibull, and q-Weibull distributions: an empirical analysis, *Physica A: Statistical Mechanics and its Applications*, **324**(3-4), 678–688.

[25] Ravi, S. and Saeb, A. (2012). *A note on entropies of l-max stable, p-max stable, generalized Pareto and generalized log-Pareto distributions.* In "ProbStat Forum", **5**, pp. 62–79.

[26] Mathai, A.M. (2005). A pathway to matrix-variate gamma and normal densities, *Linear Algebra and Its Applications*, **396**, 317–328.

[27] Mathai, A.M. and Provost, S.B. (2006). On q-logistic and related models, *IEEE Transactions on Reliability*, **55**(2), 237–244.

[28] Nasri-Roudsari, D. (1999). Limit distributions of generalized order statistics under power normalization, *Communications in Statistics – Theory and Methods*, **28**(6), 1379–1389.

[29] Nekoukhou, V. and Bidram, H. (2015). A new four-parameter discrete distribution with bathtub and unimodal failure rate, *Journal of Applied Statistics*, **42**(12), 2654–2670.

[30] Srivastava, H.M. and Choi, J. (2012). *Zeta and q-Zeta Functions and Associated Series and Integrals*, Elsevier.

[31] De Sole, A. and Kac, V. (2003). On integral representations of q-gamma and q-beta functions, *arXiv* preprint math/0302032.

REVSTAT – Statistical Journal
Volume 22, Number 1, January 2024, 87–110
https://doi.org/10.57805/revstat.v22i1.457

Identifiability Analysis Using Data Cloning

Authors: José Augusto Sartori Junior ✉
– Institute of Mathematics and Statistics, University of São Paulo, Brazil
jsartori@ime.usp.br

Márcia D'Elia Branco
– Institute of Mathematics and Statistics, University of São Paulo, Brazil
mbranco@ime.usp.br

Received: March 2021 Revised: March 2022 Accepted: March 2022

Abstract:

- Lack of identifiability in statistical models may hinder unique inferential conclusions. Therefore, the search for parametric constraints that ensure identifiability is of utmost importance in statistics. However, for complex modeling strategies, even acquiring the knowledge that the model is unidentifiable may prove very difficult. In this paper, we investigate the use of Data Cloning, a modern algorithm for classical inference in latent variable models, as a tool for assessing model identifiability. We discuss its advantages and disadvantages and illustrate its use with a dynamic linear model.

Keywords:

- *identifiability; data cloning; dynamic models; MCMC algorithms.*

AMS Subject Classification:

- 62A99, 62F10.

✉ Corresponding author.

1. INTRODUCTION

The specification of identifiable statistical models is an extremely important step in statistical inference. Data-driven decisions in unidentifiable models may be non-unique, *i.e.* it may not be possible to choose a single optimal decision based solely on the data at hand (San Martín 2018 [22]). Therefore, be it in the classical or Bayesian framework, unidentifiability may lead to severely wrong answers to scientific inquiries.

In classical statistics, for instance, lack of identifiability implies there does not exist a consistent estimator for some or all of the model parameters (Paulino and Pereira 1994 [15]). In other words, no matter how large the sample size, with an unidentifiable model we will never be able to distinguish the true parameter value from at least one other alternative value.

Unidentifiability can also cause problems in the Bayesian setting. When using flat prior distributions for unidentifiable parameters, the resulting posterior can still be flat. Moreover, if the prior is improper, then the posterior may also be improper (Lindley 1971 [14]). However, even if informative priors are used in this situation, it is not very clear what inferences can be drawn *a posteriori*. San Martín (2018) [22] argues for inference on the sufficient parameter (see meaning therein) and shows how the influence of the prior distribution never vanishes for unidentifiable parameters.

Most of the models in the statisticians' basic toolkit enjoy solid theoretical foundations. However, the recent advances in computational power have led to the appearance of more complex models for which identifiability is not always guaranteed. Earlier, the theoretical development of statistics was followed by the studies of computational feasibility of the models, however, now a lot of theoretical work is to understand the properties of the newer modeling strategies.

Due to the difficulty in answering the question of whether or not a particular statistical model is identifiable, there is a sizable literature suggesting a diverse range of methods. There have been approaches using differential geometry (Villaverde *et al.* 2019 [25]), differential algebra (Bellu *et al.* 2007 [3]), measure theory (San Martín 2015 [21]) etc. However, sometimes theory alone does not end the issue and computational strategies are called upon to provide empirical evidence of model identifiability.

A closely related concept is that of estimability or practical identifiability. It may be the case that the statistical model is identifiable, but the data available is of poor quality or the model has been incorrectly specified. This may hinder the ability to estimate the model parameters and the uncertainty associated with such estimates, which can affect both inferential and predictive tasks. In other words, estimability deals with the question of whether the data at hand can reliably estimate the desired quantities. Identifiability, however, is concerned with the existence, for any two distinct parameter values, of a hypothetical data set which can differentiate between them. As such, lack of estimability does not imply lack of identifiability, although the converse is always true (Paulino and Pereira 1994 [15]).

Lack of estimability is commonly caused by low signal-to-noise ratio in the data, low sample size, or inappropriate sampling scheme (Lele *et al.* 2010 [13]). These problems often result in a likelihood function of the parameters that have many local maxima or an almost flat region.

In these scenarios, any given estimation algorithm might result in parameter estimates which yield considerably distinct inferential conclusions. Furthermore, confidence regions may present one or more coordinates assuming unreasonably large values.

Recently, Lele *et al.* (2007) [12] proposed an algorithm for maximum likelihood estimation, called Data Cloning, which is of particular relevance in latent variable models. The method is quite intuitive and draws motivation from the idea of replicability of experiments in frequentist statistics. To be more specific, the data cloning algorithm starts from a prior distribution on the parameter space and sequentially updates it using the same data set until some diagnostic measures reach specified thresholds.

As Lele *et al.* (2010) [13] show, if the model is unidentifiable then convergence issues can be easily spotted with the tools for diagnosing convergence of the algorithm. Data cloning has been used earlier for studying model estimability. Campbell and Lele (2014) [4]) proposed an ANOVA test of estimability based on data cloning and Peacock *et al.* (2017) [16] employed data cloning to assess estimability of a spatio-temporal model under distinct study designs based on an observed data set.

Our objective in this paper is to introduce data cloning as a practical tool for the assessment of identifiability of statistical models. For this, we show how to plan and perform a simulation study that can shed light on possible problems in the structure of the statistical model. Our idea is somewhat similar to that of Peacock *et al.* (2017) [16] in that we also employ simulated data. However, instead of exploring possible alternative study designs based on observed data, we advocate the exploration of a multitude of possible data based on as many as possible parameter values to study the structure of the model itself.

In Section 2 we present the data cloning algorithm and its main diagnostic measures that can be used to study model identifiability. In Section 3 we present the formal definitions of identifiability of statistical models, relate them to data cloning, and show, theoretically, how the identifiability issue reveals itself in the Gaussian dynamic linear model. Finally, in Section 4 we present a simulation study using the package *dclone* (Solymos 2010 [23]) from *R* (R Core Team 2020 [19]) and *JAGS* (Plummer 2017 [18]), and discuss the evidence it brings about identifiability in the adopted model.

2. DATA CLONING

In the subjective realm of Bayesian inference, a great deal of discomfort in the prior specification vanishes for highly informative data. An important result in Bayesian asymptotics, due to Walker (1969) [26], shows that, under some regularity conditions, for large n the posterior distribution $\pi(\boldsymbol{\theta}|y_1, ..., y_n)$ is approximately Gaussian with mean $\hat{\boldsymbol{\theta}}$, the maximum likelihood estimate of $\boldsymbol{\theta}$, and covariance matrix $\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$, the inverse of the Fisher information evaluated at this maximum. For this, see also Turkman *et al.* (2019), Sec. 8.1 [24].

Suppose we performed an experiment k times independently and happened to observe the exact same realization $\mathbf{y}^{(j)} = \mathbf{y} = (y_1, ..., y_n)$ for all $j \in \{1, ..., k\}$ with probability density function $f(\mathbf{y}|\boldsymbol{\theta})$ for each experiment. Let $\pi_k(\boldsymbol{\theta}|\mathbf{y})$ denote the posterior distribution updated with samples for k such experiments. Since the k experiments are independent, Bayes theorem says this distribution is

$$\pi_k(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}^{(1)}, ..., \mathbf{y}^{(k)}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \prod_{j=1}^{k} f(\mathbf{y}|\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta})\{f(\mathbf{y}|\boldsymbol{\theta})\}^k \quad . \tag{2.1}$$

Let $L(\boldsymbol{\theta}; \mathbf{y}^{(1)}, ..., \mathbf{y}^{(k)})$ and $l(\boldsymbol{\theta}; \mathbf{y}^{(1)}, ..., \mathbf{y}^{(k)})$ denote the likelihood and log-likelihood functions of these experiments, respectively. If $\hat{\boldsymbol{\theta}}_{(n)}$ is maximum likelihood estimate under any one of the k experiments, then it follows immediately that

$$\hat{\boldsymbol{\theta}} = \arg\sup_{\theta\in\Theta} L(\boldsymbol{\theta}; \mathbf{y}^{(1)}, ..., \mathbf{y}^{(k)}) = \arg\sup_{\theta\in\Theta} [L(\boldsymbol{\theta}; \mathbf{y})]^k = \arg\sup_{\theta\in\Theta} L(\boldsymbol{\theta}; \mathbf{y}) = \hat{\boldsymbol{\theta}}_{(n)} \tag{2.2}$$

and if we let $\mathbb{V}(\mathbf{X})$ denote the covariance matrix of a random vector $\mathbf{X}$, then

$$\mathcal{I}(\hat{\boldsymbol{\theta}}) = \mathbb{V}\left(\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{Y}^{(1)}, ..., \mathbf{Y}^{(k)})}{\partial \boldsymbol{\theta}}\right)\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbb{V}\left(\sum_{j=1}^{k} \frac{\partial \ell(\boldsymbol{\theta}; \mathbf{Y}^{(j)})}{\partial \boldsymbol{\theta}}\right)\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = k\mathcal{I}(\hat{\boldsymbol{\theta}}_{(n)}) \quad , \tag{2.3}$$

in which the last equality follows from the independence of the experiments. Therefore, for a fixed n and k arbitrarily large, our posterior distribution would be well approximated by a Gaussian distribution with mean $\hat{\boldsymbol{\theta}}_{(n)}$ and covariance matrix $\frac{1}{k}\mathcal{I}^{-1}\left(\hat{\boldsymbol{\theta}}_{(n)}\right)$.

Note that we have not made a single comment about what prior $\pi(\boldsymbol{\theta})$ we started with. In fact, the previous results are valid as long as the prior distribution and the likelihood function satisfy some mild regularity conditions. In other words, for any two such priors π_1 and π_2 over Θ, there is a number of experiments, k, for which the posterior distributions would be arbitrarily close to each other (Lele *et al.* 2010 [13]).

Similarly, with minor modifications, the results above can be applied to latent variable models. Since the experiments are performed independently, realizations of the hidden stochastic process $\{\mathbf{X}^{(j)}\}$, $j \in \{1, ..., k\}$, are also assumed to have occurred k times independently. We begin by assigning a joint prior distribution $\pi(\boldsymbol{\theta}, \mathbf{x}) = \pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ for the parameters and latent variables. The resulting posterior distribution is given by

$$\pi_k(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta}) \prod_{j=1}^{k} f(\mathbf{y}^{(j)}|\mathbf{x}^{(j)}, \boldsymbol{\theta})\pi(\mathbf{x}^{(j)}|\boldsymbol{\theta})}{f(\mathbf{y}^{(1)}, ..., \mathbf{y}^{(k)})} \quad . \tag{2.4}$$

For inference on the parameter vector, it suffices to marginalize on $\mathbf{x}$, which is made easier by the assumption of independence:

$$\begin{aligned}\pi_k(\boldsymbol{\theta}|\mathbf{y}) &= \int_{\mathcal{X}} \pi(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y}^{(j)}, ..., \mathbf{y}^{(k)}) d\mathbf{x} \\ &= \frac{\left\{\int_{\mathcal{X}} \prod_{j=1}^{k} f(\mathbf{y}|\mathbf{x}^{(j)}, \boldsymbol{\theta})\pi(\mathbf{x}^{(j)}|\boldsymbol{\theta}) d\mathbf{x}^{(j)}\right\}\pi(\boldsymbol{\theta})}{f(\mathbf{y}^{(1)}, ..., \mathbf{y}^{(k)})} \\ &= \frac{\left\{\prod_{j=1}^{k} L(\boldsymbol{\theta}; \mathbf{y})\right\}\pi(\boldsymbol{\theta})}{f(\mathbf{y}^{(1)}, ..., \mathbf{y}^{(k)})} \\ &= \frac{\{L(\boldsymbol{\theta}; \mathbf{y})\}^k \pi(\boldsymbol{\theta})}{\int_{\Theta} \{L(\boldsymbol{\theta}; \mathbf{y})\}^k \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}} \quad . \end{aligned} \tag{2.5}$$

In summary, if we obtained such an odd data set over a very large number of independent experiments, from (almost) arbitrary initial prior distributions, we could perform frequentist inference within the Bayesian setting. All that is required is we take samples from the posterior distribution of $\boldsymbol{\theta}$ and compute the mean vector and covariance matrix.

At first glance, it may seem that we replaced the high-dimensional integral required for maximum likelihood estimation, namely $L(\boldsymbol{\theta}; \mathbf{y})$, with a possibly much more complicated one in the denominator of (2.5). However, commonly employed Bayesian software for probabilistic sampling avoid the integration procedure altogether, which surely is an important reason why there is increasing adoption of the subjective paradigm amongst researchers dealing with complex latent variable models (Lele *et al.* 2010 [13]).

2.1. The Algorithm

Admittedly, repeating the same experiment may be just as infeasible as simply increasing the sample size. Thus, given a dataset, Lele *et al.* (2007) [12] propose that we clone it k times, with k as large as is computationally possible, and then draw samples from the k-times cloned posterior π_k. Although what we are using is in fact fake data, the machine on which the sampling algorithm will run cannot tell the difference.

As before, let $\mathbf{y} = (y_1, ..., y_n)$ denote the available realization of a measurement process generated by a hidden latent process. Suppose $\boldsymbol{\theta}$ is a continuous random vector, $p(\boldsymbol{\theta})$ is a proposal distribution, and define the burn-in length $N_{burn} < N_{sim}$, the simulation length for the Metropolis-Hastings algorithm. Algorithm 1 provides a way to sample from π_k; for the regularity conditions we direct the reader to Lele *et al.* (2010) [13].

Algorithm 1: Data Cloning Metropolis-Hastings (Lele *et al.* 2007 [12])

```
Generate $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$ and $\mathbf{x}^{*(1)}, ..., \mathbf{x}^{*(k)} \sim \pi(\mathbf{x}|\boldsymbol{\theta}^*)$;
for $l \in \{1, ..., N_{sim}\}$ do
    Compute $q^* = \prod_{j=1}^{k} f(\mathbf{y}|\mathbf{x}^{*(j)}, \boldsymbol{\theta}^*)$;
    Generate $\boldsymbol{\theta}^\# \sim p(\boldsymbol{\theta})$ and $\mathbf{x}^{\#(1)}, ..., \mathbf{x}^{\#(k)} \sim \pi(\mathbf{x}|\boldsymbol{\theta}^\#)$;
    Compute $q^\# = \prod_{j=1}^{k} f(\mathbf{y}|\mathbf{x}^{\#(j)}, \boldsymbol{\theta}^\#)$;
    Generate $U \sim \text{Uniform}(0, 1)$;
    if $U < \min\{1, q^\#/q^*\}$ then
        Set $(\boldsymbol{\theta}, \mathbf{x}^{(1)}, ..., \mathbf{x}^{(k)})_l = (\boldsymbol{\theta}^\#, \mathbf{x}^{\#(1)}, ..., \mathbf{x}^{\#(k)})$;
    else
        Set $(\boldsymbol{\theta}, \mathbf{x}^{(1)}, ..., \mathbf{x}^{(k)})_l = (\boldsymbol{\theta}^*, \mathbf{x}^{*(1)}, ..., \mathbf{x}^{*(k)})$;
    end
end
Discard $(\boldsymbol{\theta}, \mathbf{x}^{(j)}, ..., \mathbf{x}^{(k)})_1, ..., (\boldsymbol{\theta}, \mathbf{x}^{(j)}, ..., \mathbf{x}^{(k)})_{N_{burn}-1}$;
```

As long as the number of clones k is large enough and the regularity conditions are satisfied, the mean of the samples drawn from Algorithm 1 is a numerical approximation to the maximum likelihood estimate. Also, their covariance matrix is the inverse of the k-times scaled observed Fisher information matrix. As pointed out in Lele *et al.* (2007) [12], increasing the number of clones provides better numerical accuracy in the estimates. However, this does increase the computational cost of the algorithm considerably, since for each new clone we must simulate the latent process generating it. For models in which the number of latent variables grows exponentially with the sample size, the data cloning algorithm will surely demand an unreasonable amount of computational time. Nevertheless, it performs incredibly well for longitudinal and time series data, since the number of latent variables is usually on the order of the sample size.

2.2. Convergence Diagnostics

Another great feature of data cloning is the plethora of diagnostic measures available. On one hand, since we are using probabilistic sampling algorithms, it is mandatory to diagnose convergence of the sampling algorithm itself. For this, measures such as the potential scale reduction factor $\hat{R}$ (Gelman and Rubin 1992 [8]) and the effective sample size N_{eff}, for example, assess convergence of the Markov chains and the within-chains autocorrelations, respectively. For these and other measures and their implementation, see for instance Turkman *et al.* (2019), Ch. 9 [24].

On the other hand, for likelihood-based inference, we need to ensure the posterior distribution is well approximated by a Gaussian distribution and also that this distribution degenerates at a point. Indeed, only then we can be confident that the influence of the choice of prior distribution has vanished and the mean of the posterior samples is a good approximation of the desired maximum likelihood estimate. Specific to the data cloning algorithm, we need to check whether the posterior distribution

(i) has become nearly degenerate and
(ii) nearly Gaussian.

The number of clones required for these behaviors depends heavily on the likelihood function and prior distribution chosen. Fortunately, the diagnostic measures recommended by Lele *et al.* (2010) [13] are simple to compute and allow the selection of an adequate number of clones for the problem at hand.

If the assumptions of the data cloning algorithm are satisfied, then the Fisher information matrix is positive definite in a neighborhood around the maximum likelihood estimate. Furthermore, recall from Equation (2.3) that, in the k-times cloned posterior π_k, the estimate of the inverse Fisher information matrix from the posterior samples is scaled by the inverse of k. Therefore, as we increase the number of clones, the eigenvalues of the estimated covariance matrix from the samples should decrease at approximately a rate k^{-1}.

For a positive definite matrix, the Courant-Fischer Theorem ensures that the greatest eigenvalue provides an upper bound on the elements of the main diagonal (Horn and Johnson 2012 [10]). Hence, since the greatest eigenvalue should decrease at the rate k^{-1}, we have an upper bound on the rate at which the elements of the main diagonal of the estimated inverse Fisher information matrix must decrease. This enables us to measure the rate at which the posterior distribution is degenerating to a point mass probability measure on the maximum likelihood estimate, since the elements in the main diagonal of the said matrix represent an estimate of the posterior variances for the model parameters.

Let $\lambda_{max,k}$ denote the maximum eigenvalue of the k-times cloned posterior covariance matrix. Then, for large k and under regularity conditions,

$$\lambda^S_{max,k} = \frac{\lambda_{max,k}}{\lambda_{max,1}} \approx \frac{1}{k} \quad . \tag{2.6}$$

Lele *et al.* (2010) [13] call $\lambda^S_{max,k}$ the scaled maximum eigenvalue for k clones. The authors suggest monitoring this quantity to assess its rate of convergence to zero as we increase the number of clones. The closer this measure is to zero, the more mass the posterior distribution assigns to small neighborhoods of the (possibly) maximum likelihood estimate.

For the second item, the normality of the posterior distribution, Lele *et al.* (2010) [13] suggest computing two statistics from the posterior samples. As before, let p be the dimension of the parameter vector $\boldsymbol{\theta}$ and N denote the number of samples obtained from Algorithm 1 after discarding the ones from the burn-in period. Let E_i denote the $(i-0.5)/N$ quantile of a χ^2_p distribution, $i \in \{1, ..., N\}$. Define

$$O_i = (\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})^\top \widehat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}}) \tag{2.7}$$

for $i \in \{1, ..., N\}$, with $\widehat{\boldsymbol{\Sigma}}$ the estimated posterior covariance matrix and let $O_{(i)}$ denote their ordered values. Since $O_{(i)}$ are simply estimates of E_i, the statistics

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (O_{(i)} - E_{(i)})^2 \tag{2.8}$$

and

$$r^2 = 1 - \hat{\rho}^2 \big(O_{(1)}, ..., O_{(N)}; E_1, ..., E_N\big) \quad , \tag{2.9}$$

in which $\hat{\rho}$ denotes the estimated Pearson correlation coefficient, approach zero as the number of clones k increases.

Solymos (2010) [23] provides an implementation of data cloning for *R* (R Core Team 2020 [19]). The package allows the use of common probabilistic sampling software amongst Bayesian practitioners, such as *JAGS* (Plummer 2017 [18]) and Stan (Carpenter *et al.* (2017) [5]) to perform sampling from the cloned posterior distribution and includes all diagnostic measures described above. The paper by Solymos (2010) [23] and also the original papers by Lele *et al.* (2007) [12] and Lele *et al.* (2010) [13] provide plenty of examples to get acquainted with data cloning.

3. IDENTIFIABILITY

Let $\mathcal{Y}$ denote a sample space, $\mathcal{A}$ a σ-algebra of subsets of $\mathcal{Y}$ and $\mathcal{M}(\mathcal{Y}, \mathcal{A})$ denote the set of probability measures on $(\mathcal{Y}, \mathcal{A})$. In statistical theory, the inferential procedure is enabled by equipping the measurable space $(\mathcal{Y}, \mathcal{A})$ with a family of probability measures $\mathcal{F} \subset \mathcal{M}(\mathcal{Y}, \mathcal{A})$. For practical purposes, this family is defined through a known map $\boldsymbol{\Phi} : \Theta \to \mathcal{M}(\mathcal{Y}, \mathcal{A})$, with Θ being the parameter space in the parametric scenario. Specifically, a statistical model is a triple $\mathcal{E} = (\mathcal{Y}, \mathcal{A}, \mathcal{F} = \{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\})$, in which $\mathcal{F}$ is a family of probability measures on $(\mathcal{Y}, \mathcal{A})$ indexed by the parameter space Θ.

Notice the definition of a statistical model imposes no restrictions on Θ, allowing for parametric, semiparametric and nonparametric structures (San Martín 2018 [22]). We assume in this paper that:

(i) $\Theta \subset \mathbb{R}^p$ is an Euclidean space;

(ii) the sample space $\mathcal{Y}$ is equipped with a topology and $\mathcal{A} = \mathcal{B}(\mathcal{Y})$ is the Borel σ-algebra obtained from the topology on $\mathcal{Y}$;

(iii) the probability measures in $\mathcal{F}$ are absolutely continuous with respect to some measure on $(\mathcal{Y}, \mathcal{A})$.

The latter constraint allows for much simplification in the discussion since now we can represent $\mathcal{F} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ as a family of probability density functions. We are now ready to define identifiability of a parameter in the sampling theoretical framework.

Definition 3.1. Let $\mathcal{E}$ be a statistical model. A parameter $\boldsymbol{\theta} \in \Theta$ is said to be **identifiable** if for any $\boldsymbol{\theta}^* \in \Theta$, $f_{\boldsymbol{\theta}}(\mathbf{y}) = f_{\boldsymbol{\theta}^*}(\mathbf{y})$ occurs, for all $\mathbf{y} \in \mathcal{Y}$, if and only if $\boldsymbol{\theta}^* = \boldsymbol{\theta}$.

The concept above is referred to as global identifiability (Koopmans and Reiersol 1950 [11]). This is done to distinguish it from local identifiability, which we define below. However, in this paper, we refer to the former property as simply *identifiability*.

Definition 3.2. Let $\mathcal{E}$ be a statistical model. A parameter $\boldsymbol{\theta} \in \Theta$ is said to be **locally identifiable** if there exists $\epsilon > 0$ and a neighborhood $N_\epsilon(\boldsymbol{\theta}) \subset \Theta$ of $\boldsymbol{\theta}$ such that for any $\boldsymbol{\theta}^* \in N_\epsilon(\theta)$, $f_{\boldsymbol{\theta}}(\mathbf{y}) = f_{\boldsymbol{\theta}^*}(\mathbf{y})$ occurs, for all $\mathbf{y} \in \mathcal{Y}$, if and only if $\boldsymbol{\theta}^* = \boldsymbol{\theta}$.

We can see thus that local identifiability is weaker than (global) identifiability. In fact, a globally identifiable parameter is always locally identifiable. The converse, however, need not be true.

So far the definitions allow us to talk only about a single point in the parameter space. For the inferential procedure to be satisfactory, we would like to know whether all parameter values $\theta \in \Theta$ are identifiable. Fortunately, the definitions above are easily extended to the entire parameter space Θ.

Definition 3.3. A statistical model $\mathcal{E}$ is said to be identifiable (locally identifiable) if for all $\boldsymbol{\theta} \in \Theta$, $\boldsymbol{\theta}$ is identifiable (locally identifiable).

Parameters which yield the same likelihood function are said to be *observationally equivalent*, and it is possible to construct an equivalence relation using the concept of identifiability; see for example Picci (1977) [17] or Florens and Simoni (2011) [6] and references therein. This relation $\sim$ is such that, for any $\boldsymbol{\theta}, \boldsymbol{\theta}^* \in \Theta$, $\boldsymbol{\theta} \sim \boldsymbol{\theta}^*$ if, and only if, $f_{\boldsymbol{\theta}}(\mathbf{y}) = f_{\boldsymbol{\theta}^*}(\mathbf{y})$ for all $\mathbf{y} \in \mathcal{Y}$. Through the equivalence relation defined above we obtain the quotient space $\tilde{\Theta} = \Theta/\sim$. The elements of the quotient spaces are the equivalence classes induced by $\sim$ on Θ. Thus, there always exists a canonical statistical model $\mathcal{E}_{\tilde{\Theta}} = (\mathcal{Y}, \mathcal{A}, \mathcal{F} = \{f_{[\boldsymbol{\theta}]} : [\theta] \in \tilde{\Theta}\})$ which is set identifiable, *i.e.* the family $\mathcal{F}$ is indexed by the equivalence classes.

An easy-to-prove property of equivalence classes which makes them very convenient for studying identifiability is that they are disjoint. Thus, it is sufficient for statistical identifiability to define a function which maps each equivalence class to a single element in that class, since then the equivalence classes will be reduced to singletons. Florens and Simoni (2011) [6] call such functions *sections*. Let $[\boldsymbol{\theta}] \in \tilde{\Theta}$ denote the class of equivalence of $\boldsymbol{\theta} \in \Theta$, *i.e.* $[\boldsymbol{\theta}] = \{\boldsymbol{\theta}^* \in \Theta : \boldsymbol{\theta}^* \sim \theta\}$.

Definition 3.4. A **section** is a function $\sigma : \tilde{\Theta} \rightarrow \Theta$ such that for all $[\boldsymbol{\theta}] \in \tilde{\Theta}$, $\sigma([\boldsymbol{\theta}]) \in [\boldsymbol{\theta}]$.

As previously mentioned, the equivalence classes being disjoint leads us to an identifiable statistical model $\mathcal{E}_\sigma = (\mathcal{Y}, \mathcal{A}, \mathcal{F} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \sigma(\tilde{\Theta})\})$, in which σ is any one section. Moreover, the choice of section is really irrelevant, as Paulino and Pereira (1994) [15] point out, since for any two sections σ and σ^*, there exists a bijective function $h : \sigma(\tilde{\Theta}) \rightarrow \sigma^*(\tilde{\Theta})$.

3.1. Implications to Data Cloning

Another important consequence of identifiability which we exploit in this paper is that, if $\mathcal{E}$ happens to be unidentifiable, then the maximum likelihood estimate of $\boldsymbol{\theta}$ for any given sample (if it exists) is not unique. In fact, given any point estimate $\hat{\boldsymbol{\theta}}(y) \in \Theta$, there is $\hat{\boldsymbol{\theta}}^*(y) \in [\hat{\boldsymbol{\theta}}]$ such that $L(\hat{\boldsymbol{\theta}}; y) = L(\hat{\boldsymbol{\theta}}^*; y)$. It becomes clear now why consistency of the maximum likelihood procedure is no longer guaranteed. Under model unidentifiability thus there is a class of equivalence of undistinguishable candidates to the maximum likelihood estimate

$$(3.1) \qquad [\hat{\boldsymbol{\theta}}] = \arg \sup_{[\boldsymbol{\theta}] \in \tilde{\Theta}} L([\boldsymbol{\theta}]; y) = \left\{ \boldsymbol{\theta} \in \Theta : \boldsymbol{\theta} = \arg \sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; y) \right\} \quad .$$

Lele *et al.* (2010) [13] discuss the behavior of the k-times cloned posterior distribution under model unidentifiability. In this scenario, let $[\hat{\boldsymbol{\theta}}]$ be the equivalence class of the maximum likelihood estimate. The authors show that if $[\hat{\boldsymbol{\theta}}]$ is not a singleton, then

$$(3.2) \qquad \pi_k(\boldsymbol{\theta}|\mathbf{y}) \xrightarrow{\mathcal{L}} \frac{\pi(\boldsymbol{\theta})}{\int_{[\hat{\boldsymbol{\theta}}]} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}} \quad , \quad \forall \boldsymbol{\theta} \in [\hat{\boldsymbol{\theta}}] \quad .$$

Therefore, it seems we can, in theory, use data cloning to investigate the identifiability of complex statistical models. If the posterior samples generated using data cloning, for increasing values of k, do not seem normally distributed and/or seem to degenerate at a set of values, then the model may be unidentifiable.

In reality, however, we must not hurry to conclusions. Identifiability of statistical models can only be assessed using analytical techniques and it is a mathematical question in general. It precedes statistical inference (Koopmans and Reiersol 1950 [11]). What we can study with data cloning diagnostics is model estimability under a particular data set. However, detecting estimability issues over many distinct data sets would lead us to question the very structure of the statistical model we are employing. Thus, a general guideline for a simulation study for assessing identifiability using data cloning is:

(i) for various sample sizes, simulate several data sets from a postulated statistical model;

(ii) for each data set, fit the model using data cloning with distinct prior distributions and with increasing values of the number of clones k;

(iii) analyze the posterior samples to study the behavior of the algorithm.

Data cloning diagnostics should always reveal convergence issues when the algorithm is used to estimate the parameters of an unidentifiable model. However, under possibly very informative priors and weakly informative data, it may happen that even for large values of k the sampling algorithm will output samples from π_k which do not indicate problems. This is resonant with the arguments of Lindley (1971) [14] on the irrelevance of the question of identifiability within the subjective Bayesian paradigm.

Furthermore, even if the statistical model is identifiable, there are plenty of other ways in which the data cloning algorithm may fail to converge. It is no coincidence that advances in Bayesian computational methods have been accompanied by developments of techniques for diagnosing convergence issues. Since data cloning uses MCMC algorithms for likelihood-based inference, the same convergence issues of sampling algorithms that Bayesian analyses face must also be considered.

3.2. Unidentifiability of the Gaussian DLM

The Gaussian dynamic linear model is particularly convenient for our purposes since it illustrates many aspects of identifiability in a simple manner.

Definition 3.5. *The dynamic model with state and observation equations*

$$\begin{cases} Y_t = FX_t + \nu_t & , \quad \nu_t \overset{iid}{\sim} \mathcal{N}(0,V) \\ X_t = GX_{t-1} + \omega_t & , \quad \omega_t \overset{iid}{\sim} \mathcal{N}(0,W) \\ X_0 \sim \mathcal{N}(m_0, C_0) \end{cases} \tag{3.3}$$

is called a ***univariate Gaussian dynamic linear model*** *with parameter vector* $\boldsymbol{\theta} = (F,G,V,W) \in \mathbb{R}^2 \times \mathbb{R}^2_+$ *and initial information set* $D_0 = \{m_0, C_0\}$.

The statistical model of Definition 3.5 is not identifiable as it is. This is a well known result in the literature of dynamic models and some identifiability constraints for the multivariate scenario can be found in Harvey (1989) [9]. The usual path to a proof of the unidentifiability of the dynamic linear model employs a change of variables in its defining observation and process equations. Under Gaussian errors it is easy to see that for any real number $s \neq 0$

$$\begin{cases} Y_t = FX_t + \nu_t \\ \\ X_t = GX_{t-1} + \omega_t \end{cases} \iff \begin{cases} Y_t = (Fs^{-1})(sX_t) + \nu_t \\ \\ sX_t = G(sX_{t-1}) + s\omega_t \end{cases} \iff \begin{cases} Y_t = F^*X_t^* + \nu_t \\ \\ X_t^* = GX_{t-1}^* + \omega_t^* \end{cases}, \tag{3.4}$$

in which $F^* = Fs^{-1}$, $\omega_t^* = s\omega_t$ and $X_t^* = sX_t$, for all $t \in \mathcal{T}$. Notice now the process equation random error is distributed as $\omega_t^* \overset{iid}{\sim} \mathcal{N}(0, W^*)$, with $W^* = s^2W$.

Therefore, this change of variables implies that, if $\boldsymbol{\theta} = (F,G,V,W)$ is the original parameter vector and $\boldsymbol{\theta}^* = (F^*,G,V,W^*)$ is the parameter vector that results from the transformation proposed, then it follows that for all $(y_1,...,y_T) \in \mathcal{Y}$ we have $f_{\boldsymbol{\theta}}(y_1,...,y_T) = f_{\boldsymbol{\theta}^*}(y_1,...,y_T)$ for any $s \neq 0$.

The structural equations define the density functions uniquely and the arguments above are sufficient to prove this model is unidentifiable. The general suggestions for enforcing model identifiability in this context are to

(i) fix the parameter F to a known non-zero constant or

(ii) fix the process variance W to a known positive constant and constrain F to be either strictly positive or strictly negative (Harvey 1989 [9]).

Thus, in the identifiable statistical model, one must choose between estimating F or estimating W.

There are, however, an infinite number of other restrictions, or sections, on the parameter space which lead to the same statistical model up to a bijective function. Firstly, note that for the Gaussian dynamic linear model, the equivalence classes are readily built from the proof in (3.4). As a matter of fact, we know that $\boldsymbol{\theta} \sim \boldsymbol{\theta}^*$ if there exists $s \in \mathbb{R}$ such that

$$\boldsymbol{\theta} = (F, G, V, W) \sim (F/s, G, V, s^2 W) = (F^*, G, V, W^*) = \boldsymbol{\theta}^* \quad .$$

This, in turn, implies $\boldsymbol{\theta} \sim \theta^*$ whenever $F^2 W = (F^*)^2 W^*$. If we let $(a, b, c) \in \mathbb{R} \times \mathbb{R}_+^2$, then we can write the quotient space as

$$\tilde{\Theta} = \bigcup_{(a,b,c) \in \mathbb{R} \times \mathbb{R}_+^2} \big\{\{(F, G, V, W) \in \Theta : G = a, V = b \text{ and } F^2 W = c\}\big\} \quad . \tag{3.5}$$

We recall once again that the equivalence classes are disjoint. Therefore, once we build them, it is sufficient for model identifiability that we find a section $\sigma : \tilde{\Theta} \to \Theta$ such that the equivalence classes of $\sigma(\tilde{\Theta}) \subset \Theta$ are singletons (Paulino and Pereira 1994 [15]). For clarity of exposition, let $[(a, b, c)] = \{(F, G, V, W) \in \Theta : G = a, V = b, F^2 W = c\} \in \tilde{\Theta}$ denote the equivalence classes on Θ for all $(a, b, c) \in \mathbb{R} \times \mathbb{R}_+^2$. The general identifiability constraints can now be stated as

Proposition 3.1. *Let $\mathcal{E}$ be the Gaussian dynamic linear model as in Definition 3.5. A sufficient condition for the function $\sigma : \tilde{\Theta} \to \Theta$ to be a section on $\tilde{\Theta}$ is that for all $(a, b, c) \in \mathbb{R} \times \mathbb{R}_+^2$, the set function $\sigma : [(a, b, c)] \mapsto (u_1(a, b, c), G, V, u_2(a, b, c))$, with $u_1^2 u_2 : (a, b, c) \mapsto c$.*

Proof: We need to show that σ is injective and $\sigma([(a, b, c)] \in [(a, b, c)]$ for all such equivalence classes. The latter follows immediatly from the fact that if $\theta = (F, G, V, W)$ is such that $G = a$, $V = b$ and $F^2 W = c$, then $\boldsymbol{\theta} \in [(a, b, c)]$. Moreover, since equivalence classes are disjoint this implies σ is injective. Therefore, taking $F = u_1(a, b, c)$ and $W = u_2(a, b, c)$, the proof is complete. □

We can now write the commonly suggested restrictions for the Gaussian DLM as sections on the parameter space. Fixing $F = s$, for some real constant $s \neq 0$, is equivalent to conducting inference over the section $\sigma_F : [(a, b, c)] \mapsto (s, G, V, c/s^2)$. Also, fixing $W = s$, for some $s \in \mathbb{R}_+$ is equivalent to adopting the section $\sigma_W : [(a, b, c)] \mapsto (\sqrt{c/s}, G, V, s)$.

Nevertheless, there is nothing wrong with using an unidentifiable statistical model as long as inference (or prediction) is conducted on identifiable quantities. An example, suggested to us by one of the reviewers, is that of a linear model with rank defficient design matrix:

even though some, or all, of the regression parameters are unidentifiable, the mean response can always be estimated uniquely from the data. We emphasize, however, that we would refer to the previous problem as an unidentifiability problem only when the design matrix is always rank defficient no matter what data we observe, such as in high-dimensional scenarios. In case some exploratory variables collected are perfectly (or highly) correlated only for a particular data set, we would refer to it as a problem of model estimability.

4. SIMULATION STUDY

In this section we present and discuss the results of several simulation studies in which the data cloning algorithm is employed to assess identifiability of the Gaussian dynamic linear model. Ideally, data cloning should not present any convergence issues when used for maximum likelihood estimation in an identifiable model. On the other hand, we would expect to see clear failures in all of the convergence measures available whenever data cloning is employed in an unidentifiable model. Before proceeding to the results there are some important points that need to be discussed so that the motivation behind the simulation study is clear.

Firstly, our main objective is to show how to use data cloning as a tool to assess identifiability statistical models. Our choice to illustrate the procedure through the dynamic linear model is justified by the fact that it is a latent variable model for which the identifying constraints are known. Hence, we can perfectly discern convergence issues due to model unidentifiability from those due to poor performance of the sampling algorithms.

Secondly, Lele *et al.* (2010) [13] advise using distinct prior distributions when performing data cloning. A simulation study for identifiability analysis should also take this into account as we need to be sure that there exists a number of clones for which the influence of any prior distribution vanishes. The more prior distributions we test the better is the study. As proposed by Lele *et al.* (2010) [13], we adopt three prior setups: uninformative, informative and disinformative. The first is simply as vague as possible, the second puts most of its probability mass around the true parameter value and the third is also an informative prior distribution, but most of its probability mass is allocated somewhat far from the true parameter values.

Lastly, we recommend employing both varying sample sizes and parameters. The former allows a view of the convergence of the maximum likelihood estimator, while the latter allows us to explore regions of the parameter space that may be of practical interest.

Data cloning is computationally demanding, although for a single data set setting the number of clones to a high value may not be a problem. For our purposes, we will be fitting the same model under multiple distinct settings and it is just not feasible to use a high number of clones. It does not matter, however, because we are not interested in finding the maximum likelihood estimate, but in gathering evidence of whether or not it can be found uniquely. Multiple starting values plus strong diagnostic measures of convergence allow us to gather solid evidence of model identifiability and issues thereof.

4.1. Simulation Parameters

We have chosen to simulate time series of sizes 25, 50, and 100. The simulated states are in Figure 1. To test distinct parameter vectors we vary the amount of noise added to the sample by the measurement process. This is done by considering the ratio between the variance of the process and measurement errors to be $W/V = 0.5$, 1, 2, and 10.

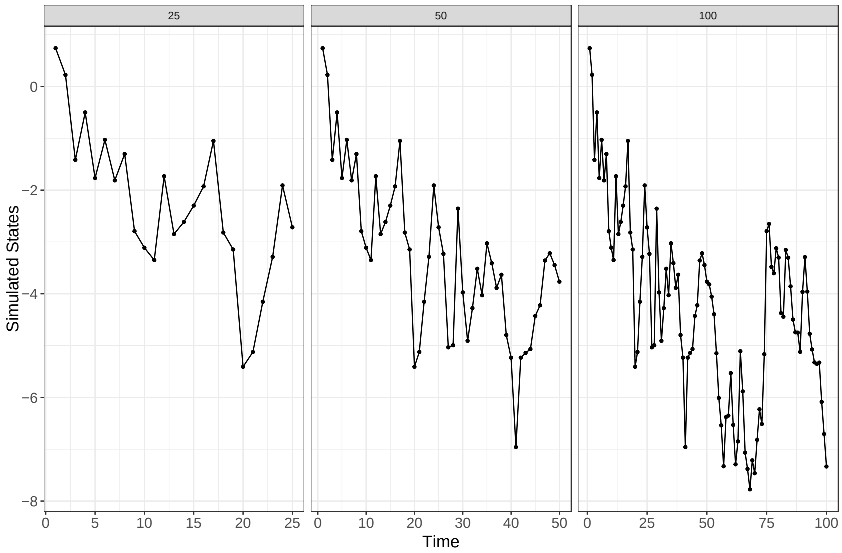


Figure 1: Simulated states for each adopted sample size.
The true parameter values used are displayed in Table 1.

We recall the parameter vector for the dynamic linear model under normality assumptions is $\theta = (F, G, V, W) \in \mathbb{R}^2 \times \mathbb{R}^2_+ = \Theta$. The four true parameter setups we use for simulating the time series are displayed in Table 1. Some illustrations of the effect of increasing the measurement error are available in Figure 2. We can see in the plots that increasing the measurement noise makes it harder to visually detect any patterns from the hidden states. We would expect that noisy and/or small datasets would be very challenging for data cloning since the likelihood function might not be well-behaved around the (possibly non-unique) maximum likelihood estimate. Nevertheless, similar situations would be challenging for most alternative estimation methods as well.

Table 1: True parameter values for simulation of the Gaussian dynamic linear model.

Setup	F	G	V	W
$W/V = 1/2$	1	1	2	1
$W/V = 1$	1	1	1	1
$W/V = 2$	1	1	0.5	1
$W/V = 10$	1	1	0.1	1

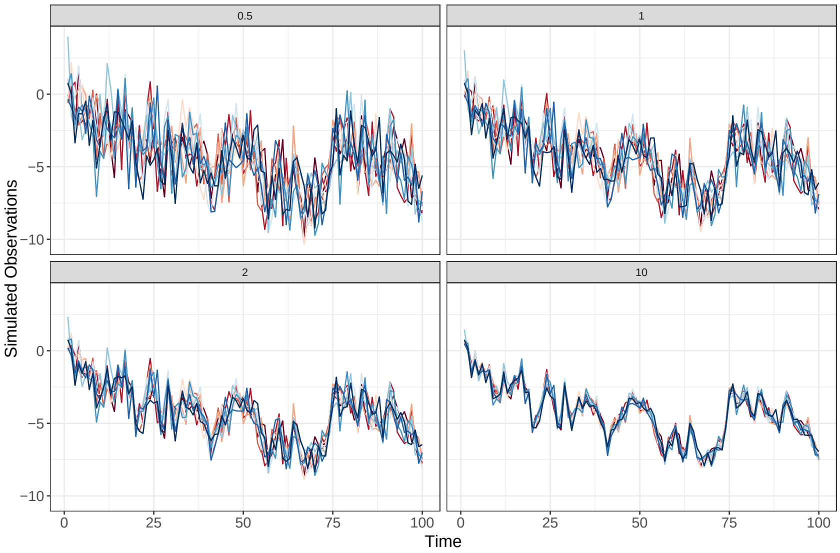


Figure 2: Plot of 10 simulated time series of length 100 arising from the hidden states in Figure 1. Each panel represents a signal-to-noise ratio W/V as presented in Table 1.

For the standard deviations $\sqrt{W}$ and $\sqrt{V}$, we adopted half-Cauchy prior distributions with scale equal to 10 as uninformative priors. This prior distribution is recommended by Gelman (2006) [7] for hierarchical models as an ideal alternative to the widely used gamma prior with small hyperparameters. Since our true parameter values are quite small compared to the tails of these prior distributions, we expect their added information to be insignificant compared to the data.

The F parameter receives a $\mathcal{N}(0, 10^4)$ in the uninformative setup. If we were to be faithful to the identifiability constraints required for this model, we would have to employ prior distributions which assign zero mass to negative values for F. However, the model is unidentifiable whether or not we restrict this parameter to the positive real line. Nonetheless, when performing some pre-tests for the simulation study, sampling F from priors on $\mathbb{R}_+$ resulted in running times up to three times longer than when using priors on $\mathbb{R}$.

The parameter G regulates the autoregressive behavior of the hidden states. The data we simulate assumes that these latent variables behave as a Gaussian random walk. We know that for values of G outside the open interval $(-1, 1)$ the latent process is non-stationary (Harvey 1989 [9]). Lele *et al.* (2007) [12] use a uniform prior distribution on the interval $(-1, 1)$ for this parameter, enforcing stationarity of the latent stochastic process. In our simulations, the data is clearly non-stationary. Therefore, we consider a $\mathcal{N}(0, 10^4)$ as the uninformative prior setup for G. This prior allows the process to present highly explosive growth behaviors if the data behaves as such. It is highly unlikely that this prior distribution would be used in a purely Bayesian framework, but data cloning allows us to use such largely uninformative prior distributions with ease.

In Table 2 we present the uninformative prior setup just discussed together with the informative and disinformative ones. The choice of the latter two, as previously discussed, simply aims to assign more probability mass closer or further (respectively) from the true parameter values. Notice that since the parameterization of the Gaussian distribution in

JAGS (Plummer 2007 [18]) is given by the mean and precision (*i.e.* the inverse of the variance), the prior distributions in Table 2 are provided in terms of the precision parameters $1/W$ and $1/V$ instead of W and V, respectively.

Table 2: Prior distributions chosen to represent the uninformative, informative and disinformative prior setups. The notation $\mathcal{HC}^{-2}$ indicates the distribution of the inverse of the square of a Half-Cauchy random variable, while λ denotes the signal-to-noise ratio W/V.

Prior Setup	F	G	$1/V$	$1/W$
Uninformative	$\mathcal{N}(0, 10^4)$	$\mathcal{N}(0, 10^4)$	$\mathcal{HC}^{-2}(0, 10)$	$\mathcal{HC}^{-2}(0, 10)$
Informative	$\mathcal{N}(1, 1)$	$\mathcal{N}(1, 1)$	$\Gamma(4^{-1}, (4\lambda)^{-1})$	$\Gamma(4^{-1}, 4^{-1})$
Disinformative	$\mathcal{N}(10, 5)$	$\mathcal{N}(-1, 1)$	$\Gamma(1, 5^{-1})$	$\Gamma(1, 5^{-1})$

4.2. Data Cloning Diagnostics

We begin our study of identifiability through the scaled maximum eigenvalue, $\lambda^S_{max,k}$, of the posterior covariance matrix, which should decay at about the theoretical rate of $1/k$, in which k denotes the number of clones used. In Figure 3 we display this measure for the case of the unidentifiable dynamic linear model. Since some of the resulting eigenvalues presented very high values, the graph with all of the observed measures is uninteresting due to the scaling of the ordinate axis.

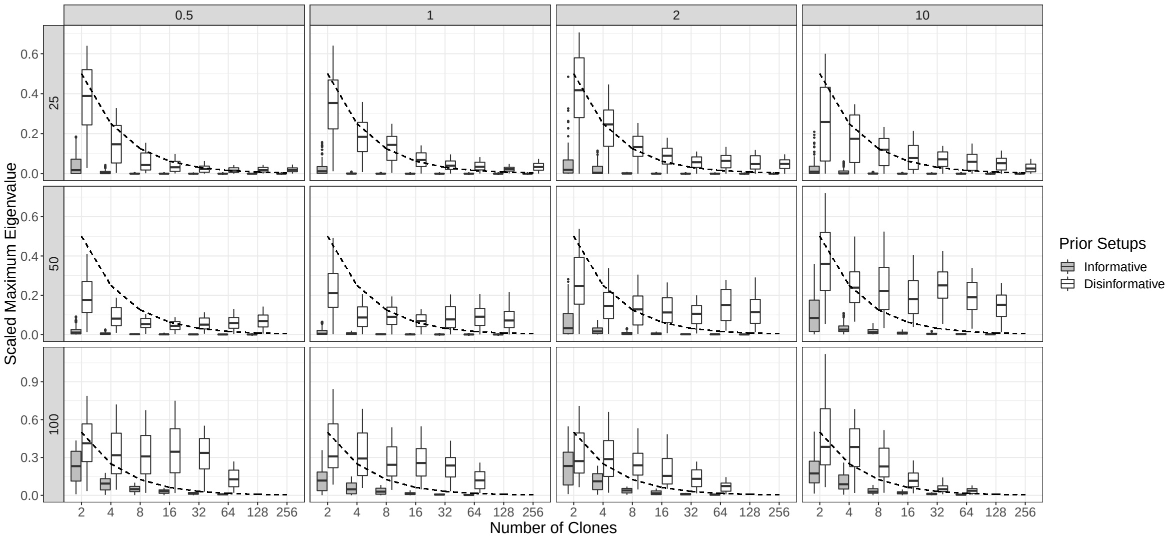


Figure 3: Box-plots of the scaled maximum eigenvalues of the posterior covariance matrix for the unidentifiable dynamic linear model. The dashed line represents the theoretical rate of convergence.

We have chosen thus to display in Figure 3 only the 50% smallest scaled eigenvalues obtained from the simulations using the informative and disinformative setup. The posteriors arising from uninformative prior distributions presented extremely high scaled eigenvalues even at the maximum number of clones for each sample size, which is a very strong evidence of identifiability issues. Nevertheless, the quantities in Figure 3 should still follow the theoretical convergence rate as long as the model is identifiable (which we know it is not).

It is easy to see in Figure 3 that $\lambda^S_{max,k}$ does not decay at the theoretical rate for the disinformative prior distributions. On the other hand, the informative prior setup seems to yield reasonable values for this diagnostic measure as the number of clones increases. However, we can see some odd behavior, particularly in the lower sample sizes with a low signal-to-noise ratio. Note that many $\lambda^S_{max,k}$ are already much below the theoretical rate of convergence in the first steps of the data cloning. This may be a sign that the variance of the unidentifiable parameters is being held low by the variance of the prior distribution.

Summing up, the scaled maximum eigenvalues observed from fitting an unidentifiable model resulted in undesired convergence properties and, in the case of the uninformative prior distributions, unreasonably high values for the eigenvalues of the posterior covariance matrix. This is in accordance with what we would expect from an unidentifiable model and indicates that data cloning is pointing towards identifiability issues when they are indeed present.

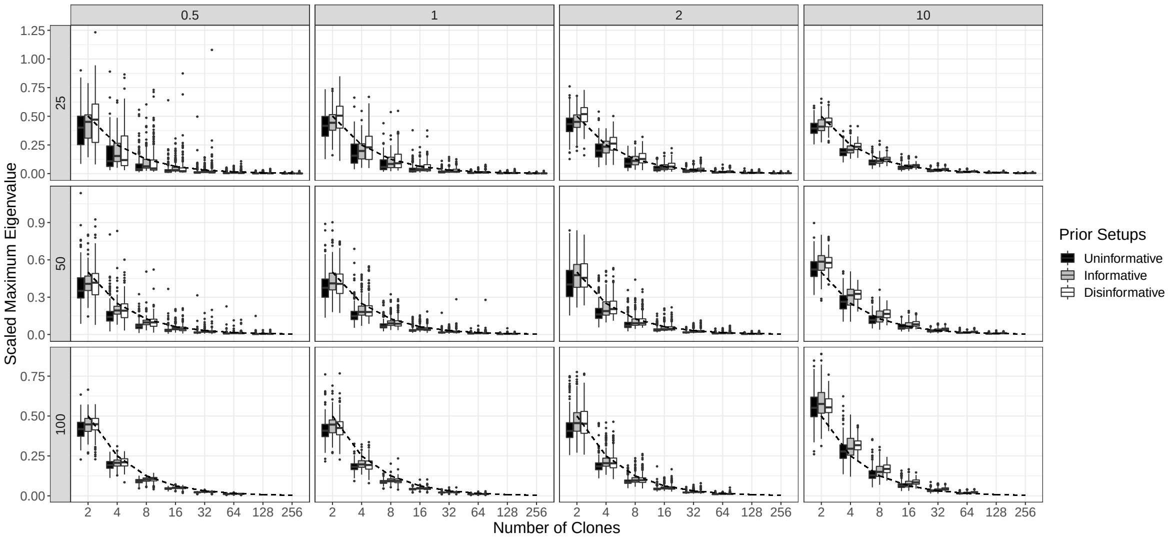


Figure 4: Box-plots of the scaled maximum eigenvalues of the posterior covariance matrix for the identifiable dynamic linear model in which we have set F to its true value. The dashed line represents the theoretical rate of convergence.

In Figure 4 we present the $\lambda^S_{max,k}$ drawn from an identifiable model in which we have fixed the parameter F to its true value. As the number of clones increases the theoretical rate of convergence is followed tightly by the quantities resulting from all three prior distributions. This is a clear indicator that the posterior distributions are becoming increasingly degenerated at the expected rate. Another important consideration is that all three prior setups differ only within small clone numbers, which indicates the influence of the prior distribution is indeed vanishing. Once again, therefore, the main diagnostic for data cloning has provided satisfactory results for the dynamic linear model. It exhibited no convergence issues when there are indeed no identifiability problems. This indicates constraining the parameter space by fixing the parameter F has led to a statistical model for which all parameters can be reliably estimated.

Table 3 presents the quartiles for $\lambda_{max,k}$ taken at the best-case scenario: the sample size is 100 and the number of clones k is 64. We can see the quartiles in the identifiable model are very close to each other, while the ones for the unidentifiable model are not. This is yet another strong evidence that there is still a considerable influence of the prior distributions on the joint posterior distribution of the parameters. There is an issue, however, because the quartiles for the maximum eigenvalues are quite small under the informative and disinformative setups even under unidentifiability. This indicates the posterior variance is decreasing as we increase the number of clones, which obviously should not happen since the model is unidentifiable. Nonetheless, for the uninformative setup, the maximum eigenvalues are of an order 10^4 higher than those from the informative setup.

Table 3: Quartiles for the maximum eigenvalues of the posterior covariance matrix of the parameters of the dynamic linear model. The values are for both the unidentifiable and identifiable models at the sample size of 100 and with number of clones equal to 64.

W/V	Prior	Identifiable			Unidentifiable		
		P_{25}	P_{50}	P_{75}	P_{25}	P_{50}	P_{75}
0.5	Uninformative	0.0035	0.0045	0.0055	32.1684	65.6898	140.4389
	Informative	0.0035	0.0044	0.0054	0.0048	0.0079	0.0115
	Disinformative	0.0034	0.0044	0.0054	0.0795	0.1493	0.3428
1	Uninformative	0.0019	0.0024	0.0032	40.4107	76.1869	132.6472
	Informative	0.0019	0.0025	0.0032	0.0024	0.0037	0.0066
	Disinformative	0.0019	0.0024	0.0032	0.1070	0.1670	0.2772
2	Uninformative	0.0015	0.0019	0.0024	38.6173	61.5385	113.3972
	Informative	0.0015	0.0019	0.0024	0.0017	0.0029	0.0048
	Disinformative	0.0015	0.0018	0.0023	0.0496	0.1045	0.2125
10	Uninformative	0.0012	0.0014	0.0016	29.9088	49.6416	72.4899
	Informative	0.0012	0.0014	0.0016	0.0014	0.0019	0.0029
	Disinformative	0.0012	0.0013	0.0016	0.0221	0.0454	0.0802

If we compare the quartiles over the three prior setups, it becomes clear there is still strong influence of the prior distribution even at 100 clones of the original dataset when the model is unidentifiable. However, under a single prior setup the conclusions related to the variance of the posterior distribution would differ considerably. In the informative prior setting, in particular, the quartiles of the maximum eigenvalues indicate no identifiability problems at all. The quartiles in this case are all small and reasonably close to each other, indicating the variance of the posterior distribution is small since the maximum eigenvalue provides an upper bound on the variances of the parameters. Therefore, for our purposes it would seem that the observed decay rate for the scaled maximum eigenvalue is the most appropriate of the two measures of degeneracy. By measuring the decay of $\lambda^S_{max,k}$, we were able to detect possible identifiability problems across all prior settings.

We can also check whether the posterior is reasonably close to a Gaussian distribution. This is done, as suggested by Lele *et al.* (2010) [13], through the measures presented in (2.8) and (2.9). In Table 4 we present the average of these diagnostics for each of the scenarios explored in the simulations. Overall, these diagnostic measures are greater, on average, in the unidentifiable than in the identifiable model. Furthermore, their averages seem to decrease in magnitude as the sample size increases, which is also to be expected.

Table 4: Diagnostic measures for normality of the samples from the posterior distribution of the parameters of the dynamic linear model.

Size	Prior Setup	Constraint	$W/V = 0.5$		$W/V = 1$		$W/V = 2$		$W/V = 10$	
			MSE	$\tilde{r}^2$	MSE	$\tilde{r}^2$	MSE	$\tilde{r}^2$	MSE	$\tilde{r}^2$
25	Uninformative	F = 1	3.173	0.088	1.718	0.068	1.590	0.065	1.891	0.078
		None	15.327	0.178	10.277	0.155	8.449	0.137	4.551	0.105
	Informative	F = 1	2.335	0.078	1.416	0.062	1.030	0.052	1.909	0.074
		None	13.662	0.175	10.891	0.159	7.148	0.130	3.779	0.091
	Disinformative	F = 1	2.430	0.083	1.398	0.062	0.989	0.052	1.194	0.054
		None	11.578	0.164	8.214	0.148	7.640	0.135	5.015	0.102
50	Uninformative	F = 1	3.699	0.092	1.600	0.057	0.624	0.039	0.393	0.030
		None	7.237	0.130	4.569	0.110	3.036	0.091	1.933	0.068
	Informative	F = 1	1.697	0.068	0.965	0.050	0.557	0.037	0.468	0.033
		None	5.467	0.112	3.947	0.101	2.704	0.086	1.301	0.055
	Disinformative	F = 1	2.556	0.083	1.229	0.055	0.678	0.040	0.409	0.031
		None	4.660	0.105	3.876	0.102	2.596	0.083	1.433	0.060
100	Uninformative	F = 1	0.591	0.037	0.364	0.028	0.253	0.021	0.224	0.020
		None	1.541	0.057	1.046	0.050	0.832	0.041	1.051	0.026
	Informative	F = 1	0.529	0.035	0.344	0.026	0.218	0.020	0.241	0.022
		None	1.425	0.059	0.915	0.047	0.578	0.034	0.345	0.024
	Disinformative	F = 1	0.623	0.038	0.394	0.029	0.255	0.022	0.177	0.017
		None	1.329	0.057	0.905	0.046	0.511	0.032	0.450	0.029

However, we would be hard-pressed to say these quantities have provided evidence of model unidentifiability (or identifiability). The values obtained under both scenarios, especially for the $\tilde{r}^2$, are satisfactory and also not very far apart from each other. For the MSE, in particular, it is to be expected that a model with one extra parameter, which is the case for the unidentifiable model, would require larger sample sizes or number of clones to achieve the same precision as a model with a lower number of parameters.

Furthermore, the quadratic form in Equation (2.7) may follow a chi-squared distribution even if the underlying probability distribution is not Gaussian. Azzalini and Valle (1996) [1], for example, show that this result holds for the quadratic form of p-variate Skew-Gaussian random variables. Therefore, these measures alone do not suffice to assess identifiability issues when using data cloning because it is possible these present reasonable values even when the posterior distribution is not Gaussian.

Therefore, although both the MSE and $\tilde{r}^2$ certainly serve their purpose when the interest is in obtaining the maximum likelihood estimates, for identifiability purposes they have not presented themselves as useful indicators of identifiability issues for this simple model and we do not advocate them to be heavily relied upon.

4.3. MCMC Diagnostics

We now move to the assessment of the quality of the posterior samples for maximum likelihood estimation. Firstly, we would like to know whether the Markov chains resulting from each of the three distinct prior setups, at the largest number of clones adopted, are targeting the same posterior distribution. For this task, we simply pretend we have run three independent chains under the same initial conditions, when in fact we used three distinct prior distributions. Within the Bayesian paradigm, individuals carrying distinct prior information about the same quantities need not arrive at the same inferential conclusions for finite samples, although asymptotic theory ensures this happens under some regularity conditions (Walker 1969 [26]). The differences in the posterior distributions for such individuals are even more pronounced whenever complex models and small and/or weakly informative data is at hand.

We emphasize yet again, however, that data cloning is a maximum likelihood estimation algorithm. Being so, it can not be affected by prior opinions. By collecting the chains from the three distinct prior setups, diagnostics such as the $\hat{R}$ can be used to check if the samples are being drawn from the same posterior distribution.

We provide in Table 5 the proportion of the simulations in which the $\hat{R}$ comparing three chains, one from each prior setup, is below the usual thresholds of 1.05 and 1.10. It is immediately clear that none of the simulations for the unidentifiable model have yielded joint posterior distributions for the parameter vector which are acceptably close enough from each other when starting from different prior distributions. For the identifiable model, the proportion starts low for the lower sample size of 25 and a low signal-to-noise ratio of 0.5 and reaches 1 for the sample size of 100. This is as to be expected, if not for the fact that many simulations do not yield close enough joint posterior distributions for some of the scenarios explored in this identifiable statistical model.

Table 5: Proportion of the simulations for which there is evidence that, starting from distinct prior distributions, the Markov chains are targeting the same posterior distribution. The values for the Gelman-Rubin diagnostic are computed at the highest number of clones for each sample size of the dynamic linear model.

Size	W/V	Identifiable		Unidentifiable	
		$\hat{R}<1.05$	$\hat{R}<1.10$	$\hat{R}<1.05$	$\hat{R}<1.10$
25	0.5	0.74	0.76	0	0
	1.0	0.90	0.91	0	0
	2.0	0.84	0.86	0	0
	10.0	0.44	0.51	0	0
50	0.5	0.51	0.55	0	0
	1.0	0.86	0.90	0	0
	2.0	0.98	0.99	0	0
	10.0	0.97	0.98	0	0
100	0.5	1.00	1.00	0	0
	1.0	1.00	1.00	0	0
	2.0	1.00	1.00	0	0
	10.0	0.94	0.98	0	0

It is possible that for some of the length 25 time series, the likelihood is very flat in a region around the maximum likelihood estimate. In this scenario, we would need a much larger number of clones to see that the Markov chains target the same joint posterior distribution. Nevertheless, the results from the Gelman-Rubin diagnostic point towards the clear failure of the data cloning when the model is unidentifiable, and present extremely promising results for the identifiable one. Were we unaware of the model's identifiability issues, these results, albeit not proof of unidentifiability, would surely lead us to reconsider the model structure.

There is one last point we need to verify when checking for model identifiability: the posterior means. The Gelman-Rubin diagnostic does not necessarily indicate that the posterior means, which are maximum likelihood estimates for a sufficiently large number of clones, are different. It can happen that the posterior means are very close to each other, while the posterior variance is not. Recall the Gelman-Rubin diagnostic indicates whether independent Markov chains happen to target the same posterior distribution. In other words, we could start from two or more distinct prior distributions and arrive at posterior distributions with the same mean but different variances. These are, thus, different posterior distributions and the Gelman-Rubin diagnostic will point towards convergence issues.

From Table 5 and Figure 3 we already know there is evidence of model unidentifiability. However, if the posterior means from distinct prior distributions were the same, we would have some evidence that we are able to reliably estimate the model parameters. Moreover, it might be the case that the unidentifiability issues found so far arise from poor tuning of the sampling algorithm.

In Table 6 we provide the average of the posterior means for the unidentifiable and identifiable Gaussian dynamic linear model with true signal-to-noise ratio $W/V = 1$ and sample size of 100. We also display the average effective sample size as a measure of the quality of the estimation of the posterior mean. Since we have drawn 1000 samples from each posterior distribution, we would want the effective sample size to be as close as possible to the total number of samples drawn. However, due to the very nature of MCMC algorithms it is expected that N_{eff} will be lower than the number of samples even if the model is identifiable. When using this measure, we are looking for parameters for which the N_{eff} is noticeably lower than both the total number of samples and the N_{eff} for other parameters.

If we focus on the parameters G and V, we can see that the averages of their posterior means are not considerably far apart from each other. However, we see that for the troublesome parameters W and F the posterior means under unidentifiability are heavily influenced by the choice of the prior distribution. Furthermore, the average effective sample size ranges from 1% to 5% of the total number of samples, indicating the chains for both parameters are highly autocorrelated. These results point towards extremely poor mixing of the Markov chains and, together with the diagnostics previously discussed, indicate clear failure of the model fitting procedure for the unidentifiable dynamic linear model.

However, the results for the identifiable functional F^2W, although not as good as that for the identifiable model, are still close to each other and also to the true value. This illustrates our previous comment that there is no harm in using unidentifiable statistical models, as long as the inferences are based on identifiable quantities. Hence, if we were interested in maximum likelihood estimation of any identifiable functional of $\boldsymbol{\theta}$, data cloning would yield good numerical approximations even if we had chosen to use the unidentifiable dynamic linear model.

Table 6: Averages of posterior means and effective sample sizes for the unidentifiable and identifiable Gaussian DLM with true signal-to-noise ratio $W/V = 1$ and sample size of 100. The true value for all parameters is 1 (see Table 1).

Parameter	Prior Setup	Identifiable		Unidentifiable	
		Mean	N_{eff}	Mean	N_{eff}
F	Uninformative	—	—	0.09	3.81
	Informative	—	—	1.89	3.69
	Disinformative	—	—	8.34	2.69
G	Uninformative	1.01	507.96	1.01	500.96
	Informative	1.01	504.54	1.01	503.79
	Disinformative	1.01	494.54	1.01	507.68
V	Uninformative	1.23	369.44	1.23	370.97
	Informative	1.23	284.53	1.23	282.88
	Disinformative	1.23	291.20	1.23	274.98
W	Uninformative	0.76	302.04	110.5	16.81
	Informative	0.76	208.40	0.22	16.87
	Disinformative	0.76	213.76	0.01	8.37
F^2W	Uninformative	0.76	302.04	0.76	311.37
	Informative	0.76	208.40	0.77	210.41
	Disinformative	0.76	213.76	0.80	218.34

Furthermore, as expected, the behavior within the identifiable model, in which we set $F = 1$, is exactly what we would want to see if we were using data cloning for maximum likelihood estimation. The posterior means, which we would like to call maximum likelihood estimates, seem to be independent of the choice of the prior distribution at the largest number of clones. The effective sample sizes are all satisfactory and indicate that the chains may be adequately exploring the posterior distribution. Given that the $\hat{R}$ diagnostics in Table 5 revealed the posterior distribution seems to be independent of the choice of prior distribution in the identifiable model at the highest number of clones adopted, we could gather the samples from all three chains to increase the effective sample size even further. Doing so would reduce the Monte Carlo variance of the numerical approximation to the maximum likelihood estimate and, consequently, improve the estimation of the Fisher information matrix.

5. FINAL COMMENTS

In this paper, we have explored the capabilities of data cloning as a tool for identifiability analysis of statistical models through a simulation study with the Gaussian dynamic linear model. Through an example, we have shown how such a simulation study can be planned and performed to gather evidence of possible model unidentifiability and how to interpret the most relevant diagnostic measures for the data cloning algorithm.

We found the bounds on the posterior covariance matrix of the parameters, its maximum eigenvalue $\lambda_{max,k}$, to be a good indicator of model identifiability. Its scaled version, $\lambda^{S}_{max,k}$, also yielded strong results since it exhibited convergence problems when they existed, while also indicating proper convergence of the algorithm in the identifiable model. The measures of normality did not present results as interesting as did $\lambda^{S}_{max,k}$. Both $\tilde{r}^2$ and MSE, suggested by Lele *et al.* (2010) [13], did not show significantly distinct values under either the identifiable or unidentifiable model. If we also consider the fact that these diagnostics can be satisfactory for quadratic forms of distributions other than the Gaussian, then our conclusion is that they are unreliable for identifiability analysis. However, for the purpose of maximum likelihood estimation using data cloning they must not be overlooked.

By exploiting distinct prior distributions, we were able to find clear parameter identifiability issues through the Gelman-Rubin diagnostic $\hat{R}$. Together with the data cloning diagnostics and the posterior means of the parameters, the evidence gathered through the diagnostics led us to the correct conclusion that the unconstrained Gaussian dynamic linear model is unidentifiable. Nonetheless, it also allowed us to conclude that fixing the parameter F to a known constant was enough to ensure statistical identifiability.

Overall, we find the results from the simulation study are very promising and indicate data cloning can (and should) be used as a tool for identifiability analysis, although some care must be taken as to how to do it properly. We emphasize here, once more, the importance of employing distinct prior distributions, parameter values and sample sizes in the simulation study to ensure that the evidence of identifiability, or lack of it, are consistent across an as wide as possible range of real possibilities.

There are also models for which MCMC algorithms either perform poorly or are simply too computationally demanding, for example those involving stochastic partial differential equations. As mentioned by one of the reviewers, the integrated nested Laplace approximation (Rue *et al.* 2009 [20]), or INLA for short, employs deterministic approximations to posterior distributions and has been paired up with data cloning for maximum likelihood estimation (see Baghishani *et al.* 2012 [2]). Although not as widely applicable as MCMC algorithms, INLA has been shown to be both extremely fast and precise when compared to the former. Furthermore, we are unaware of any studies on the usage of INLA and data cloning specifically for identifiability analysis and this may be an interesting venture within this topic.

Finally, we must also emphasize that identifiability cannot in general be proved based on simulation studies. After all, identifiability is a structural property of statistical models and it is impossible to exhaust the possible combinations of parameters and infinite sample sizes in a simulation study. Therefore, we are restricted to finite samples and a few points of interest in the parameter space. This implies that, at best, we can gather evidence of local identifiability in a region of practical interest of the postulated parameter space. The enterprise is nevertheless worth the effort since any evidence even of local unidentifiability in a statistical model can indicate undesired behavior of inferential procedures.

ACKNOWLEDGMENTS

We would like to thank the Brazilian research funding agencies CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for financial support. We also thank the valuable comments by the two reviewers and the Editor, which substantially contributed to the improvement of the paper.

REFERENCES

[1] Azzalini, A. and Valle, A.D. (1996). The multivariate skew-normal distribution, *Biometrika*, **83**(4), 715–726.

[2] Baghishani, H.; Rue, H. and Mohammadzadeh, M. (2012). On a hybrid data cloning method and its application in generalized linear mixed models, *Statistics and Computing*, **22**(2), 597–613.

[3] Bellu, G.; Saccomani, M.P.; Audoly, S. and D'Angiò, L. (2007). DAISY: A new software tool to test global identifiability of biological physiological systems, *Computer methods and Programs in Biomedicine*, **88**(1), 52–61.

[4] Campbell, D. and Lele, S. (2014). An ANOVA test for parameter estimability using data cloning with application to statistical inference for dynamic systems, *Computational Statistics and Data Analysis*, **70**, 257–267.

[5] Carpenter, B.; Gelman, A.; Hoffman, M.D.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M.; Guo, J.; Li, P. and Riddel, A. (2017). Stan: a probabilistic programming language, *Journal of Statistical Software*, **76**(1).

[6] Florens, J.P. and Simoni, A. (2011). Bayesian identification and partial identification (Unpublished Paper). Available at:
`https://cdn.uclouvain.be/public/Exports%20reddot/core/documents/Simoni.pdf`

[7] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis*, **1**(3), 515–533.

[8] Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences, *Statistical Science*, **7**(4), 457–472.

[9] Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.

[10] Horn, R.A. and Johnson, C.R. (2012). *Matrix Analysis*, Cambridge University Press, Cambridge.

[11] Koopmans, T.C. and Reiersol, O. (1950). The identification of structural characteristics, *The Annals of Mathematical Statistics*, **21**(2), 165–181.

[12] Lele, S.R.; Dennis, B. and Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Mote Carlo methods, *Ecology Letters*, **10**(7), 551–563.

[13] Lele, S.R.; Nadeem, K. and Scmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning, *Journal of the American Statistical Association*, **105**, 1617–1625.

[14] Lindley, D.V. (1971). *Bayesian Statistics: A Review*, Society for Industrial and Applied Mathematics, Philadelphia.

[15] Paulino, C.D.M. and Pereira, C.A.B. (1994). On identifiability of parametric statistical models, *Journal of the Italian Statistical Society*, **3**(1), 125–151.

[16] Peacock, S.J.; Krkošek, M.; Lewis, M.A. and Lele, S. (2017). Study design and parameter estimability for spatial and temporal ecological models, *Ecology and Evolution*, **7**(2), 762–770.

[17] Picci, G. (1977). Some connections between the theory of sufficient statistics and the identifiability problem, *SIAM Journal on Applied Mathematics*, **33**(3), 383–398.

[18] Plummer, M. (2017). *JAGS Version 4.3.0 User Manual.*

[19] R Core Team (2020). *R: A Language and Environment for Statistical Computing*, Vienna, Austria.

[20] Rue, H.; Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, *Journal of the Royal Statistical Society, Series B*, **71**(2), 319–392.

[21] San Martín, E.; González, J. and Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3PL model, *Psychometrika*, **80**(2), 450–467.

[22] San Martín, E. (2018). Identifiability of structural characteristics: how relevant is it for the Bayesian approach?, *Brazilian Journal of Probability and Statistics*, **32**(2), 346–373.

[23] Solymos, P. (2010). dclone: data cloning in R, *The R Journal*, **2**(2), 29–37.

[24] Turkman A.; Paulino, C.D. and Müller, P. (2019). *Computational Bayesian Statistics – An Introduction*, Cambridge University Press, Cambridge.

[25] Villaverde, A.F.; Tsiantis, N. and Banga, J.R. (2019). Full observability and estimation of unknown inputs, states and parameters of nonlinear biological models, *Journal of the Royal Society Interface*, **16**.

[26] Walker, A.M. (1969). On the asymptotic behaviour of posterior distributions, *Journal of the Royal Statistical Society, Series B*, **31**(1), 80–88.

REVSTAT – Statistical Journal
Volume 22, Number 1, January 2024, 111–132
https://doi.org/10.57805/revstat.v22i1.455

Analysis of Antibody Data Using Skew-Normal and Skew-t Mixture Models

Authors: Tiago Dias Domingues ✉
– CEAUL, Faculdade de Ciências Universidade de Lisboa,
Portugal
tmdomingues@fc.ul.pt

Helena Mouriño
– CMAFcIO, Faculdade de Ciências, Universidade de Lisboa,
Portugal
mhnunes@fc.ul.pt

Nuno Sepúlveda
– Faculty of Mathematics and Information Science, Warsaw University of Technology,
Poland
– CEAUL, Faculdade de Ciências, Universidade de Lisboa,
Portugal
N.Sepulveda@mini.pw.edu.pl

Received: December 2020 Revised: March 2022 Accepted: March 2022

Abstract:

- Gaussian mixture models, which assume a Normal distribution for each component, are popular in antibody (or serological) data analysis to help determining antibody-positive and antibody-negative individuals. In this work, we advocate using finite mixture models based on Skew-Normal and Skew-t distributions for serological data analysis. These flexible mixing distributions have the advantage of describing right and left asymmetry often observed in the distributions of known antibody-negative and antibody-positive individuals, respectively. We illustrate the application of these alternative mixture models in a data set on the role of human herpesviruses in the Myalgic Encephalomyelitis/Chronic Fatigue Syndrome.

Keywords:

- *finite mixture models; Skew-Normal; Skew-t; seropositivity.*

AMS Subject Classification:

- 62H30, 62P10.

✉ Corresponding author.

1. INTRODUCTION

Antibodies are proteins produced by B cells upon recognition of an antigen derived from an infectious agent. In general, they contribute to microbial clearance and, if maintained in the body over time, they translate into a quicker and more efficient immune response upon repeated exposure to the same infection. In turn, autoantibodies bind to antigens from the body and they are usually present in autoimmunity diseases, such as multiple sclerosis and rheumatoid arthritis.

In routine laboratories, antibodies (or autoantibodies) against a specific antigen are quantified by the enzymatic-linked immunosorbent assays (ELISA) using serum samples. The readout of these assays is a light intensity, also known as optical density, which is converted into a concentration or a titre using a calibration curve of known antibody concentrations. In practice, these assays are easily standardized, widely available, and ideal for high-throughput analysis of antibodies against a single antigen [1]. Such advantages make them suitable for large-scale serological surveys where one aims to estimate the prevalence of exposure to a given pathogen in the population [1, 2, 3].

With the development of high-throughput technologies, antibody quantification is shifting from the ELISA to microarray, luminex, or cytometry bead assays, where many antibodies can be evaluated in the same serum sample. However, these technologies are still being optimized before their wide use.

Antibody (or serological) statistical analysis of antibody (or serological) data often assumes the existence of multiple latent populations each one representing a distinct level of exposure to a given antigen. This basic assumption calls for the use of finite mixture models. In general, these models can be more or less complex, depending on the number of mixing distributions used to describe the data [4]. In routine serological applications, one assumes a model with only two latent populations: seronegative and seropositive individuals or, equivalently, antibody-negative and antibody-positive individuals [5, 6, 7]. Models comprising more than two serological populations are also used in practice [8, 9, 10, 11, 12], but their interpretation is not straightforward [13].

A common choice for the mixing distribution is the Lognormal distribution in the original scale of the measurements or, equivalently, the Normal distribution after applying the logarithmic transformation to the data [6, 8]. Gamma and Weibull are other choices among textbook probability distributions [7, 11].

Less-trivial mixture models can be also used in the analysis. For example, a mixture of two truncated Normal distributions was used to describe data where observations could fall below the lower limit of detection or above the upper limit of detection of the assay [9]. Another alternative model was the mixture of a Normal distribution and a combination of half-Normal distributions for the seronegative and seropositive populations, respectively [5]. The rationale behind this model is that antibody levels decrease over time and, therefore, the seropositive populations should have left-skewed distributions [8]. Similarly, seronegative populations should have right-skewed distribution due to the detection of non-specific antibodies at lower concentrations of the target antibodies. Notwithstanding the suitability of these alternative models to tackle specific characteristics of serological data, none of the above models shows sufficient flexibility in terms of skewness and flatness of each mixing distribution that could be used serological data analysis and its automation in the context of high-throughput data.

We then propose using finite mixture models based on Skew-Normal and Skew-t distributions scale in routine serological data analysis. These alternative families of distributions are highly flexible due to three parameters that control the location, the scale, and the skewness of the resulting distribution. In the case of the Skew-t distribution, further flexibility can be achieved by an additional parameter that controls the weight of tails. These distributions also have the advantage of including the Normal distribution, the Generalized Student's t-distribution, and its skewed version as special cases [14]. As an example of application, we use these models to analyse a data set of 6 antibody responses to herpesviruses in the context of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) [15].

2. DATA UNDER ANALYSIS

ME/CFS is a multifactorial disease whose patients experience persistent fatigue that cannot be alleviated by rest, or they suffer from post-exertional malaise upon minimal physical and mental activity [16]. The cause of the disease remains unknown, but it is often linked to infections by herpesviruses.

The data set under analysis is part of the United Kingdom ME/CFS biobank, and it was published in a recent study with the aim of investigating the immunological component of the disease [15]. In the data set, there is a total of 406 individuals divided into three main groups: healthy controls (HC, $n = 107$; 26.4%), patients with ME/CFS ($n = 250$; 61.8%), and patients with multiple sclerosis (MS, $n = 49$; 12.1%). The group of patients with ME/CFS was subdivided into 196 patients with mild or moderate symptoms (ME-M) and 54 severely affected patients who are home- or even bed-bound (ME-S).

The data set comprises six serological antibody concentrations measured by commercial ELISA kits and related to the following common herpesviruses: human cytomegalovirus, CMV; Epstein-Barr virus, EBV; human herpesvirus-6, HHV-6; types 1 and 2 herpes simplex viruses, HSV-1 and HSV-2, respectively; and varicella-zoster virus, VZV. Note that the tested antibodies against EBV were specific to the viral-capsid antigen.

The concentration of the antibodies was expressed in arbitrary units per ml (U/ml). According to the kit manufacturers, individuals with antibody concentration ≤ 8 U/ml or ≥ 12 U/ml should be classified as seronegative or seropositive, respectively, for all antibodies except for the one against HHV-6. For antibodies against HHV-6, seronegative and seropositivity should be defined as ≤ 10.5 U/ml or ≥ 12.5 U/ml, respectively. Samples with concentrations between the above limits were considered equivocal.

3. STATISTICAL ANALYSIS OF SEROLOGICAL DATA

3.1. Finite mixture models

Let $G_1, ..., G_g$ be the partition from a superpopulation G (sample space) and $\pi_1, ..., \pi_g$ the probabilities of sampling an individual belonging to each latent population (with the usual restriction of $\sum_{k=1}^{g} \pi_k = 1$ and $0 \le \pi_k \le 1$). A random variable Z is a finite mixture of independent random variables $Z_1, Z_2, ..., Z_g$ if the probability density function (pdf) of Z is given by

$$f(z) \;=\; \sum_{k=1}^{g} \pi_k \, f_{Z_k}(z; \boldsymbol{\theta}_k)\,, \tag{3.1}$$

where $f_{Z_k}(z; \boldsymbol{\theta}_k)$ is the mixing probability density function (pdf) of Z_k associated with the k-th latent population and parameterized by the vector $\boldsymbol{\theta}_k = \{\theta_1, ..., \theta_g\}$.

A common choice for the mixing distribution in the serological analysis is the Normal distribution which is symmetric around the mean, and it is a mesokurtic distribution (with kurtosis of 3 irrespective of the mean and standard deviation). Alternatively, the Generalized Student's t can be used as the mixing distribution because it has heavier tails than the Normal distribution. However, data from malaria seroepidemiological studies show long tails and marked right asymmetry in each latent population even after applying a logarithmic transformation [7]. In such cases, one aims to incorporate asymmetry and heavy tails in the finite mixture modelling. This is the purpose of using the Skew-Normal and Skew-t as mixing distributions [17, 18]. These alternative distributions are members of the so-called scale mixtures of Skew-Normal (SMSN) distributions [14]. This class of probability distributions is defined as follows.

Let Z_k be a random variable following a SMSN distribution with μ_k, σ_k^2, and α_k as the location, scale, and skewness parameters, respectively, and $H_k(\,\cdot\,; \boldsymbol{v}_k)$ as the mixing distribution parameterized by θ_k. Then, it can be written as

$$Z_k \;=\; \mu_k + \frac{W_k}{\sqrt{U_k}}\,, \tag{3.2}$$

where U_k is a random variable with distribution function $H_k(\,\cdot\,; \boldsymbol{v}_k)$ and $W_k \sim \mathcal{SN}(0, \sigma_k^2, \alpha_k)$, and W_k and U_k are two independent random variables [14]. See Appendix A in the Supplementary Material for additional theoretical discussion about this class of distributions.

3.1.1. Skew-Normal as a mixing distribution

Let W_k be a random variable with a Skew-Normal distribution with location parameter μ_k, scale parameter σ_k^2 and skewness parameter α_k (denoted as $W_k \sim \mathcal{SN}(\mu_k, \sigma_k^2, \alpha_k)$). The corresponding pdf is given by

$$\begin{aligned} f_{W_k}(w) \;&=\; 2\,\frac{1}{\sqrt{2\pi}\sigma_k}\; e^{-\frac{(w-\mu_k)^2}{2\sigma_k^2}} \;\times \int_{-\infty}^{\alpha_k \frac{(w-\mu_k)}{\sigma_k}} \frac{1}{\sqrt{2\pi}}\; e^{-\frac{x^2}{2}}\, dx \\ &=\; 2\phi\!\left(\frac{w-\mu_k}{\sigma_k}\right)\Phi\!\left(\frac{\alpha_k(w-\mu_k)}{\sigma_k}\right), \qquad w, \mu_k, \alpha_k \in \mathbb{R}\,, \quad \sigma_k \in \mathbb{R}^+, \end{aligned}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denotes the pdf and the cumulative distribution function (cdf) of the standard Normal distribution, respectively [14, 19, 20].

When $\alpha_k = 0$, the above formula recreates the pdf of the Normal distribution. In this case, the Fisher information matrix of the Skew-Normal is singular, thus, influencing the asymptotic properties of the maximum likelihood estimators in the vicinity of zero. A detailed discussion about this topic can be found elsewhere [21, 22, 23].

When $\alpha_k = \infty$, the limiting distribution is the half-Normal distribution [21]. In this case, the location parameter μ_k determines the support of the distribution. This property makes the Skew-Normal distribution particularly useful to model data with a lower or an upper bound.

Note that the Skew-Normal distribution can be obtained from (3.2) when $H(\cdot\,;\theta_k)$ is a degenerate mixing distribution. Alternatively, the Skew-Normal distribution is a special case of the skew Normal-Normal [24] and the skew Student-t-Normal distribution [25]. These two flexible distributions are members of the so-called skew scale mixtures of Normal distributions [25]. This class of probability distributions differs from the class of SMSN in terms of the respective stochastic representation and dependence between skewness and kurtosis coefficients; see Ferreira *et al.* [25] for more details. In theory, distributions from this class can be seen as alternative candidates to the SMSN ones for the choice of mixing distributions. However, in practice, there are no estimation algorithms available for the context of finite mixture models.

3.1.2. Skew-t as a mixing distribution

Let Z_k be a random variable that follows a Skew-t distribution with location parameter μ_k, scale parameter σ_k^2, skewness parameter α_k, and v_k degrees of freedom. Then, its pdf is given by

$$f_{Z_k}(z) \;=\; \frac{2}{\sigma_k}\; t(d; v_k)\; T\!\left(A\sqrt{\frac{v_k+1}{d+v_k}}\,;\; v_k+1\right),$$

where $d = (z-\mu_k)/\sigma_k$, $A = \alpha_k(z-\mu_k)/\sigma_k$, $t(\cdot\,;v)$ and $T(\cdot\,;v)$ are the pdf and cdf of the standard Student's t distribution with v degrees of freedom, respectively [14].

When $\alpha_k = 0$, the above distribution converts to the Generalized Student's t-distribution with location parameter μ_k, scale parameter σ_k and v_k degrees of freedom. When $v_k = 1$, one obtained the Skew-Cauchy distribution. Finally, when the degrees of freedom v_k tend to infinity, one obtains the Skew-Normal as the limiting distribution [14, 19, 20].

Note that the Skew-t distribution can be derived from (3.2) when U_k is a Gamma distribution with parameters $\alpha = v_k/2$ and $\beta = v_k/2$ [14]. As an additional note, Theodossiou [26] introduced the skew generalized t distribution with five parameters: location, scale, skewness, and two shape parameters. It can be derived from a ratio between a generalized gamma distribution and an appropriate transformation of a skew exponential power distribution, but it cannot be expressed as an SMSN distribution. As such, this alternative distribution has different skewness and kurtosis when compared to the above Skew-t distribution. See Arslan and Genç [27] and the references therein for more information.

3.2. Estimation of Skew-Normal and Skew-t mixture models

Let $X_1, ..., X_n$ be a random sample that represents the measured antibody levels in n individuals. In general, it is difficult to determine the maximum likelihood (ML) estimates of a finite mixture model by direct maximization of the log-likelihood function. To overcome this problem, one can use the Expectation-Maximization (EM) algorithm given that the latent serological status of each individual is unknown and, thus, serological data are incomplete in that sense.

An EM-type algorithm for estimating SMSN mixture models is fully described elsewhere [14]. Briefly, the E-step is the same as in Gaussian mixture models, which has been largely studied in the literature [14, 17, 28]. Replacing the classical M-step with a sequence of conditional maximization steps (CM-steps), one obtains closed form expressions for the parameter estimates and the Fisher's information matrix [25]. To ensure convergence to the global maximum of the likelihood function, one should initiate the algorithm with different values for the parameter estimates. The final parameter estimates should be the ones that provide the highest value of the log-likelihood among all the different runs of the algorithm. Note that, for Gaussian mixture models, there are modifications of the classical EM algorithm that do not require the use of initial conditions and jointly determine the optimal number of the mixture components [29, 30]. These characteristics of the proposed algorithms reduces the computational time of analyses including a large number of screened antibodies. However, similar modifications remain to be done for the context of SMSN mixture models.

To obtain confidence intervals (CIs) for the model parameters, one can simply use the Wald's CIs. In the case of skewness parameters α_k's, the respective CIs are given by

$$\hat{\alpha}_k \pm \Phi^{-1}_{(\gamma+1)/2}\, se(\hat{\alpha}_k)\,,$$

where $\hat{\alpha}_k$ is the ML estimate of α_k, γ is the confidence level, and $\Phi^{-1}_{(\gamma+1)/2}$ is the probit function evaluated at $(\gamma+1)/2$. However, according to Zeller *et al.* [40], Wald's intervals for these parameters tend to inflate the underlying uncertainty in the case of a single Skew-Normal distribution. Such inflation can be derived from a poor quadratic approximation of the profile likelihood (PL) taken as a function of α [41]; see Pawitan for a more general discussion [42]. In addition, the PL is expected to show an inflexion point at $\alpha = 0$, which affects the asymptotic normal approximation for the distribution of the respective ML estimator [21]. Similar argument is expected to hold when estimating the same parameter of a single Skew-t distribution. In these cases, the PL can be used to determine a more accurate CI for α:

$$2\left\{l(\hat{\alpha}) - l(\alpha)\right\} < \chi^2_{\gamma,1}\,,$$

where $\hat{\alpha}$ is the ML estimate of α, $l(\alpha)$ is the PL taken as a function of α, and $\chi^2_{\gamma,1}$ is the γ quantile of the χ^2 distribution with one degree of freedom. See Zeller *et al.* [40] and Montenegro *et al.* [41] for the application of this CI to non-serological data. In the context of SMSN finite models, the PL approach is not a viable solution due to the presence of different subpopulations with their own skewness parameter.

3.3. Model selection

Model selection aims to determine the best mixture model for the data in terms of the number of the constituent components, g, and the respective mixing distributions. With this purpose, one can use information criteria based on penalized forms of the log-likelihood function: the Akaike's Information Criterion (AIC) [31], the Integrated Complete Likelihood (ICL) [32], the Bayesian Information Criterion (BIC) [33] and its modified versions [34, 35].

However, AIC tends to overestimate g in Gaussian mixture models even when n is very large [36]. This overestimation can be explained by a weak penalization of AIC to complex models with spurious mixing components that can arise from unbounded likelihood functions or from the presence of multiple local maximizers of the log-likelihood function [37]. In the case of serological applications, the overestimation of g compromises interpretability of a mixture model with more than 2 components [13]. In contrast, ICL tends to underestimate g and it is more adequate when the mixture components are well separated [32]. Finally, In this regard, BIC offers a higher penalization of models with a higher components when compared to AIC. However, the regularity conditions for using BIC do not necessarily hold in analysing finite mixture models [33, 35]. However, simulation studies suggested a satisfactory performance of this criterion (or its modified versions) in determining the true number of Gaussian mixture components [29, 35]. Therefore, at this stage, BIC seems the recommended measure when comparing different mixture models. Simulation studies should be conducted in the future to confirm this recommendation.

To complement the analysis based on information criteria, one can also carry out the likelihood ratio test (LRT) for determining the optimal number of mixture components, g [4]. However, the regularity conditions for the asymptotic χ^2 approximation of the test statistic are not met in finite mixture models, because the null hypothesis is specified in the boundary of the parameter space [4]. To overcome this problem, one can use a parametric Bootstrap approach to estimate the p-value of this non-standard LRT [38, 39], as described below.

Consider the test for confronting $H_0 \colon g = g_0$ versus $H_1 \colon g = g_1$ where $g_0 < g_1$. Let $\boldsymbol{\psi}_0$ and $\boldsymbol{\psi}_1$ be the parameter vectors of the mixture models under H_0 and H_1, respectively; $\boldsymbol{x} = (x_1, ..., x_n)$ the observed data and $T(\boldsymbol{x}; \boldsymbol{\psi}_0, \boldsymbol{\psi}_1)$ the test statistic of LRT. The bootstrap approach is given by the following algorithm [39]:

1. Use the EM algorithm to estimate the $\boldsymbol{\psi}_0$ and $\boldsymbol{\psi}_1$ estimates under the H_0 and H_1 hypotheses, respectively. Calculate $T(\boldsymbol{x}; \hat{\boldsymbol{\psi}}_0, \hat{\boldsymbol{\psi}}_1)$;
2. Simulate $N = 10,000$ independent samples $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_n^*$ using the mixture model under H_0 and parameterized by $\hat{\boldsymbol{\psi}}_0$;
3. For each bootstrap sample i, calculate $T(\boldsymbol{x}_i^*; \hat{\boldsymbol{\psi}}_{0_i}, \hat{\boldsymbol{\psi}}_{1_i})$, where $\hat{\boldsymbol{\psi}}_{0_i}$ and $\hat{\boldsymbol{\psi}}_{1_i}$ are the estimated parameter vectors for the bootstrap sample i under the H_0 and H_1 hypotheses, respectively;
4. Estimate the p-value as $\frac{1}{N}\sum_{i=1}^{N} I\Big\{T(\boldsymbol{x}_i^*; \hat{\boldsymbol{\psi}}_{0_i}, \hat{\boldsymbol{\psi}}_{1_i}) > T(\boldsymbol{x}; \hat{\boldsymbol{\psi}}_0, \hat{\boldsymbol{\psi}}_1)\Big\}$, where $I\{\cdot\}$ is the indicator function.

Finally, the estimated models should be assessed in terms of their goodness of fit. For a matter of simplicity, one can simply used the Pearson's χ^2 test [43, 44]. To apply this test, one can divide the data into bins defined by the respective 5%-quantiles or deciles. Alternatively, one can use the Kolmogorov-Smirnov, Anderson-Darling, and Walton's test among others [45].

3.4. Estimation of seroprevalence

After determining the best finite mixture model for the data, the next step of the analysis is usually to estimate the seroprevalence, that is, the prevalence of antibody-positive individuals in the population (or, the probability of an individual being antibody-positive).

Seropositivity is traditionally defined by a cutoff, denoted by c, in the respective antibody distribution above which individuals would be considered seropositive. In the context of finite mixture models, cutoff determination requires the interpretation of each latent population in terms of seronegativity and seropositivity. To do that, one typically assumes the seronegative population as the one with lowest average while the remaining components are interpreted as different levels of seropositivity upon recurrent infections. In this scenario, the seropositivity of i-th individual can be seen as resulting from a Bernoulli random variable $Y_i \sim \mathrm{Ber}(p)$ where $p = P[X_i \geq c]$ and X_i $(i = 1, ..., n)$ represents the random variable representing the underlying antibody concentration. The probability p is also called seroprevalence and it embodies the probability of exposed individuals to a given antigen in the population. According to the maximum likelihood method, seroprevalence can be estimated as the proportion of seropositive individuals in the sample. Therefore, different estimates for the seroprevalence can be obtained according to the methods used to determine the cutoff.

In this work, we consider the following three different methods for determining the seropositivity cutoff:

- **Method 1:** It is based on the 99.9%-quantile associated with the estimated seronegative population. This method is the most popular in sero-epidemiology [13, 46]. It is often called as the 3σ rule, because the 99.9%-quantile is given by the mean plus 3 times the standard deviation of a normally distributed seronegative population;

- **Method 2:** It relies on the minimum of the density mixture functions. In the case of two latent populations, the cutoff corresponds to the absolute minimum, and in the case of three or more latent populations the cutoff corresponds to the lowest relative minimum. This point can be calculated using the Dekker's algorithm [47]. It should be noted that the minimum of the mixing function is not expected to coincide with the point of intersection of the probability densities of each individual subpopulation;

- **Method 3:** It imposes a threshold in the the so-called conditional classification curves [13]. Under the assumption that all components but the first one refer to seropositive individuals, the conditional classification curve of seropositive individuals given the antibody level x is defined as

$$p_{+|x} = \frac{\sum_{k=2}^{g} \pi_k \, f_k(x; \boldsymbol{\theta}_k)}{\sum_{k=1}^{g} \pi_k \, f_k(x; \boldsymbol{\theta}_k)} \,.$$

In turn, the classification curve of seronegative individuals is given by

$$p_{-|x} = 1 - p_{+|x} \,.$$

After calculating these curves, one can impose a minimum value for the classification of each individual. In this case, two cut-off values arise in the antibody distribution, one for the seronegative individuals and another for seropositive individuals. Mathematically, the classification rule is given as follows

$$C_i = \begin{cases} \text{seronegative}\,, & \text{if} \;\; x_i \leq c_- \,, \\ \text{equivocal}\,, & \text{if} \;\; c_- < x_i < c_+ \,, \\ \text{seropositive}\,, & \text{if} \;\; x_i \geq c_+ \,, \end{cases}$$

where c_- and c_+ are the cutoff values in the antibody distribution that ensure a minimum classification probability, say 90%. To calculate these cutoff values in practice, one can use the bisection method providing an initial interval where they might be located [13].

Note that the cutoff values based on the above methods are dependent on the data under analysis and, therefore, they should be seen as random realizations of the respective estimator distributions. In other words, they have some uncertainty associated with them due to random sampling. However, this uncertainty is typically neglected in serological data analysis. This topic will be discussed elsewhere in the near future.

3.5. R packages

We used the package `mixsmsn` to fit different SMSN mixture models [48]. In the EM algorithm, the tolerance value for the norm of the difference between parameter estimates from two consecutive iterations was 10^{-5} with a maximum of 10,000 iterations. For each model and antibody under analysis, the EM algorithm was started with 100 random initial guesses for the parameter estimates. The reported estimates were the ones that led to the maximum of the likelihood function among all the runs of the algorithm. For fitting the Generalized Student's t-distribution, we considered the R package `extraDistr` [49], namely, the functions `dlst` and `plst` to calculate its pdf and cdf, respectively. The estimation of the Skew-Normal and Skew-t distributions was done in the package `sn` [50]. See Appendix B in the Supplementary Material for a detailed discussion about the computational costs of the proposed methodology.

4. RESULTS

Serological data refer to positive quantities bounded by an upper limit of detection. In theory, the Skew-Normal or the Skew-t distributions can describe bounded data by setting the respective skewness parameter close to infinity. However, this situation reduces model flexibility by forcing the analysis to be done with SMSN mixture models composed of highly asymmetric mixing distributions. Besides that, it is possible to obtain a good fit of the Gaussian mixture models to serological data after a data transformation [7]. To avoid reducing model flexibility while checking the appropriateness of Gaussian normal models, we applied the logarithmic transformation to the data. For an intuitive interpretation of the resulting data, we used the base 10 logarithmic transformation.

4.1. Exploratory data analysis

In this preliminary data analysis, we aimed to demonstrate the necessity of using alternative mixture models beyond the ones based on the Normal distribution. For this purpose, we partitioned each data set according to the cutoff values suggested by the manufacturers of the commercial kits (see Section 2). We assumed that antibody values below and above these values reflected somehow the distributions of the seronegative and seropositive populations, respectively. We then calculated the empirical skewness and excess kurtosis coefficients in each subset of data (Supplementary Table 2). Note that negative and positive estimates of the excess kurtosis indicated distributions with lighter or heavier tails than the Normal distribution, respectively.

As expected, the putative seropositive populations tended to have a skewness close to zero (HHV-6 and HSV-2) or a negative skewness (CMV, EBV, HSV-1, and VZV) of the respective antibody distribution. Similar evidence could be taken by a visual inspection of the histograms of the data (Figure 1A and B). The empirical estimates of the excess kurtosis were in most cases negative, which suggested distributions with lighter tails than the Normal distribution. However, these negative estimates might have simply resulted from dividing the data into two parts, and such a division limits the "size" of the tails associated with each serological population.

With respect to the putative seronegative populations, the skewness estimates were close to zero in the case of CMV, HSV-1, and HSV-2. For the remaining cases (EBV, HHV-6, and VZV), the skewness estimates were unexpectedly negative. The estimates of the excess kurtosis suggested similar weights of the tails for HSV-2 and VZV. For the remaining, the tails seemed to be lighter or heavier than the Normal distribution.

Finally, there was no evidence based on skewness and excess kurtosis alone for an antibody distribution in which both the seronegative and seropositive populations were similar to the Normal distribution. This suggested the necessity of considering finite mixture models based on families of probability distributions, such as the Skew-Normal or Skew-t, in which skewness and the weight of tails can be modelled appropriately.

4.2. Serological data analysis using Skew-Normal and Skew-t mixture models

To avoid selecting mixture models with difficult biological interpretation due to a high number of components g, we restricted our analysis to models with $g = 1$ (data exclusively composed of a single population, seronegative or seropositive), $g = 2$ (presence of both seronegative and seropositive populations), and $g = 3$. When fitting the Skew-t mixture models, the package `mixsmsn` only allowed to estimate models with the same degree of freedom for all the mixing distributions (*i.e.*, $v_1 = \cdots = v_g = v$).

Before fitting different SMSN mixture models, we first conducted a preliminary analysis based on Gaussian mixture models. In this analysis, we applied an alternative EM algorithm in which there was no need for setting initial values for the parameter estimates while simultaneously determining the optimal number of the components, $\hat{g}$ [30]. The criterion for determining $\hat{g}$ was the maximization of the likelihood function penalized by entropy. For the antibodies against EBV, HSV-2 and VZV viruses, the best Gaussian mixture models were composed of two serological populations. These populations could be interpreted as putative seropositive and seronegative populations. For the remaining antibodies, the best models suggested the presence of three serological populations in the respective data. In this case, the biological interpretation of the respective serological populations is not straightforward, as discussed elsewhere [13].

When compared to our preliminary analysis, the best SMSN mixture models according to BIC tended to require a lesser number of components. In particular, antibodies could be divided into three major classes:

(i) antibodies against HHV-6 and VZV in which data suggested the presence of a single serological population (Table 1 and Figure 1A);

- **(ii)** antibodies against CMV, EBV, and HSV-2 for which there was evidence for two serological populations (Table 2 and Figure 1B);
- **(iii)** antibodies against HSV-2 in which the optimal mixture model is composed of three serological populations (Table 2 and Figure 1B).

Table 1: Analysis of antibody data with evidence for a single serological population, where g represents the number of serological populations, p is the respective number of model parameters, $\mathcal{L}_{\max}$ is the value of the maximized log-likelihood function, p_{gof} is the maximum p-value for the goodness-of-fit test when dividing data into deciles or 5%-quantiles, and p_{boot} is the Bootstrap p-value for testing H_0: $g=1$ versus H_1: $g=2$. Best models according to BIC and the goodness-of-fit tests are written in bold.

Virus	**SMSN**	g	p	$\mathcal{L}_{\max}$	**AIC**	**BIC**	p_{gof}	p_{boot}
HHV-6	Normal	1	2	−129.46	263.00	270.94	0.064	0.064
		2	5	−116.97	244.13	263.97	0.169	
		3	8	−110.43	241.51	268.91	0.462	
	Skew-Normal	**1**	**3**	**−121.35**	**248.80**	**260.71**	**0.140**	**0.027**
		2	7	−117.35	249.03	276.75	0.084	
		3	11	−109.40	241.22	284.87	0.152	
	Student’s t	1	3	−124.38	254.86	266.77	0.157	0.042
		2	6	−117.14	246.55	270.32	0.122	
		3	9	−105.36	229.06	264.78	0.254	
	Skew-t	1	4	−118.81	245.78	261.65	0.148	0.409
		2	8	−116.83	253.54	281.71	0.076	
		3	12	−104.00	234.73	282.36	0.001	
VZV	Normal	1	2	−108.76	221.58	229.53	<0.001	0.000
		2	5	−7.28	24.72	44.60	0.159	
		3	8	−1.70	19.95	51.45	0.153	
	Skew-Normal	1	3	−23.94	53.99	65.90	<0.001	0.180
		2	7	−0.11	14.69	42.27	0.406	
		3	11	0.10	16.87	65.87	0.068	
	Student’s t	1	3	−61.90	129.88	141.80	<0.001	0.000
		2	6	−7.41	26.99	50.86	0.082	
		3	9	−1.68	21.98	57.42	0.113	
	Skew-t	**1**	**4**	**−7.89**	**24.29**	**39.81**	**0.076**	**0.375**
		2	8	−0.05	16.76	48.16	0.211	
		3	12	5.47	25.31	62.14	0.134	

Data of antibodies against HHV-6 and VZV were best described by the Skew-Normal and the Skew-t distributions, respectively. The estimated distributions showed left asymmetry (Figure 1A) with the respective skewness parameter estimated at −1.87 and −5.14 for HHV-6 and VZV datasets, respectively. Accordingly, the Wald’s and the PL 95%s CIs provided negative values for this parameter in the case of the HHV6 data: (−2.44; −1.02) and (−2.57; −1.25), respectively. In this case, the likelihood ratio based on the PL can be roughly approximated by a quadratic function, and, therefore, these two CIs did not substantially differ from each other (Figure 2A). According to the theoretical findings of Chiogna [21], this function showed an inflexion point at $\alpha = 0$. At the level of 5%, there was evidence for a single Skew-Normal against a mixture of two Skew-Normal distributions ($p_{\text{boot}} = 0.027$).

In the case of VZV antibody data, the Wald's and the PL 95% CIs also agreed in terms of a negative skew: $(-6.94; -2.14)$ and $(-8.00; -3.32)$, respectively. However, the likelihood ratio based on the profile likelihood was far from a quadratic function and, therefore, the Wald's CI is not expected to produce reliable results for these data. Finally, there was strong evidence for a single Skew-t distribution compared to a mixture of two Skew-t distributions ($p_{\text{boot}} = 0.375$).

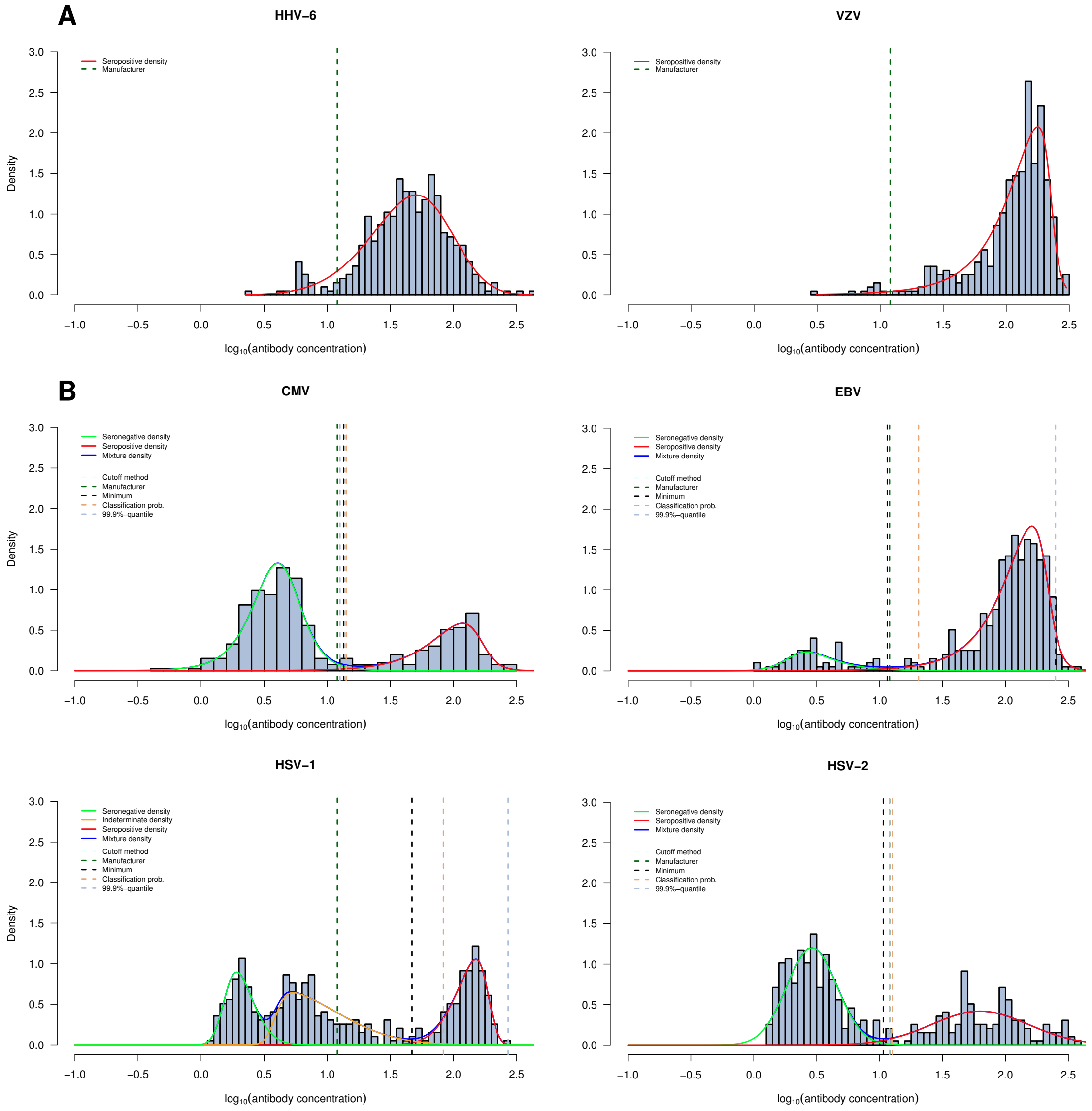


Figure 1: Carry forward and percentage change indices.
Both indices tend to approximate in the months with less prices.

In terms of the respective serological interpretation, a single population for antibodies against HHV-6 and VZV is consistent with a seropositive population, given that HHV-6 and VZV are usually acquired during childhood, and more than 95% of the adult populations show the presence of antibodies against these viruses [51]. In addition, the core values of these distributions are higher than the cutoff for seropositivity suggested by the lab protocol. Finally, a left skewness is also predicted for a hypothetical seropositive population because the antibodies should decay over time in the absence of repeated infections [8].

Table 2: Analysis of antibody data with evidence for more than one serological population. See Table 1 for further details.

Virus	**SMSN**	g	p	$\mathcal{L}_{\max}$	**AIC**	**BIC**	p_{gof}
CMV	Normal	1	2	−409.11	822.29	830.24	< 0.001
		2	5	−245.75	501.66	521.54	0.016
		3	8	−233.70	483.64	515.45	0.018
	Skew-Normal	1	3	−357.61	721.30	733.23	< 0.001
		2	7	−233.82	482.66	509.69	0.038
		3	11	−226.64	489.78	519.35	0.146
	Student’s t	1	3	−410.14	826.36	838.29	< 0.001
		2	6	−238.54	489.27	513.12	0.038
		3	9	−231.23	480.81	516.59	0.046
	Skew-t	1	4	−357.71	723.55	739.45	< 0.001
		2	**8**	**−231.55**	**479.34**	**511.45**	**0.072**
		3	12	−226.93	478.22	525.93	0.324
EBV	Normal	1	2	−342.30	688.67	696.62	< 0.001
		2	5	−152.66	315.48	335.36	< 0.001
		3	8	−129.30	274.84	306.65	0.173
	Skew-Normal	1	3	−226.42	458.93	470.86	< 0.001
		2	7	−130.57	275.34	303.17	0.084
		3	11	−128.02	278.51	322.10	0.054
	Student’s t	1	3	−240.21	486.50	498.43	< 0.001
		2	6	−151.61	315.39	339.26	< 0.001
		3	9	−129.41	277.09	312.88	0.117
	Skew-t	1	4	−173.14	354.40	370.31	< 0.001
		2	**8**	**−125.63**	**267.65**	**299.32**	**0.248**
		3	12	−126.29	280.61	324.66	0.087
HSV-1	Normal	1	2	−442.27	888.61	896.56	< 0.001
		2	5	−291.59	593.34	613.22	< 0.001
		3	8	−264.94	546.14	577.94	0.003
	Skew-Normal	1	3	−394.55	806.62	807.11	< 0.001
		2	7	−260.74	538.10	563.52	0.003
		3	**11**	**−252.32**	**527.39**	**570.70**	**0.104**
	Student’s t	1	3	−443.73	893.55	905.48	< 0.001
		2	7	−291.73	595.65	619.51	< 0.001
		3	9	−264.98	548.23	584.02	0.002
	Skew-t	1	4	−395.43	812.55	814.88	< 0.001
		2	8	−260.88	541.64	569.82	0.001
		3	12	−251.86	528.84	575.79	< 0.001
HSV-2	**Normal**	1	2	−427.29	858.63	866.59	< 0.001
		2	**5**	**−277.62**	**565.39**	**585.27**	**0.516**
		3	8	−269.24	565.92	586.54	0.007
	Skew-Normal	1	3	−337.36	684.60	692.74	< 0.001
		2	7	−264.32	544.79	570.68	0.013
		3	11	−257.19	550.71	580.45	0.003
	Student’s t	1	3	−428.40	862.88	874.81	< 0.001
		2	6	−277.84	567.85	591.71	0.688
		3	9	−269.60	557.52	593.26	0.004
	Skew-t	1	4	−337.79	687.68	699.60	< 0.001
		2	8	−264.52	547.40	577.10	0.007
		3	12	−257.38	562.77	586.83	0.001

Note that most of the SMSN mixture models could also provide a good fitting of the data of these two antibodies. This is the case of the mixture of two or three Normal distributions ($p_{\text{gof}} = 0.169$ and 0.462 for antibodies against HHV-6 and $p_{\text{gof}} = 0.159$ and 0.153), which are typically used in serological data analysis. Therefore, although not being the best models for HHV-6 and VZV-related antibodies, these models could have been used for subsequent serological analyses.

For the remaining antibodies, the respective data analysis was not straightforward because the model with lowest BIC estimate could not fit the data well according to the Pearson's goodness-of-fit test at 5% significance level (Table 2). This occurred for the mixtures of two Skew-Normal distributions for the antibodies against CMV ($\text{BIC} = 509.69$ and $p_{\text{gof}} = 0.038$), HSV-1 ($\text{BIC} = 563.52$ and $p_{\text{gof}} = 0.003$), and HSV-2 ($\text{BIC} = 570.68$ and $p_{\text{gof}} = 0.013$). For these antibodies, the best models were considered to be a mixture of two Skew-t distributions ($\text{BIC} = 511.45$ and $p_{\text{gof}} = 0.072$), a mixture of three Skew-Normal distributions ($\text{BIC} = 570.70$ and $p_{\text{gof}} = 0.104$), and a mixture of two Normal distributions ($\text{BIC} = 585.27$ and $p_{\text{gof}} = 0.516$), respectively, because they were the first models ranked by BIC with a good fit for the data (Figure 1B). Interestingly, for the HSV-2-related antibody data, when the mixture of two Normal distributions was compared to the mixture of two Skew-Normal distribution by a likelihood ratio test, the first model was strongly rejected ($p < 0.0001$), which suggested the asymmetry of at least one of the components. This inconsistency between this test and the selected model can be explained by the unavailability of fitting a mixture of a Normal distribution and a Skew-Normal distribution in the package `smsn`. For the antibody against EBV, the best model was a mixture of two Skew-t distributions, which also had a good fit for the data ($\text{BIC} = 299.32$ and $p_{\text{gof}} = 0.248$; Figure 1B).

With respect to the biological interpretation of each component, there was evidence of putative seronegative and seropositive populations for antibodies against CMV, EBV, and HSV-2 (Figure 1B). This interpretation was supported by the observation that the cutoff value suggested by the commercial kits lies between these hypothetical serological populations. In the case of antibodies against HSV-1, the respective interpretation was not so obvious, because

(i) the best mixture model was composed of three components and

(ii) the cutoff suggested by the commercial kits lies in the middle of the intermediate distribution, which shows right asymmetry.

In theory, the distribution of a putative seronegative population is expected to have right asymmetry [8] and, if so, this intermediate component should be interpreted accordingly. However, one cannot rule out that there are two seronegative populations resulting from distinct background signals in the absence of antibodies. Without additional information about the serological data, this intermediate component was considered to represent a putative seronegative population.

Finally, we performed a similar model selection using AIC instead. Again, we selected the best models with the lowest AIC estimates and with a good fit to the data ($p_{\text{gof}} > 0.05$) at the same time. In contrast with BIC results, this alternative model selection could not provide evidence for a single serological population in the data of HHV-6 and VZV (Table 1).

In these two cases of HHV-6, the best models were mixtures of three Generalized Student-t distributions (AIC = 229.06 and $p_{\text{gof}} = 0.254$) and of two Skew-Normal distributions (AIC = 14.69 and $p_{\text{gof}} = 0.406$), respectively. For antibody against CMV, the best model was a mixture of three Skew-t distributions (AIC = 478.2 and $p_{\text{gof}} = 0.32$), which reflected an increase in the number of components compared to model selection using BIC (Table 2). For the remaining antibodies, it was selected the same model (Table 2). In summary, AIC tended to select models with an increased number of components required to explain the data of each antibody.

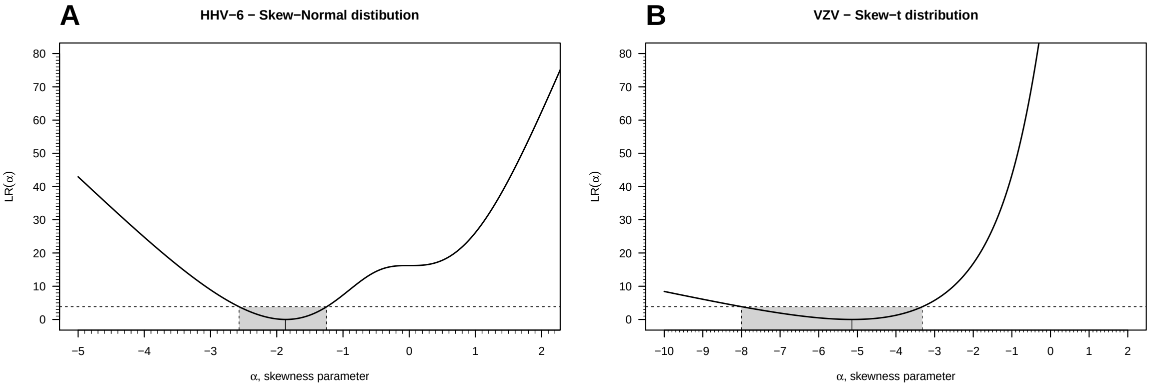


Figure 2: Likelihood ratio (LR) based on the profile likelihood as a function of the skew parameter α, when fitting the Skew-Normal and Skew-t distributions to HHV-6 (A) and VZV (B) data, respectively. The horizontal dashed lines represent the 95% quantile of a χ^2 distribution with one degree of freedom. The grey rectangles represent the 95% CI for α according to this method.

4.3. Estimation of cutoff for seropositivity

After fitting the mixture models to the data, the following step of the analysis was to estimate a cutoff value for seropositivity and the subsequent seroprevalence in the different study groups (Table 3).

For CMV and HSV-2 antibody data, the cutoff values did not vary substantially from one method to another. Interesting, the cutoff values estimated by method 1 (the 3σ rule) almost perfectly matched with the ones suggested by the commercial kits (12.6 U/ml and 12.0 U/ml for CMV and HSV-2 respectively versus 12.0). This good matching between estimates could be explained by a good approximation of the Normal distribution for the seronegative population (Figure 1B) and, therefore, we could infer that the cutoff value suggested by the commercial kits was derived from the 3σ rule; this information was absent from the original study [15]. Since the seronegative and seropositive populations were separated well in these antibody distributions, the seroprevalence estimates across the different study groups were almost invariant with respect to the cutoff value used.

With respect to the EBV antibody data, the hypothetical seronegative population is asymmetric to the right ($\alpha_1 = 1.74$; 95% CI = $(-1.30; 4.80)$; bootstrap 95% CI = $(0.04; 7.90)$; Figure 1B) with heavy tails ($v = 4.52$; 95% CI = $(0.79; 8.26)$; bootstrap 95% CI = $(3.00; 14.88)$).

Consequently, the cutoff value of 249.5 U/ml derived from method 1 was quite different from the one suggested by the commercial kit. However, this cutoff value was considered non-informative because it was well located within the seropositive population and implied seroprevalence estimates close to zero for the different study groups. In contrast, the cutoff values from the remaining methods were in the same order of magnitude of the one suggested by the commercial kits. Therefore, the subsequent seroprevalence estimates of each study group did not differ substantially among these methods. Again, the consistency of the resulting seroprevalence estimates was due to the fact that the seronegative and seropositive populations were well separated in these data.

Table 3: Seroprevalence (%) by cutoff method for seropositivity and by study group. c_- and c_+ are on the linear scale (U/ml). Seroprevalence was calculated based on c_+. The method denoted by "M" refers to the cutoff suggested by the protocol of the commercial kit. The confidence intervals (CI) refer to the Clopper-Pearson exact confidence interval for a proportion.

Virus	Method	c_-	c_+	**Seroprevalence (95% CI)**				
				Global	**HC**	**ME-M**	**ME-S**	**MS**
CMV	M	8.0	12.0	33.5 (28.9–38.4)	37.4 (28.2–47.3)	28.6 (22.4–35.4)	33.3 (21.1–47.5)	36.7 (23.4–51.7)
	1	—	12.6	33.5 (28.9–38.4)	37.4 (28.2–47.3)	28.6 (22.4–35.4)	33.3 (21.1–47.5)	36.7 (23.4–51.7)
	2	—	13.5	33.2 (28.6–38.1)	37.4 (28.2–47.3)	28.6 (22.4–35.4)	31.5 (19.5–45.6)	36.7 (23.4–51.7)
	3	9.4	14.1	32.9 (28.4–37.9)	37.4 (28.2–47.3)	28.1 (21.9–34.9)	31.5 (19.5–45.6)	36.7 (23.4–51.7)
EBV	M	8.0	12.0	87.3 (83.6–90.4)	87.9 (80.1–93.4)	86.2 (80.6–90.7)	81.5 (68.6–90.7)	75.5 (61.1–86.7)
	1	—	249.5	2.0 (0.09–3.9)	1.9 (0.02–6.6)	1.5 (0.03–4.4)	0.0 (0.0–6.6)	6.1 (1.3–16.9)
	2	—	11.5	87.3 (83.6–90.4)	87.9 (80.1–93.4)	86.2 (80.6–90.7)	81.5 (68.6–90.7)	75.5 (61.1–86.7)
	3	5.6	20.4	85.5 (81.7–88.9)	87.9 (80.1–93.4)	82.7 (76.6–87.7)	81.5 (68.6–90.7)	75.5 (61.1–75.5)
HSV-1	M	8.0	12.0	45.2 (40.2–50.2)	42.1 (32.6–51.9)	41.8 (34.8–49.1)	51.9 (37.8–65.6)	46.9 (32.5–61.7)
	1	—	271.0	0.0 (0.0–0.1)	0.0 (0.0–3.4)	0.0 (0.0–1.2)	0.0 (0.0–6.6)	0.0 (0.0–7.3)
	2	—	46.9	34.5 (29.8–39.4)	28.0 (19.8–37.5)	34.7 (28.1–41.8)	38.9 (25.9–53.1)	34.7 (21.7–49.6)
	3	42.7	83.2	30.7 (26.2–35.5)	24.3 (16.5–33.5)	32.1 (25.7–39.2)	33.3 (21.1–47.5)	28.6 (16.6–43.3)
HSV-2	M	8.0	12.0	38.1 (33.3–43.1)	33.6 (24.8–43.4)	38.8 (31.9–45.9)	40.7 (27.6–54.9)	32.7 (19.9–47.5)
	1	—	12.0	38.1 (33.3–43.1)	33.6 (24.8–43.4)	38.8 (31.9–45.9)	40.7 (27.6–54.9)	32.7 (19.9–47.5)
	2	—	10.7	38.8 (33.9–43.8)	33.6 (24.8–43.4)	39.3 (32.4–46.5)	40.7 (27.6–54.9)	36.7 (23.4–51.7)
	3	7.1	12.6	37.8 (33.0–42.8)	33.6 (24.8–43.4)	38.8 (31.9–45.9)	40.7 (27.6–54.9)	30.6 (18.3–45.4)

The largest differences in the cutoff values for seropositivity were observed for the HSV-1 antibody data. Coincidentally, this was the data set where the best mixture model was composed of three components. As discussed earlier in this paper, the intermediate component was considered a second hypothetical seronegative population, which resulted in a shift in the calculation of seropositivity towards higher values. As such, the cutoff seropositive based on the commercial kit led to the highest seroprevalence estimates for all study groups with a global estimate of 45.2% (95% CI = (40.2%; 50.2%). As an extreme case, the 3σ rule produced a too-high cutoff value again due to the right asymmetry of both seronegative populations. Such unrealistic cutoff value led to a zero seroprevalence estimates and rendered the respective analysis useless.

Finally, although not being the main objective of this study, the comparison of the four study groups suggested that, given a method for determining seropositivity and antibody under analysis, the seroprevalence of patients with ME/CFS did not appear to differ significantly from the one of healthy controls and patients with multiple sclerosis alike.

5. CONCLUSIONS

This study aimed to review the Skew-Normal and Skew-t mixture models and recommend their routine use in serological data analysis. Such recommendation sets its foundation in the high flexibility of these models in describing different data patterns, as illustrated with the data analysis of antibodies against 6 herpesviruses. In particular, high modelling flexibility is desirable given that right and left asymmetry can emerge from seronegative and seropositive populations, respectively. In this regard, most popular distributions used in Statistics are not able to exhibit either left or right asymmetry depending on the parameters specified. A less-known family of distributions that shows such stochastic property is the Generalized Tukey's λ distribution [54, 55]. This distribution offers a great variety of shapes owing to four parameters controlling the location, the scale, the skewness, and the flatness of the resulting distribution. However, the Generalized Tukey's λ distribution is only defined in terms of its quantile function and, hence, its estimation is cumbersome. This distribution has already been proposed for mixture modelling, but there are only theoretical and computational developments for the case of two components [52, 53]. This limits the application of these alternative models in data sets where there is evidence for more than two serological populations, such as the case of the antibodies against HSV-1 here analyzed or against the influenza virus reported elsewhere [11]. Therefore, Skew-Normal and Skew-t mixture models would appear the most general and flexible approach for analysing serological data.

For data analysis, we recommend using the package `mixsmsn` for estimating the finite mixture models [48]. Notwithstanding this recommendation, the package only estimates SMSN mixture models where all mixing distributions belong to the same family of SMSN probability distributions. Hence, it can only fit 4 different models per number of components. In theory, there are $4^2 = 16$ possible two-component mixture models resulting from the combination of Normal, Skew-Normal, Generalized Student's t, and Skew-t distributions as mixing distributions. Note that these possible models are nested in each other by imposing parametric restrictions to the most general mixture model based on the Skew-t distribution. For three-component mixture models, the number of possible models increases to $4^3 = 64$.

Therefore, the package mixsmsn excludes a vast number of possible models, which ultimately affects the detection of the most parsimonious model for the data; this model could be a combination of probability distributions from different families. The same limitation could also explain some inferential inconsistencies in the example of application. For instance, a single Skew-Normal distribution was considered the best model for the antibodies against HHV-6. However, the hypothesis of a single Skew-Normal distribution against a mixture of two Skew-Normal distributions could be rejected by bootstrap at the 5% significance level. A possible explanation for this statistical inconsistency is that the best model for these data could be a mixture of a Normal distribution for the seronegative population and a Skew-Normal distribution for the seropositive population. Therefore, there is a research opportunity to extend the package allowing each mixing component to be described by different families of SMSN distributions.

Another limitation of using mixsmsn package is that, for mathematical tractability, the mixtures of generalized Student t and Skew-t distributions were assumed to have the same degrees of freedom in all the mixing distributions. In theory, this assumption could be relaxed so this parameter could vary from one component of the mixture to another. This modelling option was available in the package EMMIXuskew for the mixture of Skew-t distributions [56]. However, this package is currently discontinued. In practice, we expect some degree of numerical instability when estimating different degrees of freedom in data where the serological populations overlap substantially with each other. In this regard, future research could be conducted to determine the stochastic and sampling conditions in which different degrees of freedom could infer from different components.

The problem of determining the optimal cutoff value for seropositivity has been intensively investigated, discussed, and revisited over the years [46, 57, 58, 59]. In this regard, the most popular cutoffs for seropositivity are simply defined by the mean plus a given number of times the standard deviation of the hypothetical seronegative population without checking the Normality assumption of the hypothetical seronegative population. The resulting cutoffs are associated with high-order quantiles of the Normal distribution, such as 97.7% or 99.9% for the 2σ and 3σ rules, respectively. In practice, these cutoffs imply a high specificity but show an arbitrary sensitivity for the respective serological classification. When the hypothetical seronegative population shows a right-skewed distribution, similar cutoffs can be obtained by calculating the same high quantiles of the estimated SMSN, as done here. The reverse argument can be made when analysing antibodies where seropositivity could be considered the default serological state of an individual, such as the case of antibodies against HHV-6 and VZV here analyzed or vaccine-related antibodies in populations where vaccination is mandatory. Similar cutoffs can be determined for these antibodies by the mean minus a given number of times the standard deviation of the hypothetical seropositive population assumed to be normally distributed. For a left-skewed seropositive population, the cutoff values for seropositivity are now calculated using the low order quantiles (*e.g.*, 2.3% and 0.1%-quantiles for the 2σ and 3σ rules, respectively). Inversely, these cutoffs generate a high sensitivity but an arbitrary specificity for the respective serological classification. It is worth noting that it is up to the analyst to decide on what she/he wants to control, whether specificity, sensitivity, or both with respect to the resulting serological classification. A similar decision problem occurs in analyses based on the Receiver Operating Characteristic curve. Given the multiplicity of criteria for estimating this cutoff and its uncertainty, several authors advocate a free-cutoff approach for serological analysis [6, 60]. However, a detailed discussion about the advantages and disadvantages of free-cutoff approaches was out of the scope of this study.

In summary, the mixture models based on Skew-Normal and Skew-t distributions show promise to become a routine tool for serological data analysis. They have the advantage of including the Gaussian mixture models as special cases. However, given the statistical complexity of these models and some inferential problems highlighted throughout the paper, their application should be done in a closer collaboration between biomedical researchers who generate the serological data and biostatisticians who have in principle the knowledge and skills to fit and compared these mode properly.

ACKNOWLEDGMENTS

The authors would like to thank two anonymous reviewers for their constructive comments and suggestions. The authors would also like to thank Eliana Lacerda, Luis Nacul, and Jackie-Cliff from the London School of Hygiene & Tropical Medicine (LSHTM) for sharing the data, and João Malato for helping to proof-read the manuscript. This work was partially funded by Fundação para a Ciência e a Tecnologia, Portugal (ref: UIDB/00006/2020 and UIDB/04561/2020) and resulted from a short-term scientific mission of TDS to the LSHTM, which was funded by the EUROMENE Cost Action (CA15111) from the European Union.

REFERENCES

[1] Wang, S.S.; Schiffman, M.; Shields, T.S.; Herrero, R.; Hildesheim, A.; Bratti, M.C.; Sherman, M.E.; Rodriguez, A.C.; Castle, P.E.; Morales, J.; Alfaro, M.; Wright, T.; Chen, S.; Clayman, B.; Burk, R.D. and Viscidi, R.P. (2003). Seroprevalence of human papillomavirus-16, -18, -31, and -45 in a population-based cohort of 10000 women in Costa Rica, *British Journal of Cancer*, **89**(7), 1248–1254.

[2] Cook, J.; Kleinschmidt, I.; Schwabe, C.; Nseng, G.; Bousema, T.; Corran, P.H.; Riley, E.M. and Drakeley, C.J. (2011). Serological markers suggest heterogeneity of effectiveness of malaria control interventions on Bioko Island, Equatorial Guinea, *Plos One*, **6**(9), e25137.

[3] Hsiang, M.S.; Hwang, J.; Kunene, S.; Drakeley, C.; Kandula, D.; Novotny, J.; Parizo, J.; Jensen, T.; Tong, M.; Kemere, J.; Dlamini, S.; Moonen, B.; Angov, E.; Dutta, S.; Ockenhouse, C.; Dorsey, G. and Greenhouse, B. (2012). Surveillance for malaria elimination in Swaziland: a national cross-sectional study using pooled PCR and serology, *PloS One*, **7**(1), e29550.

[4] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*, John Wiley & Sons, New York.

[5] Gay, N.J. (1996). Analysis of serological surveys using mixture models: application to a survey of parvovirus B19, *Statistics in Medicine*, **15**, 1567–1573.

[6] Chis Ster, I. (2012). Inference for serological surveys investigating past exposures to infections resulting in long-lasting immunity – an approach using finite mixture models with concomitant information, *Journal of Applied Statistics*, **39**(11), 2523–2542.

[7] Rogier, E.; Wiegand, R.; Moss, D.; Priest, J.; Angov, E.; Dutta, S.; Journel, I.; Jean, S.E.; Mace, K.; Chang, M.; Lemoine, J.F.; Udhayakumar, V. and Barnwell, J.W. (2015). Multiple comparisons analysis of serological data from an area of low Plasmodium falciparum transmission, *Malaria Journal*, **14**, 436.

[8] Parker, R.A.; Erdman, D.D. and Anderson, L.J. (1990). Use of mixture models in determining laboratory criterion for identification of seropositive individuals: application to parvovirus B19 serology, *Journal of Virological Methods*, **27**(2), 135–144.

[9] Baughman, A.L.; Bisgard, K.M.; Lynn, F. and Meade, B.D. (2006). Mixture model analysis for establishing a diagnostic cut-off point for pertussis antibody levels, *Statistics in Medicine*, **25**, 2994–3010.

[10] Rota, M.C.; Massari, M.; Gabutti, G.; Guido, M.; De Donno, A. and Ciofi degli Atti, M.L. (2008). Measles serological survey in the Italian population: interpretation of results using mixture model, *Vaccine*, **26**(34), 4403–4409.

[11] Nhat, N.; Todd, S.; de Bruin, E.; Thao, T.; Vy, N.; Quan, T.M.; Vinh, D.N.; van Beek, J.; Anh, P.H.; Lam, H.M.; Hung, E.; Lien, N.; Hong, T.; Farrar, J.; Simmons, C.P.; Chau, N.; Koopmans, M. and Boni, M.F. (2017). Structure of general-population antibody titer distributions to influenza A virus, *Scientific Reports*, **7**(1), 6060.

[12] Moreira da Silva, J.; Prata, S.; Domingues, T.D.; Leal, R.O.; Nunes, T.; Tavares, L.; Almeida, V.; Sepúlveda, N. and Gil, S. (2020). Detection and modeling of anti-Leptospira IgG prevalence in cats from Lisbon area and its correlation to retroviral infections, lifestyle, clinical and hematologic changes, *Veterinary and Animal Science*, **10**, 100144.

[13] Sepúlveda, N.; Stresman, G.; White, M.T. and Drakeley, C.J. (2015). Current mathematical models for analyzing anti-malarial antibody data with an eye to malaria elimination and eradication, *Journal of Immunology Research*, **2015**, 738030.

[14] Basso, R.M.; Lachos, V.H.; Cabral, C.R.B. and Gosh, P. (2010). Robust mixture modelling based on scale mixtures of skew-normal distributions, *Computational Statistics and Data Analysis*, **54**, 2926–2941.

[15] Cliff, J.M.; King, E.C.; Lee, J.S.; Sepúlveda, N.; Wolf, A.S.; Kingdon, C.; Bowman, E.; Dockrell, H.M.; Nacul, L.; Lacerda, E. and Riley, E.M. (2019). Cellular immune function in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS), *Frontiers in Immunology*, **10**, 796.

[16] Cortes Rivera, M.; Mastronardi, C.; Silva-Aldana, C.T.; Arcos-Burgos, M. and Lidbury, B.A. (2019). Myalgic encephalomyelitis/chronic fatigue syndrome: a comprehensive review, *Diagnostics*, **9**, 91.

[17] Lin, T.I.; Lee, J.C. and Yen, S.Y. (2007). Finite mixture modelling using the skew-normal distribution, *Statistica Sinica*, **17**, 909–927.

[18] Azzalini, A. and Capitanio, A. (2014). *The Skew-Normal and Related Families*, Cambridge University Press, Cambridge.

[19] Azzalini, A. (1985). A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics*, **12**, 171–178.

[20] Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution, *Journal of the Royal Statistical Society B*, **65**, 367–389.

[21] Chiogna, M. (2005). A note on the asymptotic distribution of the maximum likelihood estimator for the scalar skew-normal distribution, *Statistical Methods & Applications*, **14**, 331–341.

[22] Hallin, M. and Ley, C. (2012). Skew-symmetric distributions and Fisher information – a tale of two densities, *Bernoulli*, **18**, 747–763.

[23] Hallin, M. and Ley, C. (2014). Skew-symmetric distributions and Fisher information: the double sin of the skew-normal, *Bernoulli*, **20**, 1432–1453.

[24] Gómez, H.W.; Venegas, O. and Bolfarine, H. (2007). Skew-symmetric distributions generated by the distribution function of the normal distribution, *Environmetrics*, **18**, 395–407.

[25] Ferreira, C.S.; Bolfarine, H. and Lachos, V.H. (2011). Skew scale mixtures of normal distributions: properties and estimation, *Statistical Methodology*, **8**, 154–171.

[26] Theodossiou, P. (1998). Financial data and the skewed generalized t distribution, *Management Science*, **44**, 1650–1661.

[27] Arslan, O. and Genç, A.I. (2009). The skew generalized t distribution as the scale mixture of a skew exponential power distribution and its applications in robust estimation, *Statistics*, **43**, 481–498.

[28] Lachos Dávila, V.H.; Zeller, C.B. and Cabral, C.R.B. (2018). *Finite Mixture Of Skewed Distributions*, Springer.

[29] Figueiredo, M.A.T. and Jain, A.K. (2002). Unsupervised learning of finite mixture models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**, 381–396.

[30] Yang, M.; Lai, C. and Lin, C. (2012). A robust EM clustering algorithm for Gaussian mixture models, *Pattern Recognition*, **45**, 3950–3961.

[31] Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, 716–723.

[32] Biernacki, C.; Celeux, G. and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 719–725.

[33] Fraley, C. and Raftery, A.E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis, *The Computer Journal*, **41**, 578–588.

[34] Zhao, J.; Jin, L. and Shi, L. (2015). Mixture model selection via hierarchical BIC, *Computational Statistics & Data Analysis*, **88**, 139–153.

[35] Mehrjou, A.; Hosseini, R. and Araabi, B.N. (2016). Improved Bayesian information criterion for mixture model selection, *Pattern Recognition Letters*, **69**, 22–27.

[36] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions, *Psychometrika*, **52**, 345–370.

[37] Kim, Daeyoung and Seo, Byungtae (2014). Assessment of the number of components in Gaussian mixture models in the presence of multiple local maximizers, *Journal of Multivariate Analysis*, **125**, 100–120.

[38] Feng, Z.D. and McCullogh, C.E. (1996). Using bootstrap likelihood ratios in finite mixture models, *Journal of the Royal Statistical Society B*, **58**, 609–617.

[39] Yu, Y. and Harvill, J.L. (2019). Bootstrap likelihood ratio test for Weibull mixture models fitted to grouped data, *Communications in Statistics – Theory and Methods*, **48**(18), 4550–4568.

[40] Zeller, C.B.; Labra, F.V.; Lachos, V.H. and Balakrishnan, N. (2010). Influence analyses of skew-normal/independent linear mixed models, *Computational Statistics & Data Analysis*, **54**, 1266–1280.

[41] Montenegro, L.C.; Lachos, V.H. and Bolfarine, H. (2010). Inference for a skew extension of the Grubbs model, *Statistical Papers*, **51**, 701–715.

[42] Pawitan, Y. (2000). A reminder of the fallibility of the Wald statistic: likelihood explanation, The American Statistician, **54**, 54–56.

[43] Bunge, J. and Barger, K. (2008). Parametric models for estimating the number of classes, *Biometrical Journal*, **50**, 971–982.

[44] Ultsch, A.; Thrun, M.; Hansen-Goos, O. and Lötsch, J. (2015). Identification of molecular fingerprints in human heat pain thresholds by use of an interactive mixture model R toolbox (AdaptGauss), *International Journal of Molecular Sciences*, **16**, 25897–25911.

[45] Wichitchan, S.; Yao, W. and Yang, G. (2019). Hypothesis testing for finite mixture models, *Computational Statistics & Data Analysis*, **132**, 180–189.

[46] Saraswati, K.; Phanichkrivalkosil, M.; Day, N. and Blacksell, S.D. (2019). The validity of diagnostic cut-offs for commercial and in-house scrub typhus IgM and IgG ELISAs: a review of the evidence, *PLoS Neglected Tropical Diseases*, **13**(2), e0007158.

[47] Brent, R.P. (1973). *Algorithms For Minimization Without Derivatives*, Prentice-Hall, Englewood Cliffs, New Jersey, 73–76.

[48] Prates, M.O.; Lachos, V.H. and Cabral, C. (2013). Fitting finite mixture of scale mixture of skew-normal distributions, *Journal of Statistical Software*, **54**, 1–20.

[49] Wolodzko, T. (2020). Additional Univariate and Multivariate Distributions, *R CRAN*, `https://cran.r-project.org/web/packages/extraDistr/index.html`

[50] Azzalini, A. (2020). The Skew-Normal and Related Distributions Such as the Skew-t, *R CRAN*, `https://cran.r-project.org/web/packages/sn/sn.pdf`

[51] Braun, D.K.; Dominguez, G. and Pellett, P.E. (1997). Human herpesvirus 6, *Clinical Microbiology Reviews*, **10**(3), 521–567.

[52] Su, S. (2007). Fitting single and mixture of generalized lambda distributions to data via discretized and maximum likelihood methods: GLDEX in R, *Journal of Statistical Software*, **21**(9), 1–17.

[53] Su, S. (2011). Maximum log likelihood estimation using EM algorithm and partition maximum log likelihood estimation for mixtures of generalized lambda distributions, *Journal of Modern Applied Statistical Methods*, **10**, 17.

[54] Ramberg, J. and Schmeiser, B. (1974). An approximate method for generating asymmetric random variables, *Communications of the Association for Computing Machinery*, **17**, 78–82.

[55] Freimer, M.; Mudholkar, G.; Kollia, G. and Lin, C. (1988). A study of the generalised Tukey lambda family, *Communications in Statistics – Theory and Methods*, **17**, 3547–3567.

[56] McLachlan, G. and Lee, S. (2013). EMMIXuskew: an R package for fitting mixtures of multivariate skew t distributions via the EM algorithm, *Journal of Statistical Software*, **55**(12), 1–22.

[57] Ridge, S.E. and Vizard, A.L. (1993). Determination of the optimal cutoff value for a serological assay: an example using the Johne's Absorbed EIA, *Journal of Clinical Microbiology*, **31**(5), 1256–1261.

[58] Kafatos, G.; Andrews, N.J.; McConway, K.J.; Maple, P.A.; Brown, K. and Farrington, C.P. (2016). Is it appropriate to use fixed assay cut-offs for estimating seroprevalence?, *Epidemiology and infection*, **144**(4), 887–895.

[59] Migchelsen, S.J.; Martin, D.L.; Southisombath, K.; Turyaguma, P.; Heggen, A.; Rubangakene, P.P.; Joof, H.; Makalo, P.; Cooley, G.; Gwyn, S.; Solomon, A.W.; Holland, M.J.; Courtright, P.; Willis, R.; Alexander, N.D.; Mabey, D.C. and Roberts, C.H. (2017). Defining seropositivity thresholds for use in trachoma elimination studies, *PLoS Neglected Tropical Diseases*, **11**(1), e0005230.

[60] Bouman, J.A.; Riou, J.; Bonhoeffer, S. and Regoes, R.R. (2021). Estimating the cumulative incidence of SARS-CoV-2 with imperfect serological tests: exploiting cutoff-free approaches, *PLoS Computational Biology*, **17**, e1008728.

REVSTAT-Statistical Journal

Aims and Scope

The aim of REVSTAT-Statistical Journal is to publish articles of high scientific content, developing Statistical Science focused on innovative theory, methods, and applications in different areas of knowledge. Important survey/review contributing to Probability and Statistics advancement is also welcome.

Background

Statistics Portugal started in 1996 the publication of the scientific statistical journal *Revista de Estatística*, in Portuguese, a quarterly publication whose goal was the publication of papers containing original research results, and application studies, namely in the economic, social and demographic fields. Statistics Portugal was aware of how vital statistical culture is in understanding most phenomena in the present-day world, and of its responsibilities in disseminating statistical knowledge.

In 1998 it was decided to publish papers in English. This step has been taken to achieve a larger diffusion, and to encourage foreign contributors to submit their work. At the time, the editorial board was mainly composed by Portuguese university professors, and this has been the first step aimed at changing the character of *Revista de Estatística* from a national to an international scientific journal. In 2001, the *Revista de Estatística* published a three volumes special issue containing extended abstracts of the invited and contributed papers presented at the 23rd European Meeting of Statisticians (EMS). During the EMS 2001, its editor-in-chief invited several international participants to join the editorial staff.

In 2003 the name changed to REVSTAT-Statistical Journal, published in English, with a prestigious international editorial board, hoping to become one more place where scientists may feel proud of publishing their research results.

Editorial policy

REVSTAT-Statistical Journal is an open access peer-reviewed journal published quarterly, in English, by Statistics Portugal.

The editorial policy of REVSTAT is mainly placed on the originality and importance of the research. The whole submission and review processes for REVSTAT are conducted exclusively online on the journal's webpage revstat.ine.pt based in Open Journal System (OJS). The only working language allowed is English. Authors intending to submit any work must register, login and follow the guidelines.

There are no fees for publishing accepted manuscripts that will be made available in open access.

All articles consistent with REVSTAT aims and scope will undergo scientific evaluation by at least two reviewers, one from the Editorial Board and another external. Authors can suggest an editor or reviewer who is expert on the paper subject providing her/his complete information, namely: name, affiliation, email and, if possible, personal URL or ORCID number.

All published works are Open Access (CC BY 4.0) which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. Also, in the context of archiving policy, REVSTAT is a *blue* journal welcoming authors to deposit their works in other scientific repositories regarding the use of the published edition and providing its source.

Journal prints may be ordered at expenses of the author(s), and prior to publication.

Abstract and Indexing services

REVSTAT-Statistical Journal is covered *by Journal Citation Reports - JCR (Clarivate); DOAJ-Directory Of Open Access Journals; Current Index to Statistics; Google Scholar; Mathematical Reviews® (MathSciNet®); Zentralblatt für Mathematic; Scimago Journal & Country Rank; Scopus*

Author guidelines

The whole submission and review processes for REVSTAT are conducted exclusively online on the journal's webpage https://revstat.ine.pt/ based in Open Journal System (OJS). Authors intending to submit any work must *register, login* and follow the indications choosing *Submissions*.

REVSTAT - Statistical Journal adopts the COPE guidelines on publication ethics.

Work presentation

- the only working language is English;
- the first page should include the name, ORCID iD (optional), Institution, country, and mail-address of the author(s);
- a summary of fewer than one hundred words, followed by a maximum of six keywords and the MSC 2020 subject classification should be included also in the first page;
- manuscripts should be typed only in black, in double-spacing, with a left margin of at least 3 cm, with numbered lines, and a maximum of 25 pages;
- the title should be with no more than 120 characters (with spaces);
- figures must be a minimum of 300dpi and will be reproduced online as in the original work, however, authors should take into account that the printed version is always in black and grey tones;
- authors are encouraged to submit articles using LaTeX which macros are available at *REVSTAT style*;
- citations in text should be included in the text by name and year in parentheses, as in the following examples: § article title in lowercase (Author 1980); § This

theorem was proved later by AuthorB and AuthorC (1990); § This subject has been widely addressed (AuthorA 1990; AuthorB et al. 1995; AuthorA and AuthorB 1998).

- references should be listed in alphabetical order of the author's scientific surname at the end of the article;
- acknowledgments of people, grants or funds should be placed in a short section before the References title page. Note that religious beliefs, ethnic background, citizenship and political orientations of the author(s) are not allowed in the text;
- authors are welcome to suggest one of the Editors or Associate Editors or yet other reviewer expert on the subject providing a complete information, namely: name, affiliation, email and personal URL or ORCID number in the Comments for the Editor (submission form).

Accepted papers

After final revision and acceptance of an article for publication, authors are requested to provide the corresponding LaTeX file, as in REVSTAT style.

Supplementary files may be included and submitted separately in .tiff, .gif, .jpg, .png .eps, .ps or .pdf format. These supplementary files may be published online along with an article, containing data, programming code, extra figures, or extra proofs, etc; however, REVSTAT is not responsible for any supporting information supplied by the author(s).

Copyright Notice

Upon acceptance of an article, the author(s) will be asked to transfer copyright of the article to the publisher, Statistics Portugal, in order to ensure the widest possible dissemination of information.

According to REVSTAT's *archiving policy*, after assigning the copyright form, authors may cite and use limited excerpts (figures, tables, etc.) of their works accepted/published in REVSTAT in other publications and may deposit only the published edition in scientific repositories providing its source as REVSTAT while the original place of publication. The Executive Editor of the Journal must be notified in writing in advance.