ISSN 1645-6726

REVSTAT

Statistical Journal

vol. 24 - n. 2 - April 2026

Instituto Nacional de Estatística
Statistics Portugal

REVSTAT — Statistical Journal, vol.24, n. 2 (April 2026)

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

Continues: Revista de Estatística = ISSN 0873-4275.

ISSN 1645-6726 ; e-ISSN 2183-0371

Editorial Board (2025-2026)

Editor-in-Chief – *Manuel SCOTTO*

Co-Editor – *Cláudia NUNES*

Associate Editors

Abdelhakim AKNOUCHE
Andrés ALONSO
Barry ARNOLD
Wagner BARRETO-SOUZA
Francisco BLASQUES
Paula BRITO
Rui CASTRO
Valérie CHAVEZ-DEMOULIN
David CONESA
Charmaine DEAN
Fernanda FIGUEIREDO
Jorge Milhazes FREITAS
Stéphane GIRARD
Sónia GOUVEIA
Victor LEIVA
Artur LEMONTE
Shuangzhe LIU
Raquel MENEZES
Fernando MOURA
Cláudia NEVES
John NOLAN
Carlos OLIVEIRA
Paulo Eduardo OLIVEIRA
Pedro OLIVEIRA
Rosário OLIVEIRA
Gilbert SAPORTA
Alexandra M. SCHMIDT
Lisete SOUSA
Jacobo de UÑA-ÁLVAREZ
Christian WEIß

Executive Editor – *Olga BESSA MENDES*

Publisher – *Statistics Portugal*

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

Digital edition

Creative Commons Attribution 4.0 International (CC BY 4.0)
© Statistics Portugal, Lisbon. Portugal, 2026

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

INDEX

On General Weighted Cumulative Past Extropy

Pradeep Kumar Sahu and *Nitin Gupta* 115

Novel Robust Estimators for the Linear Regression Model with Multicollinearity and Outlier Problems

Murat Erişoğlu, Aydın Karakoca and *Abdullah Yurtaslan* 135

Testing the Validity of Lindley Model Based on Nonparametric Probability Density Functions of Entropy Estimators

Hadi Alizadeh Noughabi and *Mohammad Shafaei Noughabi* 157

Incorporating Direct Responses into Optional Randomized Response Models Without Compromising Respondents' Privacy

Muhammad Azeem 175

Comparative Study on Probability Density Estimators of Sample Maximum and Data Transformation

Taku Moriyama 191

Non-parametric Inference of Window-Observation Recurrent Event Data with Multiple Causes of Failure

P. G. Sankaran and *S. Hari* 213

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 24, Number 2, April 2026, 115–133
https://doi.org/10.57805/revstat.v24i2.633

On General Weighted Cumulative Past Extropy

Authors: Pradeep Kumar Sahu ✉
– Department of Mathematics, Indian Institute of Technology Kharagpur, West Bengal 721302, India
pradeep.maths@kgpian.iitkgp.ac.in

Nitin Gupta
– Department of Mathematics, Indian Institute of Technology Kharagpur, West Bengal 721302, India
nitin.gupta@maths.iitkgp.ac.in

Received: May 2023 Revised: January 2024 Accepted: February 2024

Abstract:

- In this paper, we study some properties and characterization of the general weighted cumulative past extropy (n-WCPJ). Many results including some bounds, inequalities, and effects of linear transformations are obtained. We study the characterization of n-WCPJ based on the largest order statistics. Conditional WCPJ and some of its properties are discussed. We study weighted cumulative past extropy (WCPJ) and weighted negative cumulative extropy (WNCEX) and its application for testing symmetry of dataset.

Keywords:

- *cumulative residual entropy; cumulative residual extropy; weighted cumulative past extropy; conditional weighted past extropy; weighted negative cumulative extropy.*

AMS Subject Classification:

- 62B10, 62N05, 94A15.

✉ Corresponding author

1. INTRODUCTION

We encounter phenomena or events which are associated with uncertainty. Uncertainty emerges since we have less information than the total information required to describe a system and its environment. Uncertainty and information are closely associated. Since in an experiment the information provided is equal to the amount of uncertainty removed. Entropy which is a measure of uncertainty was first introduced by Shannon (1948) in communication theory. It is useful to estimate the probabilities of rare events (large deviation theory) and in the study of likelihood-based inference principles. Shannon entropy is defined as the average amount of information that we receive per event and was the first defined entropy. For continuous case, it is given by

$$\mathcal{H}(X) = -\int_0^\infty f_X(x) \log f_X(x)\, dx,$$

where X is a non-negative absolutely continuous random variable with probability density function (pdf) f_X, cumulative distribution function (cdf) F_X and survival function (sf) $\bar{F}_X = 1 - F_X$. Shannon entropy has various applications in communication theory, mathematical, physical, engineering, biological and social sciences as well. For further details on entropy one may refer Ash (1990) and Cover and Thomas (2006).

Rao *et al.* (2004) introduced the notion of cumulative residual entropy (CRE) as:

$$\mathcal{E}(X) = -\int_0^\infty \bar{F}_X(x) \log \bar{F}_X(x)\, dx.$$

Some general results regarding this measure have been studied by Drissi *et al.* (2008) and Rao (2005). CRE finds applications in image alignment and in the measurement of similarity between images. Di Crescenzo and Longobardi (2009) proposed an entropy called cumulative past entropy (or cumulative entropy) i.e. CPE (or CE) as:

$$\mathcal{CE}(X) = -\int_0^\infty F_X(x) \log F_X(x)\, dx.$$

Asadi and Zohrevand (2007), Di Crescenzo and Longobardi (2013), Khorashadizadeh *et al.* (2013) and Navarro *et al.* (2010) investigated many aspects of CRE (CPE).

Lad *et al.* (2015) defined an alternative measure of the uncertainty of a random variable called extropy. For continuous non-negative random variable X the extropy is defined as:

$$J(X) = -\frac{1}{2}\int_0^\infty f_X^2(x)\, dx = -\frac{1}{2} E(f_X(X)). \tag{1.1}$$

Some results and properties of the extropy of order statistics and record values are given by Qiu (2017). Qiu and Jia (2018) derived some of the results of the residual extropy of order statistics. Yang *et al.* (2018) studied the bounds on extropy with a variational distance constraint. Qiu *et al.* (2019) examined certain extropy properties of mixed systems. To find out more about extropy, one may refer to Krishnan *et al.* (2020), Noughabi and Jarrahiferiz (2019) and Raqab and Qiu (2019).

Jahanshahi *et al.* (2020) introduced the cumulative residual extropy (CRJ). For continuous non-negative random variable X the cumulative residual extropy (CRJ) as:

$$\xi J(X) = -\frac{1}{2}\int_0^\infty \bar{F}_X^2(x)\,dx. \tag{1.2}$$

Kundu (2023) proposed an extropy called cumulative past extropy (CPJ). For continuous non-negative random variable X the cumulative past extropy (CPJ) is defined as:

$$\bar{\xi} J(X) = -\frac{1}{2}\int_0^\infty F_X^2(x)\,dx. \tag{1.3}$$

The idea behind this is to replace the density function by distribution function in extropy (1.1). One can only define the cumulative past extropy for random variables that have a limited range of possible values, as this measure would be infinitely negative for any random variable with an unlimited range of values. Kundu (2023) studied extreme order statistics on cumulative residual (past) extropy.

Hashempour *et al.* (2022) proposed a new measure called weighted cumulative residual extropy. Mohammadi and Hashempour (2022) studied a new measure called interval weighted cumulative residual extropy. For continuous non-negative random variable X the weighted cumulative residual extropy (WCRJ) is defined as:

$$\xi^1 J(X) = -\frac{1}{2}\int_0^\infty x\bar{F}_X^2(x)\,dx. \tag{1.4}$$

Tahmasebi and Toomaj (2022) introduced negative cumulative extropy. For continuous non-negative random variable X the negative cumulative extropy (NCEX) can be defined as:

$$\mathcal{CJ}(X) = \frac{1}{2}\int_0^\infty (1 - F_X^2(x))\,dx. \tag{1.5}$$

Chaudhary *et al.* (2023) proposed a new measure on general weighted negative cumulative extropy. For absolutely continuous non-negative random variable X, with weight $w(x) > 0$, the general weighted negative cumulative extropy (GWNCJ) is defined as:

$$\bar{\eta}^w J(X) = \frac{1}{2}\int_0^\infty w(x)\,(1 - F_X^2(x))\,dx. \tag{1.6}$$

This paper is organized in the following manner. In Section 2 we introduce the generalised weighted cumulative past extropy (m-WCPJ) and study some of its properties. In Section 3 some bounds and inequalities are derived. In Section 4, we study the characterization of m-WCPJ based on the largest-order statistic. Conditional WCPJ and some of its properties are discussed in Section 5. Section 6, provides the empirical estimator of GWCPJ. Application of empirical estimator was discussed in Section 7. In Section 8 a statistical hypothesis testing procedure is performed to assess the goodness of fit of a sample dataset to a standard uniform distribution. Power of test is observed in Section 9. In Section 10 we introduce the weighted negative cumulative extropy and relationship of symmetric distribution with weighted negative cumulative extropy (WNCEX) and weighted cumulative past extropy (WCPJ) was derived. Application of estimator for symmetric test of dataset was discussed in Section 11. In the end, Section 12 concludes this paper.

2. WEIGHTED CUMULATIVE PAST EXTROPY

Balakrishnan *et al.* (2020) and Bansal and Gupta (2021) independently introduced the weighted extropy. Mohammadi and Hashempour (2022) studied a new measure called interval weighted cumulative past extropy. Weighted cumulative past extropy (WCPJ) is an information measure, which is a generalization of cumulative past extropy. In this section, we study the properties of WCPJ.

Definition 2.1. Let X be a non-negative absolutely continuous random variable with cdf F_X. We define the m-WCPJ of X by

$$\bar{\xi}^m J(X) = \frac{-1}{2}\int_0^\infty x^m F_X^2(x)\,dx. \tag{2.1}$$

The introduction explains that, like the cumulative past extropy (CPJ), the weighted cumulative past extropy (WCPJ) also has a value of negative infinity for any random variable with an unbounded range of possible values. Therefore, we must limit the definition of WCPJ to random variables with a limited range of values. If we consider a non-negative random variable X with a bounded range of possible values denoted by B, then the WCPJ of X is defined as follows:

$$\bar{\xi}^m J(X) = \frac{-1}{2}\int_0^{\sup B} x^m F_X^2(x)\,dx. \tag{2.2}$$

Let us consider some examples.

Example 2.1. Let X has $U[a,b]$ distribution. Then CRJ and WCRJ of the uniform distribution are

$$\xi J(X) = -\frac{b-a}{6}, \qquad \text{and,} \qquad \xi^1 J(X) = \frac{a-b}{24}(3a+b),$$

respectively. Then CPJ and m-WCPJ of the uniform distribution are

$$\begin{aligned}\bar{\xi} J(X) &= -\frac{b-a}{6}, \quad \text{and,}\\ \bar{\xi}^m J(X) &= -\Big(b^{m+1}\cdot(b-a)^2 m^2 + b^{m+1}\cdot(b-a)(3b-5a)\,m\\ &\quad + 2b^{m+1}\cdot\left(b^2-3ab+3a^2\right) - 2a^{m+3}\Big)\Big/\Big(2(b-a)^2(m+1)(m+2)(m+3)\Big),\end{aligned}$$

respectively. Note that for $m=1$,

$$\bar{\xi}^1 J(X) = \left(\frac{a+3b}{4}\right)\bar{\xi} J(X) = \left(\frac{E(X)+b}{2}\right)\bar{\xi} J(X).$$

If $\frac{E(X)+b}{2} > 1$, then $\bar{\xi}^1 J(X) < \bar{\xi} J(X)$, and if $\frac{E(X)+b}{2} < 1$, then $\bar{\xi}^1 J(X) > \bar{\xi} J(X)$.

Example 2.2. Let X follow power-law distribution with pdf $f_X(x) = \lambda x^{\lambda-1}$, $x \in (0,1)$, $\lambda > 1$. The CRJ and $WCRJ$ of the distribution are

$$\xi J(X) = -\frac{\lambda^2}{(\lambda+1)(2\lambda+1)}, \qquad \text{and,} \qquad \xi^1 J(X) = -\frac{\lambda^2}{4(\lambda+1)(\lambda+2)},$$

respectively. Note that

$$\xi^1 J(X) = \left(\frac{2\lambda+1}{4(\lambda+2)}\right)\xi J(X) = \left(\frac{2(\lambda+1)\,E(X)+1}{4(\lambda+2)}\right)\xi J(X).$$

For $\lambda = -\frac{7}{2}$, $\xi^1 J(X) = \xi J(X)$. The CPJ and $WCPJ$ of the distribution are

$$\bar{\xi} J(X) = -\frac{1}{2(2\lambda+1)}, \qquad \text{and,} \qquad \bar{\xi}^m J(X) = -\frac{1}{4\lambda+2m+2}.$$

We conclude that for $m=1$, $\bar{\xi}^1 J(X) = \frac{2\lambda+1}{2\lambda+2}\,\bar{\xi} J(X)$. Also, $\bar{\xi}^1 J(X) = -\frac{\lambda+2}{4\lambda(\lambda+1)}\,E(X^2)$.

Theorem 2.1. *Let X be a non-negative continuous random variable with bounded support B for m-WCPJ, $\bar{\xi}^m J(X)$. Then we have*

$$\bar{\xi}^m J(X) = \frac{-1}{2}\,E(G_F(X)),$$

where $G_F(t) = \int_t^{\sup B} x^m F_X(x)\,dx$.

Proof: Using equation (2.2) and Fubini's theorem, we have

$$\begin{aligned}\bar{\xi}^m J(X) &= \frac{-1}{2}\int_0^{\sup B} x^m F_X^2(x)\,dx = \frac{-1}{2}\int_0^{\sup B} x^m F_X(x)\left(\int_0^x f_X(t)\,dt\right)dx \\ &= \frac{-1}{2}\int_0^{\sup B} f_X(t)\left(\int_t^{\sup B} x^m F_X(x)\,dx\right)dt = \frac{-1}{2}\int_0^{\sup B} f_X(t)\,G_F(t)\,dt \\ &= \frac{-1}{2}\,E(G_F(X)).\end{aligned}$$

□

Now we see the effect of linear transformation on WCPJ in the following proposition.

Proposition 2.1. *Let X be a non-negative random variable. If $Y = aX + b$, $a > 0$, $b \geq 0$, then*

$$\bar{\xi}^m J(Y) = \sum_{i=0}^{m}\binom{m}{i} a^i b^{m-i}\,\bar{\xi}^{m-i} J(X).$$

Proof: The proof holds using (2.2) and noting that $F_Y(y) = F_X\left(\frac{y-b}{a}\right)$, $y > b$. □

Here we provide an upper bound for WCPJ in terms of extropy.

Theorem 2.2. *Let X be a random variable with pdf $f_X(\cdot)$ and extropy $J(X)$, with bounded support B. Then*

$$\bar{\xi}^m J(X) \leq C^* \exp\{2J(X)\}, \tag{2.3}$$

where $C^* = \frac{-1}{2}\exp\{E\big(\log(X^m F_X^2(X))\big)\}$, and $\exp\{x\} := e^x$.

Proof: Using the log-sum inequality, we have

$$\int_0^{\sup B} f_X(x) \log\left(\frac{f_X(x)}{x^m F_X^2(x)}\right) dx \geq -\log\left(\int_0^{\sup B} x^m F_X^2(x)\, dx\right).$$

Then it follows that

$$\int_0^{\sup B} f_X(x) \log f_X(x)\, dx - \int_0^{\sup B} f_X(x) \log\big(x^m F_X^2(x)\big)\, dx \geq -\log\left(\int_0^{\sup B} x^m F_X^2(x)\, dx\right).$$

Note that $\log f < f$, hence

$$\begin{aligned} -\int_0^{\sup B} f_X^2(x)\, dx + \int_0^{\sup B} f_X(x) \log\big(x^m F_X^2(x)\big)\, dx &= 2J(X) + E\big(\log\big(X^m F_X^2(X)\big)\big) \\ &\leq \log\big(-2\bar{\xi}^m J(X)\big). \end{aligned} \tag{2.4}$$

Exponentiating both sides of (2.4), we have

$$\bar{\xi}^m J(X) \leq \frac{-1}{2} \exp\big\{2J(X) + E\big(\log\big(X^m F_X^2(X)\big)\big)\big\}$$

Hence the result. □

Theorem 2.3. *X is degenerate, if and only if, $\bar{\xi}^m J(X) = 0$.*

Proof: Suppose X be degenerate at point c, then by using the definition of degenerate function and $\bar{\xi}^m J(X)$, we have $\bar{\xi}^m J(X) = 0$. Now consider $\bar{\xi}^m J(X) = 0$, i.e., $\int_0^\infty x^m F_X^2(x)\, dx = 0$. Noting that the integrand in the above integral is non-negative, we have $F_X(x) = 0$, for almost all $x \in S$, where S denotes the support of random variable X, i.e., it is 0 in $\inf S$ and then 1. □

3. SOME INEQUALITIES

This section deals with obtaining the lower and upper bounds for WCPJ.

Remark 3.1. Consider X to be a non-negative random variable. Then

$$\bar{\xi}^m J(X) \geq \frac{-1}{2} \int_0^{\sup B} x^m F_X(x)\, dx. \tag{3.1}$$

Proposition 3.1. *Consider a non-negative continuous random variable X having cdf $F_X(\cdot)$ and support $[a, \sup B)$, $a > 0$. Then*

$$\bar{\xi}^m J(X) \leq a^m \bar{\xi} J(X). \tag{3.2}$$

Proof: Note that

$$\begin{aligned}\int_a^{\sup B} x^m F_X^2(x)\,dx &\geq a^m \int_a^{\sup B} F_X^2(x)\,dx,\\ \frac{-1}{2}\int_a^{\sup B} x^m F_X^2(x)\,dx &\leq \frac{-a^m}{2}\int_a^{\sup B} F_X^2(x)\,dx,\\ \bar{\xi}^m J(X) &\leq a^m \bar{\xi} J(X).\end{aligned}$$

□

Corollary 3.1. *Let X be a continuous random variable with cdf F that takes values on $[0, b]$ where b is finite. Then,*

1. $\bar{\xi}^m J(X) \leq \frac{-1}{2(m+1)}\left(b^{m+1} - E(X^{m+1})\right)\left[\log\left(\frac{b^{m+1}-E(X^{m+1})}{b^{m+1}}\right) - 1\right]$,
2. $\bar{\xi}^m J(X) \geq b^m \bar{\xi} J(X)$.

Proof: Using log-sum inequality, we have

$$\begin{aligned}\int_0^b F_X(t)\,t^m \log(F_X(t))\,dt &\geq \int_0^b F_X(t)\,t^m\,dt\ \log\left(\frac{\int_0^b F_X(t)\,t^m\,dt}{\int_0^b t^m\,dt}\right)\\ &= \left(\frac{b^{m+1} - E(X^{m+1})}{m+1}\right)\log\left(\frac{b^{m+1} - E(X^{m+1})}{b^{m+1}}\right).\end{aligned}$$

Also note that $\log F_X(t) \leq F_X(t) - 1$, then

$$\begin{aligned}\int_0^b F_X(t)\,t^m \log F_X(t)\,dt &\leq -2\bar{\xi}^m J(X) - \int_0^b t^m F_X(t)\,dt\\ &= -2\bar{\xi}^m J(X) - \left(\frac{b^{m+1} - E(X^{m+1})}{m+1}\right).\end{aligned}$$

Now, using the above two inequalities, the first part follows. The second part can be verified easily. □

Consider two random variables X and Y having cdfs F and G, respectively. Then $X \leq_{\text{st}} Y$ whenever $F(x) \geq G(x)$, $\forall x \in \mathbb{R}$, where the notation $X \leq_{\text{st}} Y$ means that X is less than or equal to Y in usual stochastic order. One may refer Shaked and Shanthikumar (2007) for detail of stochastic ordering. In the following proposition, we show the ordering of WCPJ is implied by the usual stochastic order.

Proposition 3.2. *Let X_1 and X_2 be non-negative continuous random variables. If $X_1 \leq_{\text{st}} X_2$, then $\bar{\xi}^m J(X_1) \leq \bar{\xi}^m J(X_2)$.*

Proof: Using $X_1 \leq_{\text{st}} X_2$ and (2.2), the result follows. □

4. WCPJ BASED ON LARGEST-ORDER STATISTIC

Let $X_1, ..., X_n$ be a random sample from a random variable X having absolutely continuous cdf $F_X(x)$ and pdf $f_X(x)$. Then $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ be the ordered statistics to random sample $X_1, ..., X_n$. In the following, we obtain the WCPJ of the largest-order statistic. The WCPJ of the n-th-order statistic is

$$\bar{\xi}^m J(X_{n:n}) = -\frac{1}{2}\int_0^{\sup B} x^m F^2_{X_{n:n}}(x)\, dx, \tag{4.1}$$

where $F^2_{X_{n:n}}(x) = F^{2n}_X(x)$. Using transformation $u = F_X(x)$ in (4.1),

$$\bar{\xi}^m J(X_{n:n}) = -\frac{1}{2}\int_0^1 \frac{u^{2n}[F^{-1}_X(u)]^m}{f_X(F^{-1}_X(u))}\, du, \tag{4.2}$$

where $F^{-1}_X(x)$ is the inverse function of $F_X(x)$.

Example 4.1. Let X have the uniform distribution on (0,1) with pdf $f_X(x) = 1$, $x \in (0,1)$. Then $F^{-1}_X(u) = u$, $u \in (0,1)$ and $f_X(F^{-1}_X(u)) = 1$, $u \in (0,1)$, hence $\bar{\xi}^m J(X_{n:n}) = \dfrac{-1}{2(2n+m+1)}$.

Example 4.2. Let X follow power-law distribution with pdf $f_X(x) = \lambda x^{\lambda-1}$, $\lambda > 1$, $x \in (0,1)$. Then $F^{-1}_X(u) = u^{\frac{1}{\lambda}}$, $u \in (0,1)$ and $f_X(F^{-1}_X(u)) = \lambda u^{\frac{\lambda-1}{\lambda}}$, $u \in (0,1)$, hence $\bar{\xi}^m J(X_{n:n}) = \dfrac{-1}{2(2n\lambda+m+1)}$.

Remark 4.1. Consider $\Lambda = \bar{\xi}^m J(X_{n:n}) - \bar{\xi}^m J(X)$. Since $F^{2n}_X(x) \leq F^2_X(x)$, hence $\Lambda \geq 0$.

For the proof of Theorem 4.1, we need the following lemma.

Lemma 4.1 (Lemma 4.1 of Hashempour *et al.*, 2022)**.** *Let g be a continuous function with support $[0,1]$, such that $\int_0^1 g(y)y^m dy = 0$, for $m \geq 0$, then $g(y) = 0$, $\forall\ y \in [0,1]$.*

Theorem 4.1. *Let $X_1, ..., X_n$ and $Y_1, ..., Y_n$ be two non negative random samples from continuous cdfs $F(x)$ and $G(x)$, respectively with common bounded support. Then $F(x) = G(x)$ if and only if $\bar{\xi}^m J(X_{n:n}) = \bar{\xi}^m J(Y_{n:n})$, for all n.*

Proof: The necessary condition is trivial. Hence, it remains to prove the sufficient part. If $\bar{\xi}^m J(X_{n:n}) = \bar{\xi}^m J(Y_{n:n})$, then we have

$$-\frac{1}{2}\int_0^1 u^{2n}\left(\frac{[F^{-1}(u)]^m}{f(F^{-1}(u))} - \frac{[G^{-1}(u)]^m}{g(G^{-1}(u))}\right) du = 0.$$

By using Lemma 4.1, it follows that

$$\frac{[F^{-1}(u)]^m}{f(F^{-1}(u))} = \frac{[G^{-1}(u)]^m}{g(G^{-1}(u))} \implies [F^{-1}(u)]^m \frac{dF^{-1}(u)}{du} = [G^{-1}(u)]^m \frac{dG^{-1}(u)}{du},\ u \in [0,1],$$

since $\dfrac{dF^{-1}(u)}{du} = \dfrac{1}{f(F^{-1}(u))}$. Hence it follows $F^{-1}(u) = G^{-1}(u)$, $u \in [0,1]$. Thus the proof is completed. □

5. CONDITIONAL WEIGHTED PAST EXTROPY

Now we consider the conditional weighted cumulative past extropy (CWCPJ). Consider a random variable Z on probability space $(\Omega, \mathbb{A}, P)$ such that $E|Z| < \infty$. The conditional expectation of Z given sub σ-field $\mathbb{G}$, where $\mathbb{G} \subseteq \mathbb{A}$, is denoted by $E(Z|\mathbb{G})$. For the random variable $I_{(Z\leq z)}$, we denote $E(I_{(Z\leq z)}|\mathbb{G})$ by $F_Z(z|\mathbb{G})$.

Definition 5.1. For a non-negative random variable X with bounded support B, given σ-field $\mathbb{G}$, the CWCPJ $\bar{\xi}^m J(X|\mathbb{G})$ is defined as

$$\bar{\xi}^m J(X|\mathbb{G}) = \frac{-1}{2}\int_0^{\sup B} x^m F_X^2(x|\mathbb{G})\, dx. \tag{5.1}$$

Now we assume that the random variables are continuous and non-negative.

Lemma 5.1. *If $\mathbb{G}$ is a trivial σ-field, then $\bar{\xi}^m J(X|\mathbb{G}) = \bar{\xi}^m J(X)$.*

Proof: Since here $F_X(x|\mathbb{G}) = F_X(x)$, then the proof follows. □

Proposition 5.1. *If $X \in L^p$ for some $p > 2$, then $E[\bar{\xi}^m J(X|\mathbb{G})|\mathbb{G}^*] \leq \bar{\xi}^m J(X|\mathbb{G}^*)$, provided that $\mathbb{G}^* \subseteq \mathbb{G}$.*

Proof: Consider

$$\begin{aligned}
E\big[\bar{\xi}^m J(X|\mathbb{G})|\mathbb{G}^*\big] &= \frac{-1}{2}\int_0^{\sup B} x^m E\Big(\big[P(X\leq x|\mathbb{G})\big]^2\big|\mathbb{G}^*\Big)\, dx \\
&\leq \frac{-1}{2}\int_0^{\sup B} x^m \big[E(P(X\leq x|\mathbb{G})|\mathbb{G}^*)\big]^2\, dx \\
&= \frac{-1}{2}\int_0^{\sup B} x^m \big[E\big(E(I_{(X\leq x)}|\mathbb{G})|\mathbb{G}^*\big)\big]^2\, dx \\
&= \frac{-1}{2}\int_0^{\sup B} x^m \big[E\big(I_{(X\leq x)}|\mathbb{G}^*\big)\big]^2\, dx \\
&= \frac{-1}{2}\int_0^{\sup B} x^m F_X^2(x|\mathbb{G}^*)\, dx \\
&= \bar{\xi}^m J(X|\mathbb{G}^*),
\end{aligned}$$

where the second step follows using the Jensen's inequality for convex function $\phi(x) = x^2$. Hence the result. □

In the following theorem, we investigate the relationship between conditional extropy and $\bar{\xi}^m J(X|\mathbb{G})$.

Theorem 5.1. *Let $\bar{\xi}^m J(X|\mathbb{G})$ be conditional past extropy. Then we have*

$$\bar{\xi}^m J(X|\mathbb{G}) \leq B^* \exp\{2J(X|\mathbb{G})\}, \tag{5.2}$$

where $B^* = \frac{-1}{2}\exp\big\{E\big(\log(X^m F_X^2(X))|\mathbb{G}\big)\big\}$.

Proof: The proof is on the similar lines as of Theorem 2.2, hence omitted. □

Theorem 5.2. *For a random variable X with bounded support B and σ-field $\mathbb{G}$, we have*

$$E\big(\bar{\xi}^m J(X|\mathbb{G})\big) \leq \bar{\xi}^m J(X), \tag{5.3}$$

and the equality holds if and only if X is independent of $\mathbb{G}$.

Proof: If in Proposition 5.1, $\mathbb{G}^*$ is trivial σ-field, then (5.3) can be easily obtained. Now assume that X is independent of $\mathbb{G}$, then

$$\begin{aligned} & F_X(x|\mathbb{G}) = F_X(x) \\ \implies\ & \bar{\xi}^m J(X|\mathbb{G}) = \bar{\xi}^m J(X). \end{aligned} \tag{5.4}$$

On taking expectation to both sides of (5.4), we get equality in (5.3). Conversely, assume that equality in (5.3) holds. It is sufficient to show that $F_X(x|\mathbb{G}) = F_X(x)$, to prove independence between X and σ-field $\mathbb{G}$. Take $U = F_X(x|\mathbb{G})$, and since the function $\phi(u) = u^2$ is convex hence $E(U^2) \geq E^2(U) = F_X^2(x)$, and also due to equality in (5.3), we have

$$\int_0^{\sup B} x^m E(U^2)\, dx = \int_0^{\sup B} x^m F_X^2(x)\, dx = \int_0^{\sup B} x^m E^2(U)\, dx.$$

Hence $E(U^2) = E^2(U)$. Now using the Corollary 8.1 of Hashempour *et al.* (2022), we have $F_X(x|\mathbb{G}) = F_X(x)$. Hence the proof. □

For the Markov property for non-negative random variables X, Y and Z, we have the following proposition.

Proposition 5.2. *Let $X \to Y \to Z$ be a Markov chain, then*

$$\bar{\xi}^m J(Z|X,Y) = \bar{\xi}^m J(Z|Y) \tag{5.5}$$

and

$$E\big(\bar{\xi}^m J(Z|Y)\big) \leq E\big(\bar{\xi}^m J(Z|X)\big). \tag{5.6}$$

Proof: By the definition of $\bar{\xi}^m J(Z|X,Y)$ and using the Markov property, (5.5) holds. Now letting $\mathbb{G}^* = \sigma(X)$, $\mathbb{G} = \sigma(X,Y)$ and $X = Z$ in Proposition 5.1, we have

$$\bar{\xi}^m J(Z|X) \geq E\big(\bar{\xi}^m J(Z|X,Y)|X\big). \tag{5.7}$$

Taking expectation on both sides of (5.7), we have

$$\begin{aligned} E\big(\bar{\xi}^m J(Z|X)\big) &\geq E\big(E\big(\bar{\xi}^m J(Z|X,Y)|X\big)\big) \\ &= E\big(\bar{\xi}^m J(Z|X,Y)\big) \\ &= E\big(\bar{\xi}^m J(Z|Y)\big), \end{aligned}$$

where the last equality holds using (5.5). Hence the result (5.6) holds. □

6. EMPIRICAL GWCPJ

We want to create an estimator for the general Weighted Cumulative Past Extropy (WCPJ) using the empirical m-WCPJ. To do this, we assume that we have a random sample of n non-negative, continuous, independent, and identically distributed random variables $X_1, ..., X_n$ from a population with a cumulative distribution function (cdf) $F_X(x)$. We will use the "plug-in" method to define the estimator by replacing the unknown quantities in the formula with their empirical counterparts based on the sample. Specifically, we will define the empirical general weighted cumulative past extropy as

$$\bar{\xi}_n^m J(X) = \frac{-1}{2}\int_0^\infty x^m F_n^2(x)\,dx, \tag{6.1}$$

where $F_n(x)$ is the empirical distribution function. Suppose we have a random sample of n observations, and let $X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$ be the ordered statistics to random sample $X_1, ..., X_n$. We can express the value of the empirical general weighted cumulative past extropy $\bar{\xi}_n^m J(X)$ in terms of these ordered statistics. In other words, we can rewrite the formula for $\bar{\xi}_n^m J(X)$ in a way that involves the ordered statistics of the sample.

The empirical measure of $\bar{\xi}^m J(X) = \frac{-1}{2}\int_0^\infty x^m F_X^2(x)\,dx$ is obtained as

$$\begin{aligned}\bar{\xi}_n^m J(X) &= -\frac{1}{2}\int_0^\infty x^m F_n^2(x)\,dx \\ &= -\frac{1}{2}\sum_{i=1}^{n-1}\int_{X_{i:n}}^{X_{i+1:n}}\left(\frac{i}{n}\right)^2 x^m\,dx \\ &= -\frac{1}{2(m+1)}\sum_{i=1}^{n-1}\left(X_{i+1:n}^{m+1} - X_{i:n}^{m+1}\right)\left(\frac{i}{n}\right)^2. \end{aligned} \tag{6.2}$$

7. APPLICATION

To test the uniformity of a random sample $X_1, X_2, ..., X_n$, we use the empirical weighted cumulative past extropy $\bar{\xi}_n^m J(X)$ obtained in equation (6.2) as a test statistic. Before discussing this test statistic, we first need to understand a property of the uniform distribution, which is defined on the interval $(0, 1)$. For any random variable X with a cumulative distribution function F, and for any probability value p in the interval $(0, 1)$, we define the function $\phi_p^m J(F)$ as

$$\phi_p^m J(F) = -\frac{1}{2}\int_0^p x^m F^2(x)\,dx, \qquad m \ge 0. \tag{7.1}$$

Example 7.1. If F_0 is cdf of uniform random variable on interval $(0, 1)$, then $\phi_p^m J(F_0) = -\frac{p^{m+3}}{2(m+3)}$ and in particular, when $m = 1$, $\phi_p^1 J(F_0) = -\frac{p^4}{8}$ and when $m = 0$, $\phi_p^0 J(F_0) = -\frac{p^3}{6}$.

In view of Lemma 4.1 and Example 7.1, the following proposition is obtained.

Proposition 7.1. *Let F_0 be cdf of uniform random variable on interval $(0, 1)$. Suppose that for a cdf F in the class of cdfs defined on interval $(0, 1)$, $\phi_p^m J(F) = \phi_p^m J(F_0)$. Then $F(x) = F_0(x)$, almost everywhere.*

Proof: Let $\phi_p^m J(F) = \phi_p^m J(F_0)$, using (7.1) we get

$$\int_0^p x^m \big(F^2(x) - F_0^2(x)\big)\, dx = 0, \qquad \forall\ p \in (0,1).$$

It is known that the $(0,p)$ generate the Borel σ-algebra of $\Omega = (0,1]$. Therefore, we can write

$$\int_B x^m \big(F^2(x) - F_0^2(x)\big)\, dx = 0, \qquad \forall\ B \subseteq (0,1].$$

By Lemma 4.1, we conclude that $F(x) = F_0(x)$, almost everywhere.

Thus, $\phi_p^m J(F)$ for $p \in (0,1)$ is uniquely determined by the uniform distribution in the sense that for some cdfs defined on $(0,1)$, they take a value less than $-\frac{p^{m+3}}{2(m+3)}$ and for some of them, they take more than $-\frac{p^{m+3}}{2(m+3)}$, and only for the standard uniform distribution, we have $\phi_p^m J(F_0) = -\frac{p^{m+3}}{2(m+3)}$. □

8. THE TEST OF UNIFORMITY

We aim to develop a non-parametric test for the uniform goodness of fit problem based on the GWCPJ measure. Specifically, we want to test if a random sample of size n from an unknown distribution F is uniformly distributed on the interval $(0,1)$. Our test statistic is $\bar{\xi}_n^m J(X)$, which is computed using the empirical quantiles of the sample. We choose $m = 1$ for computational simplicity, but the procedure is the same for other values of m. To determine the critical region, we need to calculate $G_1(\alpha)$ and $G_2(\alpha)$, which are the lower and upper bounds for the test statistic, respectively, based on the empirical distribution of $\bar{\xi}_n^m J(X)$ under the standard uniform distribution. If $\bar{\xi}_n^m J(X)$ falls outside this interval, we reject the null hypothesis $H_0 : F = F_0$ in favor of the alternative hypothesis $H_1 : F \neq F_0$. We estimate $G_1(\alpha)$ and $G_2(\alpha)$ using the 0.025-th and 0.975-th quantiles of the empirical distribution of $\bar{\xi}_n^m J(X)$, respectively, based on 100,000 replications. We report the values of $G_1(\alpha)$ and $G_2(\alpha)$ for different sample sizes in Table 1 at a significance level of 5% ($\alpha = 0.05$).

Table 1: Values of $G_1(\alpha)$ and $G_2(\alpha)$, for $\alpha = 0.05$.

n	20	30	40	50
$G_1(\alpha)$	−0.144891	−0.144001	−0.143339	−0.142433
$G_2(\alpha)$	−0.066024	−0.078560	−0.085851	−0.090792

Similarly we can obtain for $m = 2$. We report the values of $G_1(\alpha)$ and $G_2(\alpha)$ for different sample sizes in Table 2 at a significance level of 5% ($\alpha = 0.05$).

Table 2: Values of $G_1(\alpha)$ and $G_2(\alpha)$, for $\alpha = 0.05$.

n	20	30	40	50
$G_1(\alpha)$	−0.109538	−0.110579	−0.110441	−0.110317
$G_2(\alpha)$	−0.047356	−0.059339	−0.066270	−0.070851

9. POWER OF THE TEST

This section discusses the power of the proposed test statistic against alternative Beta$(1.5, 1.5)$ and Beta$(1.0, 1.0)$ distributions at the significance level of $\alpha = 0.05$ and $m = 1$, for different sample sizes. The Beta distribution is chosen as an alternative because it has support $(0, 1)$ like the standard uniform distribution. Previous researchers have also examined the power of the WCPJ test against the Beta$(1.5, 1.5)$ distribution for testing uniformity. The proposed test is found to be more powerful for large sample sizes, and the power increases as the sample size increases. As Beta$(1, 1)$ is equivalent to a standard uniform distribution, the power of the test is equivalent to the size of the test. The power against Beta$(1.5, 1.5)$ and Beta$(1.0, 1.0)$ for sample sizes of 40, 50, 100 and 150 are presented in Table 3.

Table 3: Power of test against different alternative for $\alpha = 0.05$ and $m = 1$.

n	40	50	100	150
Beta$(1.5, 1.5)$	0.08527	0.08509	0.08363	0.09927
Beta$(1.0, 1.0)$	0.05046	0.05131	0.051	0.04934

10. WEIGHTED NEGATIVE CUMULATIVE EXTROPY

Tahmasebi and Toomaj (2022) defined negative cumulative extropy and Chaudhary *et al.* (2023) defined general weighted negative cumulative extropy.

Definition 10.1. Let X be a non-negative absolutely continuous random variable with cdf F_X. We define the WNCEX of X by

$$\xi^w J(X) = \frac{1}{2}\int_0^\infty w(x)\,(1 - F_X^2(x))\,dx. \tag{10.1}$$

If $w(x) = x^m$, $m > 0$, then we represent $\xi^w J(X)$ as $\xi^m J(X)$.

The Lemma from Fashandi and Ahmadi (2012) serves as a foundational tool for creating various descriptions of symmetric distributions.

Lemma 10.1 (Fashandi and Ahmadi, 2012)**.** *Let X be a continuous random variable with pdf f_X and cdf F_X with support S_X. Then, the identity*

$$f_X(F_X^{-1}(u)) = f_X(F_X^{-1}(1-u)),$$

for almost all $u \in (0, \frac{1}{2})$ if and only if there exists a constant k such that $F_X(k-x)+F_X(k+x) = 1$ for all $x \in S_X$.

Let $\mathbb{C}$ denote the class of all continuous pdf f_X, having cdf F_X such that

$$f_X(F_X^{-1}(1-u)) \geq (\leq)\ f_X(F_X^{-1}(u)),$$

for all $u \in (0, \frac{1}{2})$. It can be observed using Lemma 10.1 that F is symmetric if and only if $f_X(F_X^{-1}(u)) = f_X(F_X^{-1}(1-u))$ for almost all $u \in (0, \frac{1}{2})$. The class $\mathbb{C}$ is non-empty and includes various distributions.

For any random variable X with an absolute continuous distribution function (df) $F(\cdot)$ and the Lebesgue probability density function (pdf) $f(\cdot)$. Let $l_X = \inf\{x \in \mathbb{R}\colon F(x) > 0\}$, $u_X = \sup\{x \in \mathbb{R}\colon F(x) < 1\}$ and $S_X = (l_X, u_X)$, we can define WNCEX and WCPJ with support S_X as

$$\xi^{w_1} J(X) = \frac{1}{2} \int_{S_X} w_1(x)\,(1 - F_X^2(x))\,dx, \tag{10.2}$$

and

$$\tilde{\xi}^{w_2} J(X) = -\frac{1}{2} \int_{S_X} w_2(x)\,F_X^2(x)\,dx, \tag{10.3}$$

respectively. The following theorem gives a relationship of symmetric distribution with WNCEX and WCPJ.

Theorem 10.1. *Let $w_1(x) = F_X(x)$ in (10.2) and $w_2(x) = \frac{F_X(x)(1+\bar{F}_X(x))}{\bar{F}_X(x)}$ in (10.3), then the following two statements are equivalent for any $F_X \in \mathbb{C}$:*

- **(i)** *random variable X has a symmetric distribution;*
- **(ii)** $\xi^{w_1} J(X) = -\tilde{\xi}^{w_2} J(X)$.

Proof: From (10.2), the WNCEX can be written as

$$\xi^{w_1} J(X) = \frac{1}{2} \int_{S_X} w_1(x)\,(1 - F_X^2(x))\,dx = \frac{1}{2} \int_0^1 \frac{u(1-u^2)}{f_X(F_X^{-1}(u))}\,du, \tag{10.4}$$

and from (10.3) the WCPJ can be written as

$$\tilde{\xi}^{w_2} J(X) = \frac{1}{2} \int_0^1 \frac{u(1-u^2)}{f_X(F_X^{-1}(1-u))}\,du. \tag{10.5}$$

When $\xi^{w_1} J(X) = -\tilde{\xi}^{w_2} J(X)$ holds then from (10.4) and (10.5), we have

$$\int_0^1 u(1-u^2)\,\eta(u)\,du = 0, \tag{10.6}$$

where

$$\eta(u) = \frac{1}{f_X(F_X^{-1}(u))} - \frac{1}{f_X(F_X^{-1}(1-u))}. \tag{10.7}$$

Then as $\eta(1-u) = -\eta(u)$, we can write (10.6) as

$$\int_0^{\frac{1}{2}} \eta(u)\,u(1-u)\,(2u-1)\,du = 0. \tag{10.8}$$

Note that $u(1-u)(2u-1) < 0$ for all $0 \leq u \leq \frac{1}{2}$. Since by assumption $F_X \in \mathbb{C}$, hence from (10.8) we have $f_X(F_X^{-1}(u)) = f_X(F_X^{-1}(1-u))$, so Lemma 10.1 completes the proof. □

Example 10.1. Let X have the uniform distribution on (a,b) with pdf $f_X(x)=\frac{1}{b-a}$, $x\in(a,b)$. For $u\in(a,b)$, $F^{-1}(u)=(b-a)u+a$, $F_X^{-1}(1-u)=(b-a)(1-u)+a$, and $f_X(F_X^{-1}(u))=\frac{1}{b-a}$, $f_X(F_X^{-1}(1-u))=\frac{1}{b-a}$. Then $f_X(F_X^{-1}(u))=f_X(F_X^{-1}(1-u))$. The equality is true in $\mathbb{C}$. It is symmetric.

Example 10.2. Let X follow power-law distribution with pdf $f_X(x)=\lambda x^{\lambda-1}$, $\lambda>1$, $x\in(0,1)$. Then $F_X^{-1}(u)=u^{\frac{1}{\lambda}}$, $u\in(0,1)$ and $f_X(F_X^{-1}(u))=\lambda u^{\frac{\lambda-1}{\lambda}}$, $f_X(F_X^{-1}(1-u))=\lambda(1-u)^{\frac{\lambda-1}{\lambda}}$, $u\in(0,1)$.
If $\lambda>1$ then $f_X(F_X^{-1}(u))\le f_X(F_X^{-1}(1-u))$ for $u\in(0,\frac{1}{2})$.
If $\lambda<1$ then $f_X(F_X^{-1}(u))\le f_X(F_X^{-1}(1-u))$ for $u\in(0,\frac{1}{2})$.
If $\lambda=1$ then $f_X(F_X^{-1}(u))=f_X(F_X^{-1}(1-u))$ for $u\in(0,\frac{1}{2})$.
Thus, power distribution belongs to class $\mathbb{C}$ and for $u\in(0,\frac{1}{2})$, it is symmetric for $\lambda=1$.

11. APPLICATION

Assume that $X_1,X_2,...,X_N$ is a random sample of size N taken from a population with pdf f_X and cdf F_X. As discussed in Section 10, for $w_1(x)=F_X(x)$ in (10.2) and $w_2(x)=\frac{\bar{F}_X(x)(1+\bar{F}_X(x))}{F_X(x)}$ in (10.3), the $\xi^{w_1}J(X)$ and $\tilde{\xi}^{w_2}J(X)$ can be written as

$$\xi^{w_1}J(X)=\frac{1}{2}\int_0^1\frac{u(1-u^2)}{f_X(F_X^{-1}(u))}\,du \qquad \text{and} \qquad \tilde{\xi}^{w_2}J(X)=-\frac{1}{2}\int_0^1\frac{u(1-u)(2-u)}{f_X(F_X^{-1}(u))}\,du.$$

We note from Theorem 10.1 that $\xi^{w_1}J(X)+\tilde{\xi}^{w_2}J(X)=0$ if and only if X is symmetric distribution. This motivates us that

$$\Delta=\xi^{w_1}J(X)+\tilde{\xi}^{w_2}J(X)$$

can be used to test whether X is symmetric distribution. We can write:

$$\begin{aligned}\Delta&=\frac{1}{2}\int_0^1\left[\frac{u(1-u^2)-u(1-u)(2-u)}{f_X(F_X^{-1}(u))}\right]du\\&=\frac{1}{2}\int_0^1(-2u^3+3u^2-u)\left[\frac{d}{du}F_X^{-1}(u)\right]du.\end{aligned}$$

Small or large values of Δ can be indicative of asymmetry, prompting the rejection of the null hypothesis of symmetry. Therefore, we introduce a symmetry test relying on the sample estimator of Δ. Drawing inspiration from Vasicek (1976) and adopting a methodology akin to Vasicek (1976), Park (1999), Xiong *et al.* (2021), and Jose and Sathar (2022), we formulate an estimator for Δ in the following manner:

$$\widehat{\Delta}=-\frac{1}{2N}\sum_{i=1}^{N}\left[2\left(\frac{i}{N+1}\right)^3-3\left(\frac{i}{N+1}\right)^2+\left(\frac{i}{N+1}\right)^3\right]\frac{X_{i+m:N}-X_{i-m:N}}{2m/N}.$$

Here window size m is a positive integer less than $\frac{N}{2}$ and $X_{r:N}$ denotes r-th order statistics from sample $X_1,X_2,...,X_N$. If $i+m>N$ then we consider $X_{i+m:N}=X_{N:N}$ and if $i+m<1$ then we consider $X_{i-m:N}=X_{1:N}$.

Dataset 1, sourced from Montgomery *et al.* (2021), conforms to a normal distribution, representing a symmetric model. The dataset, consists of values such as:

15.5, 23.75, 8.0, 17.0, 5.5, 19.0, 24.0, 2.5, 7.5, 11.0, 13.0,
3.75, 25.0, 9.75, 22.0, 18.0, 6.0, 12.5, 2.0, 21.5.

The data set was also utilized by Jose and Sathar (2022) and Gupta and Chaudhary (2024) in their proposed symmetry test, affirming its normal distribution and symmetry. This fact is verified by our test. The absolute value of the test statistics $\widehat{\Delta}$ is 0.03246 with an estimated p-value 0.1054 when window size $m = 2$ and sample size $N = 20$. Remarkably, our symmetry test based on $\widehat{\Delta}$ does not lead to the rejection of the null hypothesis. This outcome aligns with the dataset's adherence to a normal distribution, establishing it as a suitable model for the data.

Dataset 2, sourced from Qiu and Jia (2018), represents active repair times (in hours) for an airborne communication transceiver. The dataset comprises values such as:

0.2, 0.3, 0.5, 0.5, 0.5, 0.5, 0.6, 0.6, 0.7, 0.7, 0.7, 0.8, 0.8, 1.0, 1.0, 1.0,
1.0, 1.1, 1.3, 1.5, 1.5, 1.5, 1.5, 2.0, 2.0, 2.2, 2.5, 3.0, 3.0, 3.3, 3.3,
4.0, 4.0, 4.5, 4.7, 5.0, 5.4, 5.4, 7.0, 7.5, 8.8, 9.0, 10.3, 22.0, 24.5.

It has been employed by Xiong *et al.* (2021) in their proposed test of symmetry. As per Qiu and Jia (2018), Dataset 2 can be adequately modeled by the inverse Gaussian (IG) distribution. It's worth noting that the IG distribution is inherently asymmetric, as highlighted by Xiong *et al.* (2021) and Qiu and Jia (2018). Our conducted test affirms this asymmetry, revealing an absolute value of the test statistic $\widehat{\Delta}$ as 0.5679, along with an estimated p-value of 0.0. These computations were conducted under the conditions of a window size $m = 2$ and a sample size N of 45. Remarkably, our symmetry test based on $\widehat{\Delta}$ successfully rejects the null hypothesis. This rejection aligns with the inherent asymmetry of the dataset, as indicated by its fit to the inverse Gaussian distribution.

Now consider the Dataset 3, from Lawless (2011), which represents the quantity of 1000 cycles to failure for electrical appliances in a life test, with values such as:

0.014, 0.034, 0.059, 0.061, 0.069, 0.080, 0.123, 0.142, 0.165, 0.210, 0.381, 0.464, 0.479,
0.556, 0.574, 0.839, 0.917, 0.969, 0.991, 1.064, 1.088, 1.091, 1.174, 1.270, 1.275, 1.355,
1.397, 1.477, 1.578, 1.649, 1.702, 1.893, 1.932, 2.001, 2.161, 2.292, 2.326, 2.337,
2.628, 2.785, 2.811, 2.886, 2.993, 3.122, 3.248, 3.715, 3.790, 3.857, 3.912, 4.100.

Xiong *et al.* (2021) used this dataset for their proposed test of symmetry. The absolute value of the test statistics $\widehat{\Delta}$ is 0.0486 with an estimated p-value 0.0034 when window size $m = 2$ and sample size $N = 50$. Our test based on $\widehat{\Delta}$ succeeds in rejecting the null hypothesis that the dataset does not follow symmetry.

See Table 4 for the value of test statistics and p-value for different datasets based on the specific window size and sample size of each dataset.

Table 4: Description of models fitted.

Dataset	N	m	$\lvert\widehat{\Delta}\rvert$	p-value
Dataset 1	20	2	0.0343	0.1054
Dataset 2	45	2	0.5679	0.0
Dataset 3	50	2	0.0486	0.0034

When conducting tests at a 5% significance level, a p-value below 0.05 signals an asymmetric nature, while a p-value above 0.05 suggests symmetry in the data. Table 4 confirms that the newly proposed test effectively discerns the symmetry or asymmetry of the distribution within random samples. The observed p-values, at the 5% significance level, indicate that Datasets 2 and 3 exhibit asymmetries in the distribution of their random samples. Conversely, Dataset 1 presents moderate p-values, implying acceptance of symmetry in its distributions. Consequently, we can affirm that the test statistic accurately identifies the presence or absence of symmetry in the distribution of the random variable.

12. CONCLUSION

The introduction section of the paper highlights the various studies and generalizations of the concept of extropy. The paper then proceeds to define and analyze GWCPJ, and conditional WCPJ. To estimate GWCPJ, non-parametric estimators and empirical distribution functions have been proposed. Additionally, upper and lower bounds for WCPJ have been derived, and several results pertaining to WCPJ have been presented.

Furthermore, a new test of uniformity has also been proposed based on GWCPJ, which can be used to determine whether a given dataset follows a uniform distribution. Also we have studied the relationship of symmetric distribution with weighted negative cumulative extropy (WNCEX) and weighted cumulative past extropy (WCPJ) and formulated an estimator for symmetry test of data set.

ACKNOWLEDGMENTS

PKS would like to thank Quality Improvement Program (QIP), All India Council for Technical Education, Government of India (Student Unique Id: FP2200759) for financial assistance. The authors are thankful to the Professor Maria Isabel Fraga Alves Editor-in-Chief, Associate Editor and referees for their valuable suggestions which significantly improved this manuscript.

REFERENCES

Asadi, M. and Zohrevand, Y. (2007). On the dynamic cumulative residual entropy. *Journal of Statistical Planning and Inference*, 137(6):1931–41.

Ash, R. (1990). *Information Theory*. Dover Publications Inc., New York.

Balakrishnan, N., Buono, F., and Longobardi, M. (2020). On weighted extropies. *Communications in Statistics – Theory and Methods*, 1–31.

Bansal, S. and Gupta, N. (2021). Weighted extropies and past extropy of order statistics and k-record values. *Communications in Statistics – Theory and Methods*, 1–24.

Chaudhary, S.K., Gupta, N., and Sahu, P.K. (2023). On general weighted cumulative residual extropy and general weighted negative cumulative extropy. *Statistics*, 57(5):1117–1141.

Cover, T.M. and Thomas, J.A. (2006). *Elements of Information Theory*. 2nd ed., Wiley, Hoboken (NJ).

Di Crescenzo, A. and Longobardi, M. (2009). On cumulative entropies. *J. Statist. Plann. Inference*. 139:4072–4087.

Di Crescenzo, A. and Longobardi, M. (2013). *Stochastic comparisons of cumulative entropies*. In "Stochastic Orders in Reliability and Risk" (H. Li and X. Li, Eds.), Lecture Notes in Statistics, 208, 167–82. Springer, New York.

Drissi, N., Chonavel, T., and Boucher, J.M. (2008). Generalized cumulative residual entropy for distributions with unrestricted supports. *J. Electr. Comput. Eng.*, Article ID 790607:5.

Fashandi, M. and Ahmadi, J. (2012). Characterizations of symmetric distributions based on Rényi entropy. *Stat. Probab. Lett.*, 82(4):798–804.

Gupta, N. and Chaudhary, S.K. (2024). Some characterizations of continuous symmetric distributions based on extropy of record values. *Stat. Papers*, 65:291–308.

Hashempour, M., Kazemi, M.R., and Tahmasebi, S. (2022). On weighted cumulative residual extropy: characterization, estimation and testing. *Statistics*, 56(3):681–698.

Jahanshahi, S.M.A., Zarei, H., and Khammar, A.H. (2020). On cumulative residual extropy. *Probability in the Engineering and Informational Sciences*, 34(4):605–21.

Jose, J. and Sathar, E.I.A. (2022). Symmetry being tested through simultaneous application of upper and lower k-records in extropy. *Journal of Statistical Computation and Simulation*, 92(4):830–846.

Khorashadizadeh, M., Rezaei Roknabadi, A.H., and Mohtashami Borzadaran, G.R. (2013). Doubly truncated (interval) cumulative residual and past entropy. *Statistics and Probability Letters*, 83(5):1464–71.

Krishnan, A.S., Sunoj, S.M., and Nair, N.U. (2020). Some reliability properties of extropy for residual and past lifetime random variables. *Journal of the Korean Statistical Society*, 49(2):457–74.

Kundu, C. (2023). On cumulative residual (past) extropy of extreme order statistics. *Communications in Statistics – Theory and Methods*, 52(16):5848–5865.

Lad, F., Sanfilippo, G., and Agro, G. (2015). Extropy: complementary dual of entropy. *Statistical Science*, 30(1):40–58.

Lawless, J.F. (2011). *Statistical Models and Methods for Lifetime Data*, vol. 362. Wiley, Hoboken.

Mohammadi, M. and Hashempour, M. (2022). On interval weighted cumulative residual and past extropies. *Statistics*, 56(5):1029–1047.

Montgomery, D.C., Peck, E.A., and Vining, G.G. (2021). *Introduction to Linear Regression Analysis*, 6th Edition. John Wiley and Sons, New York.

Navarro, J., del Aguila, Y., and Asadi, M. (2010). Some new results on the cumulative residual entropy. *Journal of Statistical Planning and Inference*, 140(1):310–22.

Noughabi, H.A. and Jarrahiferiz, J. (2019). On the estimation of extropy. *Journal of Nonparametric Statistics*, 31(1):88–99.

Park, S. (1999). A goodness-of-fit test for normality based on the sample entropy of order statistics. *Stat. Prob. Lett.*, 44(4):359–363.

Qiu, G. (2017). The extropy of order statistics and record values. *Statistics and Probability Letters*, 120:52–60.

Qiu, G. and Jia, K. (2018). The residual extropy of order statistics. *Statistics and Probability Letters*, 133:15–22.

Qiu, G. and Jia, K. (2018). Extropy estimators with applications in testing uniformity. *Journal of Nonparametric Statistics*, 30(1):182–96.

Qiu, G., Wang, L., and Wang, X. (2019). On extropy properties of mixed systems. *Probability in the Engineering and Informational Sciences*, 33(3):471–86.

Rao, M., Chen, Y., Vemuri, B.C., and Wang, Fei (2004). Cumulative residual entropy: a new measure of information. *IEEE Transactions on Information Theory*, 50(6):1220–1228.

Rao, M. (2005). More on a new concept of entropy and information. *J. Theor. Probab.*, 18:967–981.

Raqab, M.Z. and Qiu, G. (2019). On extropy properties of ranked set sampling. *Statistics*, 53(1): 210–26.

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

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Tahmasebi, S. and Toomaj, A. (2022). On negative cumulative extropy with applications. *Communications in Statistics – Theory and Methods*, 51(15);5025–5047.

Vasicek, O. (1976). A test for normality based on sample entropy. *J. R. Stat. Soc. Ser. B Methodol.*, 38:54–59.

Xiong, P., Zhuang, W., and Qiu, G. (2021). Testing symmetry based on the extropy of record values. *Journal of Nonparametric Statistics*, 33(1):134–155.

Yang, J., Xia, W., and Hu, T. (2018). Bounds on extropy with variational distance constraint. *Probab. Eng. Inf. Sci.*, 33:1–19.

REVSTAT – Statistical Journal
Volume 24, Number 2, April 2026, 135–155
https://doi.org/10.57805/revstat.v24i2.686

Novel Robust Estimators for the Linear Regression Model with Multicollinearity and Outlier Problems

Authors: Murat Erişoğlu
– Necmettin Erbakan University, Faculty of Science, Statistics Department, Konya, Türkiye
merisoglu@erbakan.edu.tr

Aydin Karakoca ✉
– Necmettin Erbakan University, Faculty of Science, Statistics Department, Konya, Türkiye
akarakoca@erbakan.edu.tr

Abdullah Yurtaslan
– Necmettin Erbakan University, Faculty of Science, Statistics Department, Konya, Türkiye
abdullahyurtaslan@gmail.com

Received: October 2023 Revised: January 2024 Accepted: February 2024

Abstract:

- In this study, we introduce new robust M-estimators based on ridge estimation (M-Ridge) for data sets with both multicollinearity and outlier problems in multiple linear regression analysis. In the proposed approach, the iterative re-weighted least squares (IRLS) algorithm for parameter estimation is implemented based on ridge estimation.The proposed approach also provides a solution to the problem of the optimal ridge estimator selection with M-type estimators. The performance of the proposed estimators is evaluated against other estimators using a Monte Carlo simulation study and a real data application. The estimated mean square error (MSE) and k-fold cross validation are used as performance measures in the Monte Carlo simulation study and the real data application, respectively. The proposed M-Ridge estimators outperformed the other estimators considered in many evaluated instances in both the simulation study and the real data application.

Keywords:

- *k-fold cross-validation; M-estimator; MSE; multicollinearity; outlier; ridge regression.*

AMS Subject Classification:

- 62F35, 62J07.

✉ Corresponding author

1. INTRODUCTION

Regression analysis is a versatile statistical tool that can be used to model the relationship between a dependent variable and one or more independent variables. It is one of the most widely used statistical analyses due to its simplicity and effectiveness. Multiple linear regression model in matrix form is given by

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{1.1}$$

In equation (1.1), $\boldsymbol{y}$ is an $n \times 1$ response vector, X is an $n \times p$ design matrix of known constants, $\boldsymbol{\beta}$ is a $p \times 1$ unknown parameters vector and $\boldsymbol{\varepsilon}$ is an $n \times 1$ stochastic error vector such that $E(\boldsymbol{\varepsilon}) = 0$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$. The ordinary least squares (OLS) method is widely used for estimation of $\boldsymbol{\beta}$. OLS aims to minimize the sum of squares of the error terms. The objective function and OLS estimates are given by respectively

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min} \left(\sum_{i=1}^{n} \varepsilon_i^2 \right) = \underset{\boldsymbol{\beta}}{\arg\min} \left(\sum_{i=1}^{n} (y_i - x_i \boldsymbol{\beta})^2 \right), \tag{1.2}$$

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1} X'\boldsymbol{y}. \tag{1.3}$$

The covariance matrix of OLS estimator is $\sigma^2 (X'X)^{-1}$. The OLS estimator has the minimum variance property among all linear unbiased estimators under the model assumptions. However, the high correlations between columns of X or the nonexistence of the full rank property can result multicollinearity problem. In the presence of multicollinearity, the OLS estimator loses its minimum variance property. The ridge estimator is a widely used alternative to the OLS estimator in the presence of the multicollinearity problem. The ridge estimator proposed by Hoerl and Kennard (1970) is biased but has lower variance than OLS estimator. The normality of the error terms is another important assumption that must be met in the multiple linear regression model for statistical significance testing of the model. Under the normality assumption, the ML and OLS estimators of $\boldsymbol{\beta}$ are equal. Outliers are one of the factors that can lead to a violation of the normality assumption. M-estimators are commonly used as an alternative to OLS estimators when the normality assumption is violated. Huber (1964) proposed M-estimators, which are a generalization of ML estimators and are robust to departures from the normality assumption. Many M-estimators have been proposed in the literature (see for example De Menezes *et al.*, 2021).

In multiple linear regression analysis, outliers and multicollinearity problems can be encountered simultaneously. In the presence of both outliers and multicollinearity problems in the model, The ridge estimator based on M-estimator Ridge-M (RM) is proposed by Silvapulle (1991) and Ertaş (2018). The RM is given by

$$\widehat{\boldsymbol{\beta}}_{\mathrm{RM}} = \left(X'X + kI \right)^{-1} X'X \widehat{\boldsymbol{\beta}}_{\mathrm{M}}, \tag{1.4}$$

where $\widehat{\boldsymbol{\beta}}_{\mathrm{M}}$ denotes OLS based M-estimator. The use of OLS estimators in RM may adversely affect the performance of the estimator. From this point of view, we propose M-estimators

based on ridge estimation instead of OLS estimation when outlier and multicollinearity problems coexist. The proposed M-estimator based on ridge estimation is obtained by rearranging the Iteratively Reweighted Least Squares (IRLS) algorithm. The proposed approach also provides a solution to the problem of the optimal ridge estimator selection with M-type estimators.

This paper is organized as follows. The proposed M-estimators based on ridge estimation are defined after examining the ridge and M-estimators. The performance of the proposed estimator is examined through by a simulation study and a real data application. In the simulation study, the mean square error (MSE) is used as a comparison criterion. The proposed estimators are compared with the ridge and classical M-estimators, considering different sample sizes, outlier rates, and correlation structures. The performance of the proposed estimators is also compared in the simulation study for lognormal distributed errors. In the real data application, the Tobacco dataset, which has both multicollinearity and outlier problems, is used. Iterative k-fold cross validation is used the comparison criterion in the real data application.

2. MATERIAL AND METHOD

2.1. Ridge estimators

The ridge estimator of unknown parameter vector is given by

$$\widehat{\boldsymbol{\beta}}_{\mathrm{R}} = \left(X'X + kI\right)^{-1} X'\boldsymbol{y}, \tag{2.1}$$

where k is known as ridge (shrinkage, bias) parameter which tunes the variance-bias trade off.

The canonical form of equation (1.1) with $Z = XD$ and $\boldsymbol{\alpha} = D\boldsymbol{\beta}$ is given as

$$\boldsymbol{y} = Z\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \tag{2.2}$$

with D defined as an orthogonal eigenvector matrix such that $D'(X'X)D = \Lambda$ where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_p)$ is eigenvalues of $X'X$. The ridge estimator of $\boldsymbol{\alpha}$ is given by

$$\widehat{\boldsymbol{\alpha}}_{\mathrm{R}} = \left(Z'Z + kI\right)^{-1} Z'\boldsymbol{y}. \tag{2.3}$$

The scalar MSE of the ridge estimator is equal to

$$\mathrm{SMSE}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{R}}\right) = \sum_{i=1}^{p} \frac{\sigma^2\lambda_i}{\left(\lambda_i + k\right)^2} + k^2\boldsymbol{\beta}'\left(X'X + kI\right)^{-2}\boldsymbol{\beta} \tag{2.4}$$

or

$$\mathrm{SMSE}\left(\widehat{\alpha}_{\mathrm{R}}\right) = \sum_{i=1}^{p} \frac{\sigma^2\lambda_i}{\left(\lambda_i + k\right)^2} + \sum_{i=1}^{p} \frac{k^2\alpha_i^2}{\left(\lambda_i + k\right)^2}. \tag{2.5}$$

The first term on the right-hand side of the equations (2.4)–(2.5) is the contribution of the total variance to the MSE. The second term on the right-hand side of the equations (2.4)–(2.5) is the contribution of the sum of the square of the bias to the MSE.

Figure 1 shows the MSE values as a function of k for ridge estimation, and the comparison of MSE values between ridge and OLS. The choice of the ridge parameter has a vital role in ridge regression, so there are different ridge estimators proposed by many authors in the literature (see Kibria, 2022, and Kibria and Banik, 2016, for instance). Table 1 provides a summary of the ridge estimators investigated in this study.

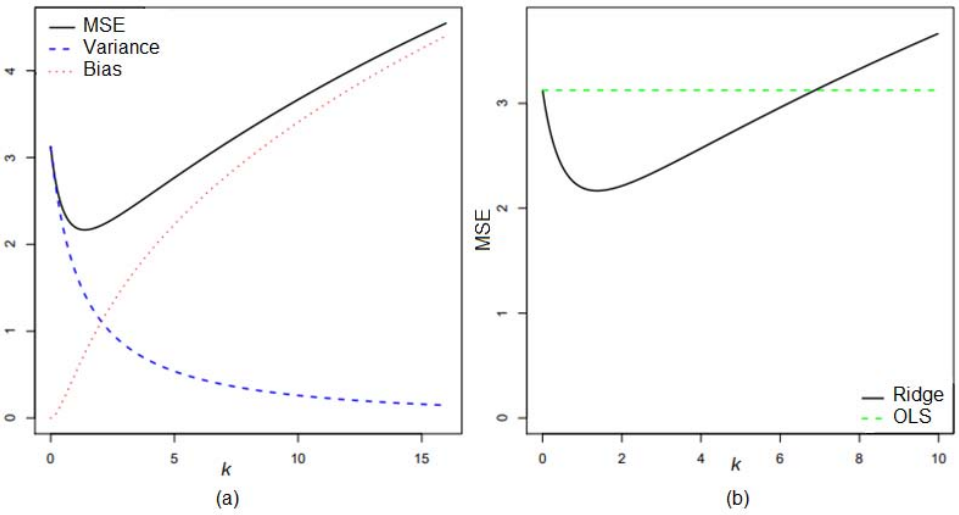

Figure 1: (a) The components of MSE for ridge estimator.
(b) The MSE comparisons of ridge and OLS estimators.

Table 1: Selected ridge parameter estimators from the literature.

Estimators	Reference
$\widehat{k}_{\mathrm{HK}} = \dfrac{\widehat{\sigma}^2}{\widehat{\alpha}^2_{\max}}$	Hoerl and Kennard (1970)
$\widehat{k}_{\mathrm{HKB}} = \dfrac{p\widehat{\sigma}^2}{\sum_{i=1}^{p}\widehat{\alpha}_i^2}$	Hoerl *et al.* (1975)
$\widehat{k}_{\mathrm{LW}} = \dfrac{p\sigma^2}{\sum_{i=1}^{p}\lambda_i\widehat{\alpha}_i^2}$	Lawless and Wang (1976)
$\widehat{k}_{\mathrm{HSL}} = \widehat{\sigma}^2\dfrac{\sum_{i=1}^{p}(\lambda_i\widehat{\alpha}_i)^2}{(\sum_{i=1}^{p}\lambda_i\widehat{\alpha}_i^2)^2}$	Hocking *et al.* (1976)
$\widehat{k}_{\mathrm{AM}} = \dfrac{1}{p}\sum_{i=1}^{p}\dfrac{\widehat{\sigma}^2}{\widehat{\alpha}_i^2}$	Kibria (2003)
$\widehat{k}_{\mathrm{GM}} = \dfrac{\widehat{\sigma}^2}{(\prod_{i=1}^{p}\widehat{\alpha}_i^2)^{\frac{1}{p}}}$	Kibria (2003)
$\widehat{k}_{\mathrm{MED}} = \mathrm{Median}\left\{\dfrac{\widehat{\sigma}^2}{\widehat{\alpha}_i^2}\right\},\ i = 1, ..., p$	Kibria (2003)
$\widehat{k}_{\mathrm{KS}} = \dfrac{\lambda_{\max}\widehat{\sigma}^2}{(n-p)\widehat{\sigma}^2 + \lambda_{\max}\widehat{\alpha}^2_{\max}}$	Khalaf and Shukur (2005)

2.2. M-estimators

The M-estimators are a class of robust estimators used as an alternative to OLS estimators for outlier or non-normality problems in regression analysis. M-estimators were first described by Huber (1964) as a generalization of MLE. Huber (1973) extended the idea of using M-estimators for solution of regression problems. In regression analysis, M-estimates are obtained by minimizing a certain objective function in equation (2.6), such as the sum of the squared error terms:

$$\widehat{\boldsymbol{\beta}}_{\mathrm{M}} = \arg\min_{\boldsymbol{\beta}} \left(\sum_{i=1}^{n} \rho(\varepsilon_i) \right). \tag{2.6}$$

In M-estimators, the objective function is the negative form of the natural logarithm of the likelihood function of the distribution of errors. The objective functions $\rho(\cdot)$ have properties following below (Martin and Yohai, 2006):

1. $\rho(0) = 0$;
2. $\rho(\varepsilon) \geq 0$;
3. $\rho(\varepsilon) = \rho(-\varepsilon)$;
4. $\rho(\varepsilon_1) < \rho(\varepsilon_2)$ for $0 < \varepsilon_1 < \varepsilon_2$;
5. $\rho(\varepsilon)$ is continuous and differentiable.

The influence function first described by Hampel (1968) is a measure of the qualitative robustness. The influence function measures the marginal effect of the data on the parameter estimator. The influence function is defined by

$$\psi(\varepsilon) = \frac{d\rho(\varepsilon)}{d\varepsilon}. \tag{2.7}$$

The influence function $\psi(\cdot)$ has the following properties (Bell, 1980):

1. $\psi(\varepsilon) \geq 0$ for $\varepsilon \geq 0$;
2. $\psi(-\varepsilon) = -\psi(\varepsilon)$;
3. $\psi'(0) = 1$;
4. $\psi''(0) = 0$;
5. $\psi'''(0) < 0$;
6. $\psi(\varepsilon)$ is continuous and partially differentiable.

Another important function in the M-estimator is the weight function obtained by dividing the influence function by ε:

$$\omega(\varepsilon) = \frac{\psi(\varepsilon)}{\varepsilon}. \tag{2.8}$$

Commonly used objective functions, influence functions, and weight functions for selected M-estimators are given in Tables 2–3.

Table 2: Objective and influence functions for the selected M-Estimators from literature.

Estimators	**Objective Function**	**Influence Function**
Huber	$\begin{cases} \frac{\varepsilon^2}{2}, & -k \le \varepsilon \le k \\ k\lvert\varepsilon\rvert - \frac{k^2}{2}, & \varepsilon < -k \text{ or } \varepsilon > k \end{cases}$	$\begin{cases} \varepsilon, & -k \le \varepsilon \le k \\ k\,\mathrm{sign}(\varepsilon), & \varepsilon < -k \text{ or } \varepsilon > k \end{cases}$
Fair	$k_{\mathrm{F}}^2 \ln\left(\frac{\lvert\varepsilon\rvert}{k_{\mathrm{F}}} - \ln\left(1 + \frac{\lvert\varepsilon\rvert}{k_{\mathrm{F}}}\right)\right), \quad \lvert\varepsilon\rvert < \infty$	$\frac{\varepsilon}{1 + \frac{\lvert\varepsilon\rvert}{k_{\mathrm{F}}}}$
Hampel	$\begin{cases} \frac{\varepsilon^2}{2}, & \lvert\varepsilon\rvert \le a \\ a\lvert\varepsilon\rvert - \frac{a^2}{2}, & a < \lvert\varepsilon\rvert \le b \\ ab - \frac{a^2}{2} + \frac{a(c-b)}{2}\left[1 - \left(\frac{c-\lvert\varepsilon\rvert}{c-b}\right)^2\right], & b < \lvert\varepsilon\rvert \le c \\ ab - -\frac{a^2}{2} + \frac{a(c-b)}{2}, & \lvert\varepsilon\rvert > c \end{cases}$	$\begin{cases} \varepsilon, & \lvert\varepsilon\rvert \le a \\ a\,\mathrm{sign}(\varepsilon), & a < \lvert\varepsilon\rvert \le b \\ \frac{a\,\mathrm{sign}(\varepsilon)(c-\lvert\varepsilon\rvert)}{c-b}, & b < \lvert\varepsilon\rvert \le c \\ 0, & \lvert\varepsilon\rvert > c \end{cases}$
Tukey	$\begin{cases} \frac{k_{\mathrm{T}}^2}{6}\left(1 - \left\{1 - (\varepsilon/k_{\mathrm{T}})^2\right\}^3\right), & \lvert\varepsilon\rvert \le k_{\mathrm{T}} \\ \frac{k_{\mathrm{T}}^2}{6}, & \lvert\varepsilon\rvert > k_{\mathrm{T}} \end{cases}$	$\begin{cases} \varepsilon\left(1 - (\varepsilon/k_{\mathrm{T}})^2\right)^2, & \lvert\varepsilon\rvert \le k_{\mathrm{T}} \\ 0, & \lvert\varepsilon\rvert > k_{\mathrm{T}} \end{cases}$
Andrew	$\begin{cases} k_{\mathrm{A}}^2\left\{1 - \cos\left(\frac{\varepsilon}{k_{\mathrm{A}}}\right)\right\}, & \lvert\varepsilon\rvert \le k_{\mathrm{A}}\pi \\ 2k_{\mathrm{A}}^2, & \lvert\varepsilon\rvert > k_{\mathrm{A}}\pi \end{cases}$	$\begin{cases} k_{\mathrm{A}} \sin\left(\frac{\varepsilon}{k_{\mathrm{A}}}\right), & \lvert\varepsilon\rvert \le k_{\mathrm{A}}\pi \\ 0, & \lvert\varepsilon\rvert > k_{\mathrm{A}}\pi \end{cases}$
Welsch	$\frac{k_{\mathrm{W}}^2}{2}\left(1 - e^{-\left(\frac{\varepsilon}{k_{\mathrm{W}}}\right)^2}\right), \quad \lvert\varepsilon\rvert < \infty$	$\varepsilon e^{-\left(\frac{\varepsilon}{k_{\mathrm{W}}}\right)^2}$
Cauchy	$\frac{k_{\mathrm{C}}^2}{2}\ln\left(1 + \left(\frac{\varepsilon}{k_{\mathrm{C}}}\right)^2\right), \quad \lvert\varepsilon\rvert < \infty$	$\frac{\varepsilon}{1 + \left(\frac{\varepsilon}{k_{\mathrm{C}}}\right)^2}$
Talwar	$\begin{cases} \frac{\varepsilon^2}{2}, & \lvert\varepsilon\rvert \le k_{\mathrm{Tw}} \\ \frac{k_{\mathrm{Tw}}^2}{2}, & \lvert\varepsilon\rvert > k_{\mathrm{Tw}} \end{cases}$	$\begin{cases} \varepsilon, & \lvert\varepsilon\rvert \le k_{\mathrm{Tw}} \\ 0, & \lvert\varepsilon\rvert > k_{\mathrm{Tw}} \end{cases}$
Ramsay	$\frac{1 - (1 + k_{\mathrm{R}}\lvert\varepsilon\rvert)e^{-k_{\mathrm{R}}\lvert\varepsilon\rvert}}{k_{\mathrm{R}}^2}, \quad \lvert\varepsilon\rvert < \infty$	$\varepsilon e^{-k_{\mathrm{R}}\lvert\varepsilon\rvert}$
Geman-McClure	$\frac{\varepsilon^2}{k_{\mathrm{GM}}^2 + \varepsilon^2}, \quad \lvert\varepsilon\rvert < \infty$	$\frac{2\varepsilon k_{\mathrm{GM}}^2}{(k_{\mathrm{GM}}^2 + \varepsilon^2)^2}$

Table 3: Weight functions and tuning parameters for the selected M-Estimators from literature.

Estimators	Weight Function	Tuning Parameter	Reference
Huber	$\begin{cases} 1, & -k \le \varepsilon \le k \\ \frac{k}{\lvert\varepsilon\rvert}, & \varepsilon < -k \text{ or } \varepsilon > k \end{cases}$	$k = 1.5\widehat{\sigma},\ \widehat{\sigma} = 1.4826\,MAD$	Birkes and Dodge (2011)
Fair	$\frac{1}{1+\frac{\lvert\varepsilon\rvert}{k_{\mathrm{F}}}}$	$k_{\mathrm{F}} = 1.3998$	Özyurt and Pike (2004)
Hampel	$\begin{cases} 1, & \lvert\varepsilon\rvert \le a \\ \frac{a}{\lvert\varepsilon\rvert}, & a < \lvert\varepsilon\rvert \le b \\ \frac{a(c-\lvert\varepsilon\rvert)}{\lvert\varepsilon\rvert(c-b)}, & b < \lvert\varepsilon\rvert \le c \\ 0, & \lvert\varepsilon\rvert > c \end{cases}$	$a = 1.35,\ b = 2.7,\ c = 5.4$	Özyurt and Pike (2004)
Tukey	$\begin{cases} \left(1-\left(\frac{\varepsilon}{k_{\mathrm{T}}}\right)^2\right)^2, & \lvert\varepsilon\rvert \le k_{\mathrm{T}} \\ 0, & \lvert\varepsilon\rvert > k_{\mathrm{T}} \end{cases}$	$k_{\mathrm{T}} = 2$	Beaton and Tukey (1974)
Andrew	$\begin{cases} \frac{k\sin\left(\frac{\varepsilon}{k_{\mathrm{A}}}\right)}{\varepsilon}, & \lvert\varepsilon\rvert \le k_{\mathrm{A}}\pi \\ 0, & \lvert\varepsilon\rvert > k_{\mathrm{A}}\pi \end{cases}$	$k_{\mathrm{A}} = 1.339$	Rey (1983)
Welsch	$e^{-\left(\frac{\varepsilon}{k_{\mathrm{W}}}\right)^2}$	$k_{\mathrm{W}} = 2.9846$	Rey (1983)
Cauchy	$\frac{1}{1+\left(\frac{\varepsilon}{k_{\mathrm{C}}}\right)^2}$	$k_{\mathrm{C}} = 2.3849$	Özyurt and Pike (2004)
Talwar	$\begin{cases} 1, & \lvert\varepsilon\rvert \le k_{\mathrm{Tw}} \\ 0, & \lvert\varepsilon\rvert > k_{\mathrm{Tw}} \end{cases}$	$k_{\mathrm{Tw}} = 2.7955$	Holland and Welsch (1977)
Ramsay	$e^{-k_{\mathrm{R}}\lvert\varepsilon\rvert}$	$k_{\mathrm{R}} = 0.3569$	De Menezes *et al.* (2021)
Geman-McClure	$\frac{2k_{\mathrm{GM}}^2}{\left(k_{\mathrm{GM}}^2+\varepsilon^2\right)^2}$	$k_{\mathrm{GM}} = 3.787376$	De Menezes *et al.* (2021)

Iteratively reweighted least squares (IRLS) algorithm is widely used for parameter estimations in M regression. IRLS algorithm is given below:

Step 1: Initial estimates of the parameters vector $\widehat{\boldsymbol{\beta}}^0$ is obtained by OLS

$$\widehat{\boldsymbol{\beta}}^0 = (X'WX)^{-1}X'W\boldsymbol{y},$$

where $W = \mathrm{diag}(\mathbf{1})$.

Step 2: The estimates of stochastic error vector $\boldsymbol{\varepsilon}$ is calculated by

$$\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - X\widehat{\boldsymbol{\beta}}^0.$$

Step 3: The weights denoted by W^{new} are updated based on the weight function of selected M-estimator by using $\widehat{\boldsymbol{\varepsilon}}$.

Step 4: The estimates of the parameters vector are updated by

$$\widehat{\boldsymbol{\beta}}^{\text{new}} = \left(X'W^{\text{new}}X\right)^{-1}X'W^{\text{new}}\boldsymbol{y}.$$

Step 5: The convergence condition is examined for the termination of the algorithm. The convergence condition can be used as

$$\begin{cases} \widehat{\boldsymbol{\beta}}_{\text{M}} = \widehat{\boldsymbol{\beta}}^{\text{new}} & \text{and stop algorithm} \quad \text{if } \sum_{i=0}^{p}\left|\widehat{\beta}_i^{\text{new}} - \widehat{\beta}_i^{0}\right| \leq \text{tol}, \\ \widehat{\boldsymbol{\beta}}^{0} = \widehat{\boldsymbol{\beta}}^{\text{new}} & \text{and repeat Step 2–5} \quad \text{if } \sum_{i=0}^{p}\left|\widehat{\beta}_i^{\text{new}} - \widehat{\beta}_i^{0}\right| > \text{tol}, \end{cases}$$

where tol is tolerance value and set to 10^{-4} in this study.

2.3. Proposed M-Ridge estimators

In this study, we propose a novel M-estimators based on ridge estimators when multicollinearity and outlier problems coexist in the multiple linear regression model. Equation (1.4) is a common approach for dealing with multicollinearity and non-normality. In many studies, the IRLS algorithm is used to obtain estimates by using equation (1.4). M-estimators are typically obtained by IRLS algorithm using OLS errors. However, it is clear that OLS estimators are unreliable in case of multicollinearity problem. Therefore, we propose to modify the IRLS algorithm by obtaining M-estimators using the errors from ridge estimators instead of OLS. OLS method gives equal weights ($w_i = 1$) to each observation vector. On the other hand, M-estimators provide more robust estimates in case of outlier or non-normality by weighting the observation vectors according to the size of the errors. M-estimates using weights based on the errors of OLS can adversely affect the performance of RM estimators since OLS estimates are unstable in existence of multicollinearity. Based on this idea obtaining M-estimates from ridge estimates instead of OLS can increase the performance of the estimator in cases where both multicollinearity and outlier problems coexist.

In this study, M-estimators based on ridge estimators (M-Ridge) are proposed to address the coexistence of outliers and multicollinearity problems in multiple linear regression models. The estimates of the stochastic error vector are computed based on ridge estimates in the proposed approach. The IRLS algorithm is arranged as follows to obtain the proposed M-Ridge estimates:

Step 1: Initial estimates of the parameter vector $\widehat{\boldsymbol{\beta}}^{0}_{\text{M-Ridge}}$ is obtained based on ridge estimator with the selected ridge parameter estimate

$$\widehat{\boldsymbol{\beta}}^{0}_{\text{M-Ridge}} = \left(X'WX + \widehat{k}I\right)^{-1}X'W\boldsymbol{y},$$

where $W = \text{diag}(\mathbf{1})$.

Step 2: The estimates of stochastic error vector ($\boldsymbol{\varepsilon}$) is calculated by

$$\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - X\widehat{\boldsymbol{\beta}}^{0}_{\text{M-Ridge}}.$$

Step 3: The weights are updated based on the weight function of selected M-estimator by using $\widehat{\boldsymbol{\varepsilon}}$ and denoted by W^{new}.

Step 4: The estimates of the parameter vector is updated by

$$\widehat{\boldsymbol{\beta}}^{\text{new}}_{\text{M-Ridge}} = \left(X'W^{\text{new}}X + \widehat{k}I\right)^{-1} X'W^{\text{new}}\boldsymbol{y}.$$

Step 5: The convergence condition for termination of the algorithm can be used as

$$\begin{cases} \widehat{\boldsymbol{\beta}}_{\text{M-Ridge}} = \widehat{\boldsymbol{\beta}}^{0}_{\text{M-Ridge}} & \text{and stop algorithm} \quad \text{if } \sum_{i=0}^{p}\left|\widehat{\boldsymbol{\beta}}^{\text{new}}_{\text{M-Ridge}} - \widehat{\boldsymbol{\beta}}^{0}_{\text{M-Ridge}}\right| \leq \text{tol}, \\ \widehat{\boldsymbol{\beta}}^{0}_{\text{M-Ridge}} = \widehat{\boldsymbol{\beta}}^{\text{new}}_{\text{M-Ridge}} & \text{and repeat Step 2–5} \quad \text{if } \sum_{i=0}^{p}\left|\widehat{\boldsymbol{\beta}}^{\text{new}}_{\text{M-Ridge}} - \widehat{\boldsymbol{\beta}}^{0}_{\text{M-Ridge}}\right| > \text{tol}, \end{cases}$$

where tol represents tolerance value and set to 10^{-4} in this study.

3. SIMULATION STUDY

In the simulation study, the performance of the proposed method for artificial data sets that contain outliers and multicollinearity problems is examined. The explanatory variables in the artificial dataset are generated as

$$x_{\text{ij}} = \left(1-\rho^2\right)^{1/2} z_{\text{ij}} + \rho z_{\text{i(p+1)}}, \tag{3.1}$$

for $i = 1, ..., n$ and $j = 1, ..., p$, where ρ^2 denotes the correlation between explanatory variables and z_{ij}'s are random numbers from standard normal distribution (McDonald and Galarneau, 1975). Coefficient vector $\boldsymbol{\beta}$ in equation (3.2) is selected as the normalized eigenvector corresponding to the largest eigenvalue of $X'X$ so that $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$. The dependent variable vector is determined by

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{\text{ip}} + \varepsilon_i, \tag{3.2}$$

where ε_i generated from two component mixture normal distribution in order to create an outlier problem. According to the selected outlier observation rate (OOR), the errors corresponding outlier observations are produced from the normal distribution with $N(5, 0.1)$ while rest of are from a standard normal distribution. Some selected box plots of dependent variables for different OOR's are shown in Figure 2.

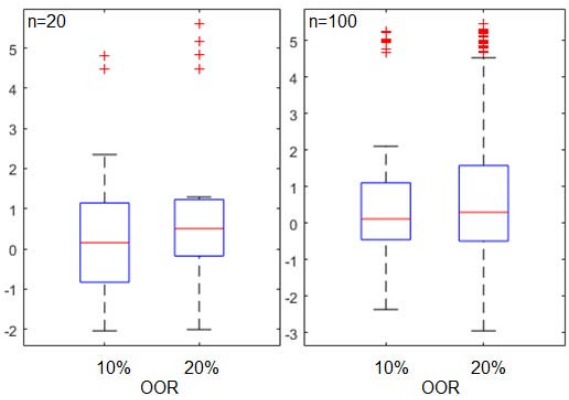

Figure 2: Box-plots of the dependent variable for a selection of artificial data sets with different sample sizes and outlier rates.

Condition index (CI) is a measure of multicollinearity and values of CI greater than 30 are indicative of strong multicollinearity (Kim, 2019). In the simulation study, CI's are computed as 41.728 and 39.844 for $n = 20$, $\rho^2 = 0.99$ and 0.999 respectively. For $n = 100$, $\rho^2 = 0.99$ and 0.999, CI's are 130.174 and 127.231 respectively.

The performance of the proposed estimator is evaluated according to the selected values of parameters for $\rho^2 = 0.99$ and 0.99, $n = 20$ and 100, OOR = 10% and 20% and $p = 4$. A Monte Carlo simulation with 5000 replications is performed to show the performance of the proposed estimator. M-Ridge is compared to OLS, ridge, M, and RM estimators in terms of the estimated MSE. Monte Carlo simulation results are given in Tables 4–11.

$$\widehat{\mathrm{MSE}}(\widehat{\boldsymbol{\beta}}) = \frac{1}{5000}\sum_{t=1}^{5000}\sum_{i=1}^{p}\left(\widehat{\beta}_{i,t} - \beta_i\right)^2. \tag{3.3}$$

According to simulation results for OOR = 10%, $n = 20$ and $\rho^2 = 0.99$, the proposed M-Ridge estimators are generally more successful than others. Cauchy M-Ridge estimator with k_{AM} ridge parameter is the best estimator with the smallest estimated MSE value of 0.3469. The M-Ridge estimators are successful 79 out of 80 comparisons against the RM estimators. Among the selected ridge parameters in robust estimators, the proposed k_{AM} ridge parameter by Kibria (2003) is the most successful ridge parameter with smaller estimated MSE. The M-Ridge estimators are successful 9 out of 10 comparisons against the RM estimators for k_{AM} ridge parameter.

When the simulation results for OOR = 20%, $n = 20$ and $\rho^2 = 0.99$ is analyzed, it is observed that the estimated MSE values are increased when the OOR's are increased as expected. Talwar M-Ridge estimator with k_{AM} ridge parameter is the best estimator with the smallest estimated MSE value of 0.400. The M-Ridge estimators are successfull 78 out of the 80 comparisons against the RM estimators.

According to results of the simulation for OOR = 10%, $n = 100$ and $\rho^2 = 0.99$, the proposed M-Ridge estimators are found more successful than RM estimators in all comparisons with RM estimators. Geman-McClure M-Ridge estimator with k_{AM} ridge parameter is the best estimator with the smallest estimated MSE value of 0.293.

According to results of the simulation for OOR = 20%, $n = 100$ and $\rho^2 = 0.99$, the proposed M-Ridge estimators are found more successful than RM estimators in all comparisons with RM estimators. Geman-McClure M-Ridge estimator with k_{HSL} ridge parameter is the best estimator with the smallest estimated MSE value of 0.198.

When the simulation results for OOR = 10%, $n = 20$ and $\rho^2 = 0.999$ are analyzed, it is observed that the estimated MSE values increase as expected when ρ^2 increases. The proposed M-Ridge estimators are generally more successful than others. The M-Ridge estimators are successful 72 out of the 80 comparisons against the RM estimators. According to results of the simulation for OOR = 20%, $n = 20$ and $\rho^2 = 0.999$, the proposed M-Ridge estimators are successful 76 out of the 80 comparisons against the RM estimators.

According to results of the simulation for OOR = 10%, $n = 100$ and $\rho^2 = 0.999$, the proposed M-Ridge estimators are found more successful than RM estimators in all comparisons with RM estimators. Geman-McClure M-Ridge estimator with k_{HSL} ridge parameter is the best estimator with the smallest estimated MSE value of 0.224.

According to results of the simulation performed for OOR = 20%, $n = 100$ and $\rho^2 = 0.999$, the proposed M-Ridge estimators are found more successful than RM estimators in all comparisons with RM estimators. Geman-McClure M-Ridge estimator with k_{HSL} ridge parameter is the best estimator with the smallest estimated MSE value of 0.186. Among the selected ridge parameters in M-Ridge estimators, the proposed k_{HSL} ridge parameter by Hocking *et al.* (1976) is the most successful ridge parameter with smaller estimated MSE values.

Table 4: Scalar MSE values of the estimators for OOR = 10%, $n = 20$ and $\rho^2 = 0.99$ (Case 1).

OLS	1507.06	Ridge	$\widehat{k}_{\text{HK}}$ 117.891	$\widehat{k}_{\text{HKB}}$ 203.820	$\widehat{k}_{\text{LW}}$ 175.044	$\widehat{k}_{\text{HSL}}$ 8.077	$\widehat{k}_{\text{AM}}$ 0.389	$\widehat{k}_{\text{GM}}$ 14.856	$\widehat{k}_{\text{MED}}$ 55.152	$\widehat{k}_{\text{KS}}$ 121.791
M		Robust	$\widehat{k}_{\text{HK}}$	$\widehat{k}_{\text{HKB}}$	$\widehat{k}_{\text{LW}}$	$\widehat{k}_{\text{HSL}}$	$\widehat{k}_{\text{AM}}$	$\widehat{k}_{\text{GM}}$	$\widehat{k}_{\text{MED}}$	$\widehat{k}_{\text{KS}}$
Huber	893.860	M-Ridge	65.026	105.234	92.032	4.379	0.353	7.331	27.022	67.014
		RM	76.065	124.456	107.344	5.159	0.362	9.094	32.482	78.446
Fair	943.136	M-Ridge	38.283	57.969	53.438	2.584	0.355	3.023	14.198	39.404
		RM	79.520	132.394	114.473	5.438	0.366	9.708	34.836	82.030
Hampel	836.801	M-Ridge	54.173	82.300	73.597	3.397	0.349	5.069	19.628	55.864
		RM	67.543	105.384	91.648	4.259	0.361	7.597	25.959	69.763
Tukey	1725.782	M-Ridge	22.249	30.281	30.816	1.812	0.478	1.482	6.477	22.991
		RM	113.010	179.605	158.260	7.309	0.547	12.915	44.879	117.495
Andrew	796.919	M-Ridge	34.516	49.614	46.638	2.183	0.350	2.647	11.226	35.558
		RM	62.959	96.741	84.938	3.931	0.356	6.884	23.387	65.036
Welsch	823.663	M-Ridge	48.872	73.823	66.670	3.091	0.3471	4.369	17.552	50.371
		RM	67.208	105.509	91.881	4.288	0.359	7.621	26.180	69.392
Cauchy	883.687	M-Ridge	58.091	92.384	81.816	3.817	0.3469	5.739	23.240	59.837
		RM	75.384	123.741	106.886	5.054	0.361	9.069	32.235	77.746
Talwar	798.563	M-Ridge	58.062	88.502	78.570	3.723	0.349	5.993	21.085	59.900
		RM	63.332	97.581	85.610	4.462	0.356	6.962	23.633	65.399
Ramsay	873.291	M-Ridge	42.236	64.156	58.646	2.781	0.350	3.506	15.571	43.483
		RM	73.695	119.985	103.895	4.921	0.363	8.779	31.011	76.031
G-Mc	838.096	M-Ridge	3.511	4.043	4.623	0.620	0.533	0.365	1.089	3.496
		RM	69.923	111.871	97.097	4.574	0.360	8.140	28.382	72.158

Table 5: Scalar MSE values of the estimators for OOR = 20%, $n = 20$ and $\rho^2 = 0.99$ (Case 2).

OLS	4694.981	Ridge	$\widehat{k}_{\text{HK}}$ 3023.756	$\widehat{k}_{\text{HKB}}$ 1657.074	$\widehat{k}_{\text{LW}}$ 676.982	$\widehat{k}_{\text{HSL}}$ 327.444	$\widehat{k}_{\text{AM}}$ 0.458	$\widehat{k}_{\text{GM}}$ 106.551	$\widehat{k}_{\text{MED}}$ 443.482	$\widehat{k}_{\text{KS}}$ 3031.023
M		Robust	$\widehat{k}_{\text{HK}}$	$\widehat{k}_{\text{HKB}}$	$\widehat{k}_{\text{LW}}$	$\widehat{k}_{\text{HSL}}$	$\widehat{k}_{\text{AM}}$	$\widehat{k}_{\text{GM}}$	$\widehat{k}_{\text{MED}}$	$\widehat{k}_{\text{KS}}$
Huber	3499.616	M-Ridge	1599.733	637.071	251.808	128.696	0.435	28.668	147.896	1605.432
		RM	2215.030	1208.167	513.362	251.770	0.446	75.058	320.177	2220.244
Fair	3013.929	M-Ridge	1002.236	337.370	121.444	69.689	0.459	10.292	68.019	1006.376
		RM	1907.099	1045.375	452.743	211.948	0.432	66.114	275.936	1911.399
Hampel	1905.393	M-Ridge	706.030	247.065	104.786	41.117	0.487	8.286	53.669	708.909
		RM	1133.244	611.446	292.873	111.803	0.685	38.936	159.321	1136.115
Tukey	2541.135	M-Ridge	286.908	68.632	27.276	15.501	0.574	2.080	13.644	288.771
		RM	1369.543	689.342	306.027	123.549	0.691	42.663	180.762	1373.767
Andrew	1689.642	M-Ridge	418.640	116.646	50.478	22.527	0.432	3.639	25.613	420.770
		RM	983.336	525.215	258.749	96.927	0.622	33.425	138.235	985.961
Welsch	1903.509	M-Ridge	665.134	223.833	93.211	39.497	0.468	7.313	47.753	667.858
		RM	1141.321	619.021	296.636	116.229	0.645	39.414	161.378	1144.114
Cauchy	2760.425	M-Ridge	1191.597	464.814	179.835	88.974	0.449	17.053	100.373	1195.715
		RM	1744.399	958.343	424.448	195.682	0.469	60.379	251.544	1748.220
Talwar	2636.496	M-Ridge	1110.995	316.013	125.896	53.185	0.400	8.258	55.988	1114.771
		RM	1602.653	856.312	392.798	192.809	0.700	51.032	224.708	1606.681
Ramsay	2493.603	M-Ridge	857.939	289.901	109.017	56.928	0.469	8.972	58.796	861.333
		RM	1562.552	857.858	385.871	173.938	0.491	54.120	224.474	1565.977
G-Mc	2105.865	M-Ridge	86.130	13.987	4.905	4.436	0.722	0.614	2.669	86.714
		RM	1297.936	711.442	332.535	139.783	0.552	45.168	185.207	1300.832

Table 6: Scalar MSE values of the estimators for OOR = 10%, $n = 100$ and $\rho^2 = 0.99$ (Case 3).

OLS	1900.100	Ridge	$\widehat{k}_{HK}$ 258.580	$\widehat{k}_{HKB}$ 486.807	$\widehat{k}_{LW}$ 407.134	$\widehat{k}_{HSL}$ 6.339	$\widehat{k}_{AM}$ 3.333	$\widehat{k}_{GM}$ 107.302	$\widehat{k}_{MED}$ 222.727	$\widehat{k}_{KS}$ 321.897
M		Robust	$\widehat{k}_{HK}$	$\widehat{k}_{HKB}$	$\widehat{k}_{LW}$	$\widehat{k}_{HSL}$	$\widehat{k}_{AM}$	$\widehat{k}_{GM}$	$\widehat{k}_{MED}$	$\widehat{k}_{KS}$
Huber	891.006	M-Ridge	93.567	181.536	154.375	2.118	0.968	33.451	75.540	114.005
		RM	113.235	219.419	185.531	2.959	1.308	44.842	93.393	138.377
Fair	988.946	M-Ridge	59.672	114.154	99.039	1.018	0.568	16.256	45.757	72.816
		RM	128.024	247.708	209.377	3.323	1.502	51.467	107.126	156.737
Hampel	772.782	M-Ridge	54.855	107.133	93.165	1.088	0.581	16.220	37.948	67.716
		RM	77.652	152.433	130.182	1.912	0.911	27.604	56.031	96.922
Tukey	1492.272	M-Ridge	25.028	48.464	46.163	0.591	0.513	5.186	16.207	31.567
		RM	129.960	251.541	216.397	2.900	1.676	46.762	98.990	167.889
Andrew	662.853	M-Ridge	33.579	63.891	57.262	0.624	0.403	8.336	21.429	41.319
		RM	63.863	123.986	106.262	1.512	0.767	21.807	44.028	80.121
Welsch	744.195	M-Ridge	50.271	97.880	85.458	0.979	0.526	14.488	35.122	61.654
		RM	77.018	151.056	128.898	1.927	0.892	27.679	56.528	95.555
Cauchy	885.819	M-Ridge	78.338	152.276	130.334	1.580	0.758	25.550	61.696	95.331
		RM	110.798	215.614	182.538	2.901	1.274	43.720	90.783	135.441
Talwar	660.464	M-Ridge	54.922	106.275	91.784	1.178	0.633	17.414	37.127	68.444
		RM	63.666	123.667	105.832	1.516	0.771	21.831	44.165	79.918
Ramsay	856.196	M-Ridge	55.031	106.159	92.334	0.992	0.531	15.402	41.355	66.862
		RM	103.387	201.764	171.137	2.703	1.183	40.237	83.454	126.539
G-Mc	777.916	M-Ridge	4.829	8.354	8.553	0.247	0.293	0.832	2.949	5.822
		RM	87.590	171.532	145.969	2.259	0.999	32.884	67.883	107.598

Table 7: Scalar MSE values of the estimators for OOR = 20%, $n = 100$ and $\rho^2 = 0.99$ (Case 4).

OLS	3140.552	Ridge	$\widehat{k}_{HK}$ 226.692	$\widehat{k}_{HKB}$ 811.123	$\widehat{k}_{LW}$ 874.977	$\widehat{k}_{HSL}$ 9.332	$\widehat{k}_{AM}$ 7.970	$\widehat{k}_{GM}$ 181.913	$\widehat{k}_{MED}$ 294.674	$\widehat{k}_{KS}$ 359.045
M		Robust	$\widehat{k}_{HK}$	$\widehat{k}_{HKB}$	$\widehat{k}_{LW}$	$\widehat{k}_{HSL}$	$\widehat{k}_{AM}$	$\widehat{k}_{GM}$	$\widehat{k}_{MED}$	$\widehat{k}_{KS}$
Huber	1474.408	M-Ridge	75.450	283.226	310.723	2.560	2.013	51.429	96.867	114.645
		RM	114.687	385.839	410.733	4.178	3.159	85.005	143.605	171.957
Fair	1507.492	M-Ridge	35.669	152.533	178.194	1.234	1.104	20.904	46.666	54.648
		RM	114.665	391.136	418.459	4.239	3.249	85.926	144.340	172.800
Hampel	1189.339	M-Ridge	29.586	111.818	128.957	0.825	0.704	15.732	34.737	43.583
		RM	63.598	221.697	240.768	2.063	1.565	42.226	74.015	95.084
Tukey	1790.016	M-Ridge	9.576	41.978	54.622	0.552	0.565	4.569	11.537	14.676
		RM	75.537	289.926	323.985	2.523	2.201	53.629	92.767	121.165
Andrew	783.499	M-Ridge	15.158	56.914	67.810	0.465	0.440	7.418	17.193	22.266
		RM	38.859	137.467	150.653	1.236	1.005	25.053	44.108	58.605
Welsch	1070.609	M-Ridge	26.181	99.490	114.792	0.736	0.628	13.944	31.093	38.397
		RM	60.273	206.505	223.232	1.974	1.452	39.910	70.273	88.976
Cauchy	1352.270	M-Ridge	50.324	196.219	219.783	1.574	1.280	31.148	64.321	75.687
		RM	101.800	338.651	360.149	3.611	2.653	73.538	125.615	150.779
Talwar	778.506	M-Ridge	27.514	99.066	111.324	0.802	0.689	15.950	31.236	40.935
		RM	38.689	136.494	149.389	1.230	1.002	24.974	44.057	58.376
Ramsay	1287.711	M-Ridge	31.230	127.685	147.766	0.970	0.842	17.486	39.782	46.736
		RM	92.675	308.265	328.386	3.238	2.349	65.859	113.328	136.630
G-Mc	1120.644	M-Ridge	1.808	7.754	11.174	0.198	0.269	0.910	2.177	2.587
		RM	72.654	243.053	260.381	2.461	1.758	49.745	86.907	106.340

Table 8: Scalar MSE values of the estimators for OOR = 10%, $n = 20$ and $\rho^2 = 0.999$ (Case 5).

OLS	14807.131	Ridge	$\widehat{k}_{\text{HK}}$ 2068.689	$\widehat{k}_{\text{HKB}}$ 1999.721	$\widehat{k}_{\text{LW}}$ 706.270	$\widehat{k}_{\text{HSL}}$ 21.945	$\widehat{k}_{\text{AM}}$ 0.300	$\widehat{k}_{\text{GM}}$ 56.045	$\widehat{k}_{\text{MED}}$ 543.728	$\widehat{k}_{\text{KS}}$ 2073.417
M		Robust	$\widehat{k}_{\text{HK}}$	$\widehat{k}_{\text{HKB}}$	$\widehat{k}_{\text{LW}}$	$\widehat{k}_{\text{HSL}}$	$\widehat{k}_{\text{AM}}$	$\widehat{k}_{\text{GM}}$	$\widehat{k}_{\text{MED}}$	$\widehat{k}_{\text{KS}}$
Huber	8764.191	M-Ridge	1111.019	1031.415	439.101	11.680	0.303	27.145	264.628	1113.431
		RM	1292.304	1221.216	520.768	13.549	0.302	34.240	319.383	1295.113
Fair	9249.708	M-Ridge	688.926	566.519	221.816	7.872	0.320	9.988	138.747	690.468
		RM	1366.085	1299.036	532.820	13.916	0.303	36.592	342.872	1369.059
Hampel	8199.155	M-Ridge	911.395	806.542	397.144	9.601	0.309	17.856	191.799	913.506
		RM	1124.288	1033.059	507.277	11.099	0.306	28.453	254.572	1126.910
Tukey	16930.813	M-Ridge	400.689	292.637	133.467	4.608	0.453	3.838	60.458	401.800
		RM	1939.736	1760.914	763.987	16.661	0.470	47.636	437.201	1945.317
Andrew	7813.175	M-Ridge	588.172	484.797	243.485	6.487	0.318	8.721	108.747	589.560
		RM	1035.535	948.363	490.190	10.272	0.305	25.642	228.576	1038.009
Welsch	8071.554	M-Ridge	830.188	723.092	348.570	8.912	0.310	15.163	171.399	832.089
		RM	1121.485	1034.472	501.415	11.180	0.305	28.555	256.846	1124.063
Cauchy	8662.866	M-Ridge	1006.607	905.638	380.991	10.934	0.305	20.375	228.035	1008.800
		RM	1285.447	1213.984	516.863	13.268	0.301	34.186	317.189	1288.227
Talwar	7831.714	M-Ridge	957.074	867.037	449.919	9.688	0.302	21.840	205.884	959.290
		RM	1039.385	956.961	483.800	11.357	0.304	25.956	230.884	1041.855
Ramsay	8559.872	M-Ridge	744.164	627.712	266.277	8.316	0.316	11.811	152.213	745.821
		RM	1252.891	1177.020	514.515	12.637	0.304	33.048	304.977	1255.644
G-Mc	8213.268	M-Ridge	62.305	37.282	16.202	1.275	0.485	0.608	8.142	62.445
		RM	1177.158	1097.163	504.450	11.864	0.304	30.579	278.798	1179.791

Table 9: Scalar MSE values of the estimators for OOR = 20%, $n = 20$ and $\rho^2 = 0.999$ (Case 6).

OLS	47401.887	Ridge	$\widehat{k}_{\text{HK}}$ 31219.846	$\widehat{k}_{\text{HKB}}$ 16777.823	$\widehat{k}_{\text{LW}}$ 21574.414	$\widehat{k}_{\text{HSL}}$ 510.123	$\widehat{k}_{\text{AM}}$ 0.297	$\widehat{k}_{\text{GM}}$ 387.819	$\widehat{k}_{\text{MED}}$ 3709.695	$\widehat{k}_{\text{KS}}$ 31226.605
M		Robust	$\widehat{k}_{\text{HK}}$	$\widehat{k}_{\text{HKB}}$	$\widehat{k}_{\text{LW}}$	$\widehat{k}_{\text{HSL}}$	$\widehat{k}_{\text{AM}}$	$\widehat{k}_{\text{GM}}$	$\widehat{k}_{\text{MED}}$	$\widehat{k}_{\text{KS}}$
Huber	35046.937	M-Ridge	16614.515	6433.902	9701.346	203.245	0.383	100.656	1201.710	16620.019
		RM	22737.765	12175.911	15828.326	398.921	0.357	270.606	2655.248	22742.587
Fair	30145.911	M-Ridge	10515.510	3417.909	5621.556	120.061	0.411	32.969	542.742	10519.477
		RM	19559.568	10531.888	13638.694	332.774	0.350	238.470	2283.153	19563.543
Hampel	18720.207	M-Ridge	7303.376	2496.267	3983.442	64.521	0.470	26.347	428.332	7306.089
		RM	11468.182	6084.823	7868.771	166.890	0.685	136.012	1282.201	11470.814
Tukey	25115.230	M-Ridge	3072.891	694.490	1377.559	29.169	0.546	5.185	102.614	3074.634
		RM	14038.131	6925.769	9098.242	191.214	0.664	150.676	1473.750	14042.051
Andrew	16616.234	M-Ridge	4398.887	1184.785	2131.911	37.422	0.403	11.202	203.191	4400.891
		RM	9958.532	5222.772	6792.274	159.023	0.616	116.154	1105.293	9960.951
Welsch	18732.179	M-Ridge	6903.999	2264.845	3692.730	62.878	0.446	23.223	380.758	6906.578
		RM	11563.878	6168.397	7977.234	176.626	0.637	138.219	1301.755	11566.440
Cauchy	27512.585	M-Ridge	12372.348	4701.652	7178.307	146.249	0.411	56.566	809.476	12376.232
		RM	17835.668	9632.454	12482.338	305.635	0.408	216.826	2071.882	17839.191
Talwar	26024.879	M-Ridge	11460.867	3189.382	5901.231	90.671	0.374	28.123	442.991	11463.500
		RM	16231.667	8519.122	11150.957	344.609	0.711	177.636	1815.415	16235.367
Ramsay	24797.853	M-Ridge	8964.764	2935.818	4811.464	95.923	0.434	28.542	468.760	8967.998
		RM	15955.039	8613.469	11160.167	270.261	0.441	193.793	1843.388	15958.193
G-Mc	20833.458	M-Ridge	943.263	138.798	339.125	7.597	0.684	1.123	17.713	943.858
		RM	13200.980	7118.483	9213.394	214.553	0.524	160.277	1507.855	13203.641

Table 10: Scalar MSE values of the estimators for OOR = 10%, $n = 100$ and $\rho^2 = 0.999$ (Case 7).

OLS	18517.999	Ridge	$\widehat{k}_{HK}$ 2504.533	$\widehat{k}_{HKB}$ 4706.645	$\widehat{k}_{LW}$ 4097.869	$\widehat{k}_{HSL}$ 1.899	$\widehat{k}_{AM}$ 1.235	$\widehat{k}_{GM}$ 467.179	$\widehat{k}_{MED}$ 2140.604	$\widehat{k}_{KS}$ 2560.969
M		Robust	$\widehat{k}_{HK}$	$\widehat{k}_{HKB}$	$\widehat{k}_{LW}$	$\widehat{k}_{HSL}$	$\widehat{k}_{AM}$	$\widehat{k}_{GM}$	$\widehat{k}_{MED}$	$\widehat{k}_{KS}$
Huber	8890.015	M-Ridge	910.818	1762.354	1564.003	0.669	0.443	139.555	729.165	928.705
		RM	1104.600	2134.951	1877.057	0.880	0.497	196.037	903.696	1126.708
Fair	9836.618	M-Ridge	577.118	1101.013	1006.757	0.419	0.386	60.238	438.211	588.751
		RM	1247.491	2407.877	2116.585	0.964	0.562	225.078	1035.408	1272.883
Hampel	7803.645	M-Ridge	536.871	1042.389	949.718	0.405	0.333	64.471	368.317	547.574
		RM	763.262	1490.965	1325.432	0.604	0.386	118.841	546.865	779.340
Tukey	15075.198	M-Ridge	239.013	462.022	470.656	0.438	0.458	16.656	152.421	244.621
		RM	1272.080	2454.062	2210.714	0.957	0.732	196.359	960.346	1304.124
Andrew	6695.389	M-Ridge	327.636	619.131	584.560	0.303	0.283	31.200	207.327	334.089
		RM	628.710	1213.243	1082.776	0.499	0.343	93.179	430.567	642.018
Welsch	7511.262	M-Ridge	491.194	951.042	870.874	0.379	0.316	56.855	340.229	500.762
		RM	756.604	1476.726	1311.232	0.600	0.377	119.490	551.231	772.158
Cauchy	8855.387	M-Ridge	761.912	1476.605	1322.699	0.532	0.398	102.518	594.711	776.775
		RM	1081.718	2099.460	1847.911	0.845	0.485	191.164	879.030	1103.333
Talwar	6670.480	M-Ridge	539.118	1037.643	935.352	0.429	0.323	72.074	361.953	550.108
		RM	626.080	1209.639	1078.217	0.522	0.345	93.196	431.408	639.413
Ramsay	8581.543	M-Ridge	534.111	1026.813	939.240	0.392	0.343	58.365	397.546	544.460
		RM	1010.555	1966.145	1734.052	0.793	0.458	175.670	809.017	1030.735
G-Mc	7829.529	M-Ridge	44.168	77.362	85.725	0.224	0.266	2.147	26.096	45.037
		RM	858.215	1674.177	1481.715	0.679	0.404	142.978	659.846	875.358

Table 11: Scalar MSE values of the estimators for OOR = 20%, $n = 100$ and $\rho^2 = 0.999$ (Case 8).

OLS	30238.893	Ridge	$\widehat{k}_{HK}$ 2118.929	$\widehat{k}_{HKB}$ 7749.269	$\widehat{k}_{LW}$ 8724.782	$\widehat{k}_{HSL}$ 3.342	$\widehat{k}_{AM}$ 3.450	$\widehat{k}_{GM}$ 749.439	$\widehat{k}_{MED}$ 2715.282	$\widehat{k}_{KS}$ 2227.345
M		Robust	$\widehat{k}_{HK}$	$\widehat{k}_{HKB}$	$\widehat{k}_{LW}$	$\widehat{k}_{HSL}$	$\widehat{k}_{AM}$	$\widehat{k}_{GM}$	$\widehat{k}_{MED}$	$\widehat{k}_{KS}$
Huber	14573.073	M-Ridge	704.728	2705.611	3119.209	1.042	1.078	196.682	891.648	736.316
		RM	1082.727	3714.399	4123.085	1.172	1.176	354.481	1334.789	1130.734
Fair	14871.272	M-Ridge	327.168	1439.940	1792.199	0.771	0.860	70.318	421.709	342.186
		RM	1081.582	3761.587	4197.642	1.224	1.234	357.826	1340.583	1130.142
Hampel	12033.327	M-Ridge	277.873	1066.741	1307.909	0.420	0.445	56.344	321.465	288.670
		RM	604.230	2141.156	2440.996	0.590	0.591	173.613	693.384	629.139
Tukey	18102.683	M-Ridge	85.479	389.947	552.495	0.474	0.511	13.505	101.956	89.391
		RM	714.260	2790.553	3290.897	0.868	0.901	214.551	865.647	749.799
Andrew	7924.098	M-Ridge	141.920	540.689	688.279	0.291	0.313	25.735	158.922	147.372
		RM	369.850	1326.828	1528.770	0.399	0.411	102.101	414.734	384.879
Welsch	10821.232	M-Ridge	245.473	948.332	1162.733	0.378	0.406	49.678	287.261	254.978
		RM	572.652	1994.584	2260.556	0.554	0.550	164.785	658.226	595.591
Cauchy	13458.812	M-Ridge	468.079	1868.520	2210.937	0.752	0.801	112.682	589.157	488.358
		RM	962.844	3264.931	3622.010	0.971	0.964	307.095	1169.258	1003.969
Talwar	7871.997	M-Ridge	260.574	950.086	1129.894	0.343	0.360	61.100	291.902	270.608
		RM	368.039	1316.858	1515.506	0.398	0.410	101.647	413.844	383.041
Ramsay	12868.287	M-Ridge	288.666	1209.728	1489.077	0.558	0.621	59.709	362.044	300.966
		RM	877.358	2973.707	3306.614	0.862	0.851	274.816	1055.887	914.148
G-Mc	11272.675	M-Ridge	13.782	67.987	109.560	0.186	0.243	2.080	16.860	14.357
		RM	689.126	2346.527	2628.418	0.657	0.646	206.975	811.716	716.889

The results from Tables 4–11 can be summarized as follows. In most cases, MSE values increased with increasing OOR and ρ^2, while controlling for other factors. The M-Ridge and RM estimators are found to be more successful than the OLS, ridge, and M-estimators according to the MSE criteria. The M-Ridge estimator is found to be superior to the RM estimator according to the MSE criteria.

Box plots of MSE values for M-Ridge and RM estimators with different ridge parameters given in Figure 3 shows that the M-Ridge estimators are more robust to the changes in ridge parameter than the RM estimators, and have smaller MSE values. When the average MSE values obtained with different ridge parameters are examined, the M-Ridge estimates are found to be more successful than the RM estimates in all cases.

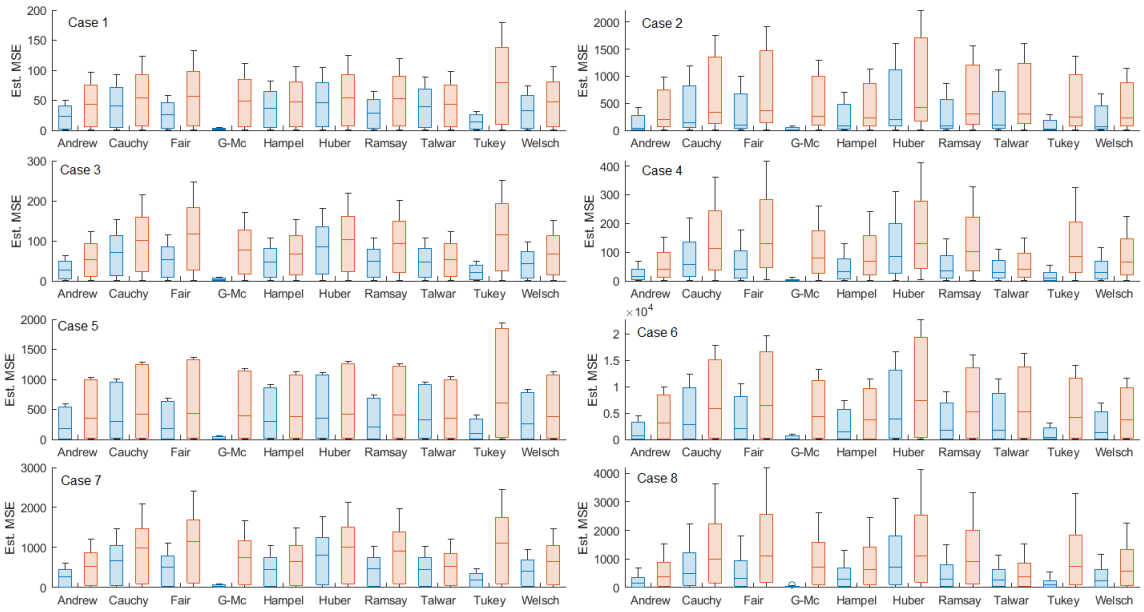

Figure 3: Box plots of MSE values for M-Ridge and RM estimators with different ridge estimators (Cases 1–8).

The proposed estimator M-Ridge has smaller MSE than the RM estimator in all cases examined except for the bias parameter k_{AM}. In comparison of M-Ridge and RM estimators with ridge parameter k_{AM}, the proposed M-Ridge estimator has smaller MSE than the RM estimator in 65 of the 80 cases. Box plot of MSE values for M-Ridge and RM estimators with ridge parameter k_{AM} is given in Figure 4. As evident from Figure 4, the proposed M-Ridge estimator outperforms the RM estimator in terms of mean squared error (MSE) for the bias parameter k_{AM}.

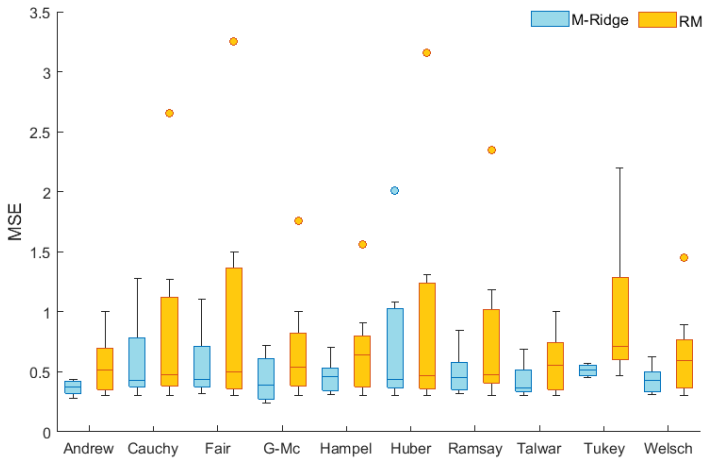

Figure 4: Box plots of MSE values for M-Ridge and RM estimators with ridge estimator k_{AM}.

In the second stage of the simulation study, the performance of the estimators is examined for the errors from log-normal distribution with parameters $\mu = 0$ and $\sigma^2 = 1$. The simulation study involves specific parameter settings, including $\rho^2 = 0.99$ and 0.999, $n = 20$ and 100, and $p = 4$. Monte Carlo simulation results based on 5000 replications are presented in Tables 12–15.

Table 12: Scalar MSE of the estimators for lognormal distributed errors, $n = 20$ and $\rho^2 = 0.99$ (Case 9).

OLS	3394.857	Ridge	$\widehat{k}_{HK}$ 916.252	$\widehat{k}_{HKB}$ 603.731	$\widehat{k}_{LW}$ 599.784	$\widehat{k}_{HSL}$ 37.128	$\widehat{k}_{AM}$ 1.277	$\widehat{k}_{GM}$ 41.289	$\widehat{k}_{MED}$ 114.869	$\widehat{k}_{KS}$ 923.979
M		Robust	$\widehat{k}_{HK}$	$\widehat{k}_{HKB}$	$\widehat{k}_{LW}$	$\widehat{k}_{HSL}$	$\widehat{k}_{AM}$	$\widehat{k}_{GM}$	$\widehat{k}_{MED}$	$\widehat{k}_{KS}$
Huber	1162.188	M-Ridge	192.488	89.897	98.023	8.009	0.442	4.535	16.129	194.639
		RM	278.480	167.505	164.175	12.759	0.562	11.868	33.977	281.253
Fair	870.719	M-Ridge	56.820	19.709	23.938	2.026	0.435	0.941	3.327	57.369
		RM	211.251	128.710	128.574	9.966	0.494	9.090	25.274	213.322
Hampel	1569.666	M-Ridge	65.604	25.721	30.567	2.646	0.397	1.206	4.050	66.543
		RM	248.562	139.487	144.229	9.720	0.552	8.410	24.123	252.698
Tukey	7023.009	M-Ridge	53.023	14.939	19.346	2.260	0.693	0.975	2.382	53.898
		RM	1563.448	1014.971	1070.368	63.501	1.962	61.597	150.988	1579.596
Andrew	1857.532	M-Ridge	49.237	16.610	21.006	1.975	0.448	0.765	2.504	49.894
		RM	296.768	171.919	179.478	10.872	0.631	10.951	31.587	301.625
Welsch	1449.204	M-Ridge	55.090	20.680	25.154	2.233	0.404	0.961	3.236	55.842
		RM	235.004	129.488	134.193	9.115	0.529	7.785	21.807	238.809
Cauchy	657.616	M-Ridge	53.355	23.447	25.361	2.451	0.381	1.237	4.320	54.083
		RM	121.514	70.437	69.866	5.810	0.397	4.793	13.782	123.219
Talwar	1647.434	M-Ridge	109.115	48.123	53.427	4.316	0.438	2.546	8.368	110.804
		RM	307.538	196.182	187.911	14.753	0.696	12.769	39.202	311.609
Ramsay	720.429	M-Ridge	33.217	12.415	14.805	1.386	0.407	0.622	2.124	33.619
		RM	126.110	73.255	75.165	5.622	0.405	4.944	13.988	127.997
G-Mc	924.968	M-Ridge	3.387	0.809	1.191	0.549	0.674	0.260	0.372	3.320
		RM	148.018	84.571	88.603	6.265	0.438	5.548	15.472	150.460

Table 13: Scalar MSE of the estimators for lognormal distributed errors, $n = 20$ and $\rho^2 = 0.999$ (Case 10).

OLS	33306.964	Ridge	$\widehat{k}_{HK}$ 8997.876	$\widehat{k}_{HKB}$ 5948.933	$\widehat{k}_{LW}$ 4765.248	$\widehat{k}_{HSL}$ 104.566	$\widehat{k}_{AM}$ 1.083	$\widehat{k}_{GM}$ 162.260	$\widehat{k}_{MED}$ 1206.539	$\widehat{k}_{KS}$ 9005.269
M		Robust	$\widehat{k}_{HK}$	$\widehat{k}_{HKB}$	$\widehat{k}_{LW}$	$\widehat{k}_{HSL}$	$\widehat{k}_{AM}$	$\widehat{k}_{GM}$	$\widehat{k}_{MED}$	$\widehat{k}_{KS}$
Huber	11407.251	M-Ridge	1887.833	881.303	843.324	25.901	0.402	15.669	160.412	1889.963
		RM	2740.443	1647.381	1379.774	40.759	0.472	45.442	346.777	2743.180
Fair	8550.317	M-Ridge	550.887	191.837	194.284	5.752	0.405	2.707	31.811	551.423
		RM	2081.149	1265.936	1050.649	31.524	0.420	34.752	257.867	2083.184
Hampel	15477.152	M-Ridge	637.706	250.228	252.280	8.641	0.362	3.724	39.164	638.621
		RM	2394.337	1374.564	1225.932	30.642	0.476	31.099	251.450	2398.295
Tukey	70204.772	M-Ridge	506.971	137.156	156.075	5.348	0.667	2.064	18.025	507.562
		RM	15823.483	10261.280	7822.940	190.337	1.673	261.502	1878.946	15839.540
Andrew	18307.094	M-Ridge	475.212	160.095	167.463	5.641	0.414	2.086	23.054	475.849
		RM	2908.757	1689.845	1508.914	32.021	0.533	40.246	322.296	2913.510
Welsch	14280.645	M-Ridge	534.852	200.797	204.341	7.167	0.369	2.873	30.971	535.587
		RM	2259.062	1275.468	1120.455	29.871	0.455	28.744	222.713	2262.735
Cauchy	6459.583	M-Ridge	514.559	228.322	238.338	7.214	0.347	3.905	41.604	515.262
		RM	1188.789	691.444	614.623	19.014	0.340	17.444	136.780	1190.450
Talwar	16168.063	M-Ridge	1042.757	469.180	483.516	14.245	0.397	7.998	81.662	1044.302
		RM	2932.190	1915.777	1739.678	45.062	0.599	48.285	404.205	2936.126
Ramsay	7079.695	M-Ridge	320.713	120.290	127.999	3.859	0.374	1.738	19.700	321.104
		RM	1241.382	719.259	623.470	18.542	0.346	17.830	139.794	1243.242
G-Mc	9098.966	M-Ridge	30.482	6.632	7.986	0.689	0.636	0.205	0.983	30.427
		RM	1444.830	831.317	722.369	20.078	0.377	19.933	155.790	1447.224

Table 14: Scalar MSE of the estimators for lognormal distributed errors, $n = 100$ and $\rho^2 = 0.99$ (Case 11).

OLS	2591.635	Ridge	$\widehat{k}_{HK}$ 573.003	$\widehat{k}_{HKB}$ 448.774	$\widehat{k}_{LW}$ 413.554	$\widehat{k}_{HSL}$ 102.524	$\widehat{k}_{AM}$ 1.180	$\widehat{k}_{GM}$ 26.348	$\widehat{k}_{MED}$ 88.479	$\widehat{k}_{KS}$ 629.044
M		Robust	$\widehat{k}_{HK}$	$\widehat{k}_{HKB}$	$\widehat{k}_{LW}$	$\widehat{k}_{HSL}$	$\widehat{k}_{AM}$	$\widehat{k}_{GM}$	$\widehat{k}_{MED}$	$\widehat{k}_{KS}$
Huber	551.834	M-Ridge	80.668	51.510	48.476	9.118	0.440	2.737	9.414	90.462
		RM	98.618	69.915	63.961	12.238	0.455	4.278	13.492	111.049
Fair	397.395	M-Ridge	29.977	15.077	15.355	3.252	0.463	0.668	2.580	33.079
		RM	75.175	53.887	49.261	9.928	0.412	3.331	10.552	84.039
Hampel	423.251	M-Ridge	29.755	17.699	18.068	3.317	0.422	0.846	2.862	34.648
		RM	52.986	37.996	36.471	6.806	0.386	2.249	6.932	63.021
Tukey	3245.259	M-Ridge	36.679	13.949	16.491	3.707	0.790	0.907	2.166	40.897
		RM	405.126	288.511	278.345	64.671	1.090	17.262	53.500	477.177
Andrew	513.681	M-Ridge	23.992	12.479	13.303	2.506	0.481	0.561	1.868	27.491
		RM	65.498	46.550	44.483	8.062	0.406	2.733	8.438	77.534
Welsch	397.044	M-Ridge	24.905	14.398	14.920	2.771	0.430	0.673	2.276	28.891
		RM	49.503	35.653	34.395	6.445	0.380	2.110	6.500	58.918
Cauchy	336.273	M-Ridge	30.534	17.975	17.966	3.347	0.406	0.897	3.131	34.839
		RM	49.484	34.726	32.953	6.161	0.377	2.163	6.694	57.297
Talwar	742.333	M-Ridge	63.221	38.948	39.548	7.171	0.476	1.927	6.432	73.963
		RM	95.081	67.508	64.573	11.798	0.476	4.040	12.171	112.141
Ramsay	298.564	M-Ridge	17.229	9.359	9.892	1.915	0.439	0.450	1.534	19.699
		RM	39.871	28.284	27.394	5.229	0.361	1.740	5.398	46.880
G-Mc	314.456	M-Ridge	2.236	0.775	1.022	0.576	0.725	0.264	0.364	2.344
		RM	39.419	28.255	27.378	5.192	0.361	1.709	5.291	46.883

Table 15: Scalar MSE of the estimators for lognormal distributed errors, $n = 100$ and $\rho^2 = 0.999$ (Case 12).

OLS	26187.711	Ridge	$\widehat{k}_{HK}$ 5765.011	$\widehat{k}_{HKB}$ 4521.224	$\widehat{k}_{LW}$ 4171.784	$\widehat{k}_{HSL}$ 363.889	$\widehat{k}_{AM}$ 1.006	$\widehat{k}_{GM}$ 100.677	$\widehat{k}_{MED}$ 886.798	$\widehat{k}_{KS}$ 5813.761
M		Robust	$\widehat{k}_{HK}$	$\widehat{k}_{HKB}$	$\widehat{k}_{LW}$	$\widehat{k}_{HSL}$	$\widehat{k}_{AM}$	$\widehat{k}_{GM}$	$\widehat{k}_{MED}$	$\widehat{k}_{KS}$
Huber	5573.883	M-Ridge	812.268	517.892	487.361	25.690	0.397	9.624	93.055	820.752
		RM	992.387	702.943	643.330	33.857	0.395	15.928	134.248	1002.994
Fair	4014.625	M-Ridge	300.611	150.789	153.540	10.356	0.430	1.866	24.279	303.397
		RM	756.714	541.853	495.544	30.149	0.359	12.483	104.993	764.332
Hampel	4273.551	M-Ridge	297.335	177.075	180.812	9.781	0.382	2.597	27.063	301.513
		RM	530.388	381.596	366.397	19.924	0.334	8.043	68.137	538.783
Tukey	32900.989	M-Ridge	362.298	133.322	160.017	11.241	0.772	1.675	15.586	366.051
		RM	4065.108	2903.287	2800.391	286.532	0.901	62.973	530.283	4125.723
Andrew	5184.126	M-Ridge	238.865	124.184	132.466	7.460	0.441	1.471	16.755	241.894
		RM	655.177	467.279	446.635	24.000	0.349	9.799	83.137	665.252
Welsch	4009.178	M-Ridge	248.654	143.897	149.126	7.927	0.391	1.976	21.212	252.068
		RM	495.659	358.134	345.584	18.386	0.329	7.543	63.862	503.526
Cauchy	3398.437	M-Ridge	306.533	180.189	180.040	9.248	0.370	2.842	30.068	310.248
		RM	497.619	349.030	331.268	16.575	0.329	7.888	66.151	504.205
Talwar	7500.811	M-Ridge	633.598	390.162	394.856	22.666	0.429	6.276	61.806	642.454
		RM	953.262	678.261	648.862	34.612	0.405	14.541	119.698	967.474
Ramsay	3017.670	M-Ridge	172.115	93.414	98.636	5.096	0.401	1.217	13.926	174.256
		RM	400.455	284.226	275.344	13.882	0.315	6.269	53.136	406.326
G-Mc	3177.000	M-Ridge	20.322	6.647	8.365	0.931	0.685	0.173	0.812	20.436
		RM	395.365	283.874	275.088	14.118	0.315	6.112	51.940	401.605

Robust estimators are found more successful in the simulation study for the log-normal distributed errors. Among the robust estimators, the proposed M-Ridge estimators generally outperformed the RM estimators. Box plots of MSE values for M-Ridge and RM estimators with different ridge parameters are given in Figure 5.

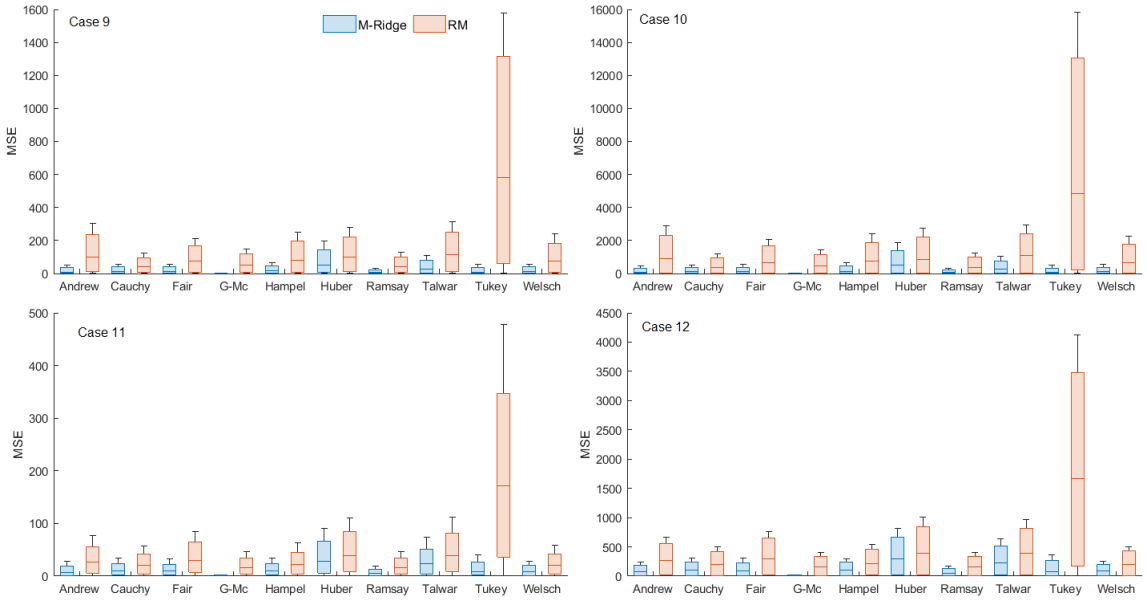

Figure 5: Box plots of MSE values for M-Ridge and RM estimators with different ridge parameters (Cases 9–12).

4. REAL DATA APPLICATION

In this section, we consider Tobacco data (Myers, 1990), which contains both multicollinearity and outlier problems. In application, the dataset centralized and standardized by proportioning the square root of the overall sum of squares. The scatter matrix for the independent variables and the box-plot of the error terms for OLS estimates are given in Figure 6.

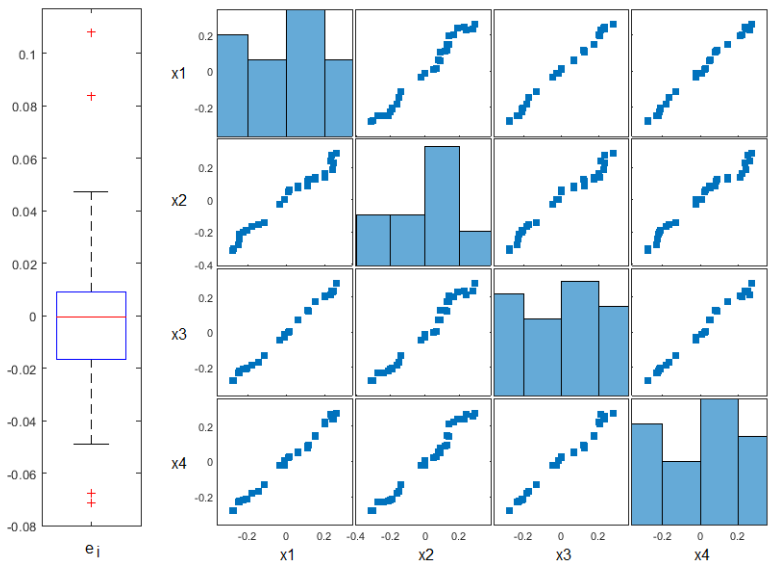

Figure 6: The scatter matrix for the independent variables and the box-plot of the error terms for the OLS estimates.

According to Figure 6, it is clearly seen that the data set contains outlier and multicollinearity problems. The calculated CI value of 43.0758 for the Tobacco dataset indicates the presence of a strong multicollinearity issue.

The performance of the proposed M-Ridge estimators evaluated with the k-fold cross validation technique in the Tobacco dataset. The original data set is randomly partitioned into k equal sized subgroups in k-fold cross validation. Of the k subgroups, a single subgroup is retained as the validation data for testing the model, and the remaining $k-1$ subgroups are used as training data. This process is then repeated k times, with each of the k subgroups used exactly once as the validation data. In k-fold cross validation, the cross validation (CV) as the mean of sum of errors for test data sets is given below:

$$E_i = \sum_{j=1}^{m} (y_{j,\text{test}} - \widehat{y}_{j,\text{test}})^2, \tag{4.1}$$

$$\text{CV} = \frac{1}{k} \sum_{i=1}^{k} E_i \tag{4.2}$$

The performances of the estimators are tested with k-fold cross validation by choosing $k = 10$. The k-fold cross validation results for the proposed estimator and other investigated estimators in the Tobacco dataset are given in Table 16.

Table 16: Comparison of estimation methods with k-fold cross validation method in Tobacco data set ($\times 10^{-4}$).

OLS	68.1432	Ridge	$\widehat{k}_{\text{HK}}$ 68.0701	$\widehat{k}_{\text{HKB}}$ 65.9109	$\widehat{k}_{\text{LW}}$ 68.7738	$\widehat{k}_{\text{HSL}}$ 59.4768	$\widehat{k}_{\text{AM}}$ 60.2571	$\widehat{k}_{\text{GM}}$ 62.8117	$\widehat{k}_{\text{MED}}$ 60.8897	$\widehat{k}_{\text{KS}}$ 68.0576
M		Robust	$\widehat{k}_{\text{HK}}$	$\widehat{k}_{\text{HKB}}$	$\widehat{k}_{\text{LW}}$	$\widehat{k}_{\text{HSL}}$	$\widehat{k}_{\text{AM}}$	$\widehat{k}_{\text{GM}}$	$\widehat{k}_{\text{MED}}$	$\widehat{k}_{\text{KS}}$
Huber	63.1867	M-Ridge	60.6909	58.3774	64.6720	57.5810	**56.3181**	56.9301	56.9256	60.7143
		RM	60.9831	60.5311	65.6347	57.4911	57.1628	58.6564	57.9275	60.9912
Fair	67.9456	M-Ridge	67.7549	65.6060	68.6012	59.3509	59.9869	62.4962	60.6256	67.7434
		RM	67.7889	65.6930	68.6143	59.3478	60.1035	62.6280	60.7455	67.7773
Hampel	68.1432	M-Ridge	68.0701	65.9109	68.7738	59.4768	60.2571	62.8117	60.8897	68.0576
		RM	68.0701	65.9109	68.7738	59.4768	60.2571	62.8117	60.8897	68.0576
Tukey	68.1401	M-Ridge	68.0502	65.8924	68.7567	59.4634	60.2362	62.7901	60.8719	68.0379
		RM	68.0540	65.9005	68.7632	59.4698	60.2493	62.8027	60.8843	68.0416
Andrew	68.1426	M-Ridge	67.9074	65.5226	68.6910	59.8668	59.9523	62.2556	60.4219	67.8935
		RM	68.0671	65.9089	68.7719	59.4755	60.2556	62.8100	60.8887	68.0546
Welsch	68.1425	M-Ridge	68.0656	65.9067	68.7700	59.4738	60.2524	62.8069	60.8857	68.0532
		RM	68.0665	65.9085	68.7714	59.4753	60.2553	62.8097	60.8885	68.0540
Cauchy	68.1425	M-Ridge	68.0656	65.9067	68.7700	59.4738	60.2524	62.8069	60.8857	68.0532
		RM	68.0665	65.9085	68.7714	59.4753	60.2553	62.8097	60.8885	68.0540
Talwar	68.1432	M-Ridge	68.0701	65.9109	68.7738	59.4768	60.2571	62.8117	60.8897	68.0576
		RM	68.0701	65.9109	68.7738	59.4768	60.2571	62.8117	60.8897	68.0576
Ramsay	68.0439	M-Ridge	67.9084	65.7540	68.6823	59.4092	60.1156	62.6480	60.7522	67.8964
		RM	67.9261	65.7997	68.6917	59.4105	60.1784	62.7178	60.8162	67.9141
G-Mc	68.0172	M-Ridge	67.0976	64.8476	73.2344	65.1371	62.7903	63.0754	62.4144	67.0648
		RM	68.0656	65.9080	68.7709	59.4749	60.2549	62.8092	60.8882	68.0532

According to results of k-fold cross validation, M-Ridge and RM estimators as robust estimators are found more successful than OLS, ridge, and M-estimators. In most cases M-Ridge is found superior to RM estimators in comparisons between the proposed M-Ridge with RM estimators. The proposed Huber's M-Ridge estimator by k_{AM} bias parameter is found the most successful estimator which has the smallest CV.

5. CONCLUSIONS

In this study, new robust estimators which named by M-Ridge are improved in case of coexistence of multicollinearity and outlier problems. The M-Ridge estimators are based on the notion that rather than converting M-estimates derived from OLS estimates into ridge estimates as proposed by Silvapulle (1991) and Ertaş (2018), M-estimates are obtained directly from errors based on ridge estimates. The performance of M-Ridge estimators is examined according to estimated MSE and k-fold cross validation in simulations and real data application, respectively. In the simulation study, M-Ridge estimators with generally smaller estimated MSE values outperformed OLS, ridge, M, and RM estimators. In the comparison of robust M-Ridge and RM estimators according to the MSE criterion, the success rates of M-Ridge estimators for normal distributed errors are 95.3125% in the small sample sizes and 100% in the large sample sizes. The success rates of M-Ridge estimators are found 96.875% in small sample sizes and 89.375% in large sample sizes for log-normal distributed errors. In the real data application, the performances of the estimators are tested with k-fold cross validation by choosing $k=10$. According to k-fold cross validation, Huber's M-Ridge estimator based on k_{AM} is found the best estimator. In future studies, new estimators can be obtained by using different biased estimators instead of ridge estimators with the proposed new approach.

ACKNOWLEDGMENTS

The authors received no financial support for this article.

REFERENCES

Beaton, A.E. and Tukey, J.W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185.

Bell, R.M. (1980). *An adaptive choice of the scale parameter for M-estimators*. Stanford Univ. CA, Dept. of Statistics.

Birkes, D. and Dodge, Y. (2011). *Alternative Methods of Regression*. John Wiley & Sons.

De Menezes, D., Prata, D.M., Secchi, A.R., and Pinto, J.C. (2021). A review on robust M-estimators for regression analysis. *Computers & Chemical Engineering*, 147:107254.

Ertaş, H. (2018). A modified ridge M-estimator for linear regression model with multicollinearity and outliers. *Communications in Statistics – Simulation and Computation*, 47(4):1240–1250.

Hampel, F.R. (1968). *Contributions to the theory of robust estimation*. University of California, Berkeley.

Hocking, R.R., Speed, F., and Lynn, M. (1976). A class of biased estimators in linear regression. *Technometrics*, 18(4):425–437.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Hoerl, A.E., Kennard, R.W., and Baldwin, K.F. (1975). Ridge regression: some simulations. *Communications in Statistics – Theory and Methods*, 4(2):105–123.

Holland, P.W. and Welsch, R.E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics – Theory and Methods*, 6(9):813–827.

Huber, P.J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Huber, P.J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799–821.

Khalaf, G. and Shukur, G. (2005). Choosing ridge parameter for regression problems. *Communications in Statistics - Theory and Methods*, 35(5):1177–1182.

Lawless, J. and Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics – Theory and Methods*, 5(4):307–323.

Kibria, B.M.G. (2003). Performance of some new ridge regression estimators, *Communications in Statistics – Simulation and Computation*, 32(2):419–435.

Kibria, B.M.G. (2022). More than hundred(100) estimators for estimating the shrinkage parameter in a linear and generalized linear ridge regression models. *Journal of Econometrics and Statistics*, 2(2):233–252.

Kibria, B.M.G. and Banik, S. (2016). Some ridge regression estimators and their performances. *Journal of Modern Applied Statistical Methods*, 15(1):206–238.

Kim, J.H. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*, 72(6):558–569.

Martin, R.D. and Yohai, V.J. (2006). *Robust Statistics: Theory and Methods*. John Wiley New York.

McDonald, G.C. and Galarneau, D.I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, 70(350):407–416.

Myers, R.H. (1990). *Classical and Modern Regression with Applications*. Duxbury press, Belmont, CA.

Özyurt, D.B. and Pike, R.W. (2004). Theory and practice of simultaneous data reconciliation and gross error detection for chemical processes. *Computers & Chemical Engineering*, 28(3):381–402.

Rey, W.J. (1983). *Introduction to Robust and Quasi-Robust Statistical Methods*. Springer, 131–133.

Silvapulle, M.J. (1991). Robust ridge regression based on an M-estimator. *Australian Journal of Statistics*, 33(3):319–333.

REVSTAT – Statistical Journal
Volume 24, Number 2, April 2026, 157–174
https://doi.org/10.57805/revstat.v24i2.671

Testing the Validity of Lindley Model Based on Nonparametric Probability Density Functions of Entropy Estimators

Authors: Hadi Alizadeh Noughabi ✉
– Department of Statistics, University of Birjand, Birjand, Iran
alizadehhadi@birjand.ac.ir

Mohammad Shafaei Noughabi
– Department of Mathematics and Statistics, University of Gonabad, Gonabad, Iran
mohamad.shafaee@gmail.com

Received: August 2023 Revised: February 2024 Accepted: March 2024

Abstract:

- The Lindley distribution is one of the fundamental models applied for reliability models and in the present article, we propose some test statistics for testing the validity of Lindley model based on correcting moments of nonparametric probability density functions of entropy estimators. Critical points and type I error of the tests are obtained and power values of the tests are computed by Monte Carlo simulation. We show that the proposed tests are more powerful than competitor tests. Finally, the proposed tests are illustrated by a real data example.

Keywords:

- *Lindley distribution; entropy; Kullback-Leibler information; critical points; test power; Monte Carlo simulation.*

AMS Subject Classification:

- 62G10, 94A15.

✉ Corresponding author

1. INTRODUCTION

Entropy, denoted as H, is a measure of uncertainty and information content associated with a continuous random variable. It is a fundamental concept in information theory and provides a measure of the average amount of information required to describe or predict the outcomes of a random variable. Introduced by Shannon (1948), entropy has been a subject of interest for many researchers in the field of statistics. In statistics, entropy estimation is a method used to quantify the uncertainty or information content associated with a random variable or a probability distribution. Entropy estimation is particularly useful when the underlying probability distribution is unknown or when it deviates significantly from known parametric distributions. Non-parametric entropy estimation methods offer a robust alternative for quantifying uncertainty in data without imposing restrictive assumptions. Various estimators for entropy have been proposed, including those by Vasicek (1976), van Es (1992), Ebrahimi *et al.* (1994), Correa (1995), Wieczorkowski and Grzegorzewsky (1999), and Alizadeh (2010).

Entropy estimation has various applications in statistics and related fields. It is widely used in information theory, machine learning, data compression, pattern recognition, and data mining. In statistical analysis, entropy estimation can help identify the complexity or variability of a dataset, assess the amount of information present in a random variable, and compare the uncertainty between different datasets or probability distributions.

In summary, entropy estimation in statistics is a non-parametric method used to quantify the uncertainty or information content associated with a random variable or a probability distribution. It provides a measure of the average amount of information required to describe or predict outcomes.

The entropy $H(f)$, of a continuous random variable X with a density function $f(x)$, is defined as

$$H(f) = -\int_{-\infty}^{\infty} f(x)\, \log f(x)\, dx\,.$$

Consider a random sample $X_1, ..., X_n$ of size n, where the order statistics are denoted as $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. Vasicek (1976) introduced an entropy estimator defined as:

$$HV_{mn} = \frac{1}{n}\sum_{i=1}^{n} \log\Big\{\frac{n}{2m}(X_{(i+m)} - X_{(i-m)})\Big\},$$

where the window size m is a positive integer smaller than $n/2$, $X_{(i)} = X_{(1)}$ if $i < 1$, $X_{(i)} = X_{(n)}$ if $i > n$. He proved the consistency of HV_{mn} as an estimator for the population entropy $H(f)$.

In statistics, a probability density function (pdf) is a function that describes the probability distribution of a continuous random variable. It assigns probabilities to different values or ranges of values that the random variable can take. Traditional parametric pdfs, such as the normal (Gaussian) distribution or the exponential distribution, are defined by a fixed set of parameters that determine their shape, location, and scale. On the other hand, non-parametric pdfs are flexible alternatives that do not rely on specific parametric assumptions about the underlying data distribution. Non-parametric pdf estimation methods aim to estimate the shape and characteristics of the pdf directly from the observed data, without imposing a predefined functional form.

Non-parametric pdf estimation provides several advantages. It allows for more flexibility in modeling complex and diverse data distributions, as it does not restrict the shape of the pdf to a specific parametric form. It also provides a more data-driven approach, relying solely on the observed data to estimate the pdf. This makes non-parametric pdf estimation suitable for a wide range of applications, including exploratory data analysis, hypothesis testing, and modeling in fields such as finance, ecology, and engineering.

Park and Park (2003) derived the nonparametric pdf of Vasicek's estimator as

$$g_v(x) = \begin{cases} 0, & x < \xi_1 \text{ or } x > \xi_{n+1}, \\ \dfrac{1}{n}\dfrac{2m}{x_{(i+m)} - x_{(i-m)}}, & \xi_i < x \le \xi_{i+1}, \quad i = 1, ..., n, \end{cases}$$

where $\xi_i = \big(x_{(i-m)} + \cdots + x_{(i+m-1)}\big)/2m$, and $x_{(i)} = x_{(1)}$ if $i < 1$, $x_{(i)} = x_{(n)}$ if $i > n$.

Ebrahimi *et al.* (1994) modified the Vasicek's estimator as

$$HE_{mn} = \frac{1}{n}\sum_{i=1}^{n} \log\left\{ \frac{n}{c_i m}(X_{(i+m)} - X_{(i-m)}) \right\},$$

where

$$c_i = \begin{cases} 1 + \dfrac{i-1}{m}, & 1 \le i \le m, \\ 2, & m+1 \le i \le n-m, \\ 1 + \dfrac{n-i}{m}, & n-m+1 \le i \le n. \end{cases}$$

They proved that $HE_{mn} \to H(f)$ as $n \to \infty, m \to \infty$, $m/n \to 0$.

The nonparametric pdf of Ebrahimi et al.'s estimator was derived by Park and Park (2003) as:

$$g_e(x) = \begin{cases} 0, & x < \eta_1 \text{ or } x > \eta_{n+1}, \\ \dfrac{1}{n}\dfrac{1}{\eta_{i+1} - \eta_i}, & \eta_i < x \le \eta_{i+1}, \quad i = 1, ..., n, \end{cases}$$

where

$$\eta_i = \begin{cases} \xi_{m+1} - \displaystyle\sum_{k=i}^{m} \frac{1}{m+k-1}\big(x_{(m+k)} - x_{(1)}\big), & \text{if } \ 1 \le i \le m, \\ \dfrac{\big(x_{(i-m)} + \cdots + x_{(i+m-1)}\big)}{2m}, & \text{if } \ m+1 \le i \le n-m+1, \\ \xi_{n-m+1} + \displaystyle\sum_{k=n-m+2}^{i} \frac{1}{n+m-k+1}\big(x_{(n)} - x_{(k-m-1)}\big), & \text{if } \ n-m+2 \le i \le n+1. \end{cases}$$

Alizadeh and Arghami (2010) proposed another entropy estimator as

$$HA_{mn} = \frac{1}{n}\sum_{i=1}^{n} \log\left\{ \frac{n}{a_i m}(X_{(i+m)} - X_{(i-m)}) \right\},$$

where

$$a_i = \begin{cases} 1, & 1 \le i \le m, \\ 2, & m+1 \le i \le n-m, \\ 1, & n-m+1 \le i \le n, \end{cases}$$

and $X_{(i-m)} = X_{(1)}$ for $i \leq m$ and $X_{(i+m)} = X_{(n)}$ for $i \geq n-m$. They proved the consistency of HA_{mn} as an estimator for $H(f)$. The nonparametric pdf of their estimator was derived as:

$$g_a(x) = \begin{cases} 0, & x < \gamma_1 \text{ or } x > \gamma_{n+1}, \\ \frac{1}{n}\frac{1}{\gamma_{i+1} - \gamma_i}, & \gamma_i < x \leq \gamma_{i+1}, \quad i = 1, ..., n, \end{cases}$$

where

$$\gamma_i = \begin{cases} \xi_{m+1} - \frac{1}{m}\sum_{k=i}^{m}\left(x_{(m+k)} - x_{(1)}\right), & \text{if } 1 \leq i \leq m, \\ \frac{\left(x_{(i-m)} + \cdots + x_{(i+m-1)}\right)}{2m}, & \text{if } m+1 \leq i \leq n-m+1, \\ \xi_{n-m+1} + \frac{1}{m}\sum_{k=n-m+2}^{i}\left(x_{(n)} - x_{(k-m-1)}\right), & \text{if } n-m+2 \leq i \leq n+1, \end{cases}$$

and $\xi_i = \left(x_{(i-m)} + \cdots + x_{(i+m-1)}\right)/2m$, $x_{(i)} = x_{(1)}$ if $i < 1$, and $x_{(i)} = x_{(n)}$ if $i > n$.

Park and Park (2003) applied the moments of the nonparametric pdfs of entropy estimators proposed by Vasicek (1976) and Ebrahimi *et al.* (1994) to construct tests for normality and exponentiality. Additionally, Alizadeh and Arghami (2013) introduced other tests for normal and exponential distributions based on the moments of the nonparametric pdfs of their estimators. In this study, we utilize the nonparametric pdfs of entropy estimators and propose goodness-of-fit tests for the Lindley distribution.

The Lindley distribution is an important statistical model for analyzing reliability data with positive support. This distribution was proposed by Lindley (1958) in the context of Bayesian statistics, as a counter example of fiducial statistics. Its density is

$$f(x;\theta) = \frac{\theta^2}{\theta+1}(1+x)e^{-\theta x}, \qquad x > 0, \ \ \theta > 0.$$

The mean and variance of this distribution are

$$\mu = E(X) = \frac{\theta+2}{\theta(\theta+1)},$$

and

$$\sigma^2 = \mathrm{Var}(X) = \frac{\theta^2+4\theta+2}{\theta^2(\theta+1)^2},$$

respectively.

Suppose $X_1, ..., X_n$ is a random sample from the Lindley distribution, both the maximum likelihood estimate (MLE) and method of moments estimate of the parameter θ coincides and is

$$\hat{\theta} = \frac{-\left(\bar{X}-1\right) + \sqrt{\left(\bar{X}-1\right)^2 + 8\bar{X}}}{2\bar{X}}, \qquad \bar{X} > 0.$$

Ghitany *et al.* (2008) conducted a detailed study about various properties of Lindley distribution including skewness, kurtosis, hazard rate function, mean residual life function, stochastic ordering, stress-strength reliability, among other things; estimation of its parameter and application to model waiting time data in a bank. They showed that the estimator $\hat{\theta}$ of θ is positively biased: $E(\hat{\theta}) - \theta > 0$, and it is consistent and asymptotically normal $\sqrt{n}\left(\hat{\theta} - \theta\right) \to N(0, 1/\sigma^2)$.

In the literature of survival analysis and reliability theory, the exponential distribution is widely used as a model of lifetime data. However, the exponential distribution only provides a reasonable fit for modeling phenomenon with constant failure rates. Distributions like gamma, Weibull and lognormal have become suitable alternatives to the exponential distribution in many practical situations. Ghitany *et al.* (2008) found that the Lindley distribution can be a better model than one based on the exponential distribution.

Shanker *et al.* (2015) discussed a comparative study of Lindley and exponential distributions for modelling various lifetime data sets from biomedical science and engineering, and concluded that even though there are lifetime data where exponential distribution gives better fit than Lindley distribution and in majority of data sets Lindley distribution gives better fit than exponential distribution.

In complete sample case, Ghitany *et al.* (2008) developed different distributional properties, reliability characteristics and some inferential procedures for the Lindley distribution. Krishna and Kumar (2011) discussed reliability estimation in Lindley distribution with progressively type II right censored sample. Gupta and Singh (2013) gave parameter estimation of Lindley distribution with hybrid censored data. Also, Al-Mutairi *et al.* (2013) studied inferences on stress-strength reliability for Lindley distribution with complete sample information. Kumar *et al.* (2015) discussed estimation of stress-strength reliability using progressively first failure censoring. These studies suggest that in many real-life situations Lindley distribution serves as a better lifetime model than the so far popular distributions like exponential, gamma, Rayleigh, Weibull etc. For more applications of the Lindley distribution one can see Dey *et al.* (2019), Khan *et al.* (2020), Nadarajah and Chan (2020), Oliveira *et al.* (2021), Al-Babtain *et al.* (2021), Ghitany and Wang (2022), Alrasheedi *et al.* (2022), Ahsan-ul-Haq *et al.* (2022), and Wang and Weib (2023).

Constructing powerful goodness-of-fit tests for the Lindley distribution holds significant importance. In line with this objective, we propose a set of distribution-free goodness-of-fit tests based on the nonparametric pdfs of entropy estimators. In Section 2, we introduce novel goodness-of-fit tests specifically designed for assessing the composite Lindley hypothesis. We present the details of these tests, including their formulation and statistical properties. Section 3 focuses on the critical values and power analysis of the newly developed tests. We provide insights into the determination of critical values and evaluate the statistical power of the tests under different scenarios. To demonstrate the practical efficacy of the proposed tests, we present two real data examples in Section 4. Through these examples, we evaluate the performance and effectiveness of the tests in practice. Finally, in Section 5, we summarize our findings and conclusions drawn from the study. We discuss the implications of the proposed tests and highlight their potential applications in assessing the goodness of fit for the Lindley distribution.

2. THE PROPOSED TESTS

The Kullback-Leibler (KL) divergence is a measure that quantifies how close a probability density $f(x)$ is to a model density $f_0(x)$. For density functions f and f_0, the KL divergence of f from f_0 is defined (Kullback and Leibler, 1951) by

$$D(f, f_0) = \int f(x) \log \frac{f(x)}{f_0(x)}\, dx\,.$$

Note that $D(f, f_0) = 0$ if and only if $f(x) = f_0(x)$ almost everywhere.

Let f represents the true density and $F_0 = \{f_0(.,\theta) : \theta \in \Omega\}$ represents a parametric model for f, where Ω is a subset of R^p. If $f \in F_0$, the minimum value, $\min_{\theta\in\Omega} D(f, f_0(.,\theta))$, of the KL divergence is zero and if $f \notin F_0$ the minimum KL divergence is strictly positive.

Let $X_1, ..., X_n$ be a random sample from an unknown continuous cumulative density function $F(x)$ with a density $f(x)$. We want to test the hypothesis

$$H_0 :\ f(x) = f_0(x;\theta) = \frac{\theta^2}{\theta+1}(1+x)e^{-\theta x}, \qquad x > 0,\ \text{ for some } \theta \in \Theta\,,$$

where θ is unspecified and $\Theta = R^+$. The alternative to H_0 is

$$H_1 :\ f(x) \neq f_0(x;\theta) \qquad \text{for any } \theta \in \Theta\,.$$

Here, we construct tests for the Lindley based on the KL information and the moments of nonparametric pdfs of the entropy estimators as follows.

The asymmetric KL distance of f from f_0 is

$$\begin{aligned} D(f, f_0) &= \int f(x) \log \frac{f(x)}{f_0(x;\theta)}\, dx \\ &= -H(f) - 2\log(\theta) + \log(\theta+1) - E_f(\log(1+X)) + \theta E_f(X)\,, \end{aligned}$$

$D(f, f_0)$ is minimum (zero) if and only if $f = f_0$, where

$$\begin{aligned} D(f, f_0) &= -H(f) - 2\log(\theta) + \log(\theta+1) - E_f(\log(1+X)) + \frac{\theta+2}{\theta+1} \\ &= \log(\theta+1) - 2\log(\theta) - E_f(\log(1+X)) + \frac{\theta+2}{\theta+1} - H(f), \end{aligned}$$

which can be estimated by

$$TV_{mn} = \log(\hat\theta_v + 1) - 2\log(\hat\theta_v) - E_{g_v}(\log(1+X)) + \frac{\hat\theta_v + 2}{\hat\theta_v + 1} - HV_{mn},$$

where

$$\hat\theta_v = \frac{-(E_{g_v}(X) - 1) + \sqrt{(E_{g_v}(X) - 1)^2 + 8E_{g_v}(X)}}{2E_{g_v}(X)},$$

and HV_{mn} is Vasicek entropy estimator. We reject H_0 for large values of TV_{mn}.

Based on the other entropy estimators, different test statistics can be obtained. We have the following test statistics:

$$TE_{mn} = \log(\hat{\theta}_e + 1) - 2\log(\hat{\theta}_e) - E_{g_e}(\log(1+X)) + \frac{\hat{\theta}_e + 2}{\hat{\theta}_e + 1} - HE_{mn},$$

$$TA_{mn} = \log(\hat{\theta}_a + 1) - 2\log(\hat{\theta}_a) - E_{g_a}(\log(1+X)) + \frac{\hat{\theta}_a + 2}{\hat{\theta}_a + 1} - HA_{mn},$$

where

$$\hat{\theta}_e = \frac{-(E_{g_e}(X) - 1) + \sqrt{(E_{g_e}(X) - 1)^2 + 8E_{g_e}(X)}}{2E_{g_e}(X)},$$

$$\hat{\theta}_a = \frac{-(E_{g_a}(X) - 1) + \sqrt{(E_{g_a}(X) - 1)^2 + 8E_{g_a}(X)}}{2E_{g_a}(X)},$$

and HE_{mn} and HA_{mn} are Ebrahimi *et al.* (1994)'s and Alizadeh and Arghami (2010)'s entropy estimator, respectively. We reject H_0 for large values of test statistics.

Here, we obtain $E(X)$ and $E(\log(1+X))$ under the distributions of g_v, g_e and g_a.

Lemma 2.1. *For the distribution of g_e, we have*

$$E_{g_e}(X) = \frac{1}{n}\left(\frac{\eta_1}{2} + \sum_{i=2}^{n}\eta_i + \frac{\eta_{n+1}}{2}\right),$$

and for g_v and g_a, η_i's replace with ξ_i's and γ_i's, respectively.

Proof: Consider the distribution g_e:

$$\begin{aligned} E_{g_e}(X) &= \int_{\eta_1}^{\eta_{n+1}} x g_e(x)dx \\ &= \sum_{i=1}^{n}\int_{\eta_i}^{\eta_{i+1}} \frac{1}{n}\frac{x}{\eta_{i+1}-\eta_i}dx = \sum_{i=1}^{n}\frac{1}{n}\frac{(\eta_{i+1}^2 - \eta_i^2)}{2(\eta_{i+1}-\eta_i)} \\ &= \frac{1}{2n}\sum_{i=1}^{n}(\eta_{i+1}+\eta_i) = \frac{1}{n}\left(\frac{\eta_1}{2} + \sum_{i=2}^{n}\eta_i + \frac{\eta_{n+1}}{2}\right). \end{aligned}$$

Similarly under other distributions $E(X)$ can be computed. □

Lemma 2.2. *For the distribution of g_e, we have*

$$E_{g_e}(\log(1+X)) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(\eta_{i+1}+1)\log(\eta_{i+1}+1) - (\eta_i+1)\log(\eta_i+1)}{\eta_{i+1}-\eta_i}\right) - 1,$$

and for g_v and g_a, η_i's replace with ξ_i's and γ_i's, respectively.

Proof: Consider the distribution g_e:

$$\begin{aligned}
E_{g_e}(\log(1+X)) &= \int_{\eta_1}^{\eta_{n+1}} \log(1+x) g_e(x) dx \\
&= \sum_{i=1}^{n} \int_{\eta_i}^{\eta_{i+1}} \frac{1}{n} \frac{\log(1+x)}{\eta_{i+1}-\eta_i} dx = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\eta_{i+1}-\eta_i} \int_{\eta_i}^{\eta_{i+1}} \log(1+x) dx \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\eta_{i+1}-\eta_i} \Big((\eta_{i+1}+1)\log(\eta_{i+1}+1) - (\eta_{i+1}+1) \\
&\quad - (\eta_i+1)\log(\eta_i+1) + (\eta_i+1) \Big) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{(\eta_{i+1}+1)\log(\eta_{i+1}+1) - (\eta_i+1)\log(\eta_i+1)}{\eta_{i+1}-\eta_i} - 1 \right\} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{(\eta_{i+1}+1)\log(\eta_{i+1}+1) - (\eta_i+1)\log(\eta_i+1)}{\eta_{i+1}-\eta_i} \right\} - 1.
\end{aligned}$$

Similar to the above argument, the others can be computed. □

Theorem 2.1. *Let F be a distribution with a positive support and F_0 be the Lindley distribution, with the unspecified parameter. Then, under H_0,*

$$TV_{mn} \to 0\,, \qquad \text{in probability as } n \to \infty,\ m \to \infty \text{ and } m/n \to 0\,,$$

and, under H_1, TV_{mn} is a consistent test.

Proof: We note that under H_0,

$$\begin{aligned}
&\log(\hat{\theta}_v+1) - 2\log(\hat{\theta}_v) - E_{g_v}(\log(1+X)) + \frac{\hat{\theta}_v+2}{\hat{\theta}_v+1} \\
&\qquad \to \quad \log(\theta+1) - 2\log(\theta) - E_{f_0}(\log(1+X)) + \frac{\theta+2}{\theta+1} = H(f_0),
\end{aligned}$$

and from Vasicek (1976) when $n \to \infty$, $m \to \infty$, and $m/n \to 0$, we have

$$HV_{mn} \to H(f_0).$$

So

$$TV_{mn} \to 0.$$

Also, under H_1,

$$HV_{mn} \to H(f).$$

Thus

$$TV_{mn} \to H(f_0) - H(f) > 0.$$

(All limits are in probability.) □

The above theorem is indeed satisfied for the other tests.

Finally, we can write the proposed test statistics as follows:

$$\begin{aligned} TV_{mn} &= 2+\log\left(\frac{\hat{\theta}_v+1}{\hat{\theta}_v^2}\right)+\frac{1}{\hat{\theta}_v+1} \\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(\xi_{i+1}+1)\log(\xi_{i+1}+1)-(\xi_i+1)\log(\xi_i+1)}{\xi_{i+1}-\xi_i}\right\}-HV_{mn}, \\ TE_{mn} &= 2+\log\left(\frac{\hat{\theta}_e+1}{\hat{\theta}_e^2}\right)+\frac{1}{\hat{\theta}_e+1} \\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(\eta_{i+1}+1)\log(\eta_{i+1}+1)-(\eta_i+1)\log(\eta_i+1)}{\eta_{i+1}-\eta_i}\right\}-HE_{mn}, \\ TA_{mn} &= 2+\log\left(\frac{\hat{\theta}_a+1}{\hat{\theta}_a^2}\right)+\frac{1}{\hat{\theta}_a+1} \\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(\gamma_{i+1}+1)\log(\gamma_{i+1}+1)-(\gamma_i+1)\log(\gamma_i+1)}{\gamma_{i+1}-\gamma_i}\right\}-HA_{mn}. \end{aligned}$$

3. CRITICAL VALUES AND POWER STUDY

The null hypothesis H_0, at the significance level α is rejected if

$$TV_{mn}(TE_{mn}, TA_{mn}) \geq C(\alpha),$$

where the critical point $C(\alpha)$ is determined by the $\alpha-$quantile of the distribution of the test statistics under the null hypothesis.

In order to obtain the critical points of the test statistics, 100,000 samples of size n were generated from the Lindley distribution with the parameter one. For each sample $(n = 10, 20, 30, 40, 50, 75, 100)$ the test statistics was computed and by using these values the critical points $C(\alpha)$, were determined. The following steps outline the methodology for determining the critical values:

1. Generate a random sample, denoted as $X_1, ..., X_n$, of size n, drawn from the Lindley(1) distribution.
2. Calculate the test statistic based on the sample $X_1, ..., X_n$.
3. Repeat steps 1 and 2 a significant number of times, creating a large number of simulated datasets and corresponding test statistic.
4. Determine the $(1-\alpha)$-th quantile of the test statistic obtained from the simulations. This quantile represents the critical value for the test at the desired significance level α.

The critical points of the statistics TV_{mn}, TE_{mn} and TA_{mn} are presented in Table 1.

Table 1: Critical values of the proposed test statistics for $\alpha = 0.05$.

n	m	TV_{mn}	TE_{mn}	TA_{mn}
10	2	0.7007	0.5277	0.4824
20	3	0.4067	0.2999	0.2710
30	4	0.3012	0.2168	0.2009
40	5	0.2464	0.1729	0.1665
50	5	0.2109	0.1531	0.1497
75	6	0.1608	0.1176	0.1209
100	7	0.1330	0.0975	0.1053

Figures 1 and 2 show the empirical probability density functions of the test statistics with Monte Carlo samples. From figures, TA_{mn} have closer values to 0 than the other statistics. Then the bias of TA_{mn} is smallest.

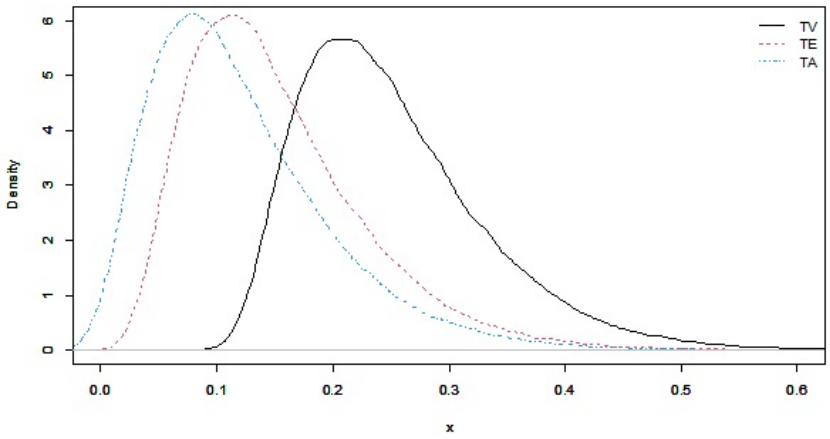

Figure 1: Empirical densities of the test statistics based on 100,000 simulations under the Lindley hypothesis and $n = 20$.

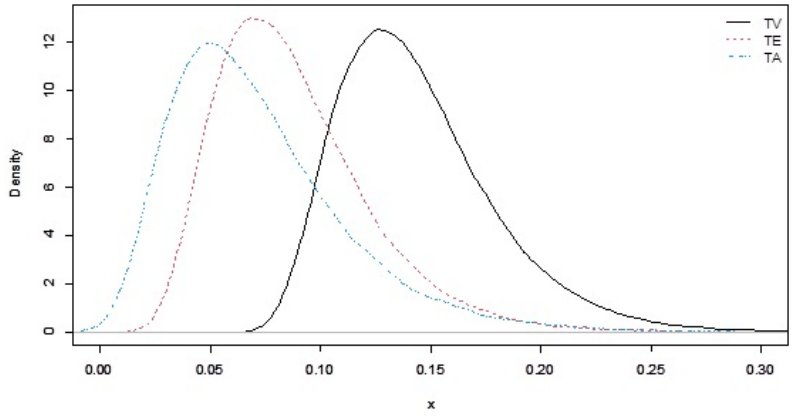

Figure 2: Empirical densities of the test statistics based on 100,000 simulations under the Lindley hypothesis and $n = 50$.

We also evaluate in Table 2 the estimated type I error control using the 0.05 percentiles of the proposed test ($\alpha = 0.05$). We generated random samples from a spectrum of Lindley populations and then obtained the actual sizes of the tests. The results are presented in Table 2.

Table 2: Type I error control of the tests for the nominal significance level $\alpha = 0.05$.

Alternative	n	TV_{mn}	TE_{mn}	TA_{mn}
Lindley(0.5)	10	0.0515	0.0529	0.0517
	20	0.0495	0.0517	0.0524
	30	0.0496	0.0524	0.0517
	50	0.0482	0.0506	0.0491
Lindley(2)	10	0.0514	0.0508	0.0504
	20	0.0485	0.0482	0.0489
	30	0.0473	0.0466	0.0460
	50	0.0491	0.0488	0.0483
Lindley(4)	10	0.0498	0.0493	0.0485
	20	0.0508	0.0491	0.0478
	30	0.0502	0.0490	0.0465
	50	0.0512	0.0506	0.0461
Lindley(4)	10	0.0515	0.0509	0.0491
	20	0.0526	0.0511	0.0463
	30	0.0515	0.0491	0.0428
	50	0.0531	0.0508	0.0416

It is evident, from Table 2, that the actual sizes of the tests are approximately equal to the nominal size 0.05. Therefore, we can conclude that the empirical percentiles presented in Table 1 provides an excellent type I error control.

In practice, a general guide for the choice of m for a fixed n is valuable to the users. Our simulations show that the optimal m (in terms of power) depends on the sample size and the alternative hypothesis. In other hand, there is no a value m that is optimal against all alternatives. Therefore, if one wants to guard against all alternatives a compromise should be made. Based our power study, we recommend for all the test statistics $m = 2$ for $n = 10$, $m = 3$ for $n = 20$, $m = 4$ for $n = 30$, and $m = 5$ for $n = 50$ as the optimal values which the tests attain good (not best) power values against all alternatives. Generally, with increasing n, an optimal choice of m also increases, while the ratio m/n tends to zero.

For power comparison, we consider the well-known tests based on the empirical distribution function (EDF) that used widely in practice. These tests are Cramer von Mises W^2, Kolmogorov-Smirnov D, Anderson-Darling A^2, Kuiper V, and Watson U^2. The test statistics of the EDF-tests are briefly described as follows. For more details about these tests, see D'Agostino and Stephens (1986). Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ are the order statistics based on the random sample $X_1, ..., X_n$.

1. The Cramer-von Mises statistic (1931):

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - F_0(X_{(i)}; \hat{\theta}) \right)^2.$$

2. The Watson statistic (1961):

$$U^2 = W^2 - n\big(\bar{P} - 0.5\big)^2,$$

where $\bar{P}$ is the mean of $F_0(X_{(i)};\hat{\theta}),\ i = 1, ..., n$.

3. The Kolmogorov-Smirnov statistic (1933):

$$D = \max(D^+, D^-),$$

where

$$D^+ = \max_{1\le i\le n}\left\{\frac{i}{n} - F_0(X_{(i)};\hat{\theta})\right\};\quad D^- = \max_{1\le i\le n}\left\{F_0(X_{(i)};\hat{\theta}) - \frac{i-1}{n}\right\}.$$

4. The Kuiper statistic (1960):

$$V = D^+ + D^-.$$

5. The Anderson-Darling statistic (1952):

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\Big\{\log F_0(X_{(i)};\hat{\theta}) + \log\Big[1 - F_0(X_{(n-i+1)};\hat{\theta})\Big]\Big\}.$$

In the above test statistics, $F_0(x)$ is the cumulative density distribution function of the Lindley distribution and $\hat{\theta}$ is the maximum likelihood estimate of the parameter θ.

We compute the power of the considered tests and the proposed tests against various distributions. In power comparison, we considered the following alternatives:

- the Weibull distribution with density $\theta x^{\theta-1}\exp\big(-x^{\theta}\big)$, denoted by $W(\theta)$;
- the gamma distribution with density $\Gamma(\theta)^{-1}x^{\theta-1}\exp(-x)$, denoted by $\Gamma(\theta)$;
- the lognormal distribution $LN(\theta)$ with density $(\theta x)^{-1}(2\pi)^{-1/2}\exp\Big(-(\log x)^2/(2\theta^2)\Big)$;
- the half-normal HN distribution with density $\Gamma(2/\pi)^{1/2}\exp\big(-x^2/2\big)$;
- the uniform distribution U with density 1, $0 \le x \le 1$;
- the modified extreme value $EV(\theta)$, with distribution function $1 - \exp\big(\theta^{-1}(1-e^x)\big)$;
- the linear increasing failure rate law $LF(\theta)$ with density $(1+\theta x)\exp\big(-x - \theta x^2/2\big)$;
- Dhillon's (1981) distribution with distribution function $1 - \exp\Big(-(\log(x+1))^{\theta+1}\Big)$;
- Chen's (2000) distribution $CH(\theta)$, with distribution function $1 - \exp\Big(2\Big(1 - e^{x^{\theta}}\Big)\Big)$.

These alternatives include densities f with decreasing failure rates (DFR), increasing failure rates (IFR) as well as models with unimodal failure rate (UFR) functions and bathtub failure rate (BFR) functions.

To assess the power values of the tests, we generate 100,000 random samples from the alternative hypothesis for different choices of sample sizes and then the test statistics are calculated. Then power of the corresponding test is computed by the frequency of the event "the statistic is in the critical region".

Here's our algorithm for calculating the power of tests:

1. Set the desired significance level (α) for your hypothesis test.
2. Choose the alternative hypothesis you want to test against.
3. Specify the sample size (n) for the test.
4. Set the number of iterations (e.g., 100,000) for generating random samples.
5. Initialize a counter for the number of times the test statistic falls into the critical region.
6. For each iteration:
   - Simulate a random sample from the alternative hypothesis distribution, with the chosen sample size (n).
   - Calculate the test statistic for the generated sample.
   - Determine the critical value corresponding to the chosen significance level (α) and the null hypothesis distribution.
   - If the test statistic falls into the critical region, increment the counter.
7. Calculate the power as the ratio of the counter to the total number of iterations.
8. Return the estimated power value.

By simulating random samples from the alternative hypothesis distribution and comparing the test statistic to the critical value, the algorithm determines the frequency at which the test statistic falls into the critical region, allowing for the estimation of the power of the test.

Tables 3 and 4 display and compares the power values of the tests for sample sizes $n = 10, 20, 30, 50$ at the significance level $\alpha = 0.05$.

Based on the power values in Table 3, it is seen that the tests based on W^2 and TV_{mn} statistics have the most power against IFR alternatives. Although for this type of alternatives the tests W^2 and TV_{mn} have the most power but the power differences of these tests with each other are small and we can select one of the tests based on W^2 or TV_{mn} statistic as a powerful test.

From Table 4, it is evident that the tests based on A^2 or TA_{mn} statistics have the most power against UFR alternatives and power differences between these tests and the other tests are substantial.

Table 4 reveals a superiority of the test based on A^2 statistic to all other tests as we can say that this test outperforms all other tests against DFR and BFR alternatives.

Although there is no uniformly most powerful test against all alternatives, the tests based on W^2, A^2, TV_{mn} and TA_{mn} statistics can be recommended in practice. Depending on type of alternative, we can say that among EDF-based tests, the tests W^2 and A^2 have the most power and among entropy-based tests, the tests TV_{mn}, TE_{mn} and TA_{mn} have the most power. In general, we can conclude that the tests W^2, A^2, TV_{mn} and TA_{mn} have a good performance and therefore can be used in practice.

Table 3: Empirical powers of the tests against IFR alternatives at significance level 5%.

Alternative	n	W^2	D	V	U^2	A^2	TV_{mn}	TE_{mn}	TA_{mn}
$W(1.4)$	10	0.1303	0.1174	0.1104	0.1170	0.0894	0.1445	0.1345	0.1183
	20	0.2258	0.1966	0.1761	0.1884	0.1917	0.2171	0.1985	0.1456
	30	0.3237	0.2691	0.2330	0.2635	0.2967	0.2920	0.2599	0.1486
	50	0.5098	0.4231	0.3736	0.4167	0.5036	0.4088	0.3629	0.1786
$\Gamma(2)$	10	0.1175	0.1028	0.1101	0.1188	0.0810	0.1486	0.1452	0.1315
	20	0.2011	0.1754	0.1772	0.1935	0.1800	0.2323	0.2226	0.1790
	30	0.2879	0.2412	0.2369	0.2687	0.2827	0.3233	0.3072	0.2078
	50	0.4745	0.4014	0.3875	0.4408	0.5104	0.4654	0.4460	0.2727
HN	10	0.0952	0.0887	0.0844	0.0875	0.0678	0.0970	0.0896	0.0773
	20	0.1364	0.1234	0.1084	0.1149	0.1076	0.1286	0.1125	0.0735
	30	0.1835	0.1552	0.1340	0.1446	0.1492	0.1579	0.1298	0.0583
	50	0.2839	0.2321	0.1960	0.2139	0.2445	0.2124	0.1667	0.0526
U	10	0.3386	0.2647	0.3088	0.2957	0.2615	0.4008	0.3658	0.3110
	20	0.6318	0.4888	0.6071	0.5477	0.5793	0.7798	0.7317	0.5952
	30	0.8309	0.6764	0.8143	0.7416	0.8056	0.9494	0.9236	0.7780
	50	0.9756	0.9000	0.9777	0.9417	0.9756	0.9987	0.9971	0.9649
$CH(1)$	10	0.0937	0.0868	0.0772	0.0789	0.0673	0.0967	0.0883	0.0750
	20	0.1364	0.1220	0.0998	0.1061	0.1074	0.1264	0.1095	0.0689
	30	0.1826	0.1557	0.1230	0.1332	0.1477	0.1596	0.1290	0.0563
	50	0.2796	0.2301	0.1810	0.1933	0.2379	0.2145	0.1654	0.0472
$CH(1.5)$	10	0.4268	0.3505	0.3359	0.3553	0.3348	0.3993	0.3714	0.3251
	20	0.7600	0.6343	0.6239	0.6480	0.7160	0.7109	0.6684	0.5439
	30	0.9200	0.8205	0.8176	0.8370	0.9071	0.8898	0.8544	0.6785
	50	0.9943	0.9684	0.9736	0.9763	0.9943	0.9880	0.9796	0.8865
$LF(2)$	10	0.1386	0.1235	0.1113	0.1187	0.0972	0.1366	0.1253	0.1069
	20	0.2282	0.1943	0.1706	0.1802	0.1851	0.1937	0.1697	0.1141
	30	0.3292	0.2723	0.2327	0.2527	0.2828	0.0.2573	0.2161	0.1053
	50	0.5133	0.4204	0.3663	0.3955	0.4662	0.3645	0.3026	0.1185
$LF(4)$	10	0.2056	0.1790	0.1594	0.1700	0.1469	0.1879	0.1732	0.1485
	20	0.3777	0.3160	0.2752	0.2980	0.3192	0.3016	0.2675	0.1884
	30	0.5308	0.4386	0.3864	0.4204	0.4758	0.4157	0.3630	0.2032
	50	0.7680	0.6595	0.6067	0.6401	0.7313	0.5868	0.5207	0.2643
$EV(0.5)$	10	0.0923	0.0861	0.0749	0.0782	0.0670	0.0951	0.0872	0.0744
	20	0.1384	0.1221	0.1020	0.1074	0.1068	0.1293	0.1115	0.0713
	30	0.1833	0.1557	0.1242	0.1345	0.1467	0.1591	0.1285	0.0537
	50	0.2779	0.2262	0.1803	0.1933	0.2378	0.2153	0.1658	0.0454
$EV(1.5)$	10	0.1681	0.1456	0.1420	0.1547	0.1170	0.1745	0.1594	0.1344
	20	0.3359	0.2706	0.2529	0.2634	0.2658	0.3060	0.2696	0.1771
	30	0.4612	0.3645	0.3618	0.3805	0.4152	0.4370	0.3734	0.1929
	50	0.7218	0.5906	0.5811	0.5943	0.6880	0.6424	0.5659	0.2656

Table 4: Empirical powers of the tests against UFR, DFR and BFR alternatives at significance level 5%.

Alternative	n	W^2	D	V	U^2	A^2	TV_{mn}	TE_{mn}	TA_{mn}
$LN(0.8)$	10	0.1413	0.1302	0.1279	0.1403	0.1068	0.1592	0.1710	0.1823
	20	0.2221	0.1968	0.2204	0.2448	0.2110	0.3197	0.3549	0.3806
	30	0.3180	0.2720	0.3268	0.3652	0.3440	0.4968	0.5451	0.5404
	50	0.5147	0.4436	0.5541	0.6054	0.6131	0.7713	0.8079	0.7577
$LN(1.5)$	10	0.5140	0.4823	0.3849	0.4001	0.5544	0.2743	0.3490	0.4208
	20	0.8027	0.7664	0.6690	0.6869	0.8197	0.6437	0.7177	0.7922
	30	0.9257	0.9020	0.8342	0.8489	0.9306	0.8459	0.8892	0.9265
	50	0.9900	0.9842	0.9642	0.9697	0.9905	0.9775	0.9861	0.9904
$DL(1)$	10	0.0877	0.0813	0.0809	0.0862	0.0629	0.1033	0.1096	0.1132
	20	0.1185	0.1064	0.1139	0.1236	0.1041	0.1503	0.1694	0.1878
	30	0.1486	0.1274	0.1445	0.1619	0.1445	0.2112	0.2433	0.2564
	50	0.2123	0.1771	0.2245	0.2533	0.2394	0.3284	0.3734	0.3649
$DL(1.5)$	10	0.1999	0.1735	0.1751	0.1937	0.1462	0.2305	0.2242	0.2063
	20	0.3844	0.3271	0.3228	0.3634	0.3601	0.3950	0.3849	0.3254
	30	0.5568	0.4783	0.4598	0.5241	0.5677	0.5569	0.5479	0.4206
	50	0.8123	0.7363	0.7129	0.7832	0.8509	0.7780	0.7685	0.6043
$W(0.8)$	10	0.1960	0.1750	0.1288	0.1366	0.2748	0.0471	0.0703	0.1019
	20	0.3570	0.3095	0.2295	0.2438	0.4417	0.1259	0.1818	0.2693
	30	0.4933	0.4319	0.3201	0.3476	0.5752	0.2312	0.3076	0.4151
	50	0.7062	0.6330	0.5093	0.5395	0.7720	0.4505	0.5335	0.6151
$\Gamma(0.4)$	10	0.5137	0.4712	0.3701	0.3914	0.7163	0.1941	0.2425	0.2961
	20	0.8109	0.7663	0.6579	0.6850	0.9222	0.5510	0.6146	0.6838
	30	0.9354	0.9074	0.8310	0.8551	0.9810	0.7874	0.8304	0.8673
	50	0.9943	0.9894	0.9697	0.9762	0.9990	0.9693	0.9766	0.9800
$CH(0.5)$	10	0.3912	0.3546	0.2711	0.2860	0.5728	0.1104	0.1449	0.1812
	20	0.6670	0.6127	0.4979	0.5281	0.8141	0.3468	0.4061	0.4780
	30	0.8331	0.7839	0.6733	0.7102	0.9251	0.5728	0.6320	0.6848
	50	0.9669	0.9464	0.8924	0.9137	0.9903	0.8670	0.8927	0.8995

4. APPLICATIONS TO REAL DATA

In this section, to show the behavior of the proposed tests in real cases, two real data sets are analyzed. In Figures 3 and 4, we depict the histogram of these data sets.

Example 4.1. The following data set gives data of the failure times of 25 ball bearings in endurance test presented by (Caroni, 2002):

17.88, 28.92, 33.00, 41.52, 42.12, 45.60, 48.48, 51.84, 51.96,
54.12, 55.56, 67.80, 67.80, 67.80, 68.64, 86.64, 68.88, 84.12,
93.12, 98.64, 105.12, 105.84, 127.92, 128.04, 173.40.

Our procedures can be used to investigate whether the data come from a Lindley distribution. The values of the proposed test statistics are

$$TV_{mn} = 0.2636,\ TE_{mn} = 0.1822,\ TA_{mn} = 0.1409,$$

and the critical values at the 5% are 0.3492, 0.2651, 0.2449, respectively.

Since the values of the test statistics are smaller than the corresponding critical values, the Lindley assumption is not rejected at the significance level of 0.05. Therefore, the null hypothesis that the failure times of ball bearings follow a Lindley distribution is not rejected.

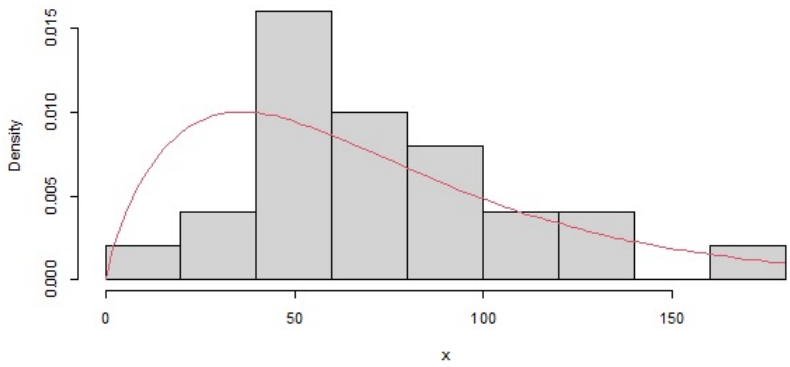

Figure 3: Histogram of failure times of ball bearings and a fitted Lindley density function.

Example 4.2. The following data set is derived from days between air-conditioning failure in Boeing 720 aircraft. These data are presented by (Pearson and Hartley, 1972):

0.013, 0.052, 0.143, 0.208, 0.234, 0.234, 0.234, 0.312, 0.404,
0.508, 0.599, 0.664, 0.703, 0.820, 0.885, 1.002, 1.041, 1.067,
1.263, 1.380, 1.445, 1.836, 1.849, 2.122, 2.486, 2.682, 2.812.

The proposed tests can be used to investigate whether these data come from a Lindley distribution. The values of the test statistics are computed as

$$TV_{mn} = 0.1842,\ TE_{mn} = 0.0822,\ TA_{mn} = 0.0311,$$

and the critical values at the significance level 5% are obtained as 0.3285, 0.2176, 0.1993, respectively.

It is evident that the values of the test statistics are smaller than the corresponding critical values and consequently the Lindley assumption is not rejected at the significance level of 0.05. Therefore, the null hypothesis that the days between air-conditioning failure in Boeing 720 a[illegible]

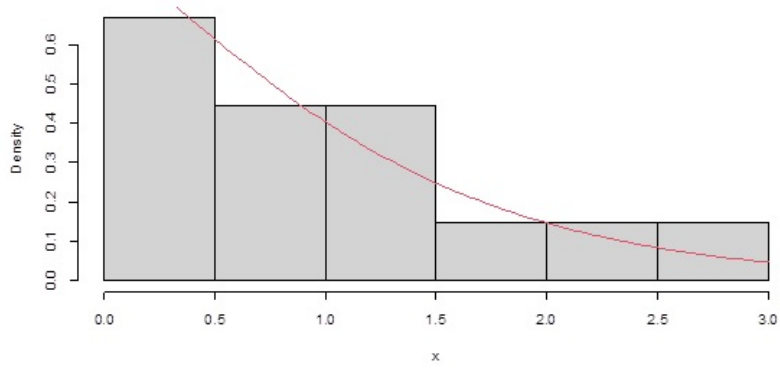

Figure 4: Histogram of days between air-conditioning failure in Boeing 720 aircraft and a fitted Lindley density function.

5. CONCLUSIONS

In this paper, we obtained the minimum discriminant information (MDI) loss estimator for the unknown parameter of the Lindley distribution. Then, based on the nonparametric distribution functions of the entropy estimators, we proposed goodness of fit test statistics for the Lindley distribution. We obtained the power values of the proposed test statistics with Monte Carlo simulation and compared them with the competing test statistics against various alternatives. Generally, we concluded that among the proposed and competing test statistics, the proposed tests have a good performance against different alternatives. Therefore, these tests can be confidently recommended in practice. Finally, we illustrated the performance of the new test statistics in real cases.

ACKNOWLEDGMENTS

The authors are grateful to two anonymous referees and the associate editor for providing some useful comments on an earlier version of this manuscript.

REFERENCES

Ahsan-ul-Haq, M., Choudhary, S.M., AL-Marshadi, A.H., and Aslam, M. (2022). A new generalization of Lindley distribution for modeling of wind speed data. *Energy Reports*, 8:1–11.

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

Alizadeh, H.N. (2010). A new estimator of entropy and its application in testing normality. *Journal of Statistical Computation and Simulation*, 80:1151–1162.

Alizadeh, H.N. and Arghami, N.R. (2010). A new estimator of entropy. *Journal of the Iranian Statistical Society*, 9:53–64.

Alizadeh, H.N. and Arghami, N.R. (2013). Goodness-of-fit tests based on correcting moments of entropy estimators. *Communications in Statistics – Simulation and Computation*, 42:499–513.

Al-Mutairi, D.K., Ghitany, M.E., and Kundu, D. (2013). Inferences on the stress-strength reliability from Lindley distributions. *Communications in Statistics – Theory and Methods*, 42:1443–1463.

Alrasheedi, A., Abouammoh, A., and Kayid, M. (2022). A new flexible extension of the Lindley distribution with applications. *Journal of King Saud University – Science*, 34:101714.

Caroni, C. (2002). The correct ball bearing data. *Lifetime Data Analysis*, 8:395–399.

Correa, J.C. (1995). A new estimator of entropy. *Communications in Statistics – Theory and Methods*, 24:2439–2449.

Dey, S., Nassar, M., and Kumar, D. (2019). Alpha power transformed inverse Lindley distribution: A distribution with an upside-down bathtub-shaped hazard function. *Journal of Computational and Applied Mathematics*, 348:130–145.

Ebrahimi, N., Pflughoeft, K., and Soofi, E. (1994). Two measures of sample entropy. *Statistics and Probability Letters*, 20:225–234.

Ghitany, M.E. and Wang, S. (2022). A note on parameter asymptotics for weighted Lindley distribution. *Communications in Statistics – Simulation and Computation*, 51:1513–1524.

Ghitany, M.E., Atieh, B., and Nadarajah, S. (2008). Lindley distribution and its application. *Mathematics Computing and Simulation*, 78:493–506.

Gupta, P.K. and Singh, B. (2013). Parameter estimation of Lindley distribution with hybrid censored data. *International Journal of System Assurance Engineering and Management*, 4:378–385.

Khan, M.J.S., Sharma, A., and Iqrar, S. (2020). On moments of Lindley distribution based on generalized order statistics. *American Journal of Mathematical and Management Sciences*, 39:214–233.

Krishna, H. and Kumar, K. (2011). Reliability estimation in Lindley distribution with progressively type II right censored sample. *Mathematics and Computers in Simulation*, 82:281–294.

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

Kumar, K., Krishna, H., and Garg, R. (2015). Estimation of $P(Y < X)$ in Lindley distribution using progressively first failure censoring. *International Journal of System Assurance Engineering and Management*, 22:330–341.

Nadarajah, S. and Chan, S. (2020). On moments of the unit Lindley distribution. *Journal of Applied Statistics*, 47:947–949.

Oliveira, R.P., Achcar, J.A., Mazucheli, J., and Bertoli, W. (2021). A new class of bivariate Lindley distributions based on stress and shock models and some of their reliability properties. *Reliability Engineering & System Safety*, 211:107528.

Park, S. and Park, D. (2003). Correcting moments for goodness of fit tests based on two entropy estimates. *Journal of Statistical Computation and Simulation*, 73:685–694.

Pearson, E.S. and Hartley, H.O. (1972). *Biometrika Tables for Statisticians*, Volume H. Cambridge University Press, Cambridge.

Shanker, R., Hagos, F. and Sujatha, S. (2015). On modeling of Lifetimes data using exponential and Lindley distributions. *Biometrics & Biostatistics International Journal*, 2:1–9.

Shannon, C.E. (1948). A mathematical theory of communications. *Bell System Technical Journal*, 27:379–423; 623–656.

van Es, B. (1992). Estimating functionals related to a density by class of statistics based on spacings. *Scandinavian Journal of Statistics*, 19:61–72.

Vasicek, O. (1976). A test for normality based on sample entropy. *Journal of the Royal Statistical Society, Series B*, 38:54–59.

Wang, S. and Weib, C.H. (2023). New characterizations of the (discrete) Lindley distribution and their applications. *Mathematics and Computers in Simulation*, 212:310–322.

Wieczorkowski, R. and Grzegorzewsky, P. (1999). Entropy estimators improvements and comparisons. *Communications in Statistics – Simulation and Computation*, 28:541–567.

REVSTAT – Statistical Journal
Volume 24, Number 2, April 2026, 175–189
https://doi.org/10.57805/revstat.v24i2.663

Incorporating Direct Responses into Optional Randomized Response Models Without Compromising Respondents' Privacy

Author: Muhammad Azeem
– Department of Statistics, University of Malakand, Chakdara, Pakistan
azeemstats@uom.edu.pk

Received: July 2023 Revised: February 2024 Accepted: March 2024

Abstract:

- Survey researchers use randomized response models for collecting data in sensitive surveys. The existing optional scrambling models lack the option of direct response. In practice, some of the respondents may be willing to provide the direct response. An efficient randomization technique is proposed where each respondent is free to provide either direct or protected response. The findings reveal that the option of direct response increases the efficiency of the models without compromising on data privacy. A real-life application of the proposed technique is also presented and the improvement in efficiency is shown for various choices of parameters.

Keywords:

- *efficiency comparison; direct response; protected response; privacy protection; sensitive surveys.*

AMS Subject Classification:

- 62D05.

1. INTRODUCTION

In data collection on sensitive characteristics such as illegal income, the number of cigarettes consumed in a certain period, cheating in an examination, and the number of violation of rules by employees, refusals and incorrect responses by the respondents is a common practice. The large number of refusals result in high rates of non-response which may badly influence the estimates of population parameter. Attempting to get reliable information from the respondents on sensitive variables, Warner (1965) devised a strategy commonly known as the randomized response technique. The Warner (1971)'s procedure was limited to sample surveys on binary sensitive variables. For estimation of the parameters of quantitative sensitive variables, Warner (1971) presented another technique based on an additive scrambling variable. Duncan (1970) studied the calculation methods for analyzing mutual information. Eichhorn and Hayre (1983) proposed a quantitative scrambling strategy where multiplicative noise is utilized in place of additive noise.

Gupta *et al.* (2002) presented a quantitative randomized response technique called optional randomized response model. In the Gupta *et al.* (2002) technique, the respondents are given the choice to either provide the true response or provide a random response. A multiplicative-type optional scrambling technique was presented by Bar-Lev *et al.* (2004). Pluim *et al.* (2003) reviewed the medical image processing literature using mutual-information registration. Kraskpv *et al.* (2004) suggested two new classes of estimators for mutual information. Yan *et al.* (2008) proposed a method for measuring the respondents' privacy level in a given quantitative randomized response model. Zamanzade and Arghami (2011) developed a new estimator of entropy for continuous variables and proved its consistency. In another study, Zamanzade and Mahdizadeh (2017) analyzed entropy estimators under ranked set sampling design. Diana and Perri (2011) presented an efficient randomized response procedure by utilizing both additive and multiplicative noises. An additive-cum-subtractive scrambling technique was introduced by Al-Sobhi *et al.* (2016). Zamanzade and Mahdizadeh (2017) studied goodness-of-fit tests under Phi-divergence.

Attempting to measure the overall quality of a given model, Gupta *et al.* (2018) presented a joint measure of privacy level and efficiency. Narjis and Shabbir (2021) introduced an efficient quantitative scrambling version of the Gjestvang and Sing (2009) technique and proved the improvement over the previous models. The research study of Khalil *et al.* (2021) demonstrated the effect of measurement errors on the mean estimators. Gupta *et al.* (2024) presented an efficient scrambling technique and proved its improvement over the Diana and Perri (2011) model. Zhou *et al.* (2022) proposed a new mutual information-driven Pan-sharpening framework. Other research studies related to various aspects of randomized response techniques were conducted by Kalucha *et al.* (2016), Young *et al.* (2019), Murtaza *et al.* (2021), Zhang *et al.* (2021), Azeem *et al.* (2024), and Azeem (2023), etc.

Recently, Azeem and Salam (2023) introduced an efficient randomized response technique which provides the following two options to the respondents:

(**i**) Report a direct response;
(**ii**) Report a scrambled response.

While the Azeem and Salam (2023) model achieved the improvement in efficiency of the previous models, one drawback of the Azeem and Salam (2023) model was that it forced all of the survey respondents to disclose to the researcher about their choice.

The Azeem and Salam (2023) technique ignored a real-life situation where some of the respondents may be willing to provide the true protected response. That is, some of the respondents may not want to disclose to the researcher about their chosen option.

The proposed technique is more versatile in the sense that it gives the following two options to each respondent:

(**i**) Report a direct response (unprotected true response);

(**ii**) Report a protected response, which further provides two options:

  (**a**) true response,

  (**b**) scrambled response.

In other words, the proposed procedure incorporates the option of true response for those who opt for protected response, which was ignored by the Azeem and Salam (2023) model.

2. Gupta *et al.* (2024) OPTIONAL RANDOMIZED RESPONSE MODEL

Let the population under consideration consists of N units and a simple random sample of n units is obtained with replacement. Let the sensitive variable under study be denoted by Y, and let the additive scrambling variable be denoted by S. Moreover, it is assumed that $E(Y_i) = \mu_y$, $E(S) = 0$, $V(Y_i) = \sigma_Y^2$, $V(S) = \sigma_S^2$. Further, let T denote a multiplicative quantitative scrambling variable with $E(T) = 1$, and $V(T) = \sigma_S^2$, where σ_Y^2, σ_T^2 and σ_S^2 are the population variances of variables Y, T, and S, respectively, whereas μ_y denotes the mean of the sensitive quantitative variable Y. Further, it is also assumed that all three variables work independently of one another.

Gupta *et al.* (2024) presented the following quantitative model:

$$Z = \begin{cases} Y, & \text{with probability } 1 - W, \\ Y + S, & \text{with probability } WA, \\ TY + S, & \text{with probability } W(1 - A), \end{cases} \tag{2.1}$$

where W is the sensitivity level, and A is a constant, $0 < A < 1$. An unbiased estimator of the mean on using the Gupta *et al.* (2024) scrambling model can be written as:

$$\hat{\mu}_G = \frac{1}{n} \sum_{i=1}^{n} Z_i. \tag{2.2}$$

The variance of $\hat{\mu}_G$ can be derived as:

$$\mathrm{Var}(\hat{\mu}_G) = \frac{1}{n}[W(1 - A)\sigma_T^2(\sigma_Y^2 + \mu_Y^2) + \sigma_Y^2 + W\sigma_S^2]. \tag{2.3}$$

The Gupta *et al.* (2024) optional model gives the following three options to the respondents:

(**i**) Report the true response (i.e., no scrambling),

(**ii**) Report the scrambled response using additive scrambling,

(**iii**) Report the scrambled response using additive and multiplicative scrambling.

3. PROPOSED MODEL

Motivated by Gupta *et al.* (2024), a quantitative randomized response technique is introduced. In the proposed technique, before asking the question on sensitive variable, the interviewer first asks the respondent to select one of the two options — the direct response and the protected response. The researcher records the choice of each response along with the reported value of the quantitative sensitive variable. After the survey is finished, the researcher knows the number of respondents opting for direct responses and the number of those who chose to provide a protected response. Let n_1 out of n be the number of respondents opting for direct response, and let $n_2 = n - n_1$ be the number of respondents opting for protected response for privacy protection. Thus, each respondent belongs to one of the two categories of the respondents:

(**i**) The n_1 respondents who choose to provide the direct response Y;

(**ii**) The n_2 respondents who choose to provide a protected response.

The second group further consists of the three types of respondents of the Gupta *et al.* (2024) technique. It only differs in that the group size is n_2 in place of n.

The proposed model offers two options — unprotected/direct response, and randomized response to each respondent before asking the sensitive question. The researcher notes the option chosen by the respondent along with his response to the question being asked. In this way, the values of n_1 and n_2 are random as they vary from sample to sample. That is, the number of respondents opting for the protected response varies from sample to sample. In one sample, 30 out of 100 respondents may choose the direct response option, however, in another sample from the same population, 40 out of 100 respondents may choose the direct response option. As opposed to the existing randomized response models where the researcher doesn't know the exact number of respondents opting for a direct response, the values of n_1 and n_1 in the proposed model are known to the researcher after the completion of the survey. Keeping values of n_1 and n_1 as random values are in line with the real-world situations.

The mean of the first group is:

$$\bar{Y} = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i. \tag{3.1}$$

The mean response based on the second group can be written as:

$$\bar{Z} = \frac{1}{n_2}\sum_{i=1}^{n_2} Z_i, \tag{3.2}$$

where Z_i are the reported responses defined in equation (2.1).

The mean estimator can be expressed as the weighted mean of the two groups of respondents. That is,

$$\hat{\mu}_G = \frac{n_1\bar{Y} + n_2\bar{Z}}{n_1 + n_2}, \tag{3.3}$$

where $n_1 + n_2 = n$.

4. MEAN AND SAMPLING VARIANCE

The mathematical proofs of the unbiasedness of the mean estimator and the derivation of variances using the proposed model, are given in the following theorems.

Theorem 4.1. *The mean estimator $\hat{\mu}_p$ is unbiased for the population mean.*

Proof: Taking expectation on equation (3.3) gives:

$$E(\hat{\mu}_p) = E(\frac{n_1\bar{Y} + n_2\bar{Z}}{n_1 + n_2}) = \frac{n_1E(\bar{Y}) + n_2E(\bar{Z})}{n_1 + n_2}. \tag{4.1}$$

Taking expectation of equation (3.1) yields:

$$E(\bar{Y}) = E(\frac{1}{n_1}\sum_{i=1}^{n_1} Y_i) = \mu_Y. \tag{4.2}$$

Taking expectation of equation (3.2) yields:

$$E(\bar{Z}) = E(\frac{1}{n_2}\sum_{i=1}^{n_2} Z_i). \tag{4.3}$$

Now,

$$E(Z_i) = (1 - W)E(Y) + (WA)E(Y + S) + W(1 - A)E(TY + S),$$

or,

$$E(Z_i) = (1 - W)\mu_y + (WA)(\mu_y + 0) + W(1 - A)(\mu_y + 0) = \mu_Y. \tag{4.4}$$

Using equation (4.4) in (4.3) yields:

$$E(\bar{Z}) = \frac{1}{n_2}\sum_{i=1}^{n_2} \mu_y = \mu_Y. \tag{4.5}$$

Using equation (4.2) and (4.5) in equation (4.1) yields:

$$E(\hat{\mu_p}) = \frac{n_1\mu_Y + n_1\mu_Y}{n_1 + n_2} = \mu_Y. \tag{4.6}$$

□

Theorem 4.2. *The variance of the mean estimators $\hat{\mu}_p$ is given by:*

$$\mathrm{Var}(\hat{\mu_p}) = \frac{\sigma_Y^2}{n} + \frac{n_2}{n^2}[W(1 - A)\sigma_T^2(\sigma_Y^2 + \sigma_T^2) + W\sigma_S^2]. \tag{4.7}$$

Proof: Applying variance on equation (3.3) gives:

$$\mathrm{Var}(\mu_G) = \frac{n_1^2\,\mathrm{Var}(\bar{Y}) + n_2^2\,\mathrm{Var}(\bar{Z})}{(n_1 + n_2)^2}, \tag{4.8}$$

$$\mathrm{Var}(\bar{Y}) = \frac{1}{n_1^2}\sum_{i=1}^{n_1}\mathrm{Var}(Y_i) = \frac{\sigma_Y^2}{n_1}. \tag{4.9}$$

Applying variance on both sides of equation (3.2) yields:

$$\text{Var}(\bar{Z}) = \frac{1}{n_2^2}\sum_{i=1}^{n_2}\text{Var}(Z_i). \tag{4.10}$$

By definition,

$$\text{Var}(Z_i) = E(Z_i^2) - [E(Z_i]^2. \tag{4.11}$$

$E(Z_i^2)$ can be simplified as:

$$E(Z_i^2) = (1-W)E(Y^2) + (WA)(Y+S)^2 + W(1-A)(TY+S)^2,$$

or,

$$E(Z_i^2) = (1-W)E(Y^2) + (WA)(Y^2+S^2+2SY) + W(1-A)(T^2Y^2+S^2+2STY).$$

Using independence of variables, and the assumptions given in Section 2, the above equation simplifies to:

$$E(Z_i^2) = (1-W)(\sigma_Y^2+\mu_Y^2) + (WA)(\sigma_Y^2+\mu_Y^2+\sigma_S^2) + [W(1-A)(\sigma_T^2+1)(\sigma_Y^2+\mu_Y^2) + \sigma_T^2 + 1].$$

On further simplification, the above equation reduces to:

$$E(Z_i^2) = W(1-A)\sigma_T^2(\sigma_Y^2+\mu_Y^2) + \sigma_Y^2 + \mu_Y^2 + W\sigma_S^2. \tag{4.12}$$

Using equation (4.4) and (4.12) in equation (4.11) yields:

$$\text{Var}(Z_i) = W(1-A)\sigma_T^2(\sigma_Y^2+\mu_Y^2) + \sigma_Y^2 + W\sigma_S^2. \tag{4.13}$$

Using equation (4.13) in equation (4.10) yields:

$$\text{Var}(Z_i) = \frac{1}{n_2}[W(1-A)\sigma_T^2(\sigma_Y^2+\mu_Y^2) + \sigma_Y^2 + W\sigma_S^2]. \tag{4.14}$$

Using equation (4.9) and (4.14) in (4.8) yields:

$$\text{Var}(\hat{\mu}_p) = \frac{n_1\sigma_Y^2 + n_2[W(1-A)\sigma_T^2(\sigma_Y^2+\mu_Y^2) + \sigma_Y^2 + W\sigma_S^2]}{(n_1+n_2)^2},$$

or,

$$\text{Var}(\hat{\mu}_p) = \frac{(n_1+n_2)\sigma_Y^2 + n_2[W(1-A)\sigma_T^2(\sigma_Y^2+\mu_Y^2) + \sigma_Y^2 + W\sigma_S^2]}{(n_1+n_2)^2},$$

or,

$$\text{Var}(\hat{\mu}_p) = \frac{\sigma_Y^2}{n} + \frac{n_2}{n^2}[W(1-A)\sigma_T^2(\sigma_Y^2+\mu_Y^2) + \sigma_Y^2 + W\sigma_S^2]. \qquad \square$$

Remark: Clearly, the $\text{Var}(\hat{\mu}_p)$ is a function of n_2, the number of respondents opting for protected responses. This implies that the less the number of respondents going for protected responses, the more efficient the model is.

5. PERFORMANCE EVALUATION

The measure of respondent-privacy introduced by Yan *et al.* (2008) for evaluation of model-performance can be written as:

$$\Delta = E[Z - Y]^2. \tag{5.1}$$

A higher value of Δ translates to a higher the level of respondent-privacy contained in a given model.

The combined measure under the Gupta *et al.* (2018) model as follows:

$$\delta = \frac{MSE}{\Delta}. \tag{5.2}$$

Equation (5.2) reveals that lower values of δ are preferable.

Based on the Gupta *et al.* (2024) quantitative technique, the measure of privacy is given by:

$$\Delta_G = (1 - A)\sigma_T^2(\sigma_Y^2 + \mu_Y^2) + \sigma_S^2. \tag{5.3}$$

The combined measure of model-efficiency and respondent-privacy for the Gupta *et al.* (2024) model can be obtained as:

$$\delta_G = \frac{\mathrm{Var}(\hat{\mu_G})}{\Delta_G} = \frac{1}{n}\left[\frac{W(1 - A)\sigma_T^2(\sigma_Y^2 + \sigma_T^2) + W\sigma_S^2}{(1 - A)\sigma_T^2(\sigma_Y^2 + \mu_y^2) + \sigma_S^2}\right]. \tag{5.4}$$

Since the proposed technique contains n_2 respondents in the second group in place of n, and since equation (5.3) is independent of the sample size n, so the measure of privacy for the proposed model produces the same quantity as in the Gupta *et al.* (2024) method. That is,

$$\Delta_P = (1 - A)\sigma_T^2(\sigma_Y^2 + \mu_Y^2) + \sigma_S^2. \tag{5.5}$$

This means that although the proposed model gives the option of direct response, there is no loss in the privacy protection level.

The joint measure of model-efficiency and respondent-privacy for the proposed technique is given as:

$$\delta_P = \frac{\mathrm{Var}(\hat{\mu_P})}{\Delta_P} = \frac{\sigma_S^2}{n[(1 - A)\sigma_T^2(\sigma_Y^2 + \mu_Y^2) + \sigma_S^2]} + \frac{n_2}{n^2}\left[\frac{W(1 - A)\sigma_T^2(\sigma_Y^2 + \sigma_T^2) + W\sigma_S^2}{(1 - A)\sigma_T^2(\sigma_Y^2 + \mu_Y^2) + \sigma_S^2}\right]. \tag{5.6}$$

Equation (5.6) is a function of, the number of respondents opting for protected response. This implies that the less the number of respondents going for protected responses, the better the quality of the model is.

6. A PRACTICAL DATA COLLECTION EXAMPLE

The suggested technique was used to estimate the true mean μ_Y of the Grade Point Average (GPA) of students. The population under study consisted of the 175 students enrolled in the Department of Statistics of the University of Malakand, Pakistan. For selection of the sample, the simple random sampling technique was applied to choose 40 students out of the total 175 students of undergraduate level. Each of the 40 selected students was given the choice to either provide the direct response or go for a protected response. In the case of opting for a protected response, the student had the choice to either provide the true GPA or report a scrambled GPA. If a respondent did not want to report the direct response, he/she was provided with a deck consisting of 100 cards and a calculator. Each card carried two random values — one each for variable T and S. Keeping in view the situation at hand, the random numbers for both of the scrambling variable S and T were generated by the researcher by utilizing a normal distribution. For generation of random numbers for the additive scrambling variable S, a normal distribution was used having mean 0 and variance 0.5. For generation of random numbers for the multiplicative scrambling variable T, a normal distribution was used having mean 1 and variance 0.5. Those respondents who chose to provide a protected response were instructed not to show the card selected by him/her to the interviewer, and hence it was ensured that the respondent's privacy was protected. Out of 40 selected students, 14 students chose to provide the direct response, and the remaining 26 students chose to go for a protected response.

The values of W and A are decided by the researcher based on his/her prior knowledge about proportion of people who feel that the question is of sensitive nature. In the absence of prior information, a pilot survey may be carried out to get an estimate of W and A. In this case, the researcher decided to choose $W = 0.4$ and $A = 0.5$, so that $1 - W = 0.6$, $WA = 0.2$, and $W(1 - A) = 0.2$. Converting these proportions into percentages, one of the following three statements were recorded on each card:

(**i**) 60 of 100 cards carried the statement:
"Report your true GPA in last exam."

(**ii**) 20 of 100 cards carried the statement:
"Add the value of S to your true GPA and report the result of the addition."

(**iii**) 20 of 100 cards carried the statement:
"Multiply the value of T with your true GPA and then add the value of S and report the result."
The students who opted for protected responses were asked to draw one card at random from the 100 cards and add or multiply the numbers as per the instruction on the selected card.

The responses reported by the 40 sampled students are presented in Table 1.

In Table 1, it may be observed that some of the protected responses exceeded 4.0 despite the fact that the students' actual GPA was based on the scale of 4.0. Any observed response greater than 4.0 clearly indicates that the respondent used some sort of scrambling, although the privacy of the respondent is protected. To estimate the true mean GPA, one may calculate the weighted mean of the two types of responses given in Table 1.

To see the accuracy of the estimator, one may collect the data of the results of all 175 students from department office, calculate the value of population mean, and observe the difference between the estimator and parameter.

Table 1: Observed Responses.

Direct Responses	Protected Responses
3.76, 2.43, 2.73, 3.16, 3.15, 2.26, 2.69, 3.30, 2.92, 3.68, 2.88, 1.76, 2.51, 3.28	1.9667, 4.3816, 3.7816, 3.3618, 3.5346, 2.4980, 1.9408, 3.7091, 2.7870, 3.6079, 4.7927, 3.8478, 1.3655, 2.9668, 4.5787, 1.5271, 2.1962, 1.4274, 1.8135, 3.9360, 2.1973, 3.3528, 3.3941, 2.8133, 4.4831, 2.6674

7. EFFICIENCY COMPARISON

The variance of the mean estimator on the basis of the Gupta *et al.* (2024) model given in equation (2.3) may be re-written in the form:

$$\text{Var}(\hat{\mu}_G) = \frac{\sigma_Y^2}{n} + \frac{n}{n^2}[W(1-A)\sigma_T^2(\sigma_Y^2 + \mu_Y^2) + W\sigma_S^2]. \tag{7.1}$$

The proposed model will be more efficient than the model suggested by Gupta *et al.* (2024) if:

$$\text{Var}(\hat{\mu_p}) \leq \text{Var}(\hat{\mu_G}),$$

or,

$$\frac{\sigma_Y^2}{n} + \frac{n_2}{n^2}[W(1-A)\sigma_T^2(\sigma_Y^2 + \mu_Y^2) + W\sigma_S^2] \leq \frac{\sigma_Y^2}{n} + \frac{n}{n^2}[W(1-A)\sigma_T^2(\sigma_Y^2 + \mu_Y^2) + W\sigma_S^2],$$

or,

$$n_2 \leq n. \tag{7.2}$$

Condition (7.2) always holds. Thus, it is clear that the suggested model is more efficient than the Gupta *et al.* (2024) model. An extreme case where the two models are equally efficient is the situation in which none of the respondents opts for direct response. Otherwise, if at least one respondent opts for direct response, then the suggested technique becomes more efficient than the Gupta *et al.* (2024) model. Since the measure of privacy protection produces the same value for both models, so the proposed model is better when the overall quality of both models is taken into account.

8. GENERALIZATION OF THE FINDINGS TO OTHER MODELS

In addition to the Gupta *et al.* (2024) optional randomized response model, the direct response option can be incorporated into any of the available optional models, and efficiency condition will always hold with no loss in the level of respondent privacy. To see this, let us consider two optional randomized response models: the Gupta *et al.* (2002) and the Bar-Lev *et al.* (2004) technique.

The observed response based on the Gupta *et al.* (2002) model is:

$$Z = \begin{cases} Y, & \text{with probability } p, \\ Y + S, & \text{with probability } 1 - p. \end{cases} \tag{8.1}$$

The mean estimator using the Gupta *et al.* (2002) model can be written as:

$$\hat{\mu}_{G1} = \frac{1}{n} \sum_{i=1}^{n} Z_i, \tag{8.2}$$

where Z is defined in equation (8.1). The sampling variance of the mean can be derived as:

$$\text{Var}(\hat{\mu}_{G1}) = \frac{\sigma_Y^2}{n} + \frac{n}{n^2}(1-p)\sigma_S^2. \tag{8.3}$$

If the option of direct responses is used, then the variance of the mean becomes:

$$\text{Var}(\hat{\mu}_{p1}) = \frac{\sigma_Y^2}{n} + \frac{n_2}{n^2}(1-p)\sigma_S^2. \tag{8.4}$$

Comparing equation (8.3) and (8.4), since $n_2 \leq n$, therefore:

$$\hat{\mu}_{p1} \leq \hat{\mu}_{G1}. \tag{8.5}$$

The measure of privacy for the Gupta *et al.* (2002) model is:

$$\Delta_{G1} = (1-p)\sigma_S^2. \tag{8.6}$$

Since equation (8.6) is independent of the sample size n, so even if the option of direct responses is given to the respondents, there will be no effect on the measure of privacy.

Now consider the Bar-Lev *et al.* (2004) model, the responses reported by the respondents are as follows:

$$Z = \begin{cases} Y, & \text{with probability } p, \\ TY, & \text{with probability } 1 - p. \end{cases} \tag{8.7}$$

The mean estimator using the Bar-Lev *et al.* (2004) model can be expressed as:

$$\hat{\mu}_B = \frac{1}{n} \sum_{i=1}^{n} Z_i,$$

where Z is defined in equation (8.7). The sampling variance can be derived as:

$$\text{Var}(\hat{\mu}_B) = \frac{\sigma_Y^2}{n} + \frac{n}{n^2}(1-p)\sigma_T^2 + (\sigma_Y^2 + \mu_Y^2). \tag{8.8}$$

If the option of direct responses is used, then the variance of the mean becomes:

$$\text{Var}(\hat{\mu}_{P2}) = \frac{\sigma_Y^2}{n} + \frac{n_2}{n^2}(1-p)\sigma_T^2 + (\sigma_Y^2 + \mu_Y^2). \tag{8.9}$$

Comparing equation (8.8) and (8.9), since $n_2 \leq n$, therefore:

$$\text{Var}(\hat{\mu_{p1}}) \leq \text{Var}(\hat{\mu_B}). \tag{8.10}$$

The measure of privacy based on the Bar-Lev *et al.* (2004) model is:

$$\Delta_B = (1-p)\sigma_T^2 + (\sigma_Y^2 + \mu_Y^2). \tag{8.11}$$

Since equation (8.11) doesn't depend on the sample size n, so even if the option of direct responses is given to the respondents, there will be no effect on the measure of privacy.

9. COMPARISON OF MODELS

Table 2 displays the variances of the mean for the proposed model with respect to the Gupta *et al.* (2024) model for different values of n_1 and n_2. The improvement in efficiency can be clearly observed over the Gupta *et al.* (2024) model. The improvement in efficiency can also be observed graphically in Figure 1, Figure 2, and Figure 3.

Table 2: Variances of the mean under different models for $\mu_Y = 20,\ \sigma_Y^2 = 3,\ n = 50$.

W	A	n_1	n_2	$\sigma_s^s = 0.8$		$\sigma_s^s = 2$		$\sigma_s^s = 3$	
				$\sigma_T^s = 1$		$\sigma_T^s = 3$		$\sigma_T^s = 4$	
				$V(\hat{\mu_G})$	$V(\hat{\mu_P})$	$V(\hat{\mu_G})$	$V(\hat{\mu_P})$	$V(\hat{\mu_G})$	$V(\hat{\mu_P})$
0.2	0.3	10	40	1.1916	0.96528	3.4532	2.77456	4.5856	3.68048
		20	30	1.1916	0.73896	3.4532	2.09592	4.5856	2.77536
		30	20	1.1916	0.51264	3.4532	1.41728	4.5856	1.87024
		40	10	1.1916	0.28632	3.4532	0.73864	4.5856	0.96512
	0.7	10	40	0.5468	0.44944	1.5188	1.22704	2.0064	1.61712
		20	30	0.5468	0.35208	1.5188	0.93528	2.0064	1.22784
		30	20	0.5468	0.25472	1.5188	0.64352	2.0064	0.83856
		40	10	0.5468	0.15736	1.5188	0.35176	2.0064	0.44928
0.5	0.3	10	40	2.889	2.3232	8.543	6.8464	11.374	9.1112
		20	30	2.889	1.7574	8.543	5.1498	11.374	6.8484
		30	20	2.889	1.1916	8.543	3.4532	11.374	4.5856
		40	10	2.889	0.6258	8.543	1.7566	11.374	2.3228
	0.7	10	40	1.277	1.0336	3.707	2.9776	4.926	3.9528
		20	30	1.277	0.7902	3.707	2.2482	4.926	2.9796
		30	20	1.277	0.5468	3.707	1.5188	4.926	2.0064
		40	10	1.277	0.3034	3.707	0.7894	4.926	1.0332
0.8	0.3	10	40	4.5864	3.68112	13.6328	10.9182	18.1624	14.54192
		20	30	4.5864	2.77584	13.6328	8.20368	18.1624	10.92144
		30	20	4.5864	1.87056	13.6328	5.48912	18.1624	7.30096
		40	10	4.5864	0.96528	13.6328	2.77456	18.1624	3.68048
	0.7	10	40	2.0072	1.61776	5.8952	4.72816	7.8456	6.28848
		20	30	2.0072	1.22832	5.8952	3.56112	7.8456	4.73136
		30	20	2.0072	0.83888	5.8952	2.39408	7.8456	3.17424
		40	10	2.0072	0.44944	5.8952	1.22704	7.8456	1.61712

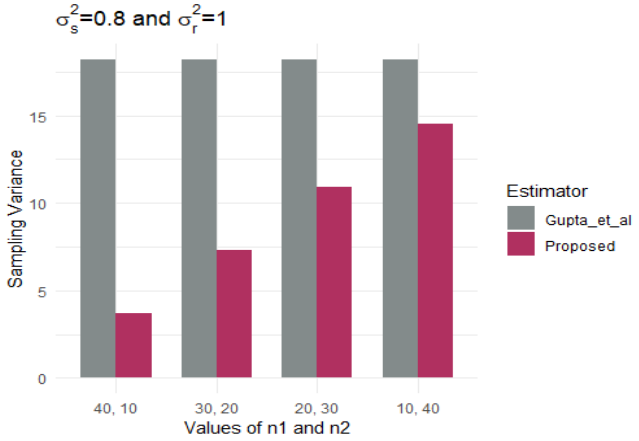

Figure 1: Variance of the Mean under the Gupta *et al.* (2024) and the Proposed Model for $W = 0.2$.

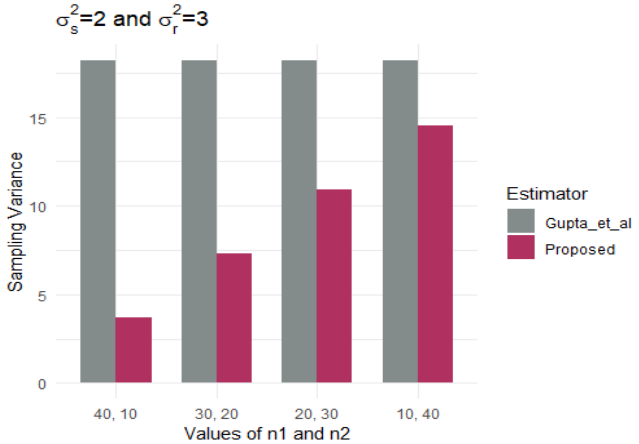

Figure 2: Variance of the Mean under the Gupta *et al.* (2024) and the Proposed Model for $W = 0.5$.

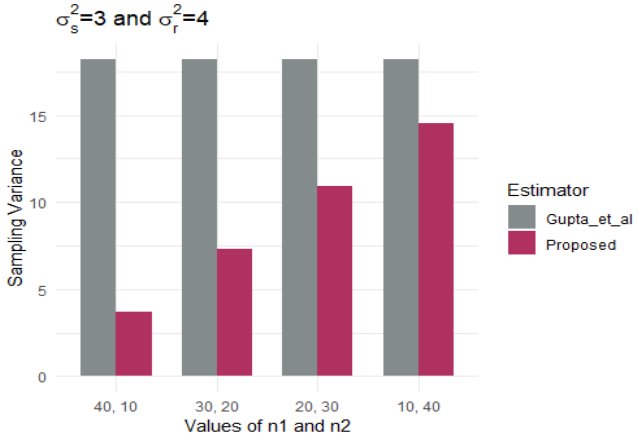

Figure 3: Variance of the Mean under the Gupta *et al.* (2024) and the Proposed Model for $W = 0.8$.

10. DISCUSSION AND CONCLUSION

A modified optional randomized response model using direct responses was presented in Section 3. The current study found that the proposed model is more efficient than the Gupta *et al.* (2024) quantitative technique, with the same level of respondent privacy as the Gupta *et al.* (2024) model. Table 2 the improvement in efficiency over the Gupta *et al.* (2024) technique. Table 2 also reveals that as n_2, the number of respondents choosing to provide a protected response, decreases, the efficiency of the proposed model increases. Thus, it is recommended to the researchers to motivate the respondents to report direct responses as far as possible. A larger number of direct responses in a given survey will result in a smaller number of protected responses, which will result in getting efficient estimates of the mean.

In order to compare the overall quality of the proposed and the Gupta *et al.* (2024) model, δ values for both models were calculated and presented in Table 3 for various choices of n_1 and n_2. The δ values for the proposed model are smaller than the Gupta *et al.* (2024) model which indicate that the suggested model is better than the Gupta *et al.* (2024) model.

Table 3: Variances of the mean under different models for $\mu_Y = 15,\ \sigma_Y^2 = 5,\ n = 50$.

W	A	n_1	n_2	$\sigma_s^s = 0.8$		$\sigma_s^s = 2$		$\sigma_s^s = 3$	
				$\sigma_T^s = 1$		$\sigma_T^s = 3$		$\sigma_T^s = 4$	
				δ_G	δ_P	δ_G	δ_P	δ_G	δ_P
0.2	0.3	10	40	0.004212	0.003412	0.004071	0.003271	0.004053	0.003253
		20	30	0.004212	0.002612	0.004071	0.002471	0.004053	0.002453
		30	20	0.003412	0.001812	0.004071	0.001671	0.004053	0.001653
		40	10	0.004212	0.001012	0.004071	0.000871	0.004053	0.000853
	0.7	10	40	0.004493	0.003693	0.004165	0.003365	0.004123	0.003323
		20	30	0.004493	0.002893	0.004165	0.002565	0.004123	0.002523
		30	20	0.004493	0.002093	0.004165	0.001765	0.004123	0.001723
		40	10	0.004493	0.001293	0.004165	0.000965	0.004123	0.000923
0.5	0.3	10	40	0.010212	0.008212	0.010071	0.008071	0.010053	0.008053
		20	30	0.010212	0.006212	0.010071	0.006071	0.010053	0.006053
		30	20	0.010212	0.004212	0.010071	0.004071	0.010053	0.004053
		40	10	0.010212	0.002212	0.010071	0.002071	0.010053	0.002053
	0.7	10	40	0.010493	0.008493	0.010165	0.008165	0.010123	0.008123
		20	30	0.010493	0.006493	0.010165	0.006165	0.010123	0.006123
		30	20	0.010493	0.004493	0.010165	0.004165	0.010123	0.004123
		40	10	0.010493	0.002493	0.010165	0.002165	0.010123	0.002123
0.8	0.3	10	40	0.016212	0.013012	0.016071	0.012871	0.016053	0.012853
		20	30	0.016212	0.009812	0.016071	0.009671	0.016053	0.009653
		30	20	0.016212	0.006612	0.016071	0.006471	0.016053	0.006453
		40	10	0.016212	0.003412	0.016071	0.003271	0.016053	0.003253
	0.7	10	40	0.016493	0.013293	0.016165	0.012965	0.016123	0.012923
		20	30	0.016493	0.010093	0.016165	0.009765	0.016123	0.009723
		30	20	0.016493	0.006893	0.016165	0.006565	0.016123	0.006523
		40	10	0.016493	0.003693	0.016165	0.003365	0.016123	0.003323

Observing Table 3, it may also be observed that δ value decreases as n_2 decreases which indicate that a survey with lesser number of protected responses, or equivalently, larger number of direct responses, is desirable. In Section 9, it was also observed that the option of direct responses can be incorporated in any of the existing optional randomized response models and will always result in a more efficient model. It was also observed that giving the respondents the option of direct responses doesn't result in the value of Δ, the measure of respondent privacy.

11. FUTURE RESEARCH

The present study analyzed the efficiency of the mean estimator when the option of direct responses is given to the respondents. Further research studies on the estimation of other parameters like population median, and variance can also be conducted under the suggested technique. Moreover, in the current study, the properties of the proposed model have been analyzed under simple random sampling design. Future researchers may use other sampling designs to evaluate the efficiency of the estimators under the proposed model.

ACKNOWLEDGMENTS

I acknowledge the valuable suggestions from the unknown referees which helped me improve the quality of the paper. I am also thankful to Ms. Irsa Sajjad, PhD Scholar at Central South University, China, for help in preparing the revised version of the paper.

REFERENCES

Al-Sobhi, M.M., Hussain, Z., Al-Zahrani, B., Singh, H.P., and Tarray, T.A. (2016). Improved randomized response approaches for additive scrambling models. *Mathematical Population Studies*, 23(4):205–221.

Azeem, M. (2023). Using the exponential function of scrambling variable in quantitative randomized response models. *Mathematical Methods in the Applied Sciences*, 46(13):13882–13893.

Azeem, M. and Salam, A. (2023). Introducing an efficient alternative technique to optional quantitative randomized response models. *Methodology*, 19(1):24–42.

Azeem, M., Hussain, S., Ijaz, M., and Salahuddin, N. (2024). An improved quantitative randomized response technique for data collection in sensitive surveys. *Quality and Quantity*, 58(1):329–341.

Bar-Lev, S.K., Bobovitch, E., and Boukai, B. (2004). A note on randomized response models for quantitative data. *Metrika*, 60(3):255–260.

Diana, G. and Perri, P.F. (2011). A class of estimators for quantitative sensitive data. *Statistical Papers*, 52(3):633–650.

Duncan, T.E. (1970). On the calculation of mutual information. *Metrika*, 19(1):215–220.

Eichhorn, B.H. and Hayre, L.S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, 7(4):307–316.

Gjestvang, C.R. and Sing, S. (2009). An improved randomized response model: Estimation of mean. *Journal of Applied Statistics*, 36(12):1361–1367.

Gupta, S., Gupta, B., and Sing, S. (2002). Estimation of sensitivity level of personal interview survey questions. *Journal of Statistical Planning and Inference*, 100(2):239–247.

Gupta, S., Mehta, S., Shabbir, J., and Khalil, S. (2018). A unified measure of respondent privacy and model efficiency in quantitative RRT models. *Journal of Statistical Theory and Practice*, 12(3):506–511.

Gupta, S., Zhang, J., Khalil, S., and Sapra, P. (2024). Mitigating lack of trust in quantitative randomized response technique models. *Communications in Statistics – Simulation and Computation*, 53(6):2624–2632.

Kalucha, G., Gupta, S., and Shabbir, J. (2016). A two-step approach to ratio and regression estimation of finite population mean using optional randomized response models. *Hacettepe Journal of Mathematics and Statistics*, 45(6):1819–1830.

Khalil, S., Zhang, Q., and Gupta, S. (2021). Mean estimation of sensitive variables under measurement errors using optional RRT models. *Communications in Statistics – Simulation and Computation*, 50(5):1417–1426.

Kraskpv, A., Stogbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69(6):066138.

Murtaza, M., Sing, S., and Hussain, Z. (2021). Use of correlated scrambling variables in quantitative randomized response technique. *Biometrical Journal*, 63(1):134–147.

Narjis, G. and Shabbir, J. (2021). An efficient new scrambled response model for estimating sensitive population mean in successive sampling. *Communications in Statistics – Simulation and Computation*, 52(11):5327–5344.

Pluim, J.P., Maintz, J.A., and Viergever, M.A. (2003). Mutual-information-based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, 22(8):986–1004.

Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69.

Warner, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66(336):884–888.

Yan, Z., Wang, J., and Lai, J. (2008). An efficiency and protection degree-based comparison among the quantitative randomized response strategies. *Communications in Statistics – Theory and Methods*, 38(3):400–408.

Young, A., Gupta, S., and Parks, R. (2019). A binary unrelated-question RRT model accounting for untruthful responding. *Involve, A Journal of Mathematics*, 12(7):1163–1173.

Zamanzade, E. and Arghami, N.R. (2011). Goodness-of-fit test based on correcting moments of modified entropy estimator. *Journal of Statistical Computation and Simulation*, 81(12):2077–2093.

Zamanzade, E. and Mahdizadeh, M. (2017). Entropy estimation from ranked set samples with application to test of fit. *Revista Colombiana de Estadística*, 40(2):223–241.

Zamanzade, E. and Mahdizadeh, M. (2017). Goodness of fit tests for Rayleigh distribution based on Phi-divergence. *Revista Colombiana de Estadística*, 40(2):279–290.

Zhang, Q., Khalil, S., and Gupta, S. (2021). Mean estimation in the simultaneous presence of measurement errors and non-response using optional RRT models under stratified sampling. *Journal of Statistical Computation and Simulation*, 91(17):3492–3504.

Zhou, M., Yan, K., Huang, J., Yang, Z., Fu, X., and Zhao, F. (2022). *Mutual information-driven pan-sharpening*. In "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", 1798-1808.

REVSTAT – Statistical Journal
Volume 24, Number 2, April 2026, 191–212
https://doi.org/10.57805/revstat.v24i2.709

Comparative Study on Probability Density Estimators of Sample Maximum and Data Transformation

Author: Taku Moriyama
– School of Data Science, Yokohama City University, Yokohama, Japan
moriyama.tak.lu@yokohama-cu.ac.jp
– Department of Management of Social Systems and Civil Engineering, Tottori University, Tottori, Japan

Received: November 2023 Revised: February 2024 Accepted: March 2024

Abstract:

- Comparative studies on estimators of the probability density function of sample maximum are conducted. This study presents a plug-in type of and a new block-maxima-based kernel density estimators as the alternatives of the parametric estimator fitting to the approximate generalized extreme value density function. Asymptotic properties of the density estimators are investigated, which shows that the optimal convergence rates depend on the extreme value index of the distribution. Furthermore, this study investigates the density estimators with data log-transformation. It is demonstrated that the log-transformation makes the estimators numerically stable in finite sample case. Finally, two illustrative examples are provided.

Keywords:

- *extreme value; data transformation; kernel-type estimator; mean squared error; nonparametric estimation; sample maximum.*

AMS Subject Classification:

- 62G32, 62G07, 60G70.

1. INTRODUCTION

Let $X_1, X_2, ..., X_n$ be independent and identically distributed random variables with a continuous distribution function F, where the density function f exists. This study considers nonparametric density estimation of sample maximum of future m observation $X_{n+1}, X_{n+2}, ..., X_{n+m}$. The sample maximum density (SMD) $f_{(m)} := mfF^{m-1}$ of the appropriately normalized random variables possibly converges to GEV. This study first surveys the accuracy of the estimator fitted to the density of the generalized extreme distribution (GEV). The accuracy of the fitting estimator fundamentally depends on the approximation to GEV, and the approximation accuracy can be very poor depending the case. Figure 1 shows an image of the convergence of a bimodal SMD the GEV, where the original density is a mixing density of two t-distributions. The fitting to the GEV with small m is neither bimodal nor good approximation to SMD (True).

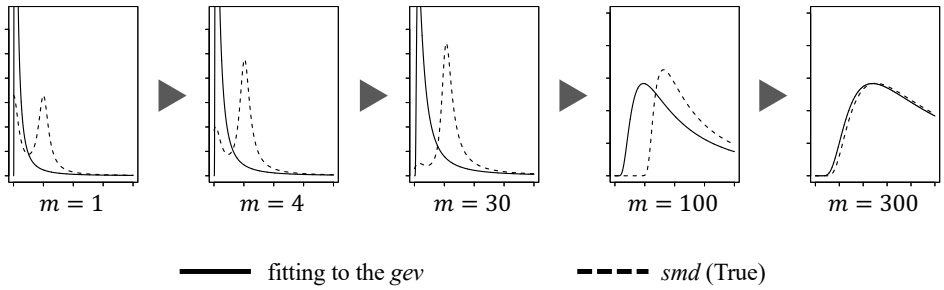

Figure 1: Convergence of SMD of a mixing density of two t-distributions to the GEV.

Nonparametric density estimator can capture properties of the underlying distribution (e.g. multi-modality and skewness). We consider some nonparametric approaches for SMD estimation in this study. Nonparametric density estimation itself is useful for visualization, estimating modes, detecting multi-modality, and so on. Beirlant and Devroye (1999) gave the lower bounds of the total variation and the supremum norm in a nonparametric setting. Using the same settings as Moriyama (2021) investigating sample maximum distribution estimation (SMD) specified below, this study obtains the asymptotic mean squared error (MSE) of the estimators at the point x under the supposition $m \to \infty$, $(m/n) \to 0$ and $x \to \infty$ as $n \to \infty$. We investigate the properties of SMD estimators and conduct comparative studies. A comparison study on extreme quantile estimators are conducted by Banfi *et al.* (2022), and the numerical accuracy of some estimators are reported.

Explicit forms are required in order to compare the asymptotic errors of different estimators, so the tail of F needs to be specified. As shown later, parameters up to the so-called second order have an effect on the first order error of the extreme-based SMD estimator. Let F belong to

- the Hall class **(i)** $\alpha > 0$, $\beta \geq 2^{-1}$, $A > 0$, $B \neq 0$ and
$$x^{\alpha+\beta}\{1 - F(x) - Ax^{-\alpha}(1 + Bx^{-\beta})\} \to 0 \quad \text{as} \quad x \to \infty$$
(see Hall and Welsh, 1984),

- the Weibull class **(ii)** $\kappa > 0$, $C > 0$ and

$$\exp(Cx^{\kappa})\{1 - F(x) - \exp(-Cx^{\kappa})\} \to 0 \quad \text{as} \quad x \to \infty,$$

- or the bounded class **(iii)** $\mu < -2$, $\sigma \leq -2^{-1}$, $D > 0$, $E \neq 0$, $x^* \in \mathbb{R}$ and

$$(x^* - x)^{\mu+\sigma}\{1 - F(x) - (x^* - x)^{-\mu}(D + E(x^* - x)^{-\sigma})\} \to 0 \quad \text{as} \quad x \uparrow x^*$$

(e.g., Stupfler, 2016).

These three classes of distributions are widely utilized as representatives of the distributions belonging to the domain of attractions of the extreme value distribution (see Beirlant *et al.*, 2004; Segers, 2005).

This study proposes a plug-in type of kernel estimator for SMD defined as $\widehat{f}_{(m)} := m\widehat{f}\widehat{F}^{m-1}$, where $\widehat{f}$ and $\widehat{F}$ are the kernel density and kernel distribution estimators of f and F, respectively. Another candidate for the SMD estimator in this study is the new block-maxima-based (BM-based) kernel estimator $\bar{f}_{(m)}$. This BM-based estimator is the kernel density estimator of SMD itself, which is a function of the block maxima. As seen from the definition, the nonparametric estimators are consistent for wide class of distributions as long as m is fixed. The nonparametric estimators are thus a promising candidate for the SMD, at least for relatively small m.

The aim of this study is to investigate the difference between the 'parametric' and nonparametric approaches in SMD estimation, where the parametric' estimator (PE) is the estimator fitted to GEV. Sections 2, 3, and 4 of this paper describe the asymptotic properties of the fitting estimator, the plug-in kernel SMD estimator $\widehat{f}_{(m)}$ and the BM-based kernel SMD estimator $\bar{f}_{(m)}$, respectively. We then prove that the optimal convergence rates of the two nonparametric estimators are different and that the plug-in estimator $\widehat{f}_{(m)}$ outperforms the BM-based estimator as long as the condition of achieving the optimal convergence rate is satisfied. In Section 5, we compare the numerical accuracies of PE and the plug-in estimator. It is demonstrated that the extreme value index severely affects both the theoretical and numerical performances in this study.

The density estimators with data log-transformation are discussed in Section 6, where the log-transformation changes the non-negative extreme indices of the distributions to zero. Section 7 provides these numerical properties, and we see that the log-transformation makes the estimators numerically much stable. Two illustrative examples of application are provided in Section 8. The proofs are given in the Supplementary file.

2. PARAMETRIC DENSITY ESTIMATION OF SAMPLE MAXIMUM

It follows from the Fisher-Tippett-Gnedenko theorem that

$$F^m(x) - G_\gamma\left(\frac{x - b_m}{a_m}\right) \to 0 \quad \text{as} \quad m \to \infty \quad \text{for} \quad 1 + \frac{x - b_m}{a_m} > 0,$$

where G_γ is the generalized extreme value distribution,

$$\gamma := \begin{cases} \alpha^{-1} & \text{for (i)} \\ 0 & \text{for (ii)} \\ \mu^{-1} & \text{for (iii)}, \end{cases}$$

$$a_m := \begin{cases} \gamma(Am)^\gamma & \text{for (i)} \\ \kappa^{-1}C^{-1/\kappa}(\ln m)^{-\theta} & \text{for (ii)} \\ -\gamma(Dm)^\gamma & \text{for (iii)}, \end{cases} \quad b_m := \begin{cases} (Am)^\gamma & \text{for (i)} \\ (C^{-1}\ln m)^{1/\kappa} & \text{for (ii)} \\ x^* - (Dm)^\gamma & \text{for (iii)}, \end{cases}$$

and $\theta := 1 - (1/\kappa)$. The setting in this study follows that the extreme value index $\gamma > -1/2$. This study employs the maximum likelihood based on the BM method in the parameter $\boldsymbol{\gamma}_m := (\gamma, a_m, b_m)$ estimation, where the block size is k and $^\exists N \in \mathbb{N}$ s.t. $n = N \times k$. k can be different from m but needs to satisfy the following assumptions.

Assumption 2.1. One of $(M_n \vee K_n) \to 0$, $(M_n \wedge K_n) \to \infty$, or $^\exists \delta > 0$ s.t. $M_n \to \delta$ and $K_n \to \delta$ holds, where

$$M_n := \begin{cases} Amx^{-\alpha} & \text{for (i)} \\ m\exp(-Cx^\kappa) & \text{for (ii)} \\ Dm(x^* - x)^{-\mu} & \text{for (iii)}, \end{cases} \quad K_n := \begin{cases} Akx^{-\alpha} & \text{for (i)} \\ k^\kappa \exp(-\kappa C^{1/\kappa}(\ln k)^\theta x) & \text{for (ii)} \\ Dk(x^* - x)^{-\mu} & \text{for (iii)}. \end{cases}$$

M_n and K_n have the following meanings $f_{(m)}(x) \sim \alpha x^{-1} M_n \exp(-M_n)$ and $g_{\boldsymbol{\gamma}_k}(x) \sim \alpha x^{-1} K_n \exp(-K_n)$ respectively. Assumption 2.1 means the block size k is asymptotically same as m. The following proposition on the convergence of a bias term of the fitting estimator PE requires the assumption.

Proposition 2.1. *Under Assumption 2.1,*

$$\widetilde{\tau}_n := f_{(m)}(x) - g_{\boldsymbol{\gamma}_k}(x) \to 0,$$

where

$$g_{\boldsymbol{\gamma}_k}(x) := \frac{1}{a_k} g_\gamma\left(\frac{x - b_k}{a_k}\right),$$

$$g_\gamma(x) := \begin{cases} \{1 + \gamma x\}^{-(1/\gamma)-1} \exp(-\{1 + \gamma x\}^{-1/\gamma}), & 1 + \gamma x > 0, \quad \textit{for} \quad \text{(i), (iii)} \\ \exp(-x)\exp(-\exp(-x)), & x \in \mathbb{R}, \quad \textit{for} \quad \text{(ii)}. \end{cases}$$

Set the MLE $\widehat{\boldsymbol{\gamma}}_k$. We note that both Assumptions 2.1 and 2.2 do not hold for the Weibull class.

Assumption 2.2. $^\exists \lambda \in \mathbb{R}$ s.t. $\lambda_n \to \lambda$, where

$$\lambda_n := \begin{cases} km^{-2\beta} & \text{for (i)} \\ k(\ln m)^{-2} & \text{for (ii)} \\ km^{2\sigma} & \text{for (iii)}. \end{cases}$$

Regarding the MSE of PE $g_{\widehat{\boldsymbol{\gamma}}_k}(x)$, the following theorem holds.

Theorem 2.1. *Under Assumptions 2.1 and 2.2,*

$$\mathbb{E}[(f_{(m)}(x) - g_{\widehat{\boldsymbol{\gamma}}_k}(x))^2] \sim (\widetilde{\tau}_n - N^{-1/2}k^{-\gamma}\lambda_n \widetilde{\boldsymbol{\eta}}_n^{\mathsf{T}} I_0^{-1}\boldsymbol{b})^2 + N^{-1}k^{-2\gamma}(\widetilde{\boldsymbol{\eta}}_n^{\mathsf{T}} I_0^{-1}\widetilde{\boldsymbol{\eta}}_n),$$

where $\widetilde{\boldsymbol{\eta}}_n := (\widetilde{s}_n, \widetilde{t}_n, \widetilde{u}_n)^{\mathsf{T}}$,

$$\begin{aligned}
\widetilde{s}_n &:= K_n^{1+\gamma}\exp(-K_n)\{(1-K_n)(1-K_n^{\gamma}+\gamma\ln K_n)+\gamma(1-K_n^{\gamma})\},\\
\widetilde{t}_n &:= K_n^{1+\gamma}\exp(-K_n)(K_n-1)\left(\frac{K_n^{\gamma}-1}{\gamma}\right),\\
\widetilde{u}_n &:= K_n^{1+\gamma}\exp(-K_n)K_n^{\gamma}(1+\gamma-K_n).
\end{aligned}$$

$\boldsymbol{b}$ *and the Fisher information matrix* I_0 *are given in Dombry and Ferreira (2019). Furthermore,*

$$N^{1/2}k^{\gamma}(\widetilde{\boldsymbol{\eta}}_n^{\mathsf{T}} I_0^{-1}\widetilde{\boldsymbol{\eta}}_n)^{-1/2}\{f_{(m)}(x) - g_{\widehat{\boldsymbol{\gamma}}_k}(x) - (\widetilde{\tau}_n - N^{-1/2}k^{-\gamma}\lambda_n\widetilde{\boldsymbol{\eta}}_n^{\mathsf{T}} I_0^{-1}\boldsymbol{b})\}$$

converges in distribution to the standard normal distribution, if the MSE converges to zero.

The symbol $\sim$ means $p_n \sim q_n \iff p_n/q_n = 1 + o(1)$. This paper utilizes the same symbols as Moriyama (2021) for clarity. We add a tilde to the symbols but are actually different. The following corollary on the convergence rate follows from Theorem 2.1. Proposition 2.1 and Theorem 2.1 are proved in the Supplementary file, and Corollary 2.1 is a direct consequence of Theorem 2.1.

Corollary 2.1. *Under the assumptions of Theorem 2.1,* $(f_{(m)}(x) - g_{\widehat{\boldsymbol{\gamma}}_k}(x))$ *converges with the rate the larger of* $f_{(m)}(x) - g_{\boldsymbol{\gamma}_k}(x)$ *and*

$$N^{-1/2}k^{-\gamma}K_n^{1+\gamma}\exp(-K_n) \times \begin{cases} \lambda_n(K_n^{\gamma}+\ln K_n)+\ln K_n & \text{for} \quad (M_n \vee K_n) \to 0\\ \lambda_n + 1 & \text{for} \quad M_n \to \delta,\ K_n \to \delta\\ \lambda_n K_n(K_n^{\gamma}+\ln K_n)+K_n^{1+\gamma} & \text{for} \quad (M_n \wedge K_n) \to \infty. \end{cases}$$

The difference of the SMD estimator and the GEV is decomposed into the three bias terms λ_n, $g_{\boldsymbol{\gamma}_m}(x) - g_{\boldsymbol{\gamma}_k}(x)$ and $f_{(m)}(x) - g_{\boldsymbol{\gamma}_m}(x)$. λ_n corresponds to the asymptotic bias of $\widehat{\boldsymbol{\gamma}}_k$, which depends on the second order parameter. The bias $g_{\boldsymbol{\gamma}_m}(x) - g_{\boldsymbol{\gamma}_k}(x)$ comes from the asymptotic difference between m and k. Although a large k yields a large variance of $\widehat{\boldsymbol{\gamma}}_k$, the precise approximation of $g_{\boldsymbol{\gamma}_m}(x)$ to $f_{(m)}(x)$ (i.e., $\widetilde{\tau}_n$) requires a large m. As the extreme value index tends to be zero, i.e., $\gamma \to 0$, the convergence rate of the approximation of $g_{\boldsymbol{\gamma}_m}(x)$ to $f_{(m)}(x)$ becomes slow. In the Weibull cases, the rate is slower than any polynomial. To sum up, the convergence rate is sensitive to k through m and the first order parameter γ. The convergence rates of the whole of PE in some example cases are given later in Table 1 in Section 5.

3. NONPARAMETRIC PLUG-IN TYPE OF SMD ESTIMATION

Applying the plug-in rule, we obtain the following kernel-type nonparametric estimator (NE1):

$$\widehat{f}_{(m)}(x; h_1, h_2) := m\widehat{f}(x; h_1)\widehat{F}^{m-1}(x; h_2),$$

where $\widehat{f}$ and $\widehat{F}$ are the kernel estimators given by

$$\widehat{f}(x; h_1) = \frac{1}{nh_1}\sum_{i=1}^{n} w\left(\frac{x - X_i}{h_1}\right), \quad \widehat{F}(x; h_2) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x - X_i}{h_2}\right).$$

Assumption 3.1. w is a symmetric and bounded density function, W is the cumulative distribution function, and h_j is bandwidth that satisfies $h_j \to 0$ for $j = 1, 2$. For the bounded class (iii), the support of w be bounded and $h_j(x^* - x)^{-1} \to 0$ for $j = 1, 2$ (it is necessary to avoid the so-called boundary bias problem).

Assumption 3.2. f is twice continuously differentiable at x.

The Hall class and the Weibull class satisfy Assumption 3.2, which requires $\mu \leq -2$ or both $\mu = -1$ and $\sigma \leq -1$ of the bounded class.

Assumption 3.3.

$$\int z^2 w(z)\mathrm{d}z < \infty, \quad \int w^2(z)\mathrm{d}z < \infty, \quad \int zW(z)w(z)\mathrm{d}z < \infty.$$

Using the asymptotic normalities of the estimators, we have the following theorem.

Theorem 3.1. *Given Assumptions 3.1–3.3, then*

$$\mathbb{E}[(f_{(m)}(x) - \widehat{f}_{(m)}(x; h_1, h_2))^2] \sim b_{n,1}^2 + v_{n,1},$$

where

$$b_{n,1} := \frac{h_1^2}{2} \exp(-M_n) M_n^{1+3\gamma} m^{-3\gamma} \psi_n \int z^2 w(z)\mathrm{d}z + \frac{h_2^2}{2} \exp(-M_n) M_n m \xi_n f(x) \int z^2 w(z)\mathrm{d}z,$$

$$v_{n,1} := \frac{m^2}{nh_1} \exp(-2M_n) f(x) \int w^2(z)\mathrm{d}z + \frac{m^3}{n} \exp(-2M_n) M_n f^2(x),$$

$$\psi_n := \begin{cases} \alpha(\alpha+1)(\alpha+2)A^{-3\gamma} & \text{for (i)} \\ \kappa^3 C^3 x^{3\kappa-3} & \text{for (ii)} \\ -\mu(\mu+1)(\mu+2)D^{-3\gamma} & \text{for (iii)}, \end{cases}$$

$$\xi_n := \begin{cases} \alpha(\alpha+1)x^{-2} & \text{for (i)} \\ \kappa^2 C^2 x^{2\kappa-2} & \text{for (ii)} \\ \mu(\mu+1)(x^* - x)^{-2} & \text{for (iii)}, \end{cases}$$

$$\omega_n := \begin{cases} A^{-1}\alpha x^{\alpha-1} & \text{for (i)} \\ \kappa C x^{\kappa-1} \exp(Cx^\kappa) & \text{for (ii)} \\ -D^{-1}\mu(x^* - x)^{\mu-1} & \text{for (iii)}. \end{cases}$$

Furthermore, $v_{n,1}^{-1/2}\{f_{(m)}(x) - \widehat{f}_{(m)}(x; h_1, h_2) + b_{n,1}\}$ *converges in distribution to the standard normal distribution, if the MSE converges to zero.*

The proof of Theorem 3.1 is provided in the Supplementary file, which shows there exists $p_{m,n}(x)$ and $q_{m,n}(x)$ s.t. $f_{(m)}(x) - \widehat{f}_{(m)}(x; h_1, h_2)$ is asymptotically given by

$$p_{m,n}(x)(f(x) - \widehat{f}(x; h_1)) + q_{m,n}(x)(F(x) - \widehat{F}(x; h_2)).$$

Hence, asymptotically optimal bandwidths of NE1 are those of $\widehat{f}(x; h_1)$ and $\widehat{F}(x; h_2)$.

Corollary 3.1. *An asymptotically optimal value of* (h_1, h_2) *is given by*

$$\left(M_n^{-2-6\gamma} m^{2+6\gamma} \psi_n^{-2} f(x) n^{-1} \left(\int z^2 w(z) \mathrm{d}z\right)^{-2} \int w^2(z) \mathrm{d}z\right)^{1/5},$$

$$\left(2\xi_n^{-2} \omega_n n^{-1} \left(\int z^2 w(z) \mathrm{d}z\right)^{-2} \int z W(z) w(z) \mathrm{d}z\right)^{1/3},$$

which are asymptotically identical to those of $\widehat{f}(x; h_1)$ *and* $\widehat{F}(x; h_2)$ *if both of the optimal values converge 0. Under the assumption of Theorem 3.1,* $f_{(m)}(x) - \widehat{f}_{(m)}(x; h_1, h_2)$ *with the optimal bandwidths is asymptotically non-degenerate normal with the following asymptotic mean:*

$$\nu_1 \exp(-M_n) M_n^{(1+3\gamma)/5} m^{(4-3\gamma)/5} \psi_n^{1/5} f^{2/5}(x) n^{-2/5} + \nu_2 \exp(-M_n) M_n m \xi_n^{-1/3} f(x) \omega_n^{2/3} n^{-2/3},$$

where

$$\nu_1 := \frac{1}{2} \left(\int z^2 w(z) \mathrm{d}z\right)^{1/5} \left(\int w^2(z) \mathrm{d}z\right)^{2/5},$$
$$\nu_2 := \left(2 \int z^2 w(z) \mathrm{d}z\right)^{-1/3} \left(\int z W(z) w(z) \mathrm{d}z\right)^{2/3}.$$

Corollary 3.2. *The asymptotically optimal bandwidths for* $M_n = O(1)$ *are of the order*

$$\left(\frac{m}{n}\right)^{1/5} \times \begin{cases} m^{\gamma} & \text{for (i), (iii)} \\ \{C^{-1} \ln m\}^{-\theta} & \text{for (ii),} \end{cases}$$

and

$$\left(\frac{m}{n}\right)^{1/3} \times \begin{cases} m^{\gamma} & \text{for (i), (iii)} \\ \{C^{-1} \ln m\}^{-\theta} & \text{for (ii).} \end{cases}$$

$\widehat{f}_{(m)}(x; h_1, h_2)$ *for* $M_n = O(1)$ *with the optimal bandwidths has the following asymptotic bias of the order:*

$$\left\{\left(\frac{m}{n}\right)^{2/5} + \left(\frac{m}{n}\right)^{2/3}\right\} \times \begin{cases} m^{-\gamma} & \text{for (i), (iii)} \\ (\ln m)^{\theta} & \text{for (ii).} \end{cases}$$

Corollary 3.2 states that the requirement for the bounded class $h_j (x^* - x)^{-1} \to 0$ for $j = 1, 2$ means $m = o(n)$ for $M_n = O(1)$. Corollary 3.2 is directly obtained from Theorem 3.1 and Corollary 3.1.

Ideally, the density estimator itself should be a density function. NE1 satisfies non-negativity but its integral does not equal those if the bandwidth values are different, i.e. $h_1 \neq h_2$. Under the assumption of $h_1 = h_2$ we cannot obtain the algebraic solution.

The nonparametric density estimator presented in the next section has only one hyper-parameter and always satisfies the requirement that the integral equals one.

4. NONPARAMETRIC BLOCK-MAXIMA-BASED ESTIMATION OF SMD

Suppose that the block size is m and $^{\exists}\bar{n} \in \mathbb{N}$ s.t. $n = \bar{n} \times m$. We present here a BM-based kernel-type of nonparametric estimator (NE2),

$$\bar{f}_{(m)}(x;h) = \frac{1}{\bar{n}h}\sum_{i=1}^{\bar{n}} w\left(\frac{x - Y_i}{h}\right),$$

where $Y_j := \max\{X_{m(j-1)+1}, X_{m(j-1)+2}, ..., X_{mj}\}$ $(j = 1, ..., \bar{n})$.

Assumption 4.1. The kernel function w is a symmetric and bounded density function. h is a bandwidth that satisfies $h \to 0$. The support of w be bounded and that $h(x^* - x)^{-1} \to 0$ for (iii) (it is necessary to avoid the so-called boundary bias problem).

The following theorem on the asymptotic normality of the naive nonparametric density estimator holds.

Theorem 4.1. *Given Assumptions 3.2, 3.3 and 4.1, if* $x^{\kappa-1}h \to 0$,

$$\mathbb{E}[(f_{(m)}(x) - \bar{f}_{(m)}(x;h))^2] \sim b_{n,2}^2 + v_{n,2},$$

where

$$\begin{aligned} b_{n,2} :=& \frac{mh^2}{2} f(x)\exp(-M_n)\phi_n \int z^2 w(z)\mathrm{d}z, \\ v_{n,2} :=& (\bar{n}h)^{-1} f_{(m)}(x) \int w^2(z)\mathrm{d}z, \\ \phi_n :=& \begin{cases} m^2 & \text{for} \quad \text{(i), (iii)} \\ (\kappa C)^2 x^{2(\kappa-1)} - 2m\kappa C x^{\kappa-1} + m^2 & \text{for} \quad \text{(ii)}. \end{cases} \end{aligned}$$

Furthermore, $v_{n,2}^{-1/2}\{f_{(m)}(x) - \bar{f}_{(m)}(x;h) + b_{n,2}\}$ *converges in distribution to the standard normal distribution, if the MSE converges to zero.*

Corollary 4.1. *The asymptotically optimal value of* h *is given by*

$$\left(n^{-1}\{f(x)\}^{-1}\phi_n^{-2}\left(\int z^2 w(z)\mathrm{d}z\right)^{-2}\int w^2(z)\mathrm{d}z\right)^{1/5}$$

if both of the optimal values converge 0. Under the assumption of Theorem 4.1, $f_{(m)}(x) - \bar{f}_{(m)}(x;h)$ *with the optimal bandwidths is asymptotically non-degenerate normal with the following asymptotic mean:*

$$\nu_1 n^{-2/5} m\{f(x)\}^{3/5}\phi_n^{1/5}.$$

Corollary 4.2. *The asymptotically optimal bandwidth for* $M_n = O(1)$ *is of the order*

$$n^{-1/5}m^{-3/5} \times \begin{cases} m^{\gamma/5} & \text{for} \quad \text{(i), (iii)} \\ (\ln m)^{-1/5} & \text{for} \quad \text{(ii)}. \end{cases}$$

Under the assumption of Theorem 4.1, $\bar{f}_{(m)}(x;h)$ has the asymptotic bias of the order

$$\left(\frac{m^2}{n}\right)^{2/5}\begin{cases} m^{-(3/5)\gamma} & \text{for (i), (iii)} \\ (\ln m)^{-(3/5)\gamma} & \text{for (ii).} \end{cases}$$

Corollary 4.2 states that the requirement for the bounded class $h(x^* - x)^{-1} \to 0$ means $m = o(n^{-1/(3+4\gamma)})$ for $M_n = O(1)$.

Theorem 4.1 is proved in the Supplementary file. Corollaries 4.1 and 4.2 are the direct consequences.

5. COMPARATIVE STUDY ON SMD ESTIMATORS

In this section, we compare the performances of three estimators: PE, NE1, and NE2. The first comparison is theoretical and the second is numerical. $M_n \equiv K_n \equiv \delta > 0$ is assumed throughout this section. Each of the MSEs of NE1 and NE2 with their respective optimal bandwidths converge on the order of $m^{-2\gamma}(m/n)^{4/5}$ and $m^{-(6/5)\gamma}(m^2/n)^{4/5}$, respectively. The following theorem is immediately evident from Corollaries 3.2 and 4.2.

Theorem 5.1. *Suppose that each of the optimal bandwidths converge to zero. Then, the convergence rate of NE1 $\widehat{f}_{(m)}(x;h_1,h_2)$ with the optimal bandwidth is faster than NE2 $\bar{f}_{(m)}(x;h)$ with the optimal bandwidth for $\gamma > -1$.*

The MSE of PE converges with the rate $m^{-2\gamma} \times (N^{-1}m^{2-4\beta} + m^{-2\gamma\beta} + m^{-2} + N^{-1})$ for the Hall class. For the Weibull class, PE does not satisfy both Assumptions 2.1 and 2.2 with any block size k and is inconsistent. For the bounded class, the convergence rate is $m^{-2\gamma} \times (N^{-1}m^{2+4\sigma} + m^{-2\gamma\sigma} + m^{-2} + N^{-1})$.

The parameters of the distributions and the convergence rates of MSE of the estimators without terms slower than any polynomial are summarized in Table 1, where the hyphen indicates that the distribution breaks the assumption of the theorem. All the target values $f_{(m)}(x)$ tend toward zero with the exception of the bounded class with $\mu = -1$ while $F^m(x) \to \exp(-\delta) > 0$ in every case. The convergence rates divided by the order of $f_{(m)}(x)$ are given in Table 2 (e.g. $(f_{(m)}(x))^{-2} \times \mathbb{E}[(f_{(m)}(x) - g_{\widehat{\gamma}_k}(x))^2]$ for PE). The hyphen in Table 2 changed from a value in Table 1 means that the density estimator becomes inconsistent by dividing $f_{(m)}(x)$. All the convergence rates of NE2 are slower than or the same as those of NE1 in Table 2, as Theorem 5.1 stated.

Comparing NE1 with PE, we see that the rates of PE become especially slow and that NE1 becomes better as γ gets small. For $m = n^{1/4}$, all the rates of NE1 are faster than PE, but the differences become small as m gets large, which basically coincides with the results of the SMD estimation reported in Moriyama (2021). For $m = n^{3/4}$, where the convergence rates of the parametric estimator and the nonparametric estimator of SMD are almost the same and $n^{-1/4}$, the convergence rates of the PE and NE1 of SMD are almost always different. Which one is better depends on γ.

Table 1: Polynomial convergence rates of MSE of the estimators.

Pareto			$m = n^{1/4}$			$m = n^{1/2}$			$m = n^{3/4}$		
ℓ	α	β	PE	NE1	NE2	PE	NE1	NE2	PE	NE1	NE2
1/2	1/2	1	$-1/2$	—	-1	-1	—	$-6/5$	$-3/2$	—	$-7/5$
1	1	1	$-1/2$	—	$-7/10$	-1	—	$-3/5$	$-3/2$	—	$-1/2$
3	3	1	$-1/3$	$-23/30$	$-1/2$	$-2/3$	—	$-1/5$	$-3/4$	—	—
10	10	1	$-1/10$	$-13/20$	$-43/100$	$-1/5$	$-1/2$	$-3/50$	$-3/10$	—	—

T			$m = n^{1/4}$			$m = n^{1/2}$			$m = n^{3/4}$		
ℓ	α	β	PE	NE1	NE2	PE	NE1	NE2	PE	NE1	NE2
1/2	1/2	2	$-3/2$	—	-1	$-5/2$	—	$-6/5$	$-13/4$	—	$-7/5$
1	1	2	$-1/2$	—	$-7/10$	$-3/2$	—	$-3/5$	$-7/4$	—	$-1/2$
3	3	2	$-1/3$	$-23/30$	$-1/2$	$-2/3$	—	$-1/5$	$-3/4$	—	—
10	10	2	$-1/10$	$-13/20$	$-43/100$	$-1/5$	$-1/2$	$-3/50$	$-3/10$	—	—

Burr			$m = n^{1/4}$			$m = n^{1/2}$			$m = n^{3/4}$		
c, ℓ	α	β	PE	NE1	NE2	PE	NE1	NE2	PE	NE1	NE2
1/2,1/2	1/4	1/2	$-5/2$	—	$-8/5$	$-9/2$	—	$-12/5$	$-25/4$	—	$-16/5$
1,1/2	1/2	1	$-1/2$	—	-1	-1	—	$-6/5$	$-3/2$	—	$-7/5$
3,1/2	3/2	3	$-2/3$	—	$-3/5$	$-7/6$	—	$-2/5$	$-5/4$	—	$-1/5$
1/2,1	1/2	1/2	$-3/2$	—	-1	$-3/2$	—	$-6/5$	$-13/4$	—	$-7/5$
1,1	1	1	$-1/2$	—	$-7/10$	-1	—	$-3/5$	$-3/2$	—	$-1/2$
3,1	3	3	$-1/3$	$-23/30$	$-1/2$	$-2/3$	—	$-1/5$	$-3/4$	—	—
1/2,3	3/2	1/2	$-1/2$	—	$-3/5$	-1	—	$-2/5$	$-5/4$	—	$-1/5$
1,3	3	1	$-1/3$	$-23/30$	$-1/2$	$-2/3$	—	$-1/5$	$-3/4$	—	—
3,3	9	3	$-1/9$	$-59/90$	$-13/30$	$-2/9$	$-23/45$	$-1/15$	$-1/3$	—	—

Fr'echet			$m = n^{1/4}$			$m = n^{1/2}$			$m = n^{3/4}$		
γ	α	β	PE	NE1	NE2	PE	NE1	NE2	PE	NE1	NE2
5	1/5	1/5	—	—	$-19/10$	—	—	—	—	—	—
2	1/2	1/2	$-3/2$	—	-1	$-3/2$	—	$-6/5$	$-13/4$	—	$-7/5$
1	1	1	$-1/2$	—	$-7/10$	-1	—	$-3/5$	$-3/2$	—	$-1/2$
1/2	2	1	$-1/2$	$-17/20$	$-11/20$	-1	—	$-3/10$	-1	—	$-1/20$
1/4	4	1	$-1/4$	$-29/40$	$-19/40$	$-1/2$	—	$-3/20$	$-5/8$	—	—

Weibull			$m = n^{1/4}$			$m = n^{1/2}$			$m = n^{3/4}$		
κ	γ	ρ	PE	NE1	NE2	PE	NE1	NE2	PE	NE1	NE2
1/2	0	0	—	$-3/5$	$-2/5$	—	$-2/5$	—	—	$-1/5$	—
1	0	0	—	$-3/5$	$-2/5$	—	$-2/5$	—	—	$-1/5$	—
3	0	0	—	$-3/5$	$-2/5$	—	$-2/5$	—	—	$-1/5$	—
10	0	0	—	$-3/5$	$-2/5$	—	$-2/5$	—	—	$-1/5$	—

inv. Burr			$m = n^{1/4}$			$m = n^{1/2}$			$m = n^{3/4}$		
c, ℓ	μ	σ	PE	NE1	NE2	PE	NE1	NE2	PE	NE1	NE2
3, 2	-6	-2	$-1/12$	$-31/60$	$-7/20$	$-1/6$	$-7/30$	—	—	—	—
1, 2	-2	-2	—	$-7/20$	$-1/4$	—	—	—	—	—	—
1/2, 2	-1	-2	—	$-1/10$	$-1/10$	—	—	—	—	—	—
3,1	-3	-1	—	$-13/30$	$-3/10$	—	$-1/15$	—	—	—	—
1, 1	-1	-1	—	$-1/10$	$-1/10$	—	—	—	—	—	—
1/2, 1	$-1/2$	-1	—	—	—	—	—	—	—	—	—
3, 1/3	-1	$-1/3$	—	—	—	—	—	—	—	—	—
1, 1/3	$-1/3$	$-1/3$	—	—	—	—	—	—	—	—	—
1/2, 1/3	$-1/6$	$-1/3$	—	—	—	—	—	—	—	—	—

Table 2: Normalized polynomial convergence rates of MSE of the estimators.

Pareto			$m = n^{1/4}$			$m = n^{1/2}$			$m = n^{3/4}$		
ℓ	α	β	PE	NE1	NE2	PE	NE1	NE2	PE	NE1	NE2
1/2	1/2	1	—	—	—	—	—	—	—	—	—
1	1	1	—	—	−1/5	—	—	—	—	—	—
3	3	1	−1/6	−3/5	−1/3	−1/3	—	—	−1/4	—	—
10	10	1	−1/20	−3/5	−19/50	−1/10	−2/5	—	−3/20	—	—

T			$m = n^{1/4}$			$m = n^{1/2}$			$m = n^{3/4}$		
ℓ	α	β	PE	NE1	NE2	PE	NE1	NE2	PE	NE1	NE2
1/2	1/2	2	−1/2	—	—	−1/2	—	—	−1/4	—	—
1	1	1	—	—	−1/5	−1/2	—	—	−1/4	—	—
3	3	1	−1/6	−3/5	−1/3	−1/3	—	—	−1/4	—	—
10	10	1	−1/20	−3/5	−19/50	−1/10	−2/5	—	−3/20	—	—

Burr			$m = n^{1/4}$			$m = n^{1/2}$			$m = n^{3/4}$		
c, ℓ	α	β	PE	NE1	NE2	PE	NE1	NE2	PE	NE1	NE2
1/2,1/2	1/4	1/2	−1/2	—	—	−1/2	—	—	−1/4	—	—
1,1/2	1/2	1	—	—	—	—	—	—	—	—	—
3,1/2	3/2	3	−1/3	—	−4/15	−1/2	—	—	−1/4	—	—
1/2,1	1/2	1/2	−1/2	—	—	—	—	—	−1/4	—	—
1,1	1	1	—	—	−1/5	—	—	—	—	—	—
3,1	3	3	−1/6	−3/5	−1/3	−1/3	—	—	−1/4	—	—
1/2,3	3/2	1/2	−1/6	—	−4/15	−1/3	—	—	−1/4	—	—
1,3	3	1	−1/6	−3/5	−1/3	−1/3	—	—	−1/4	—	—
3,3	9	3	−1/18	−3/5	−17/45	−1/9	−2/5	—	−1/6	—	—

Fréchet			$m = n^{1/4}$			$m = n^{1/2}$			$m = n^{3/4}$		
γ	α	β	PE	NE1	NE2	PE	NE1	NE2	PE	NE1	NE2
5	1/5	1/5	—	—	—	—	—	—	—	—	—
2	1/2	1/2	−1/2	—	—	—	—	—	−1/4	—	—
1	1	1	—	—	−1/5	—	—	—	—	—	—
1/2	2	1	−1/4	−3/5	−3/10	−1/2	—	—	−1/4	—	—
1/4	4	1	−1/8	−3/5	−7/20	−1/4	—	—	−1/4	—	—

Weibull			$m = n^{1/4}$			$m = n^{1/2}$			$m = n^{3/4}$		
κ	γ	ρ	PE	NE1	NE2	PE	NE1	NE2	PE	NE1	NE2
1/2	0	0	—	−3/5	−2/5	—	−2/5	—	—	−1/5	—
1	0	0	—	−3/5	−2/5	—	−2/5	—	—	−1/5	—
3	0	0	—	−3/5	−2/5	—	−2/5	—	—	−1/5	—
10	0	0	—	−3/5	−2/5	—	−2/5	—	—	−1/5	—

inv. Burr			$m = n^{1/4}$			$m = n^{1/2}$			$m = n^{3/4}$		
c, ℓ	μ	σ	PE	NE1	NE2	PE	NE1	NE2	PE	NE1	NE2
3, 2	−6	−2	−1/6	−3/5	−13/30	−1/3	−2/5	—	—	−1/5	—
1, 2	−2	−2	—	−3/5	−1/2	—	−2/5	—	—	−1/5	—
1/2, 2	−1	−2	—	−3/5	−3/5	—	−2/5	—	—	−1/5	—
3,1	−3	−1	−1/6	−3/5	−7/15	−1/6	−2/5	—	—	−1/5	—
1, 1	−1	−1	—	−3/5	−3/5	—	−2/5	—	—	−1/5	—
1/2, 1	−1/2	−1	—	—	—	—	—	—	—	—	—
3, 1/3	−1	−1/3	—	—	—	—	—	—	—	—	—
1, 1/3	−1/3	−1/3	—	—	—	—	—	—	—	—	—
1/2, 1/3	−1/6	−1/3	—	—	—	—	—	—	—	—	—

We next investigate the numerical properties of the estimators. NE1 seems theoretically better in many cases, but in the theoretically comparative study, the convergence rate of the bandwidth estimator is not taken into account. In this numerical study, we also examine the effect of bandwidth selection. By simulating the mean integrated squared error (MISE) of PE $g_{\widehat{\gamma}_k}$,

$$L_m^{-1}\int_{Q_m(0.1)}^{Q_m(0.9)}\left(g_{\widehat{\gamma}_k}(x)-f_{(m)}(x)\right)^2\mathrm{d}x,$$

and that of $\widehat{f}_{(m)}$, we investigated the numerical accuracy in small-sample cases, where $L_m := Q_m(0.9)-Q_m(0.1)$ and $Q_m(q)$ denotes the q-th quantile of the SMD. Tables 3–4 show the mean values and the standard deviation (sd) of the obtained MISE values after 1000 simulations. The underlying distributions F were Pareto distributions, T distributions, Burr distributions, Fréchet distributions, Weibull distributions, and inverse Burr distributions.

Table 3: Scaled MISE values and sd values of PE and NE1 where $n = 2^8$.

Pareto	$m=n^{1/4}$						$m=n^{1/2}$						$m=n^{3/4}$					
ℓ	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd
1/2	0.242	0.443	8.065	1.721	7.059	11.99	9.222	108.9	9.339	3.345	487.7	6491	51.01	383.8	108.6	296.2	22770	311400
1	0.037	0.041	0.353	0.393	0.771	0.422	0.510	5.696	1.474	1.156	7.401	12.05	33.36	696.2	20.49	27.17	122.8	500.6
3	0.019	0.021	0.133	0.057	0.088	0.036	0.116	0.162	0.799	0.344	0.514	0.178	54.79	1202	5.143	2.904	3.231	1.437
10	0.016	0.017	0.102	0.046	0.054	0.025	0.104	0.176	0.447	0.202	0.240	0.105	7.608	71.76	2.073	1.007	1.093	0.484

T	$m=n^{1/4}$						$m=n^{1/2}$						$m=n^{3/4}$					
ℓ	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd
1/2	0.996	0.986	7.695	0.921	4.380	2.679	6.928	53.80	9.947	0.953	146.1	3708	41.92	205.4	31.32	61.64	28650	856500
1	0.068	0.067	0.279	0.323	0.314	0.267	0.367	0.744	0.909	0.446	2.751	7.268	42.09	695.3	8.641	9.806	45.99	296.1
3	0.016	0.014	0.021	0.020	0.018	0.014	0.101	0.213	0.095	0.075	0.088	0.065	12.73	215.9	0.507	0.336	0.496	0.322
10	0.011	0.012	0.015	0.015	0.012	0.010	0.088	0.137	0.049	0.044	0.042	0.036	7.153	72.64	0.193	0.177	0.162	0.144

Burr	$m=n^{1/4}$						$m=n^{1/2}$						$m=n^{3/4}$					
c,ℓ	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd
1/2,1/2	102.4	156.6	714.1	1.879	361.2	139.4	548.3	1212	942.8	14.97	942.3	0.909	1306	6052	9436	75610	1037	0.000
1,1/2	0.217	0.416	8.115	1.693	7.056	11.85	5.507	22.83	9.281	3.120	385.1	5181	43.43	360.3	104.3	258.9	84390	2319000
3,1/2	0.023	0.029	0.097	0.068	0.113	0.093	0.178	0.341	0.754	0.386	0.866	0.485	12.23	107.1	7.402	5.880	8.760	9.997
1/2,1	0.411	0.674	11.52	1.685	10.20	27.86	7.554	46.29	10.36	4.495	1590	41270	44.30	288.4	117.4	394.4	11970	212600
1,1	0.041	0.046	0.309	0.350	0.849	0.470	0.391	1.235	1.456	1.274	8.263	12.39	17.72	160.2	20.57	26.62	159.3	738.4
3,1	0.014	0.015	0.034	0.025	0.031	0.019	0.112	0.190	0.187	0.121	0.176	0.090	7.005	39.81	1.117	0.676	1.072	0.509
1/2,3	0.072	0.072	0.242	0.209	1.355	0.698	0.291	1.140	1.266	0.976	15.32	23.85	53.22	1205	13.95	14.99	178.7	748.7
1,3	0.019	0.021	0.132	0.055	0.085	0.035	0.123	0.202	0.804	0.343	0.504	0.173	81.56	1379	5.059	2.907	3.255	1.361
3,3	0.013	0.015	0.017	0.018	0.015	0.012	0.091	0.167	0.065	0.065	0.056	0.043	18.88	437.1	0.263	0.235	0.232	0.182

Fréchet	$m=n^{1/4}$						$m=n^{1/2}$						$m=n^{3/4}$					
γ	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd
5	4449	4841	13700	1.368	5955	7608	11340	20670	13710	13690	29.60	108.4	13990	16200	74840	722000	13700	0.000
2	0.303	0.741	9.454	1.692	8.314	12.75	6.025	38.12	9.718	3.716	235.7	2446	61.86	412.0	114.2	316.4	74590	1840000
1	0.041	0.054	0.329	0.374	0.829	0.561	0.381	1.065	1.505	1.232	9.181	21.77	123.3	3281	20.28	27.15	246.2	2078
1/2	0.019	0.022	0.082	0.043	0.076	0.042	0.139	0.208	0.588	0.281	0.519	0.184	20.72	343.2	4.717	3.338	3.991	1.861
1/4	0.015	0.018	0.040	0.027	0.035	0.021	0.102	0.158	0.197	0.125	0.176	0.096	47.66	664.3	1.114	0.692	0.972	0.458

Weibull	$m=n^{1/4}$						$m=n^{1/2}$						$m=n^{3/4}$					
κ	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd
1/2	0.044	0.038	0.159	0.067	0.299	0.085	0.108	0.169	0.978	0.438	1.828	0.702	54.80	1520	5.554	2.943	10.18	7.603
1	0.017	0.020	0.085	0.045	0.044	0.024	0.092	0.148	0.333	0.178	0.167	0.087	39.12	621.0	1.285	0.712	0.669	0.343
3	0.011	0.013	0.012	0.013	0.010	0.009	0.090	0.171	0.035	0.032	0.031	0.027	18.64	242.2	0.117	0.099	0.112	0.101
10	0.011	0.012	0.012	0.011	0.009	0.008	0.088	0.127	0.034	0.030	0.027	0.023	8.291	127.5	0.126	0.091	0.098	0.073

inv. Burr	$m=n^{1/4}$						$m=n^{1/2}$						$m=n^{3/4}$					
c,ℓ	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd
3, 2	0.015	0.012	0.017	0.025	0.031	0.031	0.098	0.097	0.056	0.087	0.015	0.019	0.043	0.054	0.182	0.158	0.188	0.201
1, 2	0.015	0.011	0.019	0.022	0.032	0.023	0.144	0.100	0.078	0.092	0.015	0.023	0.039	0.077	0.258	0.204	0.121	0.126
1/2, 2	0.012	0.010	0.062	0.084	0.226	0.060	0.875	0.009	0.299	0.174	0.035	0.022	0.218	0.106	0.741	0.115	0.681	0.095
3, 1	0.175	0.152	0.533	0.229	0.572	0.160	0.684	0.021	0.672	0.084	0.363	0.225	0.682	0.024	0.689	0.005	0.674	0.000
1, 1	0.289	0.155	0.731	0.199	0.851	0.098	0.905	0.004	0.905	0.047	0.502	0.256	0.905	0.004	0.906	0.001	0.898	0.000
1/2, 1	1.524	0.184	2.400	0.192	3.027	0.022	3.260	0.000	3.043	0.014	1.921	0.471	3.260	0.000	3.260	0.000	3.340	0.000
3, 1/3	4.675	0.129	5.270	0.000	1.890	0.000	1.300	0.000	1.890	0.000	2.718	0.004	1.300	0.000	1.300	0.000	1.040	0.000
1, 1/3	56.10	0.203	56.60	0.000	30.70	0.000	26.50	0.000	30.70	0.000	37.40	0.000	26.50	0.000	26.50	0.000	25.40	0.000
1/2, 1/3	24500	0.000	24500	0.000	42700	0.000	54800	0.000	42700	0.000	33300	0.000	54800	0.000	54800	0.000	59600	0.000

The forecast periods were $m = n^{1/4}$, $m = n^{1/2}$, and $m = n^{3/4}$. All kernel functions were Epanechnikov for the bounded class and Gaussian for the other classes.

Table 4: Scaled MISE values and sd values of PE and NE1 where $n = 2^{12}$.

Pareto	$m = n^{1/4}$						$m = n^{1/2}$						$m = n^{3/4}$					
ℓ	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd
1/2	0.107	0.451	10.05	0.083	10.34	15.39	0.433	0.804	34.50	51.04	5735	173000	64.82	1061	17500	44480	10.83	0.000
1	0.004	0.004	1.682	0.043	3.820	3.328	0.042	0.070	5.073	23.29	208.8	842.2	8.390	141.0	49.05	206.5	11880	187200
3	0.002	0.002	0.034	0.009	0.043	0.041	0.017	0.020	0.515	0.179	0.653	0.621	0.964	5.703	8.932	6.583	9.987	12.90
10	0.002	0.002	0.034	0.007	0.019	0.005	0.013	0.016	0.311	0.073	0.183	0.048	0.414	1.135	3.073	1.143	1.782	0.582

T	$m = n^{1/4}$						$m = n^{1/2}$						$m = n^{3/4}$					
ℓ	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd
1/2	0.153	0.314	9.952	0.020	9.941	12.32	0.328	0.633	10.74	0.000	791.8	24500	44.45	758.3	10.83	0.000	10.83	0.000
1	0.015	0.012	1.474	0.159	2.010	1.050	0.038	0.048	15.07	60.21	124.9	425.5	3.296	20.33	224.5	1409	2846	23870
3	0.003	0.002	0.015	0.009	0.005	0.005	0.015	0.017	0.181	0.100	0.064	0.073	3.085	71.44	2.735	2.013	0.945	0.927
10	0.002	0.002	0.003	0.002	0.003	0.002	0.013	0.014	0.023	0.017	0.019	0.014	3.835	70.94	0.207	0.186	0.156	0.127

Burr	$m = n^{1/4}$						$m = n^{1/2}$						$m = n^{3/4}$					
c, ℓ	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd
1/2,1/2	457.0	250.5	845.4	0.001	790.4	24.09	549.6	257.4	1037	0.000	1037	0.000	1183	3083	161600	281100	1070	0.000
1,1/2	0.053	0.239	10.06	0.058	10.81	12.13	0.467	0.751	10.74	0.000	10.74	0.000	21.49	73.41	287.8	2015	10.83	0.000
3,1/2	0.002	0.003	0.126	0.027	0.480	0.426	0.027	0.027	0.532	0.262	12.34	15.00	0.657	0.744	19.65	45.41	361.4	1623
1/2,1	0.094	0.362	11.71	0.060	11.11	0.180	0.292	0.437	11.55	2.249	10.95	0.000	56.17	327.3	5921	32310	10.86	0.000
1,1	0.004	0.005	1.702	0.033	3.707	2.544	0.033	0.044	1.912	0.216	316.2	1121	28.81	242.3	33.34	66.91	22.28	204.4
3,1	0.002	0.002	0.008	0.004	0.009	0.007	0.016	0.016	0.124	0.043	0.131	0.119	1.263	8.059	1.842	0.618	1.983	0.715
1/2,3	0.009	0.007	0.460	0.576	4.466	3.297	0.022	0.024	2.125	15.44	360.0	1408	0.948	1.459	21.50	30.83	3822	34410
1,3	0.002	0.003	0.033	0.009	0.042	0.036	0.017	0.018	0.500	0.137	0.630	0.420	0.633	1.195	8.755	7.714	9.593	8.013
3,3	0.002	0.001	0.003	0.002	0.003	0.002	0.012	0.012	0.027	0.021	0.024	0.016	0.459	0.749	0.253	0.170	0.227	0.121

Fréchet	$m = n^{1/4}$						$m = n^{1/2}$						$m = n^{3/4}$					
γ	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd
5	11140	3586	13700	0.000	12880	631.1	10390	3020	13700	0.000	13700	0.000	31820	236400	698200	1287000	13700	0.000
2	0.108	0.511	10.80	0.075	12.45	66.10	0.389	0.636	407.6	447.7	10.84	0.060	71.17	1018	50850	96120	10.84	0.000
1	0.004	0.004	1.654	0.088	4.228	3.256	0.042	0.057	5.166	29.94	290.8	1724	4.465	32.78	138.6	1138	16710	307700
1/2	0.002	0.002	0.027	0.010	0.168	0.221	0.018	0.019	0.528	0.227	3.440	6.561	3.172	34.75	15.59	22.17	57.71	240.0
1/4	0.002	0.002	0.022	0.013	0.009	0.004	0.015	0.017	0.253	0.144	0.117	0.035	0.676	4.942	3.613	2.658	1.590	0.540

Weibull	$m = n^{1/4}$						$m = n^{1/2}$						$m = n^{3/4}$					
κ	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd
1/2	0.007	0.004	0.039	0.010	0.303	0.131	0.015	0.016	0.526	0.148	4.235	3.119	0.756	6.083	6.578	3.672	47.28	97.62
1	0.002	0.002	0.031	0.007	0.015	0.005	0.012	0.013	0.242	0.060	0.115	0.035	0.433	1.330	1.932	0.586	0.924	0.326
3	0.002	0.001	0.002	0.002	0.002	0.002	0.011	0.012	0.014	0.010	0.014	0.011	0.490	3.103	0.092	0.073	0.090	0.081
10	0.002	0.002	0.002	0.002	0.002	0.002	0.011	0.013	0.019	0.014	0.018	0.012	0.465	2.997	0.126	0.108	0.134	0.101

inv. Burr	$m = n^{1/4}$						$m = n^{1/2}$						$m = n^{3/4}$					
c, ℓ	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd	PE	sd	CV	sd	PI	sd
3, 2	0.006	0.003	0.012	0.019	0.012	0.019	0.011	0.011	0.035	0.050	0.035	0.050	0.298	0.370	0.165	0.186	0.165	0.186
1, 2	0.002	0.001	0.012	0.016	0.012	0.016	0.012	0.011	0.028	0.035	0.028	0.035	0.383	0.449	0.095	0.091	0.095	0.091
1/2, 2	0.015	0.061	0.013	0.008	0.013	0.008	0.017	0.022	0.221	0.048	0.221	0.048	0.379	0.441	0.876	0.009	0.876	0.009
3, 1	0.076	0.109	0.364	0.220	0.364	0.220	0.009	0.010	0.684	0.016	0.684	0.016	0.498	1.580	0.674	0.000	0.674	0.000
1, 1	0.226	0.246	0.493	0.272	0.493	0.272	0.023	0.035	0.904	0.006	0.904	0.006	0.467	0.845	0.898	0.000	0.898	0.000
1/2, 1	2.056	0.706	2.849	0.002	1.873	0.498	1.173	0.992	3.260	0.000	3.260	0.000	1.597	0.843	3.340	0.000	3.340	0.000
3, 1/3	2.457	0.161	2.670	0.000	2.719	0.003	0.141	0.258	1.300	0.000	1.300	0.000	0.423	1.253	1.040	0.000	1.040	0.000
1, 1/3	37.26	0.079	37.40	0.000	37.40	0.000	23.44	4.776	26.50	0.000	26.50	0.000	22.32	3.069	25.40	0.000	25.40	0.000
1/2, 1/3	33250	107.8	33300	0.000	33300	0.000	54800	0.000	54800	0.000	54800	0.000	59610	35.55	59600	0.000	59600	0.000

Both of the bandwidths h_1 and h_2 were estimated by the cross-validation approach or the plug-in approach, respectively denoted as 'CV' and 'PI' in the tables. CV means the unbiased cross-validation estimator (e.g., Silverman, 1986) and Bowman *et al.* (1998)'s cross-validation estimator. Sheather and Jones (1991)'s method and Altman and Léger (1995)'s method were utilized as the plug-in approaches. The sample sizes were $(n=)\,2^8$ in Table 3 and $(n=)\,2^{12}$ in Table 4.

Comparing CV with PI, we can see that CV performs significantly better for $1 \leq \gamma \leq 2$ while PI is much better for $\gamma \geq 4$. In the other cases, the bandwidth estimators are numerically comparable. Comparing NE1 with PE, we find that the MISE values of PE were smaller in most of the cases with $m = n^{1/4}$ and that PE worked numerically better than what we expected on the whole. For $\gamma \fallingdotseq 0$, those of NE1 were smaller as m got large. However, NE1 was not necessary better than PE in finite sample case, which is slightly different from the results of cumulative distribution estimation (see Moriyama, 2021). For the light-tailed distributions with $\gamma < 0$, the MISE values of NE1 were comparable with those of PE, although so-called boundary bias in NE1 for bounded distributions with $\gamma \geq -1$ was observed. Due to this boundary bias, the results for the distributions with large extreme indices (e.g., $\gamma \geq -1/3$) were different from the SMD estimation results reported by Moriyama (2021). Comparing the cases $(n=)\, 2^{12}$ with $(n=)\, 2^8$, we can see that PE substantially reduces the MISE values, which are smaller than those of NE1.

6. SMD ESTIMATION WITH DATA TRANSFORMATION

Ever since the Rosenblatt-Parzen kernel density estimator was invented, nonparametric density estimation has attracted the attention of both researchers and practitioners. Many researchers have tried to improve the accuracy of nonparametric density estimators. Most of them are on the choice of the bandwidth that significantly affects on the MSE. As in case of SMD estimation the kernel function does not change the convergence rates (as long as the sufficiently high-order moment exists); however, it is known that the choice possibly has effect on the numerical accuracy in finite sample case especially when the tail of the underlying distribution is heavy (e.g. Hall, 1987; Maiboroda and Markovich, 2004; Markovich, 2007). Hall (1987) proved that the KL divergence of the kernel density estimator and the performance of the likelihood cross-validation (for the bandwidth selection) depend on both the tails of the kernel and the underlying distribution.

Markovich (2018) proposed the Weibull kernel to fit the tail part of the underlying distribution. Wang *et al.* (2020) reported cross validation approaches in heavy tailed density estimation do not work well and proposed modified kernel functions. A reasonable kernel has a data-driven tail, that is the extreme value index of the kernel is fitted to that of the underlying distribution. The piecing-together approach (Markovich, 2007; MacDonald *et al.*, 2011), which fits the generalized Pareto distribution to the tail of the distribution, is related to the tail selection of the kernel.

Following the reviewer's comments, additional numerical experiments on the effect of kernel selection was conducted in this study. It was demonstrated that for heavy tailed distributions NE1 with equally heavy tailed kernel generally outperforms that of the Gaussian kernel used in Section 5. We omit the details; however, the experimental results imply that the numerical accuracy does not depend only on the level of heaviness (i.e. the extreme index) of the kernel. The experiments also clarified that the kernel being same as the underlying distribution is not necessarily best choice. More detailed studies on the choice of the kernel in SMD estimation may be important, but we postpone it to future work. This study aims at improving the nonparametric estimators without specifying the heaviness of the underlying distribution F.

Section 5 demonstrates that both of the theoretical and numerical accuracies of the nonparametric estimators of SMD heavily depends on the extreme value index $\gamma := \gamma_X$ of the underlying distribution (we don't add the subscript but note that γ is the extreme value index of F throughout the paper). When $\gamma \fallingdotseq 0$, NE functions well. However, the scaled MISE of NE explodes for large $|\gamma|$. Following these results, we consider the transformed nonparametric density estimation (see e.g. Wand *et al.*, 1991; Marron and Ruppert, 1994; Charpentier and Flachaire, 2015; Béranger *et al.*, 2019). Maiboroda and Markovich (2004), Buch-Larsen *et al.* (2005), Bolancé *et al.* (2015), Béranger *et al.* (2019) and Tokdar *et al.* (2022) intend to apply data transformation in order to change the distribution's tail to the same as that of the kernel.

By applying a data transformation $X \mapsto \varphi(X)$ to SMD estimation, the plug-in type nonparametric estimator with data transformation (TNE1):

$$\widehat{f}^{\varphi}_{(m)}(x; h_1, h_2) := m\widehat{f}^{\varphi}(x; h_1)(\widehat{F}^{\varphi})^{m-1}(x; h_2),$$

where $\widehat{f}^{\varphi}$ and $\widehat{F}^{\varphi}$ are the kernel estimators given by

$$\widehat{f}^{\varphi}(x; h_1) := \frac{\varphi'(x)}{nh_1}\sum_{i=1}^{n} w\left(\frac{\varphi(x) - \varphi(X_i)}{h_1}\right),$$
$$\widehat{F}^{\varphi}(x; h_2) := \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{\varphi(x) - \varphi(X_i)}{h_2}\right).$$

The BM-based kernel-type of nonparametric estimator with data transformation (TNE2) is

$$\bar{f}^{\varphi}_{(m)}(x; h) := \frac{\varphi'(x)}{\bar{n}h}\sum_{i=1}^{\bar{n}} w\left(\frac{\varphi(x) - \varphi(Y_i)}{h}\right).$$

This study considers the log-transformation $\varphi = \ln$, which is considered as a special version of the Box-Cox transformation. Then, the log-transformation $X \mapsto \ln(X)$ changes positive extreme value indices to zero as follows.

Corollary 6.1 (Wadsworth *et al.* (2010))**.** *If F belongs to the Hall class, the extreme value index of the distribution of $\ln(X)$ is zero. If F belongs to the Weibull class or the bounded class, the extreme value index is same as γ.*

This is seen from the Hall class of distributions becoming the Weibull class of distributions as follows:

$$\mathbb{P}[\ln X > x] \sim A\exp(-\alpha x) + AB\exp(-\alpha - \beta x),$$

which corresponds to the Weibull class with $C = \alpha$ and $\kappa = 1$.

As seen in Section 5, the nonparametric estimators NE1 and NE2 are numerically stable and work well for the Weibull class even when m is relatively large. Theoretically, the MSE is given by the following theorem.

Theorem 6.1. *Given Assumptions 3.1–3.3, then*

$$\mathbb{E}[(f_{(m)}(x) - \widehat{f}^{\ln}_{(m)}(x; h_1, h_2))^2] \sim b^2_{n,3} + v_{n,3},$$

where

$$b_{n,3} := \frac{h_1^2}{2}\exp(-M_n)M_n^{1+3\gamma}m^{-3\gamma}x^2\psi_n'\int z^2 w(z)\mathrm{d}z + \frac{h_2^2}{2}\exp(-M_n)M_n m\xi_n' x^2 f(x)\int z^2 w(z)\mathrm{d}z,$$

$$v_{n,3} := \frac{m^2}{nh_1}\exp(-2M_n)x^{-1}f(x)\int w^2(z)\mathrm{d}z + \frac{m^3}{n}\exp(-2M_n)M_n f^2(x),$$

$$\psi_n' := \begin{cases} \alpha\{1-3(\alpha+1)+(\alpha+1)(\alpha+2)\}A^{-3\gamma} & \text{for (i)} \\ \kappa^3 C^3 x^{3\kappa-3} & \text{for (ii)} \\ -\mu\{1-3(\mu+1)+(\mu+1)(\mu+2)\}D^{-3\gamma} & \text{for (iii)}, \end{cases}$$

$$\xi_n' := \begin{cases} \alpha^2 x^{-2} & \text{for (i)} \\ \kappa^2 C^2 x^{2\kappa-2} & \text{for (ii)} \\ \mu^2(x^*-x)^{-2} & \text{for (iii)}. \end{cases}$$

Furthermore, $v_{n,3}^{-1/2}\{f_{(m)}(x) - \widehat{f}_{(m)}^{\ln}(x;h_1,h_2) + b_{n,3}\}$ *converges in distribution to the standard normal distribution, if the MSE converges to zero.*

Similarly as Corollary 3.1, asymptotically optimal bandwidths of TNE1 are those of $\widehat{f}^{\ln}(x;h_1)$ and $\widehat{F}^{\ln}(x;h_2)$.

Corollary 6.2. *An asymptotically optimal value of* (h_1,h_2) *is given by*

$$x^{-1}\left(M_n^{-2-6\gamma}m^{2+6\gamma}\psi_n'^{-2}f(x)n^{-1}\left(\int z^2 w(z)\mathrm{d}z\right)^{-2}\int w^2(z)\mathrm{d}z\right)^{1/5},$$

$$x^{-1}\left(2\xi_n'^{-2}\omega_n n^{-1}\left(\int z^2 w(z)\mathrm{d}z\right)^{-2}\int zW(z)w(z)\mathrm{d}z\right)^{1/3},$$

which are asymptotically identical to those of $\widehat{f}^{\ln}(x;h_1)$ *and* $\widehat{F}^{\ln}(x;h_2)$ *if both of the optimal values converge 0. Under the assumption of Theorem 6.1,* $f_{(m)}(x) - \widehat{f}_{(m)}^{\ln}(x;h_1,h_2)$ *with the optimal bandwidths is asymptotically non-degenerate normal with the following asymptotic mean:*

$$\nu_1\exp(-M_n)M_n^{(1+3\gamma)/5}m^{(4-3\gamma)/5}\psi_n'^{1/5}f^{2/5}(x)n^{-2/5} + \nu_2\exp(-M_n)M_n m\xi_n'^{-1/3}f(x)\omega_n^{2/3}n^{-2/3},$$

whose order is same as that of NE1 $\widehat{f}_{(m)}$.

Corollary 6.3. *The asymptotically optimal bandwidths for* $M_n = O(1)$ *are of the order*

$$\left(\frac{m}{n}\right)^{1/5}\times\begin{cases}1 & \textit{for (i), (iii)} \\ \{C^{-1}\ln m\}^{-1} & \textit{for (ii)},\end{cases}$$

and

$$\left(\frac{m}{n}\right)^{1/3}\times\begin{cases}1 & \textit{for (i), (iii)} \\ \{C^{-1}\ln m\}^{-1} & \textit{for (ii)}.\end{cases}$$

The convergence rate of TNE1 $\widehat{f}^{\ln}_{(m)}$ is same as that of the non-transformed $\widehat{f}_{(m)}$. However, the order of the optimal bandwidth is different. The condition $h_j \to 0$ for $j = 1, 2$ is equivalent to $m = o(n)$ when $M_n = O(1)$, which is less restrictive for $\gamma > 0$. The proof of Theorem 6.1 is provided in the Supplementary file. Corollaries 6.2 and 6.3 are directly obtained from Theorem 6.1.

Theorem 6.2. *Given Assumptions 3.2, 3.3 and 4.1, if* $x^{\kappa-1}h \to 0$,

$$\mathbb{E}[(f_{(m)}(x) - \bar{f}^{\ln}_{(m)}(x;h))^2] \sim x^4 b^2_{n,2} + x^{-1} v_{n,2}.$$

Furthermore, $x^{1/2} v^{-1/2}_{n,2}\{f_{(m)}(x) - \bar{f}^{\ln}_{(m)}(x;h) + x^2 b_{n,2}\}$ *converges in distribution to the standard normal distribution, if the MSE converges to zero.*

Corollary 6.4. *The asymptotically optimal value of* h *is given by*

$$x^{-1}\left(n^{-1}\{f(x)\}^{-1}\phi_n^{-2}\left(\int z^2 w(z)\mathrm{d}z\right)^{-2}\int w^2(z)\mathrm{d}z\right)^{1/5}$$

if both of the optimal values converge 0. Under the assumption of Theorem 4.1, $f_{(m)}(x) - \bar{f}^{\ln}_{(m)}(x;h)$ *with the optimal bandwidths is asymptotically non-degenerate normal with the following asymptotic mean:*

$$\nu_1 n^{-2/5} m\{f(x)\}^{3/5}\phi_n^{1/5},$$

which is exactly same as that of NE2 $\bar{f}_{(m)}(x;h)$.

Corollary 6.5. *The asymptotically optimal bandwidth for* $M_n = O(1)$ *is of the order*

$$n^{-1/5}m^{-3/5} \times \begin{cases} m^{-4\gamma/5} & \text{for (i), (iii)} \\ (\ln m)^{-1/5} & \text{for (ii).} \end{cases}$$

Corollary 6.5 shows that the optimal convergence rate of TNE2 $\bar{f}^{\ln}_{(m)}$ is same as $\bar{f}_{(m)}$. The asymptotically optimal bandwidth is different from that of NE2. $h \to 0$ in Assumption 4.1 is less restrictive for $\gamma > 0$ same as TNE1. Theorem 6.2 is proved in the Supplementary file. Corollaries 6.4 and 6.5 are the direct consequences.

Corollary 6.1 states the extreme value index of $\ln(X)$ do not need to be estimated as long as γ (the index of F) is non-negative. Motivated by the facts, applying the back transformation we can consider the following transformed type of PE (TPE)

$$\widetilde{g}^{\ln}_{\widehat{\gamma}_k}(\ln x) := \frac{1}{\widehat{a}^{\ln}_k x}\exp\left(-\frac{\ln x - \widehat{b}^{\ln}_k}{\widehat{a}^{\ln}_k}\right)\exp\left(-\exp\left(-\frac{\ln x - \widehat{b}^{\ln}_k}{\widehat{a}^{\ln}_k}\right)\right)$$

for distributions with $\gamma \geq 0$, where $\widehat{a}^{\ln}_k$ and $\widehat{b}^{\ln}_k$ are the estimators of the normalizing constants of $\max\{\ln(X_i)\}^n_{i=1}$. Specifically, $a^{\ln}_k = b_k^{-1}a_k$ and $b^{\ln}_k = \ln b_k$ (Wadsworth *et al.*, 2010). However, the distributions with $\gamma = 0$ never satisfy all the assumptions of PE at the same time as described in Section 5.

For distributions with $\gamma < 0$ pointwise estimation by PE at a point $x \sim (x^* - z_n^{-1/\mu})$ changes that of by the following type of estimator

$$g^{\ln}_{\widehat{\gamma}_k}(\ln x) := \frac{1}{\widehat{a}^{\ln}_k x} g_{\widehat{\gamma}^{\ln}_k}\left(\frac{\ln x - \widehat{b}^{\ln}_k}{\widehat{a}^{\ln}_k}\right)$$

at the point $\ln x \sim (\ln x^* - z_n^{-1/\mu})$ by the transformation, which means the asymptotics of do not change since the extreme value index is same as γ. $\widehat{\gamma}_k^{\ln}$ is the estimator based on $\{\ln(X_i)\}_{i=1}^n$ of the extreme value index of $\{\ln(X)\}$, that is same as γ in this case.

The above means TPE is only applicable for the bounded class under the assumption of negativity of γ, but the log-transformation does not improve theoretical accuracy. In short, there are no theoretical reasons to employ TPE instead of PE in any case.

7. NUMERICAL STUDY ON TRANSFORMED SMD ESTIMATORS

This section surveys the numerical performances of the estimators TPE and TNE1. Suppose $M_n \equiv \delta > 0$. Then, Corollary 6.5 states that the asymptotic convergence rates of

Table 5: Scaled MISE values and sd values of TPE and TNE1 where $n = 2^8$.

Pareto	$m = n^{1/4}$						$m = n^{1/2}$						$m = n^{3/4}$					
ℓ	TPE	sd	CV	sd	PI	sd	TPE	sd	CV	sd	PI	sd	TPE	sd	CV	sd	PI	sd
1/2	4.013	0.714	**0.668**	0.456	**0.383**	0.310	**7.332**	2.885	**2.754**	2.182	**1.630**	1.568	**8.739**	3.653	**12.95**	15.47	**7.500**	9.460
1	0.070	0.034	**0.160**	0.093	**0.085**	0.054	**0.097**	0.135	**0.611**	0.407	**0.334**	0.252	**0.705**	0.649	**2.562**	1.963	**1.356**	1.045
3	0.465	0.013	**0.092**	0.050	**0.047**	0.027	0.431	0.038	**0.344**	0.184	**0.181**	0.097	**0.579**	0.167	**1.378**	0.825	**0.721**	0.428
10	0.643	0.006	**0.085**	0.043	**0.045**	0.024	0.620	0.015	**0.317**	0.170	**0.169**	0.093	**0.658**	0.049	**1.280**	0.699	**0.695**	0.394

T	$m = n^{1/4}$						$m = n^{1/2}$						$m = n^{3/4}$					
ℓ	TPE	sd	CV	sd	PI	sd	TPE	sd	CV	sd	PI	sd	TPE	sd	CV	sd	PI	sd
1/2	30.44	3.044	11.82	2.253	11.75	2.203	33.62	8.386	11.91	4.714	**11.65**	4.558	**22.87**	17.51	**16.17**	12.67	**15.65**	12.23
1	1.978	0.145	2.009	0.303	2.006	0.300	1.559	0.265	2.113	0.651	**2.056**	0.620	**1.027**	0.658	**2.687**	1.660	**2.576**	1.553
3	0.324	0.027	0.848	0.104	0.844	0.098	0.130	0.021	0.916	0.209	0.880	0.189	**0.065**	0.042	1.263	0.599	1.151	0.469
10	0.182	0.018	0.688	0.080	0.685	0.073	**0.047**	0.008	0.674	0.123	0.640	0.107	**0.017**	0.011	0.895	0.354	0.777	0.245

Burr	$m = n^{1/4}$						$m = n^{1/2}$						$m = n^{3/4}$					
c, ℓ	TPE	sd	CV	sd	PI	sd	TPE	sd	CV	sd	PI	sd	TPE	sd	CV	sd	PI	sd
1/2,1/2	4974	1196	**17.78**	19.19	**15.91**	17.39	6531	3155	**82.01**	115.8	**76.20**	113.3	5276	4927	**426.2**	857.8	**397.2**	804.2
1,1/2	6.648	1.423	**0.215**	0.211	**0.194**	0.180	7.942	2.976	**0.826**	0.939	**0.761**	0.847	**8.633**	3.691	**3.608**	5.146	**3.362**	4.835
3,1/2	0.185	0.021	**0.029**	0.025	**0.026**	0.022	0.214	0.080	**0.104**	0.088	**0.096**	0.081	**0.554**	0.362	**0.439**	0.437	**0.408**	0.397
1/2,1	24.43	6.227	**0.307**	0.327	**0.268**	0.269	11.41	4.900	**0.726**	0.793	**0.664**	0.723	**9.402**	4.217	**3.177**	4.600	**2.938**	4.275
1,1	0.067	0.038	**0.040**	0.039	**0.033**	0.032	**0.079**	0.113	**0.125**	0.141	**0.108**	0.126	**0.680**	0.646	**0.518**	0.640	**0.458**	0.560
3,1	0.326	0.014	**0.017**	0.017	**0.015**	0.013	0.400	0.043	**0.057**	0.052	**0.050**	0.044	**0.576**	0.171	**0.230**	0.239	**0.203**	0.207
1/2,3	0.717	0.262	**0.039**	0.038	**0.033**	0.031	**0.075**	0.088	**0.075**	0.076	**0.065**	0.060	**0.498**	0.438	**0.264**	0.263	**0.218**	0.201
1,3	0.072	0.014	**0.016**	0.017	**0.014**	0.013	0.230	0.051	**0.045**	0.042	**0.039**	0.037	**0.503**	0.205	**0.184**	0.183	**0.146**	0.133
3,3	0.407	0.010	**0.013**	0.012	**0.011**	0.009	0.503	0.023	**0.040**	0.033	**0.034**	0.029	**0.612**	0.068	**0.157**	0.124	**0.123**	0.097

Fréchet	$m = n^{1/4}$						$m = n^{1/2}$						$m = n^{3/4}$					
γ	TPE	sd	CV	sd	PI	sd	TPE	sd	CV	sd	PI	sd	TPE	sd	CV	sd	PI	sd
5	253600	23340	16110	3595	16000	3407	341200	101000	18350	9540	18030	9273	246000	188000	25310	27120	24300	25270
2	45.47	2.126	13.65	2.458	13.58	2.415	42.02	7.146	13.20	4.971	12.92	4.817	**26.01**	18.43	**17.09**	13.50	**16.53**	12.75
1	2.730	0.113	2.136	0.315	2.131	0.308	1.845	0.255	2.178	0.635	2.125	0.608	**1.098**	0.739	**2.713**	1.542	**2.605**	1.461
1/2	0.700	0.030	1.057	0.136	1.055	0.131	0.343	0.051	1.126	0.277	1.091	0.255	**0.175**	0.119	**1.523**	0.731	**1.413**	0.611
1/4	0.361	0.016	0.800	0.096	0.799	0.090	0.137	0.022	0.842	0.175	0.812	0.154	**0.062**	0.040	1.149	0.505	1.042	0.398

Weibull	$m = n^{1/4}$						$m = n^{1/2}$						$m = n^{3/4}$					
κ	TPE	sd	CV	sd	PI	sd	TPE	sd	CV	sd	PI	sd	TPE	sd	CV	sd	PI	sd
1/2	0.106	0.064	**0.020**	0.019	**0.016**	0.014	0.131	0.059	**0.039**	0.034	**0.035**	0.032	**0.443**	0.194	**0.123**	0.088	**0.101**	0.074
1	0.151	0.014	**0.013**	0.012	**0.011**	0.009	0.339	0.034	**0.031**	0.027	**0.027**	0.025	**0.532**	0.103	**0.116**	0.084	**0.090**	0.066
3	0.454	0.007	0.012	0.011	0.010	0.009	0.545	0.013	**0.032**	0.028	**0.028**	0.025	**0.619**	0.036	0.118	0.084	**0.092**	0.066
10	0.606	0.003	0.012	0.011	0.010	0.008	0.633	0.004	**0.031**	0.025	**0.026**	0.023	**0.657**	0.011	**0.121**	0.081	**0.091**	0.063

inv. Burr	$m = n^{1/4}$						$m = n^{1/2}$						$m = n^{3/4}$					
c, ℓ	TPE	sd	CV	sd	PI	sd	TPE	sd	CV	sd	PI	sd	TPE	sd	CV	sd	PI	sd
3, 2	0.719	0.000	0.019	0.022	**0.016**	0.017	0.678	0.000	**0.049**	0.051	0.060	0.061	0.672	0.000	**0.181**	0.130	0.244	0.119
1, 2	0.722	0.000	0.019	0.022	**0.016**	0.015	0.691	0.000	**0.053**	0.054	0.065	0.059	0.685	0.000	**0.248**	0.139	0.322	0.103
1/2, 2	0.836	0.000	**0.035**	0.028	**0.034**	0.023	**0.868**	0.000	**0.194**	0.100	0.281	0.079	0.888	0.000	**0.684**	0.134	0.801	0.044
3, 1	0.870	0.000	**0.324**	0.234	**0.180**	0.126	0.727	0.000	**0.541**	0.193	0.619	0.115	0.690	0.000	**0.661**	0.077	0.688	0.006
1, 1	1.071	0.000	**0.566**	0.263	**0.313**	0.143	**0.934**	0.000	**0.834**	0.125	0.872	0.071	0.906	0.000	**0.903**	0.012	0.906	0.000
1/2, 1	2.657	0.000	**2.244**	0.261	**1.642**	0.162	**3.048**	0.000	**3.030**	0.028	3.031	0.019	3.262	0.000	3.262	0.000	**3.262**	0.000
3, 1/3	**4.041**	0.000	**3.867**	0.050	3.232	0.194	1.770	0.000	**1.769**	0.001	**1.767**	0.001	**1.263**	0.000	**1.263**	0.000	1.263	0.000
1, 1/3	**51.58**	0.000	**51.58**	0.004	50.99	0.189	30.13	0.000	**30.13**	0.000	**30.13**	0.000	**26.42**	0.000	**26.42**	0.000	26.42	0.000
1/2, 1/3	**23800**	0.000	**23800**	0.000	**23800**	0.129	**4009**	0.000	**4009**	0.000	**4009**	0.000	**5.562**	0.000	**5.562**	0.000	**5.562**	0.000

the transformed nonparametric estimator TNE1 correspond to those of TNE1 (shown in Table 1 and Table 2). On the other hand, TPE for the distributions with $\gamma \geq 0$ cannot be consistent.

The scaled MISE values were simulated in Pareto, T, Burr, Fréchet, Weibull, and inverse Burr cases. To apply the log-transformation the distributions are truncated to the range $(0, \infty)$ or $(0, \ln x^*)$. Other technical details are same as Section 5. In Table 5 the MISE value in bold means being smaller than that of the MISE value in Table 3. That means the log-transformation improves the estimator.

Table 5 shows the log-transformation improves the estimators in many cases on the whole. The exceptions are mainly TPE and the cases where m is relatively small. Though the MSE values increase in the cases, the increments are relatively not large. For heavy-tailed data the log-transformation especially improves the accuracy of the estimators. To sum up, the log-transformation makes TNE1 numerically stable and much more accurate.

8. ILLUSTRATIVE EXAMPLES

In this section, we give two illustrative examples of applying SMD estimation. The solid and dashed lines in Figures 2 and 3 respectively show the probability density functions estimated by the nonparametric approach with cross-validated bandwidth (NE1) and by the fitting to the GEV (PE). Figure 2 shows the Potomac River peak stream flow (cfs) data for water years 1901–2000 (Oct–Sep) at Point of Rocks, Maryland. Figure 3 shows the latest $n = 2100$ losses in Danish fire insurance over one million Danish kroner (DKK) collected from 1980 to 1990. The datasets are available in the extRemes and evir packages in the R software environment, respectively. The range of the horizontal axis in Figure 2 is 0–480000, where 480000 is the maximal value observed in 1936. The minimum value was 27800 observed in 1969. The range in Figure 3 is 0–(263.25×1.5), where 263.25 is the maximum and the minimum is 1.

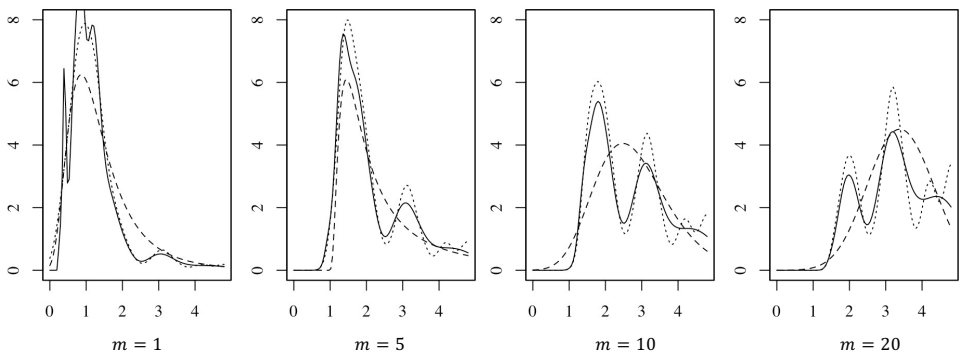

Figure 2: The estimated probability ($\times 10^6$) density functions of Annual peak flow ($\times 10^{-5}$) data of the Potomac River at Point of Rocks, MD, USA, 1901–2000, by the log-transformed approach (solid line), the naive kernel approach (dotted line) and by the fitting to the GEV (dashed line).

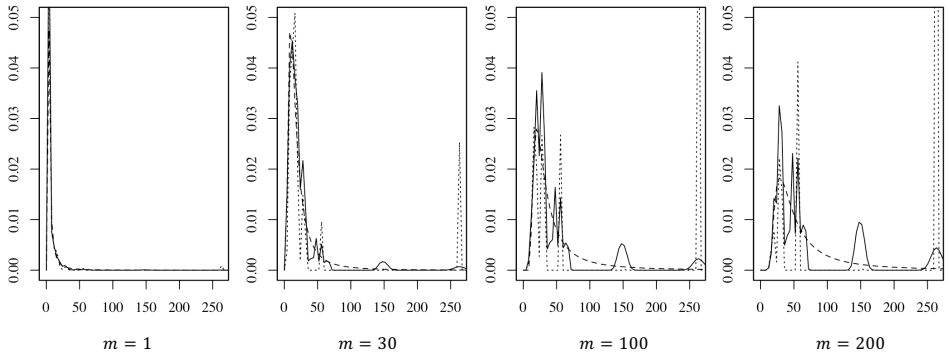

Figure 3: The estimated probability density functions of 2100 Danish data consisting of losses over one million Danish kroner (DKK) by the log-transformed approach (solid line), the naive kernel approach (dotted line) and by the fitting to the GEV (dashed line).

Figures 2 and 3 suggest the existence of multimodality. Moriyama (2021) reported that, when comparing the estimated probabilities by the nonparametric approach and by the fitting to the GEV, the estimated risk is not one-sided. This may stems partly from the multimodality of the estimated density function. Note that the nonparametrically estimated density is sensitive to bandwidth. Many studies support the heavy-tailness in the Danish datasets, and so the statistical evidence of multimodality is not strong. Since the mixture of distributions that belong to the maximum domain of attractions converges to a unimodal GEV (see Bolancé *et al.*, 2015), there are two possible reasons for the multimodality. One is the slowness of the convergence of the limitation to the GEV. The other is that the underlying distribution does not belong to the maximum domain of attractions. Future work should examine how to construct a method for testing the properties of underlying distributions including multimodality.

ACKNOWLEDGMENTS

The author appreciates the referee's valuable comments that helped me to improve this manuscript. This work was supported by the Japan Society for the Promotion of Science (JSPS) *JSPS KAKENHI Grant Numbers JP19K20223, JP23K16850.*

REFERENCES

Altman, N. and Léger, C. (1995). Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, 46:195–214.

Banfi, F., Cazzaniga, G., and De Michele, C. (2022). Nonparametric extrapolation of extreme quantiles: a comparison study. *Stochastic Environmental Research and Risk Assessment*, 36(6):1579–1596.

Beirlant, J. and Devroye, L. (1999). On the impossibility of estimating densities in the extreme tail. *Statistics and Probability Letters*, 43:57–64.

Beirlant, J., Goegebeur, Y., Teugels, J., and Segers, J. (2004). *Statistics of Extremes*. John Wiley and Sons, Ltd.

Béranger, B., Duong, T., Perkins-Kirkpatrick, S.E., and Sisson, S.A. (2019). Tail density estimation for exploratory data analysis using kernel methods. *Journal of Nonparametric Statistics*, 31(1):144–174.

Bolancé, C., Bahraoui, Z., and Alemany, R. (2015). Estimating extreme value cumulative distribution functions using bias-corrected Kernel approaches. *XREAP2015-01*.

Bowman, A.W., Hall, P., and Prvan, T. (1998). Cross-validation for the smoothing of distribution functions. *Biometrika*, 85:799–808.

Buch-Larsen, T., Nielsen, J.P., Guillén, M., and Bolancé, C. (2005). Kernel density estimation for heavy-tailed distributions using the Champernowne transformation. *Statistics*, 39(6):503–516.

Charpentier, A. and Flachaire, E. (2015). Log-transform kernel density estimation of income distribution. *L'Actualité Économique*, 91:141–159.

Cohen, J.P. (1988). Fitting extreme value distributions to maxima. *Sankhyā: The Indian Journal of Statistics, Series A*, 50(1):74–97.

Dombry, C. and Ferreira, A. (2019). Maximum likelihood estimators based on the block maxima method. *Bernoulli*, 25(3):1690–1723.

Hall, P. and Welsh, A.H. (1984). Best attainable rates of convergence for estimates of parameters of regular variation. *Annals of Statistics*, 12(3):1079–1084.

Hall, P. (1987). On Kullback-Leibler loss and density estimation. *Annals of Statistics*, 15(3):1491–1519.

MacDonald, A., Scarrott, C.J., Lee, D., Darlow, B., Reale, M., and Russell, G. (2011). A flexible extreme value mixture model. *Computational Statistics and Data Analysis*, 55(6):2137–2157.

Maiboroda, R.E. and Markovich, N.M. (2004). Estimation of heavy-tailed probability density function with application to Web data. *Computational Statistics*, 19(4):569–592.

Markovich, N. (2007). *Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice*. Wiley Series in Probability and Statistics. Wiley, Chichester.

Markovich, L. (2018). *Light- and heavy-tailed density estimation by gamma-Weibull kernel*. In "Nonparametric Statistics". ISNPS 2016. Springer Proceedings in Mathematics & Statistics, vol. 250, pp. 145–158.

Marron, J.S. and Ruppert, D. (1994). Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society. Series B*, 56:653–671.

Moriyama, T. (2021). Parametric and nonparametric probability distribution estimators of sample maximum. *arXiv*: 2111.03765.

Segers, J. (2005). Generalized Pickands estimators for the extreme value index. *Journal of Statistical Planning and Inference*, 128(2):381–396.

Sheather, S.J. and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B (Methodological)*, 53(3):683–690.

Silverman, B.W. (1986). *Density Estimation*. Chapman and Hall, London.

Stupfler, G. (2016). Estimating the conditional extreme-value index under random right-censoring. *Journal of Multivariate Analysis*, 144:1–24.

Tokdar, S.T., Jiang, S., and Cunningham, E.L. (2022). Heavy-tailed density estimation. *Journal of the American Statistical Association*, 1-13.

Wadsworth, J.L., Tawn, J.A., and Jonathan, P. (2010). Accounting for choice of measurement scale in extreme value modeling. *The Annals of Applied Statistics*, 4(3):1558–1578.

Wang, S., Li, A., Wen, K., and Wu, X. (2020). Robust kernels for kernel density estimation. *Economics Letters*, 191:109138.

Wand, P., Marron, J., and Ruppert, D. (1991). Transformations in density estimation. *Journal of the American Statistical Association*, 86:343–361.

REVSTAT – Statistical Journal
Volume 24, Number 2, April 2026, 213–229
https://doi.org/10.57805/revstat.v24i2.464

Non-parametric Inference of Window-Observation Recurrent Event Data with Multiple Causes of Failure

Authors: P. G. Sankaran ✉
– Department of Statistics, Cochin University of Science and Technology, Kerala, India
sankaran.pg@gmail.com

S. Hari
– Department of Statistics, Cochin University of Science and Technology, Kerala, India
harivlnr@yahoo.com

Received: May 2022 Revised: March 2024 Accepted: April 2024

Abstract:

- Recurrent event data with multiple causes of failure are common in survival studies. There are situations where the recurrence history of some systems is available in disconnected observation windows with gaps in between. Such data are called window-observation recurrence data, and no events can be observed during these gaps. This article discusses the non-parametric estimation of cause specific mean function for window-observation recurrence data with multiple failure modes. A test statistic is proposed to compare the effect of different causes. Simulation studies are performed to evaluate the performance of the proposed methods. An automobile warranty database is analyzed to illustrate the suggested techniques.

Keywords:

- *counting process; window observation; recurrent events; mean function; multiple causes.*

AMS Subject Classification:

- 62G05, 62G10, 62N02, 62N03.

✉ Corresponding author

1. INTRODUCTION

In survival studies, the event of interest can occur multiple times on the same subject. Such outcomes are termed as recurrent events. Recurrent event data are common in fields such as public health, medicine, reliability, social sciences, and insurance. Examples in biomedicine include the recurrence of tumors in cancer patients, repeated hospitalizations of patients with a specific disease, and recurrence of caries in oral health studies. In engineering and reliability, the recurrence of a crack in concrete structures, failure of an electronic system, and software bugs are examples. Various methods for analyzing recurrent event data are available in many literature; examples include Lawless and Nadeau (1995), Nelson (1995), and Nelson (2003). A comprehensive review of recurrent event data analysis is given in Cook and Lawless (2007).

There are situations in which a study unit may experience an event due to multiple causes. For example, an electronic system may experience a breakdown due to the failure of any of the components in that system. To model and characterize the marginal event processes that generate the particular type of events, the cause specific sub distribution functions and the cause specific hazard rate functions are generally used. The cause specific hazard rate function gives the instantaneous probability of failure due to a specific cause, while the cause specific sub-distribution function measures the probability of failure due to a specific cause before a given time. Sankaran and Anisha (2011) and Sankaran and Anisha (2012) have studied semi-parametric inferences of recurrent event data with multiple causes. A non-parametric method for estimating the system lifetime distribution with recurrent competing risks data was studied by Taylor and Pena (2014). Dauxois and Sencey (2009) discussed the non-parametric tests in the context of the recurrent event competing risks based on the mean number of events, and Sisuma and Sankaran (2022) proposed a non-parametric test for comparing cause specific cumulative incidence functions.

In many lifetime studies, the recurrence of events may not be recorded throughout the study period. Even though the event process is continuous in time, the recurrence status of some systems is available in separate observation windows with gaps in between. For example, in a clinical trial study, patients may enter and leave the study multiple times, resulting in gaps in their follow-up tenure. No events are recorded during these gaps in the follow-up window, and the exact recurrence times are recorded when the patient is under monitoring. Such data is referred to as window-observation recurrence data. Zuo *et al.* (2008) provides an example of window observation data utilizing the U.S. Army's Field Exercise Data Collection (FEDC) program, which is overseen by the Army Materiel Systems Analysis Activity (AMSAA). The FEDC program tracks the replacement rates of parts for military vehicles during field training exercises, where distinct vehicles within the fleet engage in various exercises with intervals in between. Some vehicles may not participate in every FEDC exercise, and during the intervals between exercises, they accumulate mileage from non-exercise activities and other unobserved field exercises. Consequently, recurrences may occur in these intervals but remain unrecorded. The observation windows and the gaps can vary for different observational units, and also, the length of each observation window and the gaps between the windows need not be the same.

In typical recurrence data with left and/or right censoring, each system is associated with a single observation window. However, for window-observation recurrence data, multiple disjoint observation windows can exist for each system. These windows can have random

lengths, and the length of the gaps between them may also vary randomly. Additionally, the beginning or ending time points across windows for different observational units can differ. Window-observation data differ from interval-grouped recurrent event data, where the number of events in time intervals is recorded but the exact times of recurrences are not specified, and there are no gaps between intervals. Zuo *et al.* (2008) analyzed such data in terms of mean cumulative function. They proposed non-parametric estimation methods for handling window-observation recurrence data. The asymptotic properties of the non-parametric estimator for the mean function are studied in Zuo *et al.* (2012). Kim and Jhun (2008) considered the case when information on the observation gap is incomplete, that is, the starting time of intermittent dropout is known, but the terminating time is not available, and they modeled it in terms of an interval-censored mechanism. Regression analysis of recurrent event data with repeated observation gaps with unknown termination times of observation gaps was studied by Kim (2014). Zhu *et al.* (2014) focused on window-censored data, where only events within a specific interval are documented. They developed the likelihood function for a model where the distributions of inter-recurrence intervals within a single path are not required to be identical and may be linked to covariate information.

In this study, we focus on the situation in which a unit under observation can experience more than one type of failure, and also, there are intermittent dropouts that result in observation gaps during which no recurrent events are observed. The mean cumulative function corresponding to different causes is used to model this problem, and a test is proposed to compare the effect of various causes of failures on the recurrence process.

The article is organized as follows. In Section 2, we introduce cause specific mean function for the window-observation recurrence data. We propose a non-parametric estimator for the cause specific mean function, and the asymptotic properties of the estimator are discussed. The test procedure for the mean functions due to different causes is discussed in Section 3. Section 4 discusses the simulation studies performed for assessing the finite sample behaviors of the proposed estimator. We put the empirical power to check the efficiency of the test. The methods are then illustrated using a real life data set in Section 5. Finally, Section 6 presents the major conclusion of the study.

2. NON-PARAMETRIC INFERENCE PROCEDURES

Consider a study on n individuals exposed to the recurrent events due to $\{1, 2, ..., K\}$ different causes. Let $N_k(t)$ denote the cumulative number of events due to cause k up to time t, $k = 1, 2, ..., K$. Then the cause specific mean function is defined as

$$\mu_k(t) = E[N_k(t)]; \quad k = 1, 2, ..., K. \tag{2.1}$$

Then $\mu_k(t)$ in (2.1) is interpreted as the expected number of cumulative events due to cause k up to time t, $k = 1, 2, ..., K$. Assumptions required for the non-parametric estimation of mean cumulative function were described in Nelson (2003) and the same for the window-observation scenario were covered in Zuo *et al.* (2008). These assumptions are briefed below:

(**i**) Target population should be clearly specified and the sample units are a simple random sample from this population.

(**ii**) The population cause specific mean function exists up to the greatest censoring and is zero at time zero.

(**iii**) The stochastic process that generates the recurrences and the observation windows are independent.

(**iv**) The size of the risk set must be positive up to the maximum study time.

Let m_k denote the number of unique event times due to cause k. Thus $t_1, t_2, ...,$ and t_{m_k} are the unique event times for cause k, $k = 1, 2, ..., K$. Then a non-parametric estimator of the cause specific mean function is given by

$$\hat{\mu}_k(t_j) = \sum_{l=1}^{j} \Big[\frac{\sum_{i=1}^{n} \delta_i(t_l) d_{ki}(t_l)}{\sum_{i=1}^{n} \delta_i(t_l)} \Big]; \quad j = 1, 2, ..., m_k, \ k = 1, 2, ..., K, \tag{2.2}$$

where

$$\delta_i(t_l) = \begin{cases} 1, & \text{if individual } i \text{ is under observation in a time window at } t_l, \\ 0, & \text{otherwise,} \end{cases}$$

and $d_{ki}(t_l)$ is the number of events due to cause k at time t_l for the individual i. Denote $d_{k\cdot}(t_l) = \sum_{i=1}^{n} \delta_i(t_l) d_{ki}(t_l)$, total number of events at t_l due to cause k and $\delta_\cdot(t_l) = \sum_{i=1}^{n} \delta_i(t_l)$, total number of units at risk at time t_l. Then (2.2) can be written as

$$\begin{aligned} \hat{\mu}_k(t_j) &= \sum_{l=1}^{j} \frac{d_{k\cdot}(t_l)}{\delta_\cdot(t_l)} \\ &= \sum_{l=1}^{j} \bar{d}_k(t_l); \quad j = 1, 2, ..., m_k, \ k = 1, 2, ..., K. \end{aligned} \tag{2.3}$$

The size of the risk set at time t_l is calculated by taking into account the gaps between observation windows and the censoring.

2.1. Asymptotic properties

Assume that the cause specific mean function $\mu_k(t)$ is differentiable. Then the recurrence rate at time t due to cause k is given by $\nu_k(t) = \frac{d\mu_k(t)}{dt}$, $k = 1, 2, ..., K$. Let there be w_i observation windows for the individual i, $i = 1, 2, ..., n$. If $t_{\max}$ denotes the largest observation time among all n units, then the number of units at risk, $\delta_\cdot(t) = \sum_{i=1}^{n} \delta_i(t)$, is a piece-wise constant over $(0, t_{\max}]$. To ensure a non zero risk set throughout the study period, we assume that the overlap of the set of observation windows of all the n units leaves no complete gaps in the study timeline. If $(t^*_{1L}, t^*_{1U}], (t^*_{2L}, t^*_{2U}], ..., (t^*_{bL}, t^*_{bU}]$ be the b intervals in which $\delta_\cdot(t)$ is constant, then $t^*_{1L} = 0$, $t^*_{bU} = t_{\max}$ and $t^*_{jL} = t^*_{(j-1)U}$ for $j = 2, 3, ..., b$. Let $\tau < \infty$ be the ending time for all the windows with $t_{\max} \le \tau$ for all n and $t_{\max} \to \tau$ as $n \to \infty$. Then for this window-observation recurrence data, Zuo *et al.* (2012) showed that $\delta_\cdot(t)$ follows a Binomial$(n, p(t))$ distribution, where $p(t)$ is the probability that a unit is observed at time t, with $p(t) > 0$ for all $t \in (0, t_{\max}]$.

To establish the uniform consistency and the asymptotic normality of the estimator, Theorem IV.1.1 and Theorem IV.1.2 of Andersen *et al.* (1993) are employed.

Theorem 2.1. *Let $t \in (0, \tau]$ and assume that, as $n \to \infty$,*

$$\int_0^t \frac{I[\delta.(u) > 0]}{\delta.(u)} \nu_k(u) du \xrightarrow{\mathbf{P}} 0 \tag{2.4}$$

and

$$\int_0^t \{1 - I[\delta.(u) > 0]\} \nu_k(u) du \xrightarrow{\mathbf{P}} 0. \tag{2.5}$$

Then, as $n \to \infty$,

$$\sup_{t \in [0,\tau]} |\hat{\mu}_k(t) - \mu_k(t)| \xrightarrow{\mathbf{P}} 0, \tag{2.6}$$

where $\hat{\mu}_k(t)$ is the non-parametric estimator of $\mu_k(t)$, $k = 1, 2, ..., K$.

Theorem 2.1 can be used to show that the estimator of cause specific mean function given in (2.2) is uniformly consistent on compact intervals. It can be verified that the conditions (2.4) and (2.5) are satisfied by the estimator and hence the uniform consistency can be established. The proof for the same is similar to the one given in Zuo *et al.* (2012) and can be referred.

Theorem 2.2. *Let $t \in (0, \tau]$ and assume that there exist a sequence of positive constants $\{a_n\}$, increasing to infinity as $n \to \infty$, and non-negative function $g(\cdot)$ such that $\nu_k(u)/g(u)$ is integrable over $[0, t]$ for $k = 1, 2, ..., K$. Assume the following:*

(**A**) *For each $s \in [0, t]$ and $k = 1, 2, ..., K$,*

$$a_n^2 \int_0^s \frac{I[\delta.(u) > 0]}{\delta.(u)} \nu_k(u) du \xrightarrow{\mathbf{P}} \sigma_k^2(s) \quad as \quad n \to \infty, \tag{2.7}$$

where

$$\sigma_k^2(s) = \int_0^s \frac{\nu_k(u)}{g(u)} du.$$

(**B**) *For all $\epsilon > 0$ and $k = 1, 2, ..., K$,*

$$a_n^2 \int_0^t \frac{I[\delta.(u) > 0]}{\delta.(u)} \nu_k(u) I\left\{ a_n \frac{I[\delta.(u) > 0]}{\delta.(u)} > \epsilon \right\} du \xrightarrow{\mathbf{P}} 0 \quad as \quad n \to \infty. \tag{2.8}$$

(**C**) *For $k = 1, 2, ..., K$,*

$$a_n \int_0^t \{1 - I[\delta.(u) > 0]\} \nu_k(u) du \xrightarrow{\mathbf{P}} 0 \quad as \quad n \to \infty. \tag{2.9}$$

Then

$$a_n(\hat{\mu}_k(t) - \mu_k(t)) \xrightarrow{\mathbf{D}} U_k \quad as \quad n \to \infty$$

on $D[0, \tau]$, where $U_1, U_2, ..., U_K$ are independent Gaussian martingales with $U_k(0) = 0$ and $\text{cov}(U_k(s_1), U_k(s_2)) = \sigma_k^2(s_1 \wedge s_2)$. Here $D[0, \tau]$ is the Skorohod space on $[0, \tau]$, that is, the space of right-continuous functions with left-hand limits on $[0, t]$; $s_1 \wedge s_2$ is the smaller of s_1 and s_2.

The conditions (A), (B) and (C) of Theorem 2.2 can be verified using the methods given in Zuo *et al.* (2012) and hence the asymptotic normality of the estimator in (2.2) can be established.

2.2. Non-parametric estimation in the presence of size-zero risk set intervals

Sometimes there can be time intervals in which the number of subjects at risk is zero. Such intervals can be referred to as size-zero risk set intervals. This may lead to a downward bias in the non-parametric estimation method proposed. To overcome this, Zuo *et al.* (2008) proposed a hybrid estimator for mean cumulative function, which makes use of the information from the non-zero risk set interval to estimate the increase in the mean function over the size-zero risk set interval. A similar approach can be employed for the estimation of cause specific mean function and is briefed below.

Let $(t_{1L}, t_{1U}], (t_{2L}, t_{2U}], ..., (t_{rL}, t_{rU}]$ be the r size-zero risk set intervals with $t_{1L} \geq 0$, $t_{rU} < t_{\max}$ and $t_{(j-1)U} < t_{jL}$ for $j = 2, 3, ..., r$. Then $(t_{1U}, t_{2L}], (t_{2U}, t_{3L}], ..., (t_{rU}, t_{\max}]$ are the non-zero risk set interval. The estimated cause specific mean function of cause k increases only at time points that have events due to cause k and the increase at an event time t_l is given by $\bar{d}_k(t_l) = \frac{\sum_{i=1}^{n} \delta_i(t_l) d_{ki}(t_l)}{\sum_{i=1}^{n} \delta_i(t_l)}$ as defined in (2.3). Then the estimated increase in the cause-specific mean function over the non-zero risk set interval $(t_{jU}, t_{(j+1)L}]$ is

$$\bar{d}_{k\cdot}(t_{jU}, t_{(j+1)L}) = \sum_{l: t_{jU} < t_l \leq t(j+1)L} \bar{d}_k(t_l). \tag{2.10}$$

An estimate of the cause specific recurrence rate in this interval is $\frac{\bar{d}_{k\cdot}(t_{jU}, t_{(j+1)L})}{(t_{(j+1)L} - t_{jU})}$.

To calculate the cause specific rate of a size-zero risk set interval, a weighted mean of the recurrence rates of the two neighboring non-zero risk set intervals, weighted by the length of those intervals, is taken. The recurrence rate due to cause k for the j-th size-zero risk set interval $(t_{jL}, t_{jU}]$ is given by

$$\omega_{kj} = \frac{\bar{d}_{k\cdot}(t_{(j-1)U}, t_{jL}) + \bar{d}_{k\cdot}(t_{jU}, t_{(j+1)L})}{(t_{jL} - t_{(j-1)U}) + (t_{(j+1)L} - t_{jU})}. \tag{2.11}$$

If the data begin with a size-zero risk set interval, the recurrence rate is estimated as that of the first non-zero risk set interval.

Then for the j-th size-zero risk set interval, the estimated increase in the cause specific mean function is

$$d^*_{kj} = \omega_{kj} \times (t_{jU} - t_{jL}). \tag{2.12}$$

Therefore, a non-parametric hybrid estimator for the cause specific mean function is given by

$$\hat{\mu}^*_k(t) = \begin{cases} \sum_{l: t_l \leq t} \bar{d}_k(t_l) + \sum_{j: t_{jU} \leq t} d^*_{kj} \\ \qquad \text{if } t \text{ is in a non-zero risk set interval,} \\ \sum_{l: t_l \leq t} \bar{d}_k(t_l) + \sum_{j: t_{jU} \leq t} d^*_{kj} + (\omega_{kj} \times (t - t_{iL}) \\ \qquad \text{if } t \text{ is in a size-zero risk set interval } (t_{iL}, t_{iU}], \end{cases} \quad k = 1, 2, ..., K. \tag{2.13}$$

3. TEST STATISTIC

In the study of multiple modes of failure, it is always of greater interest to compare the effect of various causes on recurrence times. We can compare the cause specific mean functions to test whether or not the effect of all causes is identical on the recurrence process. Now consider the testing problem,

$$H_0 : \mu_k(t) = \mu_l(t) \text{ for all } t > 0 \text{ and } k \neq l = 1, 2, ..., K$$

against

$$H_1 : \mu_k(t) \neq \mu_l(t) \text{ at least for some } t > 0 \text{ and for any } k \neq l = 1, 2, ..., K. \tag{3.1}$$

If we denote the overall mean function as $\mu(t)$, the above hypothesis can be rewritten as

$$H_0 : \mu_k(t) = \frac{\mu(t)}{K} \text{ for all } t > 0 \text{ and } k = 1, 2, ..., K$$

against

$$H_1 : \mu_k(t) \neq \frac{\mu(t)}{K} \text{ at least for some } t > 0 \text{ and for any } k = 1, 2, ..., K. \tag{3.2}$$

To test H_0 against H_1, the cause specific mean function $\hat{\mu}_k(t)$ defined in (2.2) is taken as the estimator of $\mu_k(t)$. The estimator for the overall mean function $\mu(t)$ can be constructed by ignoring the cause of failure information as discussed in Zuo *et al.* (2008). To develop the test statistic, consider the function

$$\eta_k(t) = \int_0^t w(u)\Big[\hat{\mu}_k(u) - \frac{\hat{\mu}(u)}{K}\Big]du \text{ for all } k = 1, 2, ..., K, \tag{3.3}$$

where $w(\cdot)$ is a data dependent weight function. The function $\eta_k(\cdot)$ is similar to the one proposed by Pepe and Mori (1993) and Lin (1997) in the context of comparing two independent cumulative incidence functions. The weight function is used to compensate for the effect of censoring and is also employed to enhance the efficiency of the test statistic and ensure its asymptotic distribution-free nature, as suggested by Pepe and Mori (1993). The weight function should be chosen such that it maximizes the power of the test.

Now we propose a test statistic $Z(\tau)$ given by

$$Z(\tau) = \eta'(\tau)\hat{\Sigma}(\tau)^{-1}\eta(\tau), \tag{3.4}$$

where τ is the largest monitoring time in the study, $\eta(\tau) = (\eta_1(\tau), ..., \eta_K(\tau))'$ and $\hat{\Sigma}(\tau)^{-1}$ is the generalized inverse of $\hat{\Sigma}(\tau)$, with $\hat{\Sigma}(\tau)$ as a consistent estimator of the variance-covariance matrix $\Sigma(\tau)$ of $\eta(\tau)$. The expression for the estimator $\hat{\Sigma}(\tau)$ is complex and thus generally requires smoothing. In practice, $\hat{\Sigma}(\tau)$ can be calculated using bootstrap resampling technique.

To find the asymptotic distribution of $Z(\tau)$, consider

$$\begin{aligned}\eta_k(t) &= \int_0^t w(u)\Big[\hat{\mu}_k(u) - \frac{\hat{\mu}(u)}{K}\Big]du \\ &= \int_0^t w(u)\Big[\hat{\mu}_k(u) - \mu_k(u)\Big]du + \int_0^t w(u)\Big[\mu_k(u) - \frac{\mu(u)}{K}\Big]du \\ &\quad + \int_0^t \frac{w(u)}{K}\Big[\mu(u) - \hat{\mu}(u)\Big]du.\end{aligned} \tag{3.5}$$

Under H_0, $\mu_k(u) = \frac{\mu(u)}{K}$ for all t. Then

$$\eta_k(t) = \int_0^t w(u)\Big[\hat{\mu}_k(u) - \mu_k(u)\Big]du + \int_0^t \frac{w(u)}{K}\Big[\mu(u) - \hat{\mu}(u)\Big]du. \tag{3.6}$$

Since for fixed t, $\hat{\mu}_k(t)$ and $\hat{\mu}(t)$ are asymptotically normal, η_k is asymptotically normal with mean zero and $\eta(\tau) = (\eta_1(\tau), ..., \eta_K(\tau))'$ is asymptotically a k-variate normal with mean zero vector and variance-covariance matrix $\Sigma(\tau)$. Therefore, under H_0, the quadratic form $Z(\tau)$ given in (3.4) follows a chi-square distribution with $K-1$ degrees of freedom. The null hypothesis given in (3.2) is rejected when $Z(\tau) \geq \chi^2_{\alpha,K-1}$ where $\chi^2_{\alpha,K-1}$ is the ordinate value of chi-square distribution with $K-1$ degrees of freedom at α level.

4. SIMULATION STUDY

An extensive simulation study is conducted to evaluate the finite sample performance of the proposed estimator and the test statistic. The study is limited to two failure mode scenario, and recurrence times are generated using Weibull and Gompertz cause-specific hazard functions. The following algorithm outlines the data simulation process:

1. Two cases are considered for the generation of recurrence times:

(**a**) Using the Weibull cause-specific hazard functions

$$\lambda_k(t) = \theta_k \alpha_k t^{\alpha_k - 1}, k = 1, 2 \text{ with } \theta_k, \alpha_k > 0.$$

(**b**) Using the Gompertz cause-specific hazard functions

$$\lambda_k(t) = \rho_k e^{\beta_k t}, k = 1, 2 \text{ with } \rho_k, \beta_k > 0.$$

For simplicity, $\alpha_1 = \alpha_2 = \alpha$ and $\beta_1 = \beta_2 = \beta$ are assumed in the simulation study.

2. Recurrence times are simulated from the distribution with the overall hazard function $\lambda_1(t) + \lambda_2(t)$.

3. To assign failure modes associated with a simulated event time T, a binomial experiment is run. The experiment decides with probability $\frac{\lambda_1(T)}{\lambda_1(T)+\lambda_2(T)}$ for failure mode 1 and $\frac{\lambda_2(T)}{\lambda_1(T)+\lambda_2(T)}$ for failure mode 2.

4. The maximum follow-up time for each unit is fixed at time 2.

5. It is assumed that the follow-up of each unit begins with an observation window, and this initiation is randomly generated from a Uniform$(0.08, 0.4)$ distribution. The lengths of the subsequent observation windows are also randomly generated from a Uniform$(0.08, 0.4)$ distribution, while the lengths of the gaps between the windows follow a Uniform$(0.08, 0.24)$ distribution.

6. Recurrences falling in gaps for each unit are not considered, and only recurrences in an observational window for each unit are included in the analysis.

We simulate 500 data with sample sizes $n = 50$ and 100 each for different combinations of $(\theta_1, \theta_2, \alpha)$ and (ρ_1, ρ_2, β). The average of the cause specific mean functions estimated across different iterations are presented in Figure 1 and Figure 2 for different sample sizes along with the actual mean functions. The absolute bias and mean squared error (MSE) of the cause specific mean functions are shown in Table 1 and Table 2, which are small and decrease as the sample size increases.

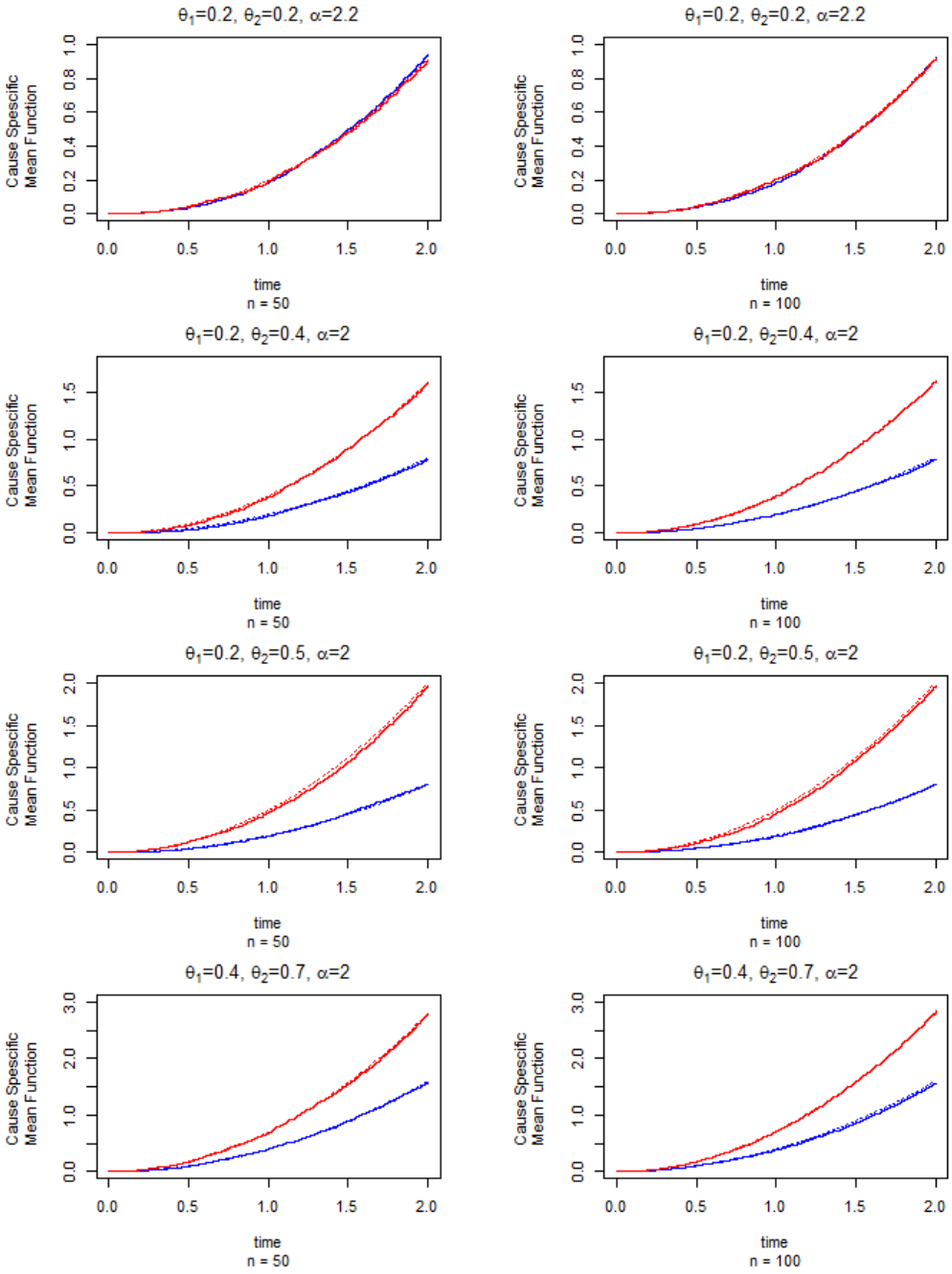

Figure 1: Estimates of cause specific mean functions of Weibull model for different values of $(\theta_1, \theta_2, \alpha)$. The blue color represents cause 1, and the red color represents cause 2. Solid lines are for the estimated values, and dotted lines are the true values.

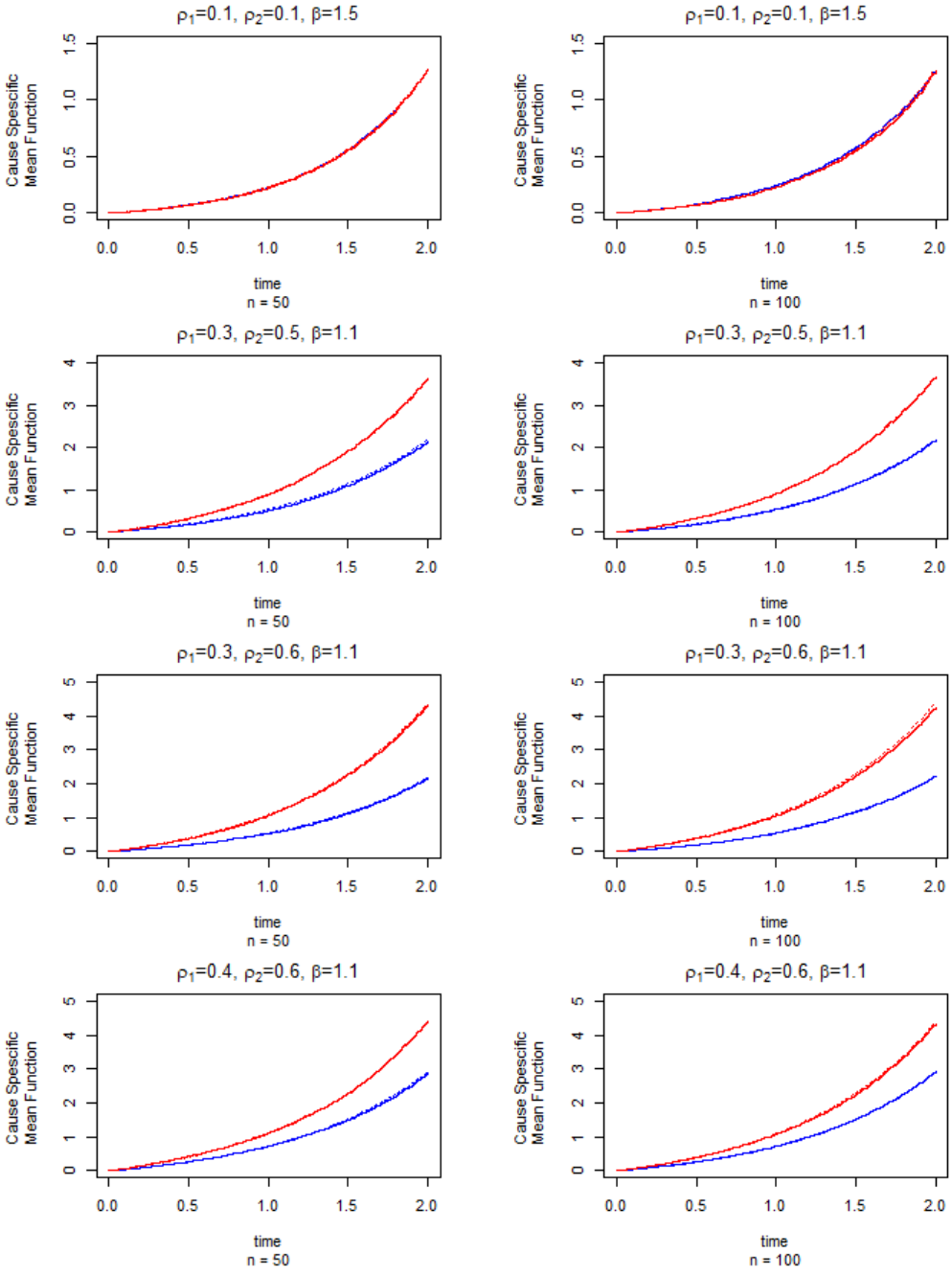

Figure 2: Estimates of cause specific mean functions of Gompertz model for different values of (ρ_1, ρ_2, β). The blue color represents cause 1, and the red color represents cause 2. Solid lines are for the estimated values, and dotted lines are the true values.

Table 1: Absolute bias and MSE of $\hat{\mu}_k; k = 1, 2$ for the Weibull model.

$(\theta_1,\theta_2,\alpha)$	Time	$n = 50$				$n = 100$			
		$\hat{\mu}_1(t)$		$\hat{\mu}_2(t)$		$\hat{\mu}_1(t)$		$\hat{\mu}_2(t)$	
		Absolute bias	MSE	Absolute bias	MSE	Absolute bias	MSE	Absolute bias	MSE
(0.2,0.2,2.2)	0.1	0.00232	0.00002	0.00303	0.00004	0.00277	0.00003	0.00285	0.00002
	0.4	0.02594	0.00096	0.02320	0.00079	0.01802	0.00050	0.01457	0.00034
	0.7	0.04400	0.00326	0.04249	0.00294	0.02755	0.00112	0.02993	0.00131
	1.0	0.05778	0.00462	0.07198	0.00720	0.03699	0.00230	0.04974	0.00410
	1.3	0.08815	0.01329	0.08670	0.01233	0.05409	0.00494	0.05948	0.00555
	1.6	0.12241	0.02142	0.12409	0.02347	0.06313	0.00592	0.06600	0.00817
	1.9	0.14646	0.03271	0.14965	0.03204	0.07500	0.00895	0.07438	0.00842
(0.2,0.4,2)	0.1	0.00337	0.00005	0.00488	0.00005	0.00292	0.00002	0.00493	0.00003
	0.4	0.02341	0.00077	0.04511	0.00292	0.01658	0.00045	0.02794	0.00121
	0.7	0.04669	0.00303	0.06383	0.00624	0.02952	0.00141	0.05062	0.00373
	1.0	0.07269	0.00782	0.10479	0.01570	0.04803	0.00375	0.06855	0.00737
	1.3	0.09093	0.01299	0.11231	0.01954	0.04789	0.00397	0.09759	0.01396
	1.6	0.10599	0.01801	0.14735	0.02900	0.06520	0.00718	0.09465	0.01402
	1.9	0.11490	0.01835	0.18701	0.05052	0.08615	0.01171	0.12482	0.02209
(0.2,0.5,2)	0.1	0.00264	0.00002	0.00722	0.00009	0.00294	0.00002	0.00561	0.00004
	0.4	0.02547	0.00104	0.04314	0.00289	0.01836	0.00054	0.02861	0.00135
	0.7	0.03862	0.00243	0.05972	0.00579	0.03343	0.00162	0.05475	0.00463
	1.0	0.06698	0.00658	0.08847	0.01269	0.04784	0.00342	0.07128	0.00770
	1.3	0.08988	0.01273	0.12212	0.02266	0.06528	0.00602	0.09917	0.01473
	1.6	0.10743	0.02123	0.16251	0.04193	0.07965	0.00878	0.11085	0.01989
	1.9	0.11565	0.01929	0.15434	0.04036	0.08725	0.01063	0.12895	0.02524
(0.4,0.7,2)	0.1	0.00618	0.00006	0.00961	0.00012	0.00553	0.00005	0.00706	0.00006
	0.4	0.03772	0.00216	0.05864	0.00491	0.03006	0.00133	0.04186	0.00271
	0.7	0.07467	0.00944	0.08599	0.01092	0.04562	0.00290	0.07009	0.00757
	1.0	0.10002	0.01520	0.11883	0.02211	0.07594	0.00737	0.09478	0.01445
	1.3	0.11585	0.02046	0.15592	0.03342	0.09810	0.01346	0.12758	0.02827
	1.6	0.15117	0.03388	0.19412	0.05414	0.12189	0.02208	0.15579	0.03901
	1.9	0.17583	0.04800	0.22552	0.07684	0.13617	0.02605	0.17251	0.04172

The empirical type I error and power of the proposed test are calculated at both 1% and 5% levels of significance. The parameter combination with $\theta_1 = \theta_2$ and $\rho_1 = \rho_2$ gives the type I error of the proposed test and all other choices of parameter combinations give the power of the test. Three different weight functions are employed, which are:

(**i**) $w(\cdot) = 1$,

(**ii**) $w(\cdot) = \sqrt{n}$,

(**iii**) $w(\cdot) = \hat{\mu}_k(\cdot)$.

Results are presented in Table 3 and Table 4, and it is clear that the empirical type I error approaches the significance level when $\theta_1 = \theta_2$ and $\rho_1 = \rho_2$. The rejection probabilities of the test at the 5% level under the Weibull model, with $\theta_1 = 0.2$ and θ_2 increasing from 0.2 to 0.5, are plotted in Figure 3. Similarly, under the Gompertz model, with $\rho_1 = 0.3$ and ρ_2 increasing from 0.3 to 0.6, the rejection probabilities are shown in Figure 4. Also, the test has very good power that increases as the difference between the parameters increases. The weight function $w(\cdot) = \hat{\mu}_k(\cdot)$ gives better power compared to others.

Table 2: Absolute bias and MSE of $\hat{\mu}_k; k = 1, 2$ for the Gompertz model.

(ρ_1,ρ_2,β)	Time	$n = 50$				$n = 100$			
		$\hat{\mu}_1(t)$		$\hat{\mu}_2(t)$		$\hat{\mu}_1(t)$		$\hat{\mu}_2(t)$	
		Absolute bias	MSE	Absolute bias	MSE	Absolute bias	MSE	Absolute bias	MSE
(0.1,0.1,1.5)	0.1	0.01201	0.00018	0.01207	0.00019	0.00886	0.00014	0.00721	0.00010
	0.4	0.03242	0.00163	0.03645	0.00201	0.02378	0.00094	0.02481	0.00089
	0.7	0.04957	0.00367	0.05051	0.00398	0.03410	0.00183	0.03873	0.00227
	1.0	0.07282	0.00804	0.06945	0.00747	0.05784	0.00472	0.04500	0.00330
	1.3	0.09463	0.01419	0.09130	0.01289	0.07351	0.00823	0.06289	0.00599
	1.6	0.11684	0.02130	0.11783	0.02205	0.08795	0.01231	0.08179	0.00992
	1.9	0.14055	0.03176	0.14181	0.03201	0.11550	0.01996	0.10227	0.01592
(0.3,0.5,1.1)	0.1	0.02066	0.00063	0.02681	0.00094	0.01293	0.00031	0.01927	0.00058
	0.4	0.05937	0.00474	0.07070	0.00825	0.03730	0.00206	0.04705	0.00385
	0.7	0.08685	0.01103	0.10038	0.01744	0.05888	0.00485	0.07130	0.00826
	1.0	0.10684	0.01600	0.15112	0.03776	0.08144	0.01007	0.10955	0.01745
	1.3	0.14177	0.03057	0.14024	0.03676	0.09694	0.01618	0.13714	0.02670
	1.6	0.15256	0.03867	0.17802	0.05433	0.11517	0.02219	0.15139	0.03627
	1.9	0.16660	0.04904	0.26433	0.09548	0.13815	0.02723	0.17773	0.04989
(0.3,0.6,1.1)	0.1	0.02073	0.00059	0.02829	0.00148	0.01433	0.00036	0.02333	0.00090
	0.4	0.05059	0.00426	0.07906	0.00921	0.03823	0.00244	0.05138	0.00429
	0.7	0.08411	0.01049	0.11910	0.02050	0.05953	0.00594	0.06841	0.00735
	1.0	0.09824	0.01711	0.14610	0.03745	0.07124	0.00850	0.10017	0.01430
	1.3	0.15494	0.03880	0.21100	0.06770	0.09728	0.01380	0.12914	0.02981
	1.6	0.17544	0.04808	0.25562	0.09666	0.11209	0.01869	0.17581	0.04214
	1.9	0.23005	0.07679	0.31608	0.14118	0.13288	0.02529	0.20030	0.06149
(0.4,0.6,1.1)	0.1	0.02242	0.00080	0.03360	0.00179	0.01661	0.00050	0.02032	0.00072
	0.4	0.06641	0.00712	0.06943	0.00705	0.04521	0.00326	0.05072	0.00427
	0.7	0.07941	0.00896	0.08502	0.01136	0.06455	0.00714	0.07980	0.00888
	1.0	0.10688	0.02057	0.10526	0.01834	0.07837	0.00997	0.09747	0.01348
	1.3	0.12311	0.02451	0.15443	0.03817	0.09829	0.01608	0.12627	0.02763
	1.6	0.17441	0.04299	0.25481	0.09228	0.11675	0.02344	0.13286	0.02868
	1.9	0.21571	0.07401	0.25492	0.10076	0.13504	0.03037	0.15555	0.04198

Table 3: Empirical type I error and power of the test in percentage at an asymptotic level of 1% and 5% for the Weibull model.

$(\theta_1,\theta_2,\alpha)$	$w(\cdot)$	1%		5%	
		$n = 50$	$n = 100$	$n = 50$	$n = 100$
(0.2,0.2,2.2)	1	1.8	1.2	5.6	5.2
	$\sqrt{n}$	1.8	1.4	5.8	5.4
	$\hat{\mu}_k(\cdot)$	1.6	1.2	5.6	4.8
(0.2,0.4,2)	1	73	90.8	88.6	96
	$\sqrt{n}$	67.6	92	88.4	95.2
	$\hat{\mu}_k(\cdot)$	74.4	94	91.4	100
(0.2,0.5,2)	1	92.4	100	98.2	100
	$\sqrt{n}$	91.2	100	98.8	100
	$\hat{\mu}_k(\cdot)$	93.6	100	100	100
(0.4,0.7,2)	1	82.4	98	95	100
	$\sqrt{n}$	80.8	96.4	94.4	98
	$\hat{\mu}_k(\cdot)$	82.8	98.4	97.2	100

Table 4: Empirical type I error and power of the test in percentage at an asymptotic level of 1% and 5% for the Gompertz model.

(ρ_1,ρ_2,β)	$w(\cdot)$	1%		5%	
		$n = 50$	$n = 100$	$n = 50$	$n = 100$
(0.1,0.1,1.5)	1	1.8	1.4	5.8	5.2
	$\sqrt{n}$	2	1.4	6.2	5.6
	$\hat{\mu}_k(\cdot)$	1.6	1.2	5.4	4.6
(0.3,0.5,1.1)	1	88.2	96.8	94.6	100
	$\sqrt{n}$	87.2	96.4	94.2	100
	$\hat{\mu}_k(\cdot)$	87.8	97	95.4	100
(0.3,0.6,1.1)	1	95.6	100	100	100
	$\sqrt{n}$	94.4	100	100	100
	$\hat{\mu}_k(\cdot)$	96.2	100	100	100
(0.4,0.6,1.1)	1	71.4	92.8	94.2	98.4
	$\sqrt{n}$	67.4	91.4	92.8	98.4
	$\hat{\mu}_k(\cdot)$	78.2	93.4	94.6	100

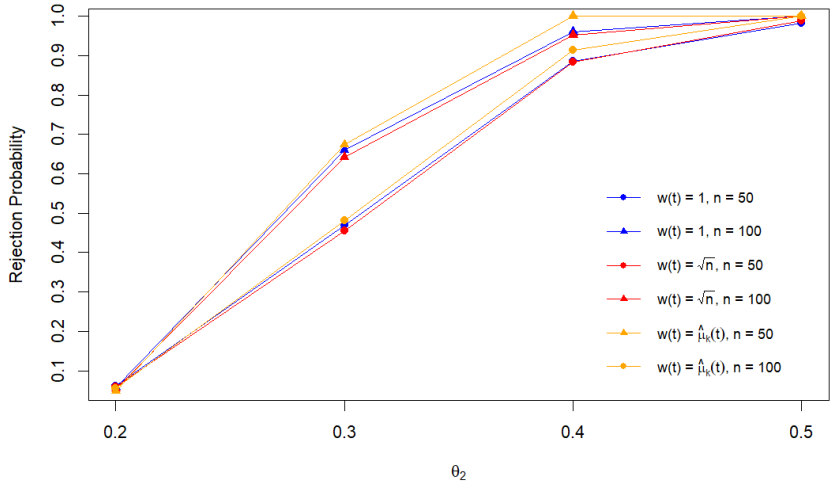

Figure 3: Rejection probabilities of the test at the 5% level under the Weibull model with $\theta_1 = 0.2$.

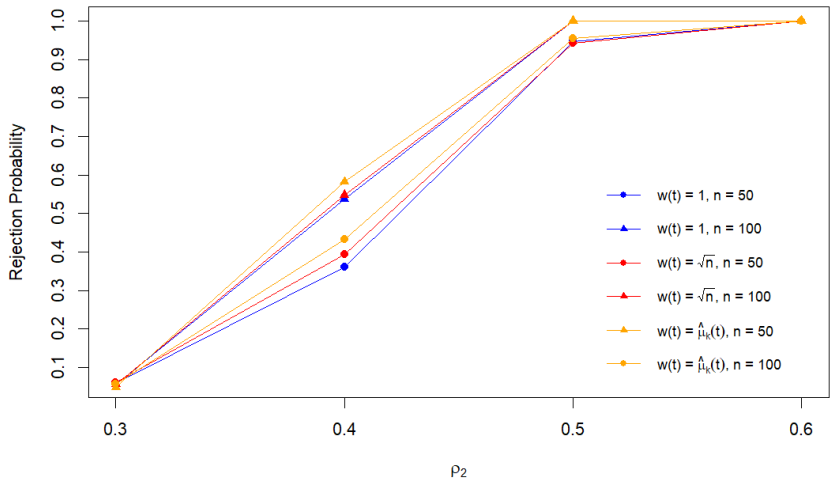

Figure 4: Rejection probabilities of the test at the 5% level under the Gompertz model with $\rho_1 = 0.3$.

5. DATA ANALYSIS

The proposed inference procedures are illustrated using an automobile warranty database given in Somboonsavatdee and Sen (2015). The data set consists of recurrent failure histories of 172 automobiles in which the outcome of interest is the repeated mileages at which the failures occur. The mileages are type I censored at 3000 miles. The failure modes are classified into three categories, FM1, FM2, and FM3, resulting in multiple modes of failure framework. We can observe from the data a total of 274 failures in which 76 failures are due to FM1, 87 are due to FM2, and 111 are due to FM3.

Since there are no intermittent observation gaps in the original data set, we generate gaps in the data records to create an automobile warranty database with simulated windows. We modify the original data set by allowing each vehicle to have a minimum of four windows of observation gaps in their study tenure. The length of the observation windows is randomly generated from a Uniform$(400, 600)$ distribution while the length of the gaps follows a Uniform$(50, 150)$ distribution. All the vehicles are assumed to begin with an observation window. In the transformed data, we observe a total of 245 failure recurrences, where 29 failures are missing due to gaps in the observation timeline. The new recurrences comprise 71 failures due to FM1, 78 due to FM2, and 96 due to FM3. Because of the presence of gaps in the follow-up, the number of vehicles at risk changes with time, and the size of the new risk set is presented in Figure 5.

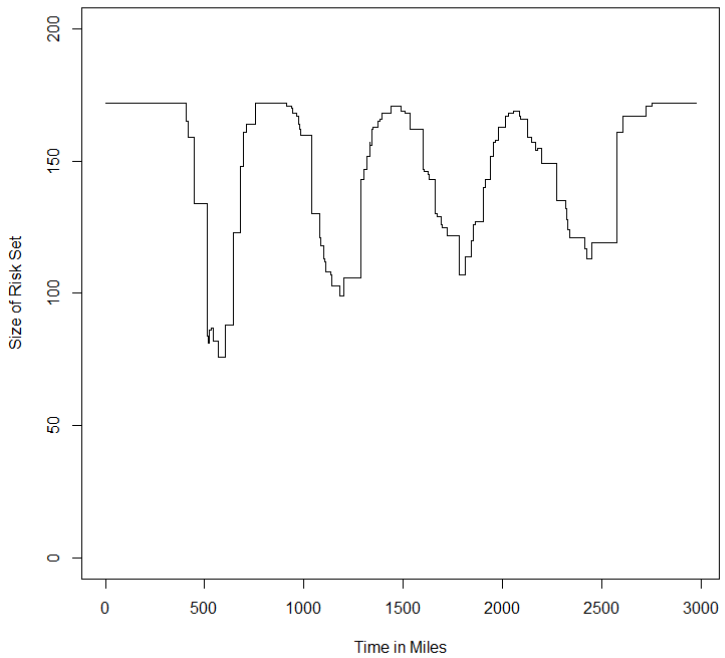

Figure 5: Risk set plot for the automobile data with observation gaps.

Our goal here is to compare the effect of three modes of failures FM1, FM2, and FM3 on the failure recurrences of automobiles. The mean functions due to three different modes

of failures are computed using the estimator proposed in (2.3) and the results are presented in Figure 6.

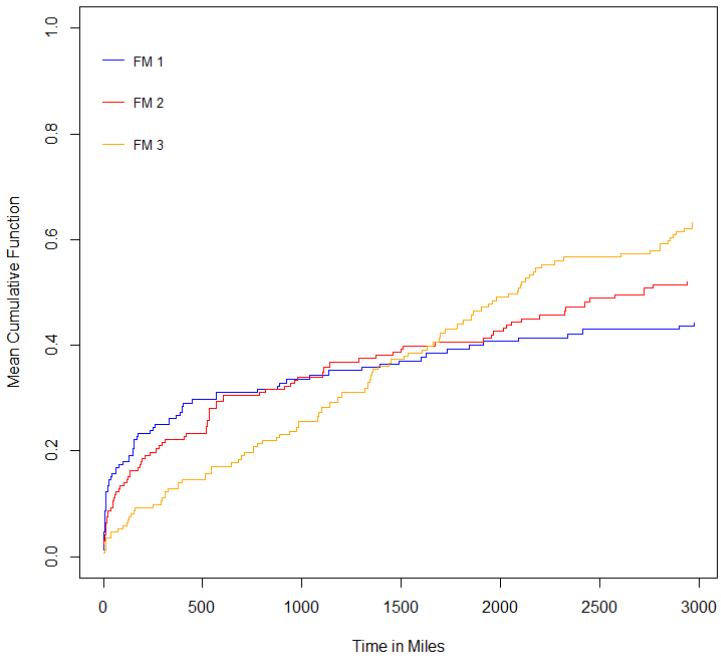

Figure 6: Estimates of cumulative cause specific mean functions of automobile data with observation gaps for different failure modes (FM1, FM2 and FM3).

From Figure 6, we observe that the mean functions for three different causes FM1, FM2, and FM3 are different. The mean cumulative functions of the causes FM1 and FM2 show an almost similar pattern and compete, but failure mode 3 exhibits a different pattern. We can see a higher failure rate at the beginning of the observation time, somewhat up to 500 miles, due to causes FM1 and FM2. The rate becomes steady beyond 1500 miles for these two causes. The earlier failures of vehicles, somewhat up to 1500 miles, are more likely to be due to failure modes 1 or 2, and failure mode 3 causes more events beyond 1500 miles than the other two.

We test statistically whether these three modes of failures namely FM1, FM2, and FM3 have the same effect on the recurrence process or not by using the method proposed in Section 3. With the help of bootstrap techniques, the variance-covariance matrix in (3.4) is estimated. Table 5 presents the test statistic values obtained and the corresponding p-values for weight functions $w(\cdot) = 1$, $w(\cdot) = \sqrt{n}$ and $w(\cdot) = \hat{\mu}_k(\cdot)$.

Table 5: Test statistic values and the corresponding p-values for different weight functions.

Weight function	Test statistic	p-value
$w(\cdot) = 1$	15.857	0.00036
$w(\cdot) = \sqrt{n}$	15.856	0.00036
$w(\cdot) = \hat{\mu}_k(\cdot)$	9.577	0.00832

From Table 5, we see that the p-value is small for each of the weight functions, and we conclude that the three cause specific mean functions are significantly different.

6. CONCLUSION

Recurrent event data with gaps in the follow-up windows are often found in survival and reliability studies. In this paper, we are extending the methods proposed by Zuo *et al.* (2008) for analyzing the window-observation recurrence data to the multiple causes of failure scenario. A non-parametric estimator for the cause specific mean function in the window-observation setup is studied and the asymptotic properties were discussed. A test statistic is proposed to test whether the mean functions due to different causes are identical or not. The results of the simulation study ensure that the proposed methods are efficient. A data analysis is also performed to illustrate the methods proposed in this paper.

In some applications, there will be additional information on the factors that may affect the recurrence process, such as medical history, demographic details, and vital signs of patients in clinical studies or model, make, and operating environment of systems in reliability studies. Allowing such covariate information to analyze multiple modes of failure in window-observation recurrent event data is an important research area, and works in these directions will be reported elsewhere.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to the unknown referee for the valuable comments and constructive suggestions, which have greatly contributed to the improvement of this research paper. The second author would like to thank the Council of Scientific & Industrial Research, Government of India for the financial support.

REFERENCES

Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Cook, R.J. and Lawless, J.F. (2007). *The Statistical Analysis of Recurrent Events*. Springer, New York.

Dauxois, J.Y. and Sencey, S. (2009). Non-parametric tests for recurrent events under competing risks. *Scandinavian Journal of Statistics*, 36(4):649–670.

Kim, Y.J. (2014). Regression analysis of recurrent events data with incomplete observation gaps. *Journal of Applied Statistics*, 41(7):1619–1626.

Kim, Y.J. and Jhun, M. (2008). Analysis of recurrent event data with incomplete observation gaps. *Statistics in Medicine*, 27(7):1075–1085.

Lawless, J.F. and Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics*, 37(2):158–168.

Lin, D.Y. (1997). Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in Medicine*, 16(8):901–910.

Nelson, W. (1995). Confidence limits for recurrence data?applied to cost or number of product repairs. *Technometrics*, 37(2):147–157.

Nelson, W. (2003). *Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications*. ASA-SIAM, Philadelphia.

Pepe, M.S. and Mori, M. (1993). Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in Medicine*, 12(8):737–751.

Sankaran, P.G. and Anisha, P. (2011). Shared frailty model for recurrent event data with multiple causes. *Journal of Applied Statistics*, 38(12):2859–2868.

Sankaran, P.G. and Anisha, P. (2012). Additive hazards models for gap time data with multiple causes. *Statistics & Probability Letters*, 82(7):1454–1462.

Sisuma, M.S. and Sankaran, P.G. (2022). Non-parametric test of recurrent cumulative incidence functions for competing risks models. *Metron*, 1–12.

Somboonsavatdee, A. and Sen, A. (2015). Parametric inference for multiple repairable systems under dependent competing risks. *Applied Stochastic Models in Business and Industry*, 31(5):706–720.

Taylor, L.L. and Pena, E.A. (2014). Nonparametric estimation with recurrent competing risks data. *Lifetime Data Analysis*, 20(4):514–537.

Zhu, Y., Yashchin, E., and Hosking, J.R.M. (2014). Parametric estimation for window censored recurrence data. *Technometrics*, 56(1):55–66.

Zuo, J., Meeker, W.Q., and Wu, H. (2008). Analysis of window-observation recurrence data. *Technometrics*, 50(2):128–143.

Zuo, J., Wu, H., and Meeker, W.Q. (2012). Asymptotic properties of mean cumulative function estimators from window-observation recurrence data. *Journal of Statistical Planning and Inference*, 142(11):2943–2952.

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 — *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 email-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, in .pdf format;
- 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, 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.

EDITORIAL BOARD 2025-2026

Editor-in-Chief

Manuel SCOTTO, University of Lisbon, Portugal

Co-Editor

Cláudia NUNES, University of Lisbon, Portugal

Associate Editors

Abdelhakim AKNOUCHE, Qassim University, Saudi Arabia
Andrés ALONSO, Carlos III University of Madrid, Spain
Barry ARNOLD, University of California, United States
Wagner BARRETO-SOUZA, University College Dublin, Ireland
Francisco BLASQUES, VU Amsterdam, The Netherlands
Paula BRITO, University of Porto, Portugal
Rui CASTRO, Eindhoven University of Technology, The Netherlands
Valérie CHAVEZ-DEMOULIN, University of Lausanne, Switzerland
David CONESA, University of Valencia, Spain
Charmaine DEAN, University of Waterloo, Canada
Fernanda FIGUEIREDO, University of Porto, Portugal
Jorge Milhazes FREITAS, University of Porto, Portugal
Stéphane GIRARD, Inria Grenoble Rhône-Alpes, France
Sónia GOUVEIA, University of Aveiro, Portugal
Victor LEIVA, Pontificia Universidad Católica de Valparaíso, Chile
Artur LEMONTE, Federal University of Rio Grande do Norte, Brazil
Shuangzhe LIU, University of Canberra, Australia
Raquel MENEZES, University of Minho, Portugal
Fernando MOURA, Federal University of Rio de Janeiro, Brazil
Cláudia NEVES, King's College London, England
John NOLAN, American University, United States
Carlos OLIVEIRA, Norwegian University of Science and Technology, Norway
Paulo Eduardo OLIVEIRA, University of Coimbra, Portugal
Pedro OLIVEIRA, University of Porto, Portugal
Rosário OLIVEIRA, University of Lisbon, Portugal
Gilbert SAPORTA, Conservatoire National des Arts et Métiers, France
Alexandra M. SCHMIDT, McGill University, Canada
Lisete SOUSA, University of Lisbon, Portugal
Jacobo de UÑA-ÁLVAREZ, University of Vigo, Spain
Christian WEIß, Helmut Schmidt University, Germany

Executive Editor

Olga BESSA MENDES, Statistics Portugal