

# REVSTAT

Statistical Journal

vol. 20 - no. 5 - October 2022



REVSTAT-Statistical Journal, vol.20, n. 5 (October 2022)

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

Continues: Revista de Estatística = ISSN 0873-4275.

ISSN 1645-6726 ; e-ISSN 2183-0371

## Editorial Board (2019-2023)

**Editor-in-Chief** – *Isabel FRAGA ALVES*

**Co-Editor** – *Giovani L. SILVA*

### Associate Editors

*Marília ANTUNES*

*Barry ARNOLD*

*Narayanaswamy BALAKRISHNAN*

*Jan BEIRLANT*

*Graciela BOENTE*

*Paula BRITO*

*Valérie CHAVEZ-DEMOULIN*

*David CONESA*

*Charmaine DEAN*

*Fernanda FIGUEIREDO*

*Jorge Milhazes FREITAS*

*Alan GELFAND*

*Stéphane GIRARD*

*Marie KRATZ*

*Victor LEIVA*

*Artur LEMONTE*

*Shuangzhe LIU*

*Maria Nazaré MENDES-LOPES*

*Fernando MOURA*

*John NOLAN*

*Paulo Eduardo OLIVEIRA*

*Pedro OLIVEIRA*

*Carlos Daniel PAULINO*

*Arthur PEWSEY*

*Gilbert SAPORTA*

*Alexandra M. SCHMIDT*

*Manuel SCOTTO*

*Lisete SOUSA*

*Milan STEHLÍK*

*María Dolores UGARTE*

**Executive Editor** – *José A. PINTO MARTINS*

**Assistant Editor** – *Olga BESSA MENDES*

**Publisher** – *Statistics Portugal*

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

**Edition** - 140 copies | **Legal Deposit Registration** - 191915/03 | **Price** [VAT included] - € 9,00



Creative Commons Attribution 4.0 International (CC BY 4.0)

© Statistics Portugal, Lisbon. Portugal, 2022

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

# INDEX

<b>Comparison of Estimates Using L- and TL-Moments and Other Robust Characteristics of Distributional Shape and Tail Heaviness</b>	
<i>Ivana Malá, Václav Sládek and Filip Habarta</i> .....	529
<b>Plug-in Estimation of Dependence Characteristics of Archimedean Copula via Bézier Curve</b>	
<i>Selim Orhun Susam and Mahmut Sami Erdoğan</i> .....	547
<b>Estimation in Weibull Distribution Under Progressively Type-I Hybrid Censored Data</b>	
<i>Yasin Asar and Reza Arabi Belaghi</i> .....	563
<b>The Destructive Zero-Inflated Power Series Cure Rate Models for Carcinogenesis Studies</b>	
<i>Rodrigo R. Pescim, Adriano K. Suzuki, Gauss M. Cordeiro and Edwin M. M. Ortega</i> .....	587
<b>Single Index Regression Model for Functional Quasi-Associated Time Series Data</b>	
<i>Salim Bouzebda, Ali Laksaci and Mustapha Mohammedi</i> .....	605
<b>On Analyzing Non-Monotone Failure Data</b>	
<i>Muhammad Mansoor, M.H. Tahir, Gauss M. Cordeiro, Edwin M.M. Ortega and Ayman Alzaatreh</i> .....	633
<b>Likelihood-Based Finite Sample Inference for Synthetic Data from Pareto Model</b>	
<i>Nutan Mishra and Sandip Barui</i> .....	655







---

---

## Comparison of Estimates Using L- and TL-Moments and Other Robust Characteristics of Distributional Shape and Tail Heaviness

---

---

- Authors: IVANA MALÁ  
- Department of Probability and Statistics, University of Economics in Prague, Czech Republic  
[malai@vse.cz](mailto:malai@vse.cz)
- VÁCLAV SLÁDEK 
- Department of Probability and Statistics, University of Economics in Prague, Czech Republic  
[vaclav.sladek@vse.cz](mailto:vaclav.sladek@vse.cz)
- FILIP HABARTA 
- Department of Probability and Statistics, University of Economics in Prague, Czech Republic  
[filip.habarta@vse.cz](mailto:filip.habarta@vse.cz)

Received: March 2020

Revised: December 2020

Accepted: December 2020

### Abstract:

- Correct identification of a probability distribution is crucial in many areas of parametric statistics, inappropriate choice of the model can result in misleading or even incorrect decisions. In the text, we study the performance of robust characteristics of skewness and kurtosis of probability distributions that are less sensitive to outliers than the characteristics based on classical product moments. We use Monte Carlo simulation to illustrate properties of various robust (mainly quantile type) characteristics of skewness and kurtosis and compare them to the L-skewness (TL-skewness) and L-kurtosis (TL-kurtosis). The bias, standard and mean squared error of estimators are compared using simulations for standard normal, Laplace, Student, gamma and beta distributions and sample sizes ranged from 10 to 500 observations. The selected distributions gain symmetric and asymmetric unimodal distributions with different tail heaviness.

### Keywords:

- *robust characteristics; L-moments; TL-moments; skewness; kurtosis.*

### AMS Subject Classification:

- 49A05, 78B26, 65C05.

---

## 1. INTRODUCTION

---

Correct identification of a probability distribution is essential in many areas of parametric statistics, from the modelling of probability distributions to the regression modelling (assuming a dependent variable distribution), multivariate statistics, extreme-value analysis or time series analysis. The assumption about distribution form is crucial for parametric statistics, the correct or at least suitable choice of distribution allows a wide range of parametric procedures to be applied; in case of inappropriate choice, the results might be misleading or even incorrect. To test such the assumption, a large spectrum of statistical goodness-of-fit tests is available. For the general information on the sample, empirical distribution (histogram of data or nonparametric kernel density estimate) can be plotted. Sample characteristics of the location, variability, shape and concentration also can be evaluated. The typical sample characteristics are (raw, centred or standardised) product moments: mean, sample variance, coefficient of skewness and coefficient of kurtosis. Theoretical and sample moments are used not only to describe the distribution but also in the choice of suitable distribution to model the data or in inferential statistics. For example, if the normal distribution of data is assumed, the absolute value of the sample coefficient of skewness is supposed to be small, and the coefficient of kurtosis close to three. A frequently used test of normality Jarque–Bera compares theoretical and sample coefficients of skewness and kurtosis. The moment matching method can be applied to estimate parameters, the equations of theoretical and sample moments are solved explicitly or numerically with respect to the unknown parameters.

In the definition of the coefficient of skewness, a finite third raw moment is needed and for a finite value of the coefficient of kurtosis, the finite fourth raw moment is required. Moreover, the sample coefficients of skewness and kurtosis are strongly dependent on the sample size and the presence of outliers in the data. With higher sample product moments, the impact of outliers becomes more substantial. If the sample is drawn from long- or heavy-tailed distributions, we expect multiple outliers in the data and the use of more robust methods is essential.

We analyse data of this type in many fields of applications. For this reason, more robust characteristics and its estimates can be preferred to describe the distribution. There are various robust characteristics whose estimates are based on sample quantiles, which are more robust than sample product moments. Robust quantile characteristics of the distribution shape were presented, e.g., by [5], [10], [17] and [9], those of tail heaviness being dealt with by [7], [19] and many others. Hosking in [11] defined L-moments as a linear combination of order statistics, a robust alternative to product moments (robust moment characteristics). According to [11], [12], [14], [2] and other authors, estimates of L-moments are more reliable than those of product moments. TL-moments, defined as a trimmed linear combination of order statistics, are even more robust than L-moments. They were applied by [8] to describe the probability distribution.

The present article is focused on the estimation of the distribution shape and tail heaviness using robust moments and quantile characteristics. We treat L- and TL-moments in comparison with product moments and robust quantile characteristics. In this paper, we consider random samples from both symmetric (Student, normal, and Laplace) and asymmetric (gamma and beta) probability distributions. The latter ones being flexible, a different com-

bination of their parameters allows us to obtain different shapes of the distribution, including asymmetric distributions. The aim of the article is to compare estimated characteristics (bias and both standard and mean squared errors) of the shape and tail heaviness characteristics of distributions depending on the distribution and size of the random sample, employing Monte Carlo simulations. The calculation was performed in the program **R** ([20]), using predefined and author-written functions (cf. [22] and [2]). Along with formulas and a short description of the considered robust moment and quantile characteristics, the following methodology section also contains the algorithm of the simulation study. Results and inferences drawn from Monte Carlo simulation are summarised in the next part of the paper, the concluding section assessing the outcomes of the simulation.

---

## 2. METHODS

---

### 2.1. L-moments

---

Let  $X$  be a continuous random variable with a cumulative distribution function  $F(x)$ , quantile function  $Q(x)$  and let  $X_{1:n}, X_{2:n}, \dots, X_{n:n}$  be an ordered sample of the size  $n$  drawn from the distribution of the random variable  $X$ . L-moments are defined in [11] as a linear combination of order statistics, the  $r$ -th L-moment  $\lambda_r$  being as follows:

$$(2.1) \quad \lambda_r = \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} EX_{r-k:r}, \quad r = 1, 2, \dots,$$

where  $EX_{r-k:r}$  is an expected value of the  $(r-k)$ -th order statistics from a sample of size  $r$ . L-moments, a robust alternative to product moments, are used to describe random variables similarly as the classical product moments. The most used L-moments are those of order  $r = 1, 2, 3$ , and 4. The  $\lambda_1$  is equal to the expected value of the variable  $X$ , describing its level,  $\lambda_2$  indicating variability,  $\lambda_3$  shape, and  $\lambda_4$  tail heaviness of the distribution. Hosking and Wallis in [14] mentioned a dimensionless version of L-moments which is independent of the distribution scale and more useful than the unbounded version. Dimensionless L-moments  $\tau$  are called L-moment ratios. They are defined as

$$(2.2) \quad \tau_r = \lambda_r / \lambda_2, \quad r = 3, 4, \dots,$$

where  $\tau_3$  (L-skewness) and  $\tau_4$  (L-kurtosis) are most widely used in selection of probability distributions. Common properties of L-moments and L-moment ratios are as follows:

- They are defined for all distributions with finite expected values (finite values of higher moments are not required);
- There are not two distributions with the same values of all L-moments;
- $-1 < \tau_r < 1$  for  $r = 3, 4, \dots$ ;
- $\frac{1}{4}(3\tau_3^2 - 1) \leq \tau_4 < 1$ ;
- $\lambda_3 = \tau_3 = 0$  for symmetric distributions.

Estimates of  $\lambda_r$  and  $\tau_r$  are based on an ordered random sample  $X_{1:n}, X_{2:n}, \dots, X_{n:n}$  drawn from the probability distribution of  $X$ . Estimates can be calculated using the following formulas:

$$(2.3) \quad \hat{\lambda}_r = \frac{1}{r} \binom{n}{r}^{-1} \sum_{i=1}^n \sum_{j=0}^{r-1} (-1)^j \binom{r-1}{j} \binom{i-1}{r-1-j} \binom{n-i}{j} X_{i:n},$$

$$(2.4) \quad \hat{\tau}_r = \hat{\lambda}_r / \hat{\lambda}_2, \quad r = 3, 4, \dots$$

Statistical characteristics of estimates are available, e.g. in [2] or [11]).

The R package `Lmoments` [18] provides functions for evaluation of both symmetric and asymmetric sample moments. R packages `lmomco` [3] and `Lmoments` [16] allow a wide range of calculations based on L-moments.

---

## 2.2. TL-moments

---

Elamir and Seheult in [8] introduced TL-moments (trimmed L-moments) as a robust version of L-moments defined by the formula

$$(2.5) \quad \lambda_r^{(t)} = \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} EX_{r+t-k:r+2t}, \quad r = 1, 2, \dots,$$

where  $t$  represents the number of trimmed expected values from both sides of the sample. Zero weight is assigned to expected (mean) values, which are thus considered as outliers. Trimming can be either symmetric or asymmetric, the choice of the respective approach depends on the nature of the data. In the asymmetric form, the ordered sample is trimmed by  $t_1$  values from left and  $t_2$  values from right. The  $EX_{r+t-k:r+2t}$  in (2.5) is then changed to  $EX_{r+t_1-k:r+t_1+t_2}$ . The TL-moment is then denoted by  $\lambda_r^{(t_1, t_2)}$ .

In this article, we focus only on symmetric trimming with  $t = 1$ ; this choice is usually applied in the literature (also in [8]) as well as in practical applications. Trimming of one value is sufficient to overcome the problem of finite values of the moments for example for Cauchy distribution and enables the existence of all TL-moments to be finite ([8]), as only the expected values of minimum and maximum are not defined and their trimming allows the calculation of all TL-moments. The existence of TL-moments depends on the existence of expected values of ordered statistics and sometimes, more trimmed values should be used. For the Pareto distribution, from the formula (6) in [1], the number of necessary trimmed values depends on the shape parameter. The smaller the parameter, the higher the number of trimmed values. TL-moments can be used for the description of the distribution of a random variable.  $\lambda_1^{(t)}$  is equal to the expected value of the random variable (if it exists), describing its level;  $\lambda_2^{(t)}$  quantifying variability,  $\lambda_3^{(t)}$  shape and  $\lambda_4^{(t)}$  tail heaviness of the distribution.

The TL-moment ratios for the shape of distribution  $\tau_3^{(t)}$  (TL-skewness) and tail heaviness  $\tau_4^{(t)}$  (TL-kurtosis) are:

$$(2.6) \quad \tau_r^{(t)} = \lambda_r^{(t)} / \lambda_2^{(t)}, \quad r = 3, 4, \dots$$



Both characteristics are location and scale invariant ([8]). Main properties of TL-moments and TL-moment ratios are as follows:

- We obtain the L-moments for  $t = 0$  (or  $t_1 = t_2 = 0$ );
- For  $r \geq 3$  applies (see [13] for the more general equation for the trimming  $(t_1, t_2)$  instead of symmetric  $(t, t)$  denoted by  $(t)$ ):

$$(2.7) \quad |\tau_r^{(t)}| \leq \frac{2(t+1)!(r+2t)!}{r(t+r-1)!(2+2t)!};$$

- $|\tau_3^{(1)}| \leq \frac{10}{9}$  and  $|\tau_4^{(1)}| \leq \frac{5}{4}$  (substituting  $t = 1$  to (2.7));
- $\tau_3^{(t)} = 0$  for symmetric distributions.

Sample counterparts of (symmetric) TL-moments  $\lambda_r^{(t)}$  and  $\tau_r^{(t)}$  are based on an ordered random sample of size  $n$ :

$$(2.8) \quad \hat{\lambda}_r^{(t)} = \frac{1}{r} \sum_{i=t+1}^{n-t} \frac{\sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} \binom{i-1}{r+t-k-1} \binom{n-i}{t+k}}{\binom{n}{t+2t}} x_{i:n},$$

$$(2.9) \quad \hat{\tau}_r^{(t)} = \hat{\lambda}_r^{(t)} / \hat{\lambda}_2^{(t)}, \quad r = 3, 4, \dots$$

The R packages `TLmoments` [18] and `lmomco` [3] provides a wide range of useful functions for application of both symmetric and asymmetric sample TL-moments. In our analysis, we applied the former one.

### 2.3. Quantile characteristics of the distributional shape

In classical parametric statistics, the product moment coefficient of skewness is used as a third standardised raw moment

$$(2.10) \quad \alpha_3 = E(X - EX)^3 / (\text{Var } X)^{3/2},$$

in the sample version  $(X_i, i = 1, 2, \dots, n)$  based on the sample moments

$$(2.11) \quad a_3 = \sum_{i=1}^n (X_i - \bar{X})^3 / \left[ \sum_{i=1}^n (X_i - \bar{X})^2 \right]^{2/3},$$

where  $\bar{X}$  is a mean of the sample.

In the paper, we use more robust characteristics of shape based on robust moments as L-skewness (2.2) and TL-skewness (2.6) and quantiles. The first characteristic of the distribution shape mentioned above is the medcouple (referred to as  $MC_F$ ); see [6]. The sample version is defined as

$$(2.12) \quad MC_F = \text{median}_{i,j; x_i < x_j} h(x_i, x_j),$$

where  $h$  is a kernel function measuring the difference in the distances of  $x_i$  and  $x_j$  to the sample median  $\tilde{x}$ . This function is given by

$$(2.13) \quad h(x_i, x_j) = [(x_j - \tilde{x}) - (\tilde{x} - x_i)] / (x_j - x_i),$$

for  $x_i \neq x_j$ . If  $x_i = \tilde{x}$ , the value of  $h(\tilde{x}, x_j) = 1$  for  $x_j > Q(0.5)$ ;  $x_j = \tilde{x}$  gives  $h(x_i, \tilde{x}) = -1$  for  $x_i < \tilde{x}$ . If  $x_j$  is infinitely larger than  $\tilde{x}$ ,  $h$  is closed to 1. On the other hand, if  $x_i$  is infinitely smaller than  $\tilde{x}$ ,  $h$  approaches  $-1$ . Thus, the medcouple is not influenced by the presence of extreme values in a random sample; there are no larger/smaller values than  $\pm 1$ . The medcouple is defined for all continuous distribution functions, existence of the expected value of any distribution moment is not needed. The functional form of the characteristics is given in [6].

Bowley in [5] introduced the characteristic of shape based solely on distribution quartiles. It is called the Bowley coefficient of skewness ( $BC$ ) and is defined as:

$$(2.14) \quad BC = \left\{ [Q(0.75) - Q(0.5)] - [Q(0.5) - Q(0.25)] \right\} / [Q(0.75) - Q(0.25)].$$

Some authors use the term “quartile skewness” instead of the Bowley coefficient.

Hinkley [10] introduced the generalisation of Bowley measure:

$$(2.15) \quad \nu_1(p) = \left\{ [Q(1-p) - Q(0.5)] - [Q(0.5) - Q(p)] \right\} / [Q(1-p) - Q(p)],$$

for  $p \in (0, 1)$ . It is obvious that the Bowley coefficient is a special case of (2.15) for  $p = 0.25$ .

If we use (in (2.15)) the first and seventh octiles  $Q(0.125)$  and  $Q(0.875)$ , we obtain the octile skewness ( $OC$ ):

$$(2.16) \quad OC = \left\{ [Q(0.875) - Q(0.5)] - [Q(0.5) - Q(0.125)] \right\} / [Q(0.875) - Q(0.125)].$$

Groeneveld and Meeden in [9] proposed the coefficient of skewness ( $GMC$ ) given by

$$(2.17) \quad GMC = [EX - Q(0.5)] / E|X - Q(0.5)|.$$

The last robust quantile characteristic considered is the Pearson coefficient ( $PC$ ) introduced by Kendall and Stuart in [17]. Its formula is based on (2.17), where instead of the expected value of the absolute deviation between  $x_i$  and the median, they employ the standard deviation  $\sqrt{\text{Var } X}$  of a distribution:

$$(2.18) \quad PC = [EX - Q(0.5)] / \sqrt{\text{Var } X}.$$

$GMC$  is defined only for distributions with a finite value  $E|X|$ , whereas  $PC$  is defined for that with a finite variance.

All robust quantile and moment characteristics of the distribution shape (2.14)–(2.18) are defined on a range of values  $[-1, 1]$  (except TL-moments,  $|\tau_3^{(1)}| \leq 1.11$ ). For symmetric distributions, they are equal to zero. This allows us to compare properties of their estimates (bias and  $MSE$ ) using an absolute value basis. The same applies for asymmetric distributions as they are functionally bounded (despite particular characteristics acquiring different values).

The robust characteristics considered are applied only to the distributions for which they are defined (this is not the case, e.g., of Student distribution with 1 degree of freedom [the Cauchy distribution] and all characteristics based on classical moments or L-moments).

Sample counterparts of characteristics (2.14)–(2.18) are obtained by substituting the mean for the  $EX$ , sample standard error or more robust median absolute deviation for the standard deviation. There is not a generally accepted method for evaluation of sample quantiles. In the present paper, we use linear interpolation of the inverse of the empirical cumulative distribution function in the form

$$(2.19) \quad \hat{Q}(p) = x_{[h]:n} + (h - [h]) (x_{[h]+1:n} - x_{[h]:n}),$$

where  $h = (n - 1)p + 1$  and  $[h]$  is the floor function.

---

#### 2.4. Quantile characteristics of the tail heaviness

---

The moment coefficient of the kurtosis is defined as the fourth standardised raw moment

$$(2.20) \quad \alpha_4 = E(X - EX)^4 / (\text{Var } X)^2,$$

in the sample version based on the sample moments

$$(2.21) \quad a_4 = \sum_{i=1}^n (X_i - \bar{X})^4 / \left[ \sum_{i=1}^n (X_i - \bar{X})^2 \right]^2.$$

This sample formula is a biased estimator to  $\alpha_4$  (as well as for  $\alpha_3$  given in (2.11)) even for a sample from the normal distribution ([15]).

The first robust characteristics of tail heaviness based on octiles is the Moors coefficient of kurtosis ( $MKC$ ); see [19]. The coefficient is defined as

$$(2.22) \quad MKC = \left\{ [Q(0.875) - Q(0.625)] + [Q(0.375) - Q(0.125)] \right\} / [Q(0.625) - Q(0.125)].$$

$MKC$  exists for any continuous distribution taking all positive values.

Crown and Siddiqui in [7] introduced their coefficient defined as

$$(2.23) \quad CKC = [Q(1 - \alpha) - Q(\alpha)] / [Q(1 - \beta) - Q(\beta)],$$

for  $\alpha, \beta \in (0, 0.5)$ . The authors recommend using  $\alpha = 0.025$  and  $\beta = 0.25$  for the normal distribution. In our study, we follow Crown and Siddiqui recommendation regarding the considered probability distributions because many of them are symmetric. The suitability of this arrangement for asymmetric distributions is also open to analysis.

The last characteristic of tail heaviness considered, defined by Schmid and Tiede in [21] as a special case of (2.23), selects  $\alpha = 0.125$  and  $\beta = 0.25$ :

$$(2.24) \quad PKC = [Q(0.875) - Q(0.125)] / [Q(0.75) - Q(0.25)].$$

All the above robust quantile characteristics of tail heaviness exist for any distribution, no finite moment values being necessary. The range of values of these characteristics is  $[0, \infty]$ . A comparison of the characteristics of tail heaviness estimates based on their absolute values (means and standard deviations) is not appropriate because of various possible ranges of values; modification of characteristics (expected value and variance) is thus applied (see (2.25)).

---

## 2.5. The algorithm of simulation and methods of comparison

---

The simulation study employs the Monte Carlo methodology, which assumes knowledge of the theoretical distribution and its parameters, random samples being drawn from the distribution and values of all characteristics of interest being computed. Let us consider samples ranging from 10 to 500 observations, because the greatest differences in estimate characteristics are expected to concentrate in small-sized samples. The samples comprising 100 and more observations are used to analyse the convergence of estimation bias and variability. In the simulation study, we have chosen three symmetric and two asymmetric distributions:

- Student distribution ( $t(\nu)$ , degrees of freedom  $\nu = 1, 2, 3$ );
- Standard normal distribution ( $N(0; 1)$ );
- Laplace distribution ( $La(\mu; b)$ ,  $\mu = 0$ , the location parameter,  $b = 10$  the scale of the distribution);
- Gamma distribution ( $Gamma(\theta; k)$ ,  $\theta = 2$  the shape parameter,  $k = 2$  the scale parameter);
- Beta distribution ( $Beta(\theta_1; \theta_2)$ ,  $\theta_1 = 2$ ,  $\theta_2 = 5$  the shape parameters).

In Table 1, all characteristics introduced in Sections 2.1–2.4 for selected distributions are given.

The Student distribution is a symmetric continuous probability distribution, bell-shaped (similar to the normal distribution) and heavy-tailed (unlike the normal distribution). With an increasing number of degrees of freedom, the shape of the Student distribution converges to that of the normal distribution. We focus on how the above affects the properties of estimates. Gamma and beta distributions are also well-established distribution families. Their shape can be modified by setting different values of parameters. We choose combinations of parameter values to obtain a positively skewed shape.

The Laplace distribution is also called the double exponential distribution. It is symmetrically shaped like Student and normal distributions. The distinction between probability density functions of Laplace and normal distributions lies in that the latter is expressed as the squared difference, while the former as the absolute difference from their means, respectively. The Laplace distribution as a result has heavier tails than the normal distribution. Some of the chosen robust quantile and moment characteristics exist only if the distribution has one or more defined raw moments (Table 1). L-moments, for example, exist only for the distribution with a finite expected value, which does not apply to the Student distribution with one degree of freedom (Cauchy distribution). The existence of the moment characteristics of skewness

(coefficient of skewness) assumes finite 3rd raw moment, for the coefficient of kurtosis, we need finite 4-th raw moment. Estimate calculations are done only if the characteristics are defined for a particular distribution.

**Table 1:** Basic characteristics of probability distributions selected in the simulation. We use “—” for non-existence.

Characteristic	$t(1)$	$t(3)$	$N(0; 1)$	$La(0; 10)$	$Gamma(2; 2)$	$Beta(2; 5)$
$EX$	—	0	0	0	4	217
$\sqrt{\text{Var } X}$	—		1	14.142	2.828	0.126
$\alpha_3$	—	0	0	0	1.414	0.596
$\alpha_4$	—	—	3	3	0.142	2.88
$\tau_3$	—	0	0	0	0.235	0.123
$\tau_3^{(1)}$	0	0	0	0	0.150	0.080
$MC_F$	0	0	0	0	0.225	0.128
$BC$	0	0	0	0	0.172	0.095
$OC$	0	0	0	0	0.287	0.160
$GMC$	0	0	0	0	0.306	0.165
$PC_{MAD}$	0	0	0	0	0.227	0.133
$PC_{SD}$	—	0	0	0	0.227	0.133
$\tau_4$	—	0.035	0.123	0.035	0.071	0.090
$\tau_4^1$	0.077	0.041	0.063	0.041	1.731	0.048
Pearson	8.630	0.547	1.706	0.547	1.262	1.659
$MKC$	8.663	0.590	1.233	0.590	3.078	1.181
$CKC$	24,628,907	5.325	2.906	5.325	3.078	2.619

Some 50,000 times random samples were generated from the considered distributions for sample sizes 10–500, point estimates and standard errors of analysed characteristics calculated as the mean and the sample standard deviation of 50,000 generated values. Bias in characteristics of shape is shown by the difference between their theoretical value and estimated expected value. We compare the variability of characteristics using estimates of their standard errors calculated as sample standard deviations and mean squared error  $MSE$ .

Because of the wide range of tail heaviness values, which acquire different ones for given distributions, a comparison using bias and standard error does not induce relevant inferences. Therefore, we use modified bias and modified  $MSE$  characteristics, which are the same as classical ones divided by the (squared) theoretical value of a robust characteristic. All the characteristics are put on the same level, the coefficient of variation being based on such an approach. Finally, we obtain the ratio of the value of bias and  $MSE$  to the theoretical value of the robust characteristic. The modified bias and modified  $MSE$  are calculated as ( $\theta$  is a parameter,  $\hat{\theta}$  is its estimate)

$$(2.25) \quad E(\hat{\theta} - \theta)/\theta, \quad \text{Var}(\hat{\theta})/\theta^2.$$

This allows for a statistical comparison between the properties of different estimates.

---

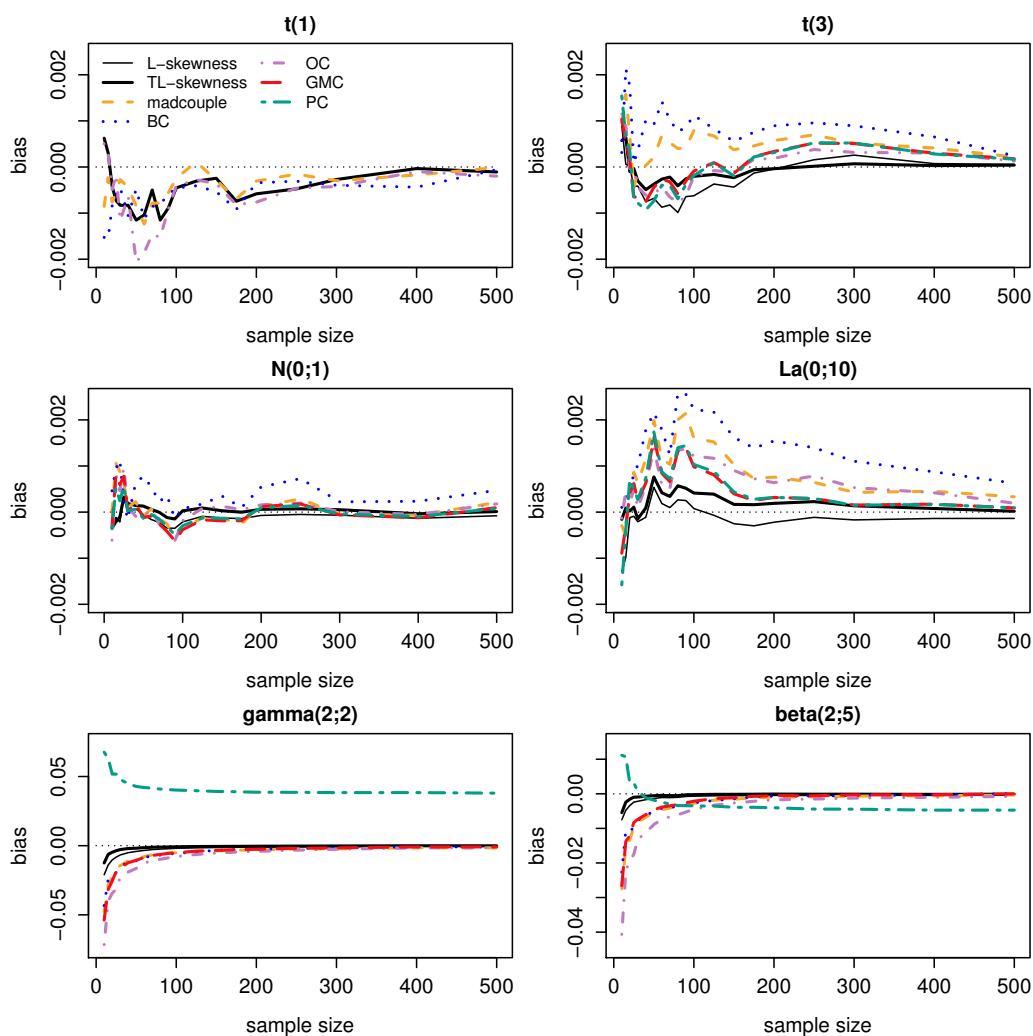
### 3. RESULTS

---

#### 3.1. Characteristics of skewness

---

The bias in estimated characteristics is low for symmetric distributions (Student, normal and Laplace). Figure 1 shows its development depending on sample sizes. For the Student distribution with one degree of freedom (Cauchy distribution), the characteristics  $\tau_3$ ,  $GMC$ , and  $PC$  are undefined ( $EX$  does not exist) and they are not included in the figure (see also Figure 2).



**Figure 1:** Estimated bias in distribution shape characteristics ((2.4), (2.9), (2.12), (2.14), (2.16)–(2.18)) for  $n = 10$ –500.

Bias curves are similar for all estimates, differing only in their levels, converging to zero (represented by the dashed line) with an increasing number of observations. Estimation bias is volatile in small samples (up to 100 observations), which show no constant development.

The figure indicates the highest bias in  $BC$  for each symmetric distribution,  $\tau_3$  and  $\tau_3^{(1)}$  being among the estimates with the lowest bias value. In absolute terms, however, it is obvious that estimation bias is generally small. Values for Student distribution are shown in Table 2. The lowest values (the best performance for the particular distribution and degrees of freedom) are highlighted in bold letters, the highest in italics (the worst performance in the block in the table). The TL-skewness is superior in bias from two degrees of freedom and performs well. For the Cauchy distribution, the L-skewness is undefined (this characteristic is equivalent to TL-skewness for  $t = 0$  (no trimmed values)), and the TL-skewness is a first possible defined value concerning the number of trimmed values.

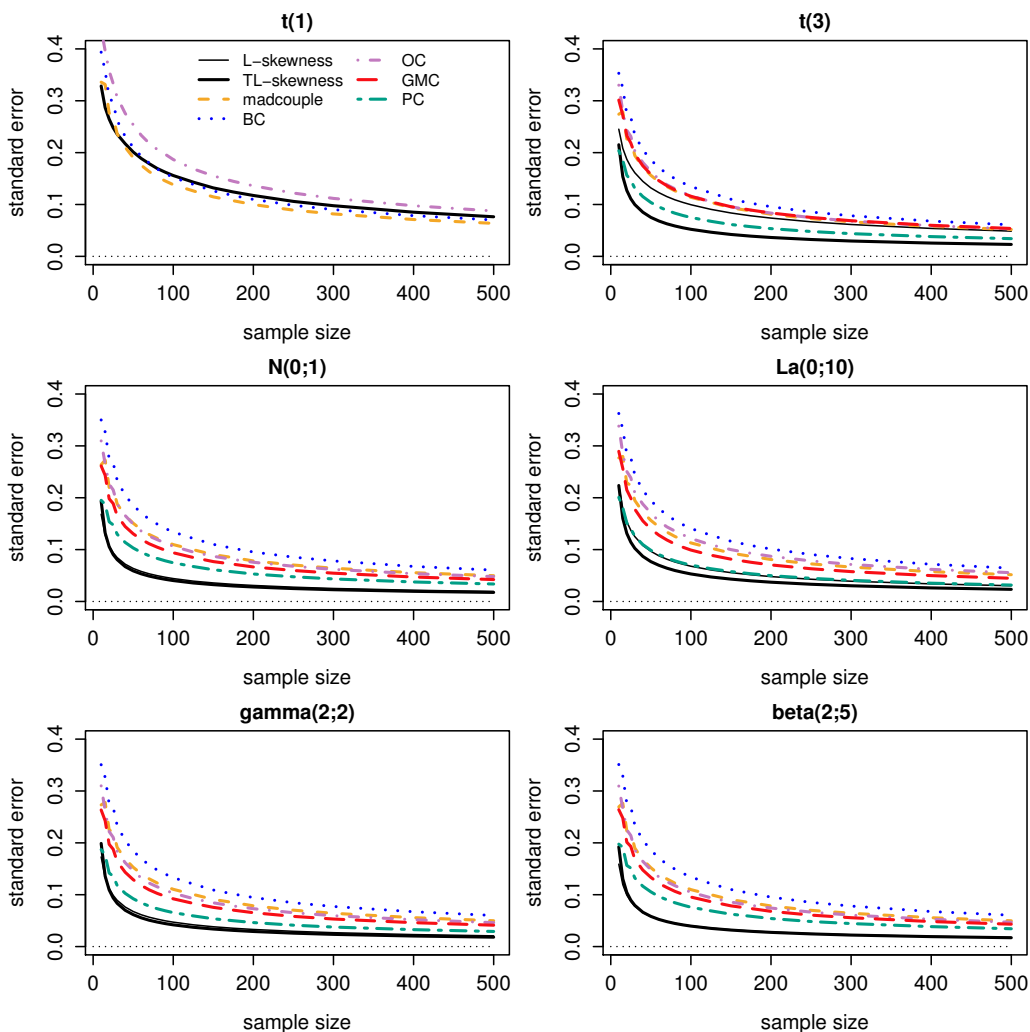
**Table 2:** Student distribution ( $\nu = 1, 2, 3; n = 10, 20, 30, 50, 70, 90, 100, 150, 200, 300$ ), characteristics of shape ((2.4), (2.9), (2.12), (2.14)–(2.18)). Estimated standard errors of estimates.

DF	Char.	10	20	30	50	70	90	100	150	200	300
1	$\tau_3^{(1)}$	<b>0.328</b>	<b>0.264</b>	<b>0.235</b>	0.201	0.178	0.162	0.155	0.132	0.117	0.097
	$MC_F$	0.335	0.276	0.236	<b>0.190</b>	<b>0.163</b>	<b>0.145</b>	<b>0.1382</b>	<b>0.114</b>	<b>0.099</b>	<b>0.081</b>
	$BC$	0.393	0.307	0.261	0.210	0.180	0.159	0.152	0.126	0.109	0.090
	$OC$	<i>0.439</i>	<i>0.357</i>	<i>0.306</i>	<i>0.253</i>	<i>0.218</i>	<i>0.196</i>	<i>0.186</i>	<i>0.155</i>	<i>0.136</i>	<i>0.111</i>
2	$\tau_3$	0.304	0.255	0.231	0.202	0.18	<i>0.177</i>	<i>0.165</i>	<i>0.146</i>	<i>0.133</i>	<i>0.115</i>
	$\tau_3^{(1)}$	<b>0.236</b>	<b>0.150</b>	<b>0.124</b>	<b>0.093</b>	<b>0.078</b>	<b>0.069</b>	<b>0.066</b>	<b>0.054</b>	<b>0.046</b>	<b>0.038</b>
	$MC_F$	0.285	0.235	0.201	0.162	0.139	0.123	0.118	0.096	0.084	0.069
	$BC$	<i>0.358</i>	<i>0.278</i>	0.237	0.189	0.161	0.143	0.135	0.111	0.097	0.079
	$OC$	0.350	0.263	0.226	0.179	0.151	0.136	0.128	0.105	0.092	0.075
	$GMC$	0.346	0.282	<i>0.249</i>	<i>0.209</i>	<i>0.186</i>	0.170	0.164	0.142	0.127	0.106
3	$\tau_3$	0.245	0.187	0.161	0.132	0.115	0.104	0.099	0.083	0.073	0.061
	$\tau_3^{(1)}$	0.215	<b>0.128</b>	<b>0.099</b>	<b>0.075</b>	<b>0.062</b>	<b>0.055</b>	<b>0.052</b>	<b>0.042</b>	<b>0.036</b>	<b>0.029</b>
	$MC_F$	0.274	0.226	0.194	0.156	0.134	0.119	0.113	0.093	0.081	0.065
	$BC$	0.353	<i>0.274</i>	<i>0.233</i>	<i>0.186</i>	<i>0.159</i>	<i>0.141</i>	<i>0.134</i>	<i>0.112</i>	<i>0.095</i>	<i>0.078</i>
	$OC$	0.330	0.243	0.203	0.163	0.138	0.123	0.116	0.095	0.083	0.068
	$GMC$	0.301	0.232	0.197	0.158	0.136	0.121	0.115	0.097	0.083	0.068
	$PC_{MAD}$	<i>0.443</i>	0.280	0.224	0.174	0.148	0.129	0.122	0.099	0.085	0.069
	$PC_{SD}$	<b>0.203</b>	0.156	0.131	0.103	0.088	0.078	0.075	0.061	0.053	0.043

For the asymmetric distributions (last row in Figure 1) the  $\tau_3$  and  $\tau_3^{(1)}$  exhibit the lowest absolute and relative values of bias for small samples (relative bias is equal to the estimation bias value divided by the actual value of the characteristic). An extremely high bias even for the samples with several hundred observations occurs in  $PC_{MAD}$ , where  $MAD$  is used as a standard deviation estimate in [19], its convergence being very slow. Both for small and large samples, bias is low in absolute terms and estimate convergences are relatively fast (except  $PC$  estimates).

Distribution shape estimates vary mostly in variability. Using the standard error of estimation, Figure 2 illustrates standard errors of estimates. For the Student distribution with one degree of freedom the characteristics  $\tau_3$ ,  $GMC$ , and  $PC$  are undefined (see also Figure 1), for this reason, no lines are included. For the normal and Laplace distributions with relatively low kurtosis and absence of outliers, the standard errors are very close for both characteristics based on L-moments. Let us first summarise the results for the Student

distribution (also see Table 3). The  $\tau_3^{(1)}$  and  $MC_F$  show lower variability (especially in the case of small samples) than other estimates with one degree of freedom. The shape of the Student distribution can be estimated using L-moments and Pearson and  $GM$  coefficients only if the number of degrees of freedom is higher than one. The  $\tau_3$  has lower variability than the quantile-based estimates but higher than  $\tau_3^{(1)}$  and  $MC_F$ . For small samples, the study outcomes confirm that the variability of  $\tau_3$  and  $\tau_3^{(1)}$  decreases more sharply with an increasing number of degrees of freedom than the variability of other estimates. The  $PC_{SD}$  and  $\tau_3^{(1)}$  are the estimates with the lowest variability for the Student distribution with three degrees of freedom, other estimates showing much higher variability.



**Figure 2:** Standard errors of estimates of distribution shape characteristics ((2.4), (2.9), (2.12), (2.14), (2.16)–(2.18)) for  $n = 10$ –500.

The order of estimates showing the lowest variability for the sample containing ten observations drawn from the standard normal distribution is  $\tau_3$ ,  $\tau_3^{(1)}$  and  $PC_{SD}$ . Other estimates show much higher variability. The difference in variability between  $\tau_3$  and  $\tau_3^{(1)}$  decreases with an increasing number of observations. The variability of  $PC_{SD}$  declines slowly



compared to  $\tau_3$  and  $\tau_3^{(1)}$ . Convergence in the variability of other estimates is not fast enough to reach the value of  $\tau_3$  and  $\tau_3^{(1)}$ . The last considered symmetric distribution is the Laplace distribution. As is the case with the normal distribution,  $\tau_3$  and  $\tau_3^{(1)}$  and  $PC_{SD}$  are estimates with the lowest variability. Also, the speed of their variability convergence seems similar. The variability of other estimates is significantly higher, their convergence not being fast enough to achieve  $\tau_3$  and  $\tau_3^{(1)}$  and  $PC_{SD}$  variability. The order of estimates arranged according to their variability is the same for gamma and beta distributions. The conclusions drawn are analogous to those concerning the normal distribution. The  $\tau_3$  and  $\tau_3^{(1)}$  and  $PC_{SD}$  show the lowest variability, that of  $PC_{SD}$  decreases more slowly compared to  $\tau_3$  and  $\tau_3^{(1)}$ . Convergence in the variability of other estimates is not as fast  $\tau_3$  and  $\tau_3^{(1)}$ .

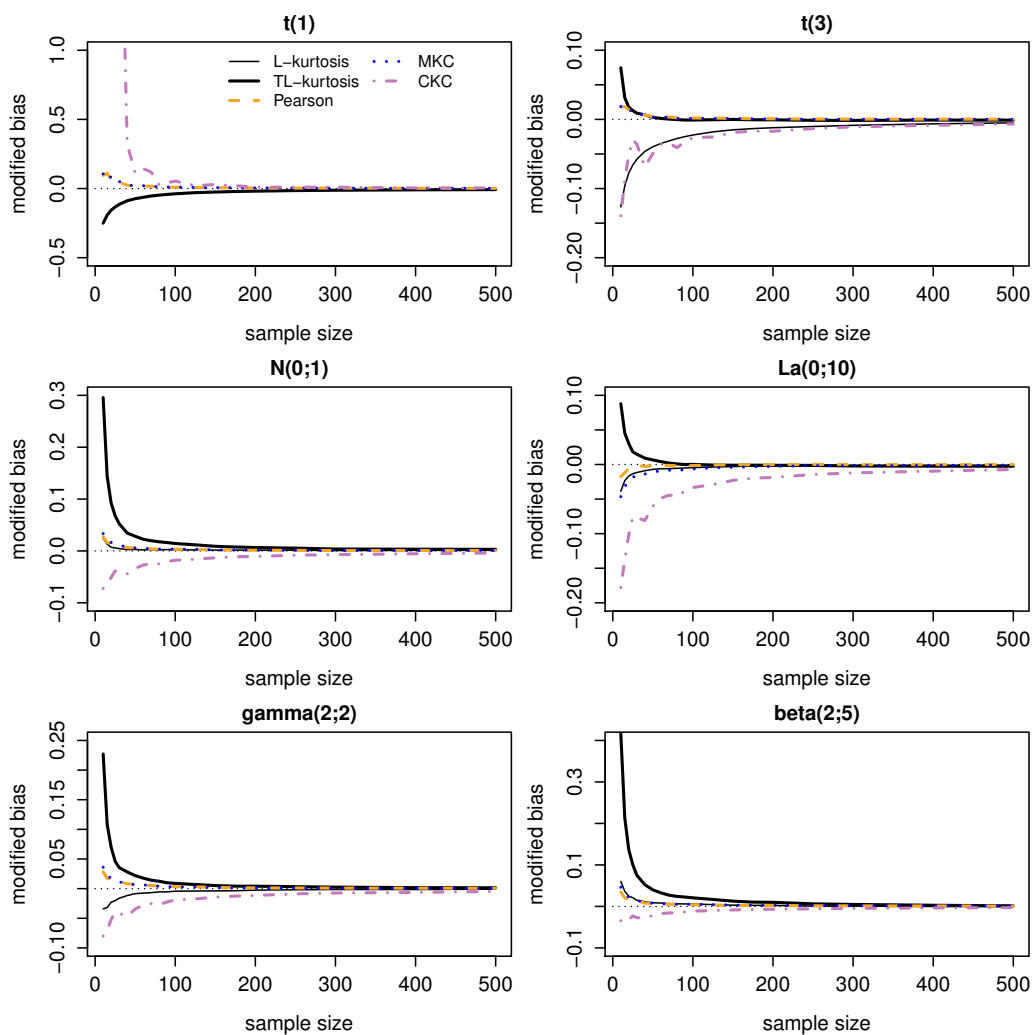
**Table 3:** Student distribution ( $\nu = 1, 2, 3; n = 10, 20, 30, 50, 70, 90, 100, 150, 200, 300$ ), modified bias of estimates (defined in (2.25)) of characteristics of kurtosis ((2.4), (2.9), (2.22)–(2.24)).

DF	Sample	10	20	30	50	70	90	100	150	200	300
1	$\tau_4^{(1)}$	-0.250	-0.193	-0.156	-0.115	-0.074	-0.054	-0.048	-0.042	-0.038	-0.030
	<i>PKC</i>	<b>0.095</b>	<b>0.110</b>	<b>0.065</b>	<b>0.045</b>	<b>0.019</b>	<b>0.016</b>	<b>0.011</b>	<b>0.010</b>	<b>0.009</b>	<b>0.008</b>
	<i>MKC</i>	0.106	0.129	0.075	0.051	0.020	0.017	0.012	0.011	0.010	0.008
	<i>CKC</i>	3.028	3.746	3.419	2.768	0.111	0.119	0.026	0.042	0.054	0.021
2	$\tau_4$	-0.224	-0.183	-0.157	-0.125	-0.093	-0.075	-0.069	-0.063	-0.059	-0.050
	$\tau_4^{(1)}$	<b>-0.010</b>	<b>-0.022</b>	-0.022	-0.017	-0.012	-0.009	-0.008	-0.007	-0.006	-0.005
	<i>PKC</i>	0.024	0.026	0.018	0.013	0.004	0.004	0.002	0.002	0.002	0.002
	<i>MKC</i>	0.025	0.030	0.020	0.013	0.002	0.003	0.001	0.001	0.002	0.002
	<i>CKC</i>	-0.118	-0.037	0.014	0.021	-0.047	-0.026	-0.041	-0.029	-0.021	-0.025
3	$\tau_4$	-0.127	-0.096	-0.077	-0.058	-0.040	-0.031	-0.028	-0.025	-0.023	-0.018
	$\tau_4^{(1)}$	0.075	0.031	0.018	<b>0.009</b>	<b>0.003</b>	<b>0.000</b>	<b>0.000</b>	-0.001	<b>-0.001</b>	<b>0.000</b>
	<i>PKC</i>	<b>0.018</b>	<b>0.018</b>	<b>0.014</b>	0.010	0.004	0.004	0.002	0.002	0.002	0.002
	<i>MKC</i>	0.019	0.021	0.016	0.010	0.003	0.003	0.001	<b>0.001</b>	0.001	0.001
	<i>CKC</i>	-0.138	-0.087	-0.051	-0.036	-0.049	-0.034	-0.041	-0.032	-0.026	-0.025

### 3.2. Characteristics of kurtosis

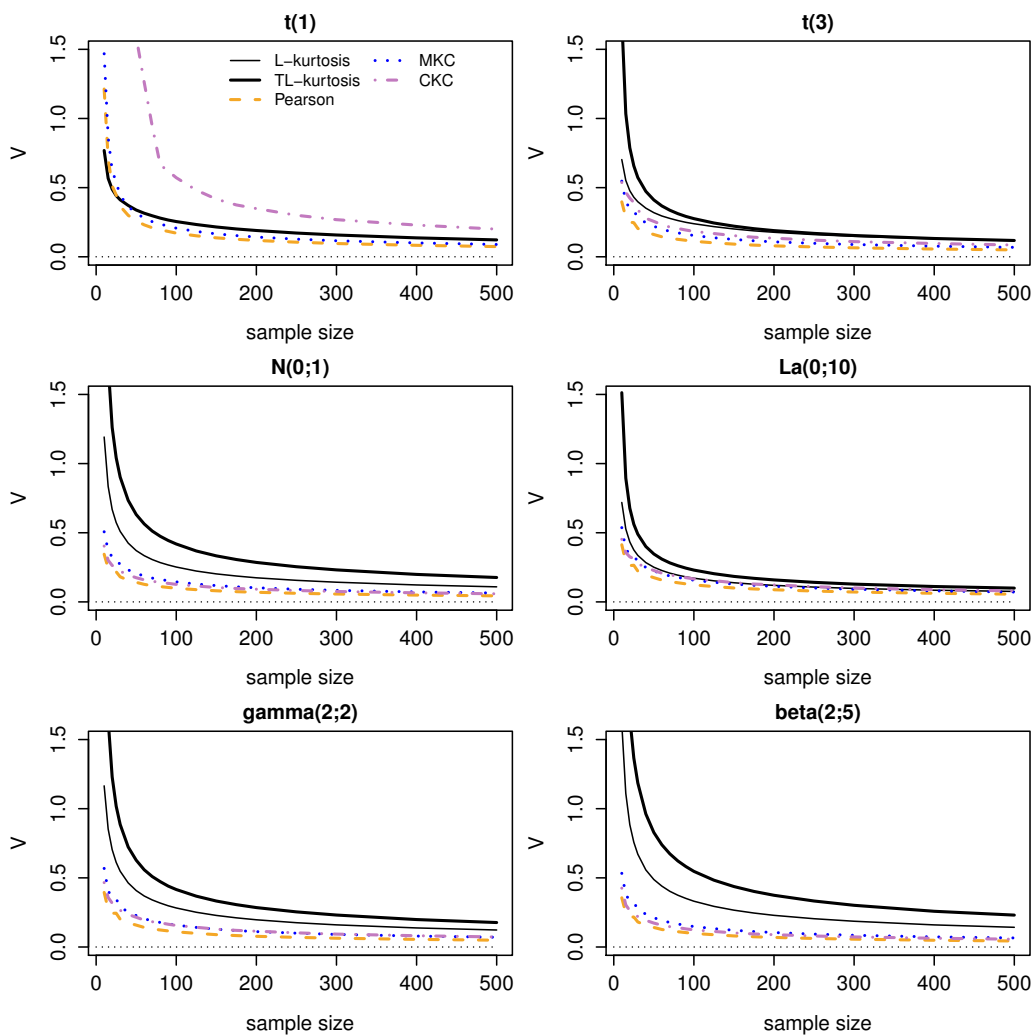
Given the inconsistent values of tail heaviness characteristics of considered distributions is shown in Figure 3 and for Student distribution in Table 3. For the Cauchy distribution  $t(1)$ , again the characteristics based on L-moments is undefined. The most biased estimate for the Student distribution with one degree of freedom is *CKC*, converging faster than the other considered estimates. It belongs to high-biased symmetric distribution estimates comparable with those for samples with 100 and more observations. The  $\tau_4^{(1)}$  has a high value of modified bias for small samples drawn from normal and Laplace distributions. The  $\tau_4$ , *MKC*, and *PKC*, on the other hand, are the least-biased estimates for all symmetric distributions (both small and large samples) considered. Their bias modification values are similar. Table 3 contains the values of the modified bias of estimates for the Student distribution. *MKC* and *PKC* are estimates with the lowest modified bias values for all degrees of freedom considered, including small samples. The modified bias of  $\tau_4^{(1)}$  is close to *MKC* and *PKC* for two and more degrees of freedom.

The  $\tau_4$  and  $CKC$  exhibit the highest modified bias for this distribution. The  $\tau_4^{(1)}$  is the most biased estimate for both distributions considered. The  $\tau_4^{(1)}$  overestimates its theoretical value (for a sample with 10 observations) in the cases of gamma and beta distributions by about 22 and 41%, respectively. Its decline is relatively sharp with an increasing number of observations, and bias is similar to that of other estimates for samples with 60 and more observations. The  $MKC$  and  $PKC$  show the lowest value of modified bias analogous to that for a symmetric distribution.  $\tau_4$  is close to  $MKC$  and  $PKC$ , and  $CKC$  estimate is less biased for asymmetric distributions than for symmetric ones.



**Figure 3:** Modified estimation bias (2.25) in distribution tail characteristics ((2.4), (2.9), (2.22)–(2.24)) for  $n = 10$ –500.

The variability of estimates is quantified using the variation coefficient. Its development for symmetric distributions is shown in Figure 4. The  $\tau_4^{(1)}$  has the lowest coefficient of variation for small samples (up to 25 observations) drawn from the Student distribution with one degree of freedom, its convergence being slower than that of  $MKC$  and  $PKC$ . If a sample consists of more than 25 observations,  $MKC$  and  $PKC$  are less variable than  $\tau_4^{(1)}$ .



**Figure 4:** Variation coefficient of estimates of tail heaviness characteristics ((2.4), (2.9), (2.22)–(2.24)) for  $n = 10$ –500.

Table 4 shows values of the variation coefficient of  $CKC$ ,  $\tau_4^{(1)}$  and  $\tau_4$  for the Student distribution dependent on degrees of freedom. Interestingly, the variability of  $CKC$  declines markedly when degrees of freedom change from one to two (from 390.565 to 1.103 for a 10-observation sample), the variability of  $\tau_4^{(1)}$  and  $\tau_4$  growing with an increase in degrees of freedom. The latter two estimates are more variable than other ones in the case of small samples generated from normal and Laplace distributions. For large samples, the variability of  $\tau_4$  is comparable with other estimates.  $PKC$  has the lowest variability for each symmetric distribution considered. For the asymmetric distributions both  $\tau_4$  and  $\tau_4^{(1)}$  show fast convergence of variability. However, even for the sample with 500 observations, their variability is several times greater than that of  $MKC$ ,  $PKC$  and  $CKC$ , the variability of  $\tau_4^{(1)}$  being the highest. Therefore, in terms of variability, neither  $\tau_4$  nor  $\tau_4^{(1)}$  are appropriate estimates of the tail heaviness of asymmetric distributions.

Because of estimation bias, comparison of estimates is made on the basis of the modified mean square error (2.25), the method taking into account both bias and variability of

estimates. The development of modified  $MSE$  is similar to that of estimates of a coefficient of variation (not given in the text).  $MSE$ -based methodology provides results similar to those yielded by variation analysis.

**Table 4:** Student distribution ( $\nu = 1, 2, 3; n = 10, 20, 30, 50, 70, 90, 100, 150, 200, 300$ ). Coefficient of variation of characteristics of kurtosis ((2.4), (2.9), (2.12), (2.14)–(2.18)).

DF	Char.	10	20	30	50	70	90	100	150	200	300
1	$\tau_4^{(1)}$	1.025	0.579	0.465	0.365	0.314	0.278	0.265	0.221	0.194	0.160
	<i>PKC</i>	<b>1.108</b>	<b>0.481</b>	<b>0.350</b>	<b>0.255</b>	<b>0.207</b>	<b>0.182</b>	<b>0.169</b>	<b>0.136</b>	<b>0.118</b>	<b>0.096</b>
	<i>MKC</i>	1.327	0.581	0.422	0.308	0.250	0.220	0.205	0.165	0.142	0.116
	<i>CKC</i>	<i>96.964</i>	<i>62.156</i>	<i>47.601</i>	<i>1.431</i>	<i>0.844</i>	<i>0.594</i>	<i>0.543</i>	<i>0.405</i>	<i>0.346</i>	<i>0.266</i>
2	$\tau_4$	0.724	0.500	0.422	0.348	<i>0.308</i>	<i>0.281</i>	<i>0.271</i>	<i>0.235</i>	<i>0.213</i>	<i>0.184</i>
	$\tau_4^{(1)}$	<i>1.384</i>	0.685	0.512	<i>0.372</i>	0.307	0.266	0.251	0.202	0.174	0.141
	<i>PKC</i>	<b>0.462</b>	<b>0.275</b>	<b>0.224</b>	<b>0.175</b>	<b>0.146</b>	<b>0.130</b>	<b>0.122</b>	<b>0.100</b>	<b>0.086</b>	<b>0.071</b>
	<i>MKC</i>	0.611	0.368	0.298	0.232	0.194	0.172	0.162	0.132	0.114	0.093
3	$\tau_4$	0.805	0.518	0.422	0.332	0.286	0.256	0.245	0.204	0.179	0.149
	$\tau_4^{(1)}$	<i>1.601</i>	<i>0.775</i>	<i>0.568</i>	<i>0.410</i>	<i>0.336</i>	<i>0.292</i>	<i>0.276</i>	<i>0.221</i>	<i>0.189</i>	<i>0.154</i>
	<i>PKC</i>	<b>0.390</b>	<b>0.247</b>	<b>0.203</b>	<b>0.160</b>	<b>0.134</b>	<b>0.119</b>	<b>0.112</b>	<b>0.092</b>	<b>0.079</b>	<b>0.065</b>
	<i>MKC</i>	0.537	0.343	0.279	0.220	0.184	0.163	0.154	0.125	0.108	0.089
	<i>CKC</i>	0.623	0.447	0.351	0.269	0.222	0.200	0.189	0.156	0.137	0.111

---

#### 4. CONCLUSION

---

Simulation results show that the bias of distribution shape estimates is low for both symmetric and asymmetric probability distributions. The main difference between estimates is in their variability (quantified by standard error).  $\tau_3$  and  $\tau_3^{(1)}$  are estimates with small variability, the best robust quantile ones in terms of variability being  $MC_F$  and  $PC_{SD}$ . The variability of  $\tau_3$  and  $\tau_3^{(1)}$  decreases more sharply than that of other estimates in the case of small samples with an increasing number of degrees of freedom of the Student distribution. The  $\tau_3$  and  $\tau_3^{(1)}$  and  $PC_{SD}$  are the most appropriate  $\tau$  estimates for symmetric (Student, normal and Laplace) and asymmetric (gamma and beta) distributions dealt with in this paper.

Some conclusions concerning tail heaviness estimates follow:  $\tau_4$  and  $\tau_4^{(1)}$  has a high value of modified bias for small samples drawn from normal and Laplace distributions,  $\tau_4^{(1)}$  is the most biased estimate for asymmetric distributions, and  $MKC$  and  $PKC$  are those with the lowest value of modified bias as far as Student distributions are concerned. We conclude that  $\tau_4$ ,  $MKC$  and  $PKC$  show the lowest bias for all the considered symmetric and asymmetric distributions (small and large samples alike), values of their modified bias being mutually comparable. The variability of  $\tau_4^{(1)}$  and  $\tau_4$  increases with increasing degrees of freedom of the Student distribution. As for small samples generated from Normal and Laplace distributions,  $\tau_4^{(1)}$  and  $\tau_4$  are more variable than other estimates. The  $\tau_4$  variability is comparable to other estimates for large samples. While  $PKC$  indicates the lowest variability for each given symmetric distribution, the variability of  $\tau_4$  and  $\tau_4^{(1)}$  is much higher than that of other tail heaviness estimates for asymmetric distributions.

Estimates of distributional shape based on L- and TL-moments possess the best characteristics (bias and variability), outperforming those yielded by a robust quantile approach in the situations considered. Our study, however, also confirms that robust quantile-based estimators produce more reliable tail heaviness estimation outcomes than those based on L- and TL-moments.

---

## ACKNOWLEDGMENTS

---

Institutional support from the funds for the long-term conceptual advancement of science and research, number IP 400 040, at the Faculty of Informatics and Statistics, Prague University of Economics and Business, Prague, Czech Republic is gratefully acknowledged. We also acknowledge valuable suggestions from the referees.

---

## REFERENCES

---

- [1] ARNOLD, B.C. (2015). *Pareto Distribution*. In: Wiley Online Library.  
<https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat01100.pub2>
- [2] ASQUITH, W.H. (2011). *Distributional analysis with L-moment Statistics using the R environment for statistical computing*, Lubbock, Texas, Create Space Independent Publishing Platform.
- [3] ASQUITH, W.H. (2020). `lmomco`: L-moments, censored L-moments, trimmed L-moments, L-comoments, and many distributions. R package version 2.3.6.  
<http://www.cran.r-project.org/package=lmomco>
- [4] BONATO, M. (2011). Robust estimation of skewness and kurtosis in distributions with infinite higher moments, *Financial Research Letters*, **8**, 77–87.
- [5] BOWLEY, A.L. (1920). *Elements of Statistics*, New York, P.S. King & son, Limited.
- [6] BRYS, G.; HUBERT, M. and STRUYF, A. (2004). A Robust measure of skewness, *Journal of Computational and Graphical Statistics*, **13**, 996–1017.
- [7] CROWN, E. and SIDDIQUI, M. (1967). Robust estimation of location, *The Statistician*, **62**, 353–389.
- [8] ELSAYED, A.H.E. and SEHEULT, A.H. (2003). Trimmed L-moments, *Computational Statistics & Data Analysis*, **43**, 299–314.
- [9] GROENEVELD, R. and MEEDEN, G. (1984). Measuring skewness and kurtosis, *The Statistician*, **33**, 391–399.
- [10] HINKLEY, D.V. (1975). On power transformations to symmetry, *Biometrika*, **62**, 101–111.
- [11] HOSKING, J.R.M. (1990). L-moments: analysis and estimation of distribution using linear combinations of order statistics, *J. R. Statist. Soc. B*, **52**, 105–124.
- [12] HOSKING, J.R.M. (1992). Moments of L-moments? An example comparing two measures of distributional shape, *The American Statistician*, **46**, 186–189.
- [13] HOSKING, J.R.M. (2007). Some theory and practical uses of trimmed L-moments, *Journal of Statistical Planning and Inference*, **137**, 3024–3039.

- [14] HOSKING, J.R.M. and WALLIS, J.R. (1997). *Regional Frequency Analysis*, Cambridge, New York, Cambridge University Press.
- [15] JOANES, D.N. and GILL, C.A. (1998). Comparing Measures of Sample Skewness and Kurtosis, *The American Statistician*, **47**, 183–189.
- [16] KARVANEN, J. (2019). **Lmoments**: L-moments and quantile mixtures. R package version 1.3-1. <http://www.cran.r-project.org/package=Lmoments>
- [17] KENDALL, M. and STUART, A. (1977). *The Advance Theory of Statistics*, London, Griffin.
- [18] LILIENTHAL, J. (2019). **tlmoments**: calculate TL-moments and convert them to distribution parameters. R package version 0.7.5. <https://cran.r-project.org/web/packages/TLMoments>
- [19] MOORS, J. (1988). A quantile alternative for kurtosis, *The Statistician*, **37**, 25–32.
- [20] R CORE TEAM (2016). **R**: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- [21] SCHMID, F. and TREDE, M. (2003). Simple tests for peakedness, fat tails and leptokurtosis based on quantiles, *Computational Statistics & Data Analysis*, **43**, 1–12.
- [22] YEE, T.W. (2016). **VGAM**: Vector Generalized Linear and Additive Models. R package version 1.0-2. <http://CRAN.R-project.org/package=VGAM>

---

---

## Plug-in Estimation of Dependence Characteristics of Archimedean Copula via Bézier Curve

---

---

Authors: SELIM ORHUN SUSAM  

– Department of Econometrics, Munzur University,  
Tunceli, Turkey  
[orhunsusam@munzur.edu.tr](mailto:orhunsusam@munzur.edu.tr)

MAHMUT SAMI ERDOĞAN 

– Department of Statistics, Istanbul Medeniyet University,  
Istanbul, Turkey  
[sami.erdogan@medeniyet.edu.tr](mailto:sami.erdogan@medeniyet.edu.tr)

Received: June 2019

Revised: January 2021

Accepted: January 2021

### Abstract:

- This article introduces measurement of the dependence between variables with a dependence structure defined by Archimedean copulas. The estimation of the dependence measure, such as Kendall's tau as well as the lower and upper tail dependence, is investigated by using estimation of the Kendall distribution function based on the Bézier curve. A Monte Carlo study is performed to measure the performance of the new estimation method. The simulation results showed that the proposed methods has good results in terms of estimation performance. The new estimators are also used to estimate the dependence coefficients for three sets of real data.

### Keywords:

- *Archimedean copula; Bernstein polynomials; tail dependence; Kendall's tau.*

### AMS Subject Classification:

- 62G32, 62G05.

---

## 1. INTRODUCTION

---

In statistical theory and applications, copula models are useful tools for determining the dependence structure between the random variables. For instance, when two random variables of  $X$  and  $Y$  with joint cumulative distribution function  $H$  and marginals of  $F$  and  $G$  are considered respectively, there exists a copula  $C$  such that  $H(x, y) = C(F(x), G(y))$ , for all  $x, y$  in  $\mathbb{R}$ . In the literature, there are many parametric copula families which have different dependence structure. The main focus of this paper was on the Archimedean copula class, which is characterized by generator function  $\varphi$ . Archimedean copula with generator function  $\varphi$  is defined by

$$(1.1) \quad C(u, v) = \varphi^{[-1]} \{ \varphi(u) + \varphi(v) \}, \quad u, v \in [0, 1],$$

where  $\varphi$  is a generator function which is continuous and strictly decreasing convex function defined from  $\mathbf{I}$  to  $[0, \infty)$  such that  $\varphi(1) = 0$ .

Genest *et al.* [15] showed that the function  $\varphi$  can be obtained by the univariate distribution function of  $K(t) = P(C(u, v) \leq t)$ . Remarkably, there is a relationship between the function  $\varphi(t)$  and  $K(t)$  as

$$(1.2) \quad K(t) = t - \frac{\varphi(t)}{\varphi'(t)}.$$

The Kendall distribution function  $K(t)$  has some important properties. These properties are summarized by Nelsen [20] as follows:

1.  $K(0) = 0$ ;
2.  $K(1) = 1$ ;
3.  $K(t) > t$ ,  $t \in (0, 1)$ ;
4.  $K'(t) > 0$ ,  $t \in (0, 1)$ .

The dependence structure of the Archimedean copula family is characterized by  $K(t)$ . Kendall's tau ( $\tau$ ) is designed to describe how large (or small) values of one random variable appear with large (or small) values of the other as defined by Genest *et al.* [13] by

$$(1.3) \quad \tau = 3 - 4 \int_0^1 K(t) dt.$$

Also, the tail dependence is related to the level of dependence in the upper-right ( $\lambda_U$ ) or lower-left ( $\lambda_L$ ) quadrant tail of a bivariate distribution. Michiels *et al.* [19] defined  $\lambda_L$  and upper  $\lambda_U$  dependence as

$$(1.4) \quad \lambda_L = 2 \lim_{t \rightarrow 0^+} (t - K(t))',$$

$$(1.5) \quad \lambda_U = 2 - 2 \lim_{t \rightarrow 1^-} (t - K(t))'.$$



Some well-known Archimedean copula functions were proposed by Clayton [4], Frank [11], and Gumbel [17]. The generator functions of  $\varphi(t)$  and Kendall distribution functions  $K(t)$  of these copulas are summarized in Table 1. And also, Kendall's Tau ( $\tau$ ), Lower  $\lambda_L$  and Upper  $\lambda_U$  tail dependence coefficients for Gumbel, Clayton, and Frank copula are listed in Table 2.

**Table 1:** Archimedean Copulas with Generator functions  $\varphi(t)$ .

Copula	$\varphi(t)$	$K(t)$	Range of $\theta$
Clayton	$\frac{t^{-\theta} - 1}{\theta}$	$t + \frac{t(1 - t^\theta)}{\theta}$	$(-1, \infty) - \{0\}$
Frank	$-\log\left(\frac{\exp(t\theta) - 1}{\exp(\theta) - 1}\right)$	$t - \frac{(\exp(t\theta) - 1) \log\left(\frac{\exp(-t\theta) - 1}{\exp(-\theta) - 1}\right)}{\theta}$	$(-\infty, \infty) - \{0\}$
Gumbel	$(-\log(t))^\theta$	$t - \frac{t \log(t)}{\theta}$	$[1, \infty)$
Independence	$-\log(t)$	$t - t \log(t)$	—

**Table 2:** Kendall's Tau ( $\tau$ ), Lower  $\lambda_L$  and Upper  $\lambda_U$  tail dependence for some Archimedean copulas.

Copula	$\tau(\theta)$	$\lambda_L$	$\lambda_U$
Clayton	$\frac{\theta}{\theta + 2}$	$2^{-\frac{1}{\theta}}$	0
Frank	$1 + 4\theta^{-1}(D_1^*(\theta) - 1)$	0	0
Gumbel	$\frac{\theta - 1}{\theta}$	0	$2 - 2^{\frac{1}{\theta}}$

$$* D_1(x) = x^{-1} \int_0^x t (\exp(t) - 1)^{-1} dt$$

Modern risk management is mainly interested in assessing the amount of Kendall's tau and tail dependence. For this reason, many minimum-variance portfolio models are based on correlation. However, correlation itself is not enough to describe a tail dependence structure and often results in misleading interpretations (Embrechts *et al.* [7]). The importance of this issue has led to some improvements in the estimation of the dependence coefficients. Kollo *et al.* [18] examined tail behavior of skew  $t$ -copula considering the bivariate case. They used the method of moments and the maximum likelihood for the estimation of the tail dependence coefficients. Ferreira [10] proposed a nonparametric estimator of the tail dependence coefficient and proved its strong consistency and asymptotic normality in the case of known marginal distribution functions. Schmidt *et al.* [21] proposed a set of nonparametric estimators for the upper and lower tail copula and established results of weak convergence and strong consistency for the tail-copula estimators. Ferreira *et al.* [9] introduced the  $s, k$ -extremal coefficients for studying the tail dependence between the  $s$ -th lower and  $k$ -th upper order statistics of a normalized random vector. Caillaud *et al.* [3] introduced nonparametric estimators for upper and lower tail dependence whose confidence intervals are obtained with the bootstrap method as they called these estimators "Naive estimators".

Goegebeur *et al.* [16] introduced a class of weighted functional estimators for the coefficient of the tail dependence in bivariate extreme value statistics while they also derived the minimum variance asymptotically unbiased estimator.

In this paper, plug-in estimations of Kendall's tau, upper tail dependence and lower tail dependence are introduced. To the author's best knowledge, this is the first study examining the estimation of the dependence coefficients using the plug-in method. The use of Bernstein–Bézier polynomials reduced the complexity of the non-parametric estimation of the tail dependence coefficients. Besides, the proposed estimation method of the dependence coefficient is flexible depending on its polynomial degree while the error of the estimation can be reduced by increasing or decreasing the degree of the polynomial.

The remainder of the study is organized as follows. In Section 2, the estimation of Kendall distribution function based on Bernstein polynomials is discussed. In Section 3, Kendall's tau and tail dependence coefficients are estimated by the plug-in principle. The performance of the new estimation methods for the dependence coefficients is investigated in Section 4. In Section 5, the new estimator of Kendall's tau and tail dependence coefficients are applied to three real data sets. Finally, the conclusion is presented in Section 6.

---

## 2. ESTIMATION OF THE KENDALL DISTRIBUTION FUNCTION

---

Before introducing the estimation of the dependence coefficients for Archimedean copulas, it is important to investigate the estimation of Kendall distribution function since the dependence coefficients of Archimedean copula are closely related to the Kendall distribution function as stated in the last section. First time in the literature, Genest *et al.* [15] investigated the empirical estimate of Kendall distribution function. For the estimation of the random variable of  $T = H(x, y)$ , univariate distribution function of  $K(t) = P(H(x, y) \leq t) = P(C(u, v) \leq t)$  should be estimated within the interval of  $[0, 1]$ . This estimation process can be accomplished by two steps:

1. Constructing the empirical bivariate distribution function of  $H_n(X, Y)$ ;
2. Obtaining the pseudo observations of  $\hat{T}_i$  by

$$\hat{T}_i = \sum_{j=1}^n \mathbf{I}(X_i < X_j, Y_i < Y_j) / (n - 1), \quad i = 1, \dots, n.$$

By using these pseudo observations,  $K(t)$  is estimated by the empirical distribution function as

$$K_n(t) = \sum_{i=1}^n \mathbf{I}(\hat{T}_i \leq t) / n.$$

Genest *et al.* [15] stated that the empirical estimation of Kendall distribution function is  $\sqrt{n}$ -consistent estimator while Barbe *et al.* [1] proved consistency of this estimator.

Generally, the classical empirical distribution function has a good performance as an estimator of the distribution function. However, estimating continuous distribution function may not be appropriate (Susam *et al.* [22, 23], Erdoğan *et al.* [8]) since it has discontinuities.

Because of this, Susam *et al.* [22] proposed a smooth estimate of Kendall distribution function  $K_{n,m}$  given by the following equation:

$$K_{n,m}(t) = \sum_{k=0}^m K_n\left(\frac{k}{m}\right) P_{k,m}(t), \quad t \in [0, 1],$$

where  $P_{k,m}(t) = \binom{m}{k} t^k (1-t)^{m-k}$  is the Binomial probability. Susam *et al.* [23] proposed the Bézier curve based estimation of Kendall distribution function of  $K_{\alpha,m}$  which has lower mean integrated squared error (MISE) scores than  $K_{n,m}(t)$ . They defined  $K_{\alpha,m}$  as it is based on a set of the control points of  $\alpha_i, i = 0, \dots, m$ , as given by the following equation:

$$K_{\alpha,m}(t) = \sum_{k=0}^m \alpha_k P_{k,m}(t), \quad t \in [0, 1].$$

Also, they state that if the following constraints defined on the control points of  $\alpha_i (i = 1, \dots, m)$  hold, then the Bézier curve based on the estimation of Kendall distribution function of  $K_{\alpha,m}$  satisfies all the properties of the Kendall distribution function.

**Theorem 2.1** (Susam *et al.* [23]). *The estimator  $K_{\alpha,m}(t)$  satisfies properties of Kendall distribution function under the following constraints hold:*

1.  $\alpha_0 = 0 < \alpha_1 < \alpha_2 < \dots < \alpha_m = 1$ ;
2.  $\alpha_i > \frac{i}{m}, i = 1, \dots, m - 1$ .

They used minimum quadratic distance estimator which is based on the empirical Kendall distribution for estimating the control points of  $\alpha_i (i = 1, \dots, m - 1)$ . Also, Susam *et al.* [24] proposed minimum distance estimator for  $K_{\alpha,m}(t)$  based on Bernstein estimate of Kendall distribution function  $K_{n,m}(t)$ . They stated that the minimum distance method based on Kendall distribution using Bernstein polynomials outperforms the method based on empirical Kendall distribution.

### 3. ESTIMATION OF DEPENDENCE COEFFICIENTS BASED ON BÉZIER CURVE ESTIMATION OF KENDALL DISTRIBUTION FUNCTION

It is possible to estimate Kendall's tau, lower and upper tail dependence by replacing  $K(t)$  with its non-parametric estimation provided in Equations (1.3), (1.4) and (1.5). For a given bivariate random sample of size  $n, (X_1, Y_1), \dots, (X_n, Y_n)$  from  $X$  and  $Y$ , plug-in estimation of Kendall's tau, lower and upper tail dependence for Archimedean copula could be derived from the following equations:

$$(3.1) \quad \hat{\tau} = 3 - 4 \int_0^1 K_{\alpha,m}(t) dt,$$

$$(3.2) \quad \hat{\lambda}_L = 2 \lim_{t \rightarrow 0^+} (t - K_{\alpha,m}(t))',$$

$$(3.3) \quad \hat{\lambda}_U = 2 - 2 \lim_{t \rightarrow 1^-} (t - K_{\alpha,m}(t))',$$

where  $K_{\alpha,m}(t)$  is the estimation of Kendall distribution function based on the Bézier curve introduced in Section 2. Then, the next lemmas are provided for the estimation of Kendall’s tau, lower and upper tail dependence for Archimedean copulas.

**Lemma 3.1.** *Let  $K_{\alpha,m}(\cdot)$  be the estimator of Kendall distribution function based on the Bézier curve while  $\hat{\alpha}_k$  ( $k = 1, \dots, m - 1$ ) estimates the control points. The estimator of Kendall’s tau for Archimedean copula is obtained by*

$$\hat{\tau} = 3 - 4 \sum_{k=0}^m \hat{\alpha}_k \binom{m}{k} \beta(k + 1, m - k + 1),$$

where  $\beta(\cdot, \cdot)$  is the beta function defined as  $\beta(v_1, v_2) = \int_0^1 t^{v_1-1} (1 - t)^{v_2-1} dt$  for  $v_1$  and  $v_2$  positive integers.

**Lemma 3.2.** *Let  $K_{\alpha,m}(\cdot)$  be the estimator of Kendall distribution function based on the Bézier curve while  $\hat{\alpha}_k$  ( $k = 1, \dots, m - 1$ ) estimates the control points. The estimation of the lower tail and the upper tail dependence for the Archimedean copula is obtained by*

$$\begin{aligned} \hat{\lambda}_L &= 2^{1-m\hat{\alpha}_1}, \\ \hat{\lambda}_U &= 2 - 2^{1-m(1-\hat{\alpha}_{m-1})}. \end{aligned}$$

**Proof:** First order derivative of Bézier curve is derived by

$$K'_{\alpha,m}(t) = m \sum_{k=0}^{m-1} (\alpha_{k+1} - \alpha_k) P_{k,m-1}.$$

From the end-point rule of the Bézier curve,  $\lim_{t \rightarrow 0^+} K'_{\alpha,m}(t)$  and  $\lim_{t \rightarrow 1^-} K'_{\alpha,m}(t)$  are equal to  $m(\alpha_1 - \alpha_0)$  and  $m(\alpha_m - \alpha_{m-1})$  respectively (see Duncan [6]). Because of  $\alpha_0 = 0$  and  $\alpha_m = 1$ , then the desired results are obtained. □

It is observed that  $\hat{\lambda}_L$  and  $\hat{\lambda}_U$  are affected by only the control points of  $\alpha_1$  and  $\alpha_{m-1}$ , respectively. The range of the dependence coefficients depending on the polynomial degree  $m$  is summarized in Table 3. The results show that the range of dependence coefficients gets wider as the degree of the polynomial increases.

**Table 3:** Interval of Kendall’s Tau ( $\tau$ ), Lower  $\lambda_L$  and Upper  $\lambda_U$  tail dependence for varying polynomial degrees of  $m$ .

Degree ( $m$ )	$\tau$	$\lambda_U$	$\lambda_L$
5	$[-0.33, 1]$	$[0, 1]$	$[0.0625, 1]$
10	$[-0.64, 1]$	$[0, 1]$	$[0.0019, 1]$
15	$[-0.75, 1]$	$[0, 1]$	$[6.1 \times 10^{-5}, 1]$
20	$[-0.81, 1]$	$[0, 1]$	$[1.9 \times 10^{-6}, 1]$

For estimating the control points of  $\alpha_i$  ( $i = 0, \dots, m - 1$ ), statistical programming language R is used. The package “nloptr” is quite handy for optimizing non-linear function. The Augmented Lagrangian algorithm (`auglag`) included in the package “nloptr” should be used. Since  $K_{\alpha,m}(\cdot)$  has a complex function for higher polynomial degree so that may cause trouble in optimization. In order to overcome such a problem, the number of maximum evaluation number (`maxeval`) is recommended to be selected as at least 50.000 in the optimization.

---

#### 4. MONTE CARLO SIMULATION

---

To determine the performance of the estimation of  $\tau$ ,  $\lambda_U$ , and  $\lambda_L$ , the Monte Carlo simulation is conducted. 1.000 Monte Carlo samples with  $n = 150$  size are generated from each type of Archimedean copulas. For instance, parameters of  $\theta = 1.11, 1.25, 1.44$  is used for Gumbel copula while parameters of  $\theta = 0.22, 0.50, 0.85$  is used for Clayton copula and  $\theta = 0.91, 1.86, 2.92$  is used for Frank copula. Each copula has different shapes and characteristics. Clayton copula exhibits strong left tail dependence. In contrast to Clayton, Gumbel has strong right tail dependence while Frank copula exhibits symmetric and weak tail dependence in both tails. Detailed information about these Archimedean copulas is provided in Nelsen [20]. In all estimation methods, the Bézier curve degrees are selected for  $m = 1, \dots, 20$ . The mean of the estimation of the dependence coefficients for  $\tau$ ,  $\lambda_U$ , and  $\lambda_L$  Archimedean copulas for the varying degrees of  $m = 5, 10, 15$  and  $20$  are summarized in Tables 4, 5, and 6.

The following results are obtained from Tables 4, 5, and 6:

- For the estimation of Kendall’s tau, the mean of the  $\tau$  estimates is closer to the true value for the polynomial degree of  $m = 5$  when the true copula belongs to Gumbel, Clayton, or Frank.
- When the true copula is Gumbel with  $\tau = 0.1, 0.2, 0.3$ , mean of the  $\lambda_U$  estimates is closer to true value for the polynomial degree of  $m = 10$  while the mean of the estimation of  $\lambda_L$  is closer to true value for the polynomial degree of  $m = 20$ .
- When the true copula is Clayton with  $\tau = 0.1, 0.2, 0.3$ , mean of the  $\lambda_U$  estimates is closer to true value for the polynomial degree of  $m = 20$  while the mean of the estimation of  $\lambda_L$  is closer to true value for the polynomial degree of  $m = 5$ .
- When the true copula is Frank with  $\tau = 0.1, 0.2, 0.3$ , while the mean of the  $\lambda_U$  estimates are closer to true value for the polynomial degree of  $m = 20$  while the mean of the  $\lambda_L$  estimates is closer to true value for the polynomial degree of  $m = 20$ .

The results obtained from Figures 1, 2, and 3 are:

- As the dependence level increases for Gumbel, Clayton, and Frank copula, the variance of the estimations of the  $\tau$ ,  $\lambda_U$ , and  $\lambda_L$  increases as well.
- When the true copula belongs to the Clayton family with  $\theta = 0.22, 0.50$  and  $0.85$ , the variance of  $\lambda_U$  estimation decreases as the degree of polynomial increases. On the contrary, the variance of  $\lambda_L$  increases as the degree of polynomial increases.

- When the true copula is Frank with  $\theta = 0.91, 1.86, 2.92$ , the variance of  $\lambda_U$  decreases as the degree of polynomial increases. On the other hand, the variance of  $\lambda_L$  does not change as the degree of polynomial increases.
- In all the estimations of dependence coefficients, the estimation of  $\tau$ ,  $\lambda_L$  and  $\lambda_U$  get closure to the real values as the polynomial degree increases.

**Table 4:** Mean of the estimation of  $\tau$  of Archimedean copulas.

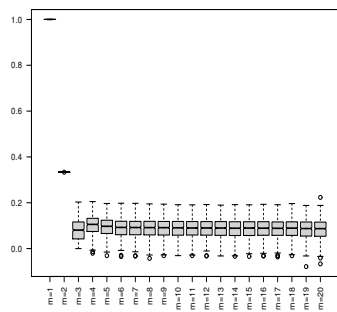
Copula	$\theta$	$\tau$	$\hat{\tau}^5$	$\hat{\tau}^{10}$	$\hat{\tau}^{15}$	$\hat{\tau}^{20}$
Gumbel	1.11	<b>0.099</b>	0.0956	0.0886	0.0876	0.0845
	1.25	<b>0.200</b>	0.1923	0.1883	0.1876	0.1858
	1.43	<b>0.300</b>	0.2909	0.2885	0.2880	0.2862
Clayton	0.22	<b>0.099</b>	0.0937	0.0896	0.0889	0.0871
	0.50	<b>0.200</b>	0.1922	0.1903	0.1898	0.1882
	0.85	<b>0.300</b>	0.2901	0.2884	0.2879	0.2869
Frank	0.91	<b>0.099</b>	0.0992	0.0897	0.0885	0.0861
	1.86	<b>0.200</b>	0.1972	0.1894	0.1879	0.1872
	2.92	<b>0.300</b>	0.2966	0.2897	0.2886	0.2875

**Table 5:** Mean of the estimation of  $\lambda_U$  of Archimedean copulas.

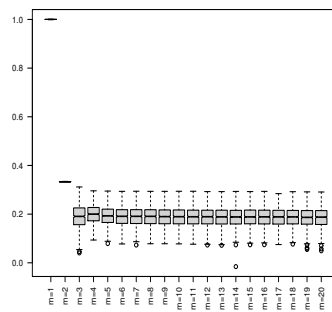
Copula	$\theta$	$\lambda_U$	$\hat{\lambda}_U^5$	$\hat{\lambda}_U^{10}$	$\hat{\lambda}_U^{15}$	$\hat{\lambda}_U^{20}$
Gumbel	1.11	<b>0.132</b>	0.0963	0.1384	0.1265	0.1326
	1.25	<b>0.258</b>	0.1862	0.2299	0.2080	0.2311
	1.43	<b>0.376</b>	0.2912	0.3207	0.2947	0.3418
Clayton	0.22	<b>0.000</b>	0.0243	0.0450	0.0346	0.0198
	0.50	<b>0.000</b>	0.0448	0.0454	0.0401	0.0251
	0.85	<b>0.000</b>	0.0563	0.0518	0.0447	0.0296
Frank	0.91	<b>0.000</b>	0.0164	0.0516	0.0395	0.0213
	1.86	<b>0.000</b>	0.0338	0.0564	0.0481	0.0344
	2.92	<b>0.000</b>	0.0476	0.0721	0.0620	0.0444

**Table 6:** Mean of the estimation of  $\lambda_L$  of Archimedean copulas.

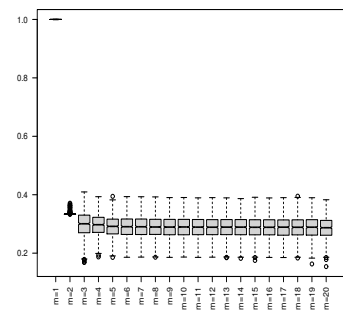
Copula	$\theta$	$\lambda_L$	$\hat{\lambda}_L^5$	$\hat{\lambda}_L^{10}$	$\hat{\lambda}_L^{15}$	$\hat{\lambda}_L^{20}$
Gumbel	1.11	<b>0</b>	0.1605	0.0876	0.0637	0.0580
	1.25	<b>0</b>	0.1932	0.1132	0.0855	0.0790
	1.43	<b>0</b>	0.2330	0.1490	0.1185	0.1082
Clayton	0.22	<b>0.04</b>	0.1964	0.1380	0.1161	0.1140
	0.50	<b>0.25</b>	0.2985	0.2558	0.2302	0.2405
	0.85	<b>0.44</b>	0.4240	0.3943	0.3744	0.4010
Frank	0.91	<b>0</b>	0.1694	0.0901	0.0651	0.0578
	1.86	<b>0</b>	0.2107	0.1168	0.0852	0.0737
	2.92	<b>0</b>	0.2565	0.1484	0.1101	0.0949



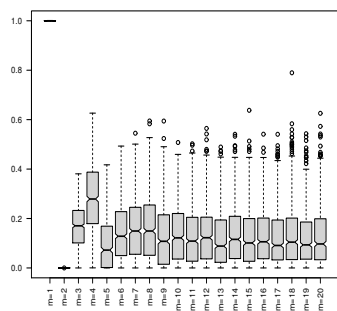
(a)  $\tau$  estimation for  $\theta = 1.11$  ( $\tau = 0.1$ ).



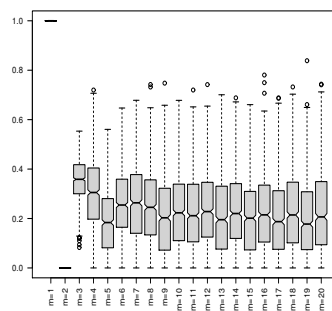
(b)  $\tau$  estimation for  $\theta = 1.25$  ( $\tau = 0.2$ ).



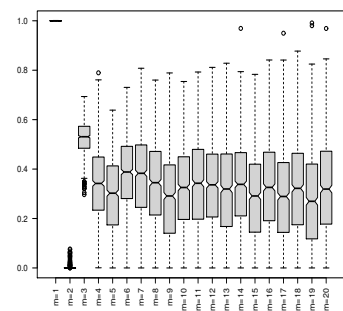
(c)  $\tau$  estimation for  $\theta = 1.43$  ( $\tau = 0.3$ ).



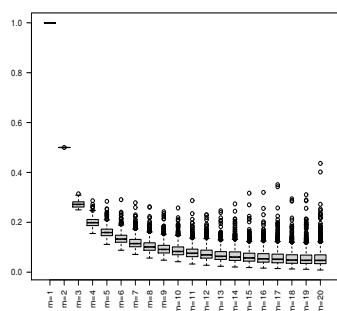
(d)  $\lambda_U$  estimation for  $\theta = 1.11$  ( $\lambda_U = 0.13$ ).



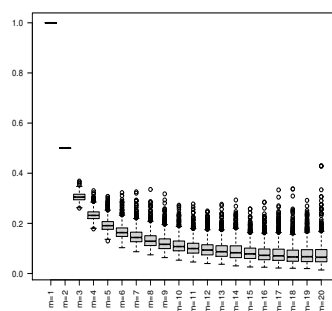
(e)  $\lambda_U$  estimation for  $\theta = 1.42$  ( $\lambda_U = 0.25$ ).



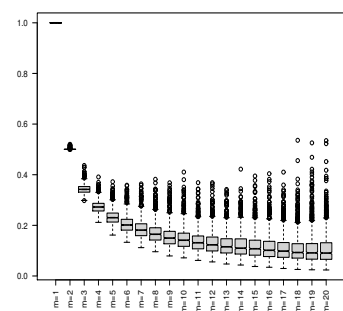
(f)  $\lambda_U$  estimation for  $\theta = 2$  ( $\lambda_U = 0.37$ ).



(g)  $\lambda_L$  estimation for  $\theta = 1.11$  ( $\lambda_L = 0$ ).

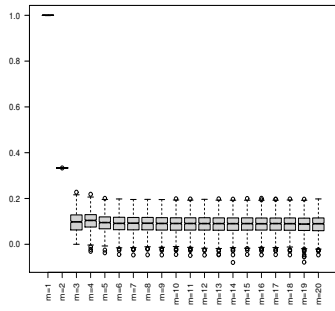


(h)  $\lambda_L$  estimation for  $\theta = 1.42$  ( $\lambda_L = 0$ ).

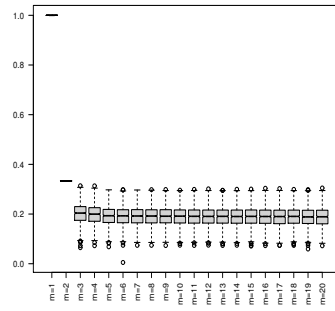


(i)  $\lambda_L$  estimation for  $\theta = 2$  ( $\lambda_L = 0$ ).

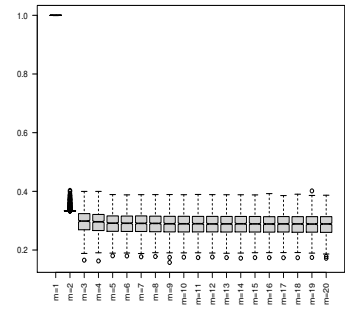
**Figure 1:** Box-plots of the estimation of the dependence coefficients of Gumbel copula with parameters of  $\theta = 1.11, 1.25, 1.43$ .



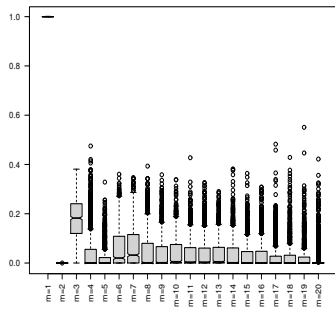
(a)  $\tau$  estimation for  $\theta = 0.22$  ( $\tau = 0.1$ ).



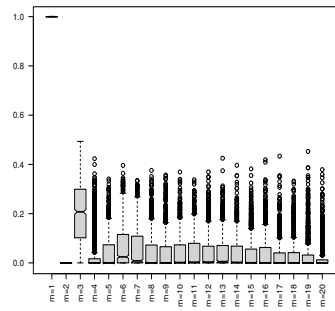
(b)  $\tau$  estimation for  $\theta = 0.50$  ( $\tau = 0.2$ ).



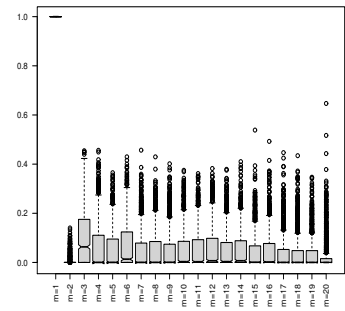
(c)  $\tau$  estimation for  $\theta = 0.85$  ( $\tau = 0.3$ ).



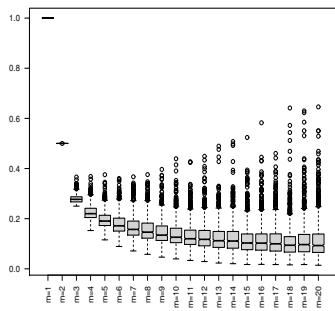
(d)  $\lambda_U$  estimation for  $\theta = 0.22$  ( $\lambda_U = 0$ ).



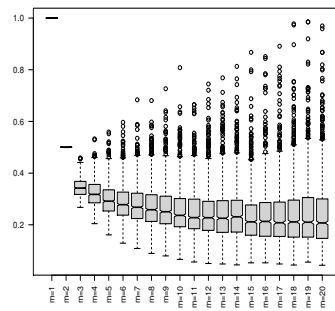
(e)  $\lambda_U$  estimation for  $\theta = 0.50$  ( $\lambda_U = 0$ ).



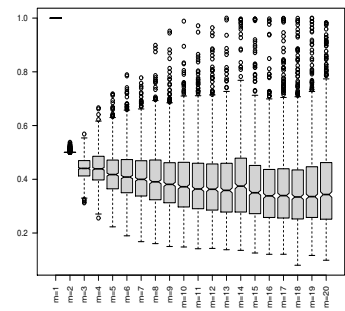
(f)  $\lambda_U$  estimation for  $\theta = 0.85$  ( $\lambda_U = 0$ ).



(g)  $\lambda_L$  estimation for  $\theta = 0.22$  ( $\lambda_L = 0.04$ ).



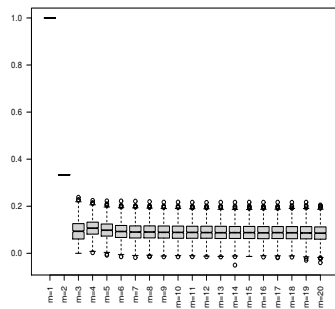
(h)  $\lambda_L$  estimation for  $\theta = 0.50$  ( $\lambda_L = 0.25$ ).



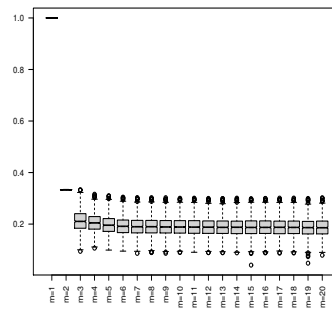
(i)  $\lambda_L$  estimation for  $\theta = 0.85$  ( $\lambda_L = 0.44$ ).

**Figure 2:** Box-plots of the estimation of the dependence coefficients Clayton copula with parameters of  $\theta = 0.22, 0.50, 0.85$ .

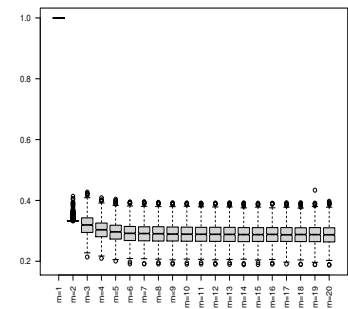




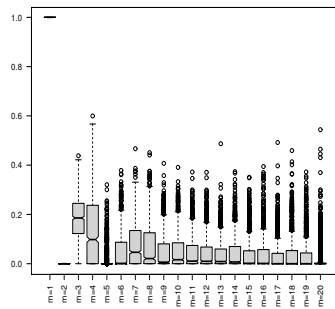
(a)  $\tau$  estimation for  $\theta = 0.91$  ( $\tau = 0.1$ ).



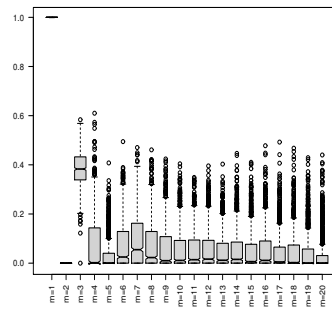
(b)  $\tau$  estimation for  $\theta = 1.86$  ( $\tau = 0.2$ ).



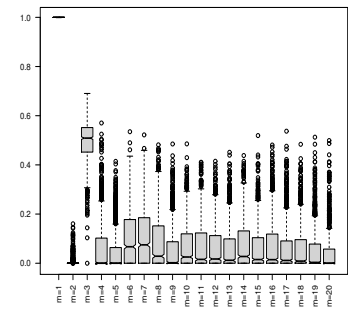
(c)  $\tau$  estimation for  $\theta = 2.92$  ( $\tau = 0.3$ ).



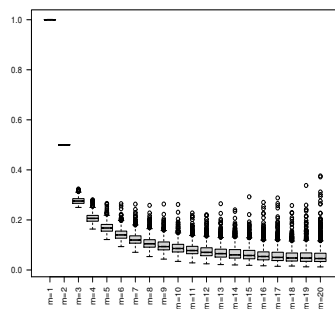
(d)  $\lambda_U$  estimation for  $\theta = 0.91$  ( $\lambda_U = 0$ ).



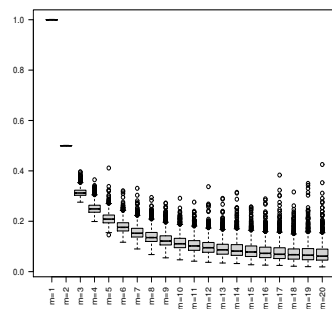
(e)  $\lambda_U$  estimation for  $\theta = 1.86$  ( $\lambda_U = 0$ ).



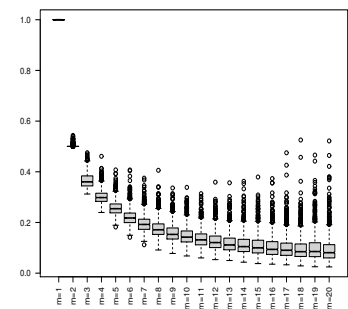
(f)  $\lambda_U$  estimation for  $\theta = 2.92$  ( $\lambda_U = 0$ ).



(g)  $\lambda_L$  estimation for  $\theta = 0.91$  ( $\lambda_L = 0$ ).



(h)  $\lambda_L$  estimation for  $\theta = 1.86$  ( $\lambda_L = 0$ ).



(i)  $\lambda_L$  estimation for  $\theta = 2.92$  ( $\lambda_L = 0$ ).

**Figure 3:** Box-plots of the estimation of the dependence coefficients Frank copula with parameters of  $\theta = 0.91, 1.86, 2.92$ .

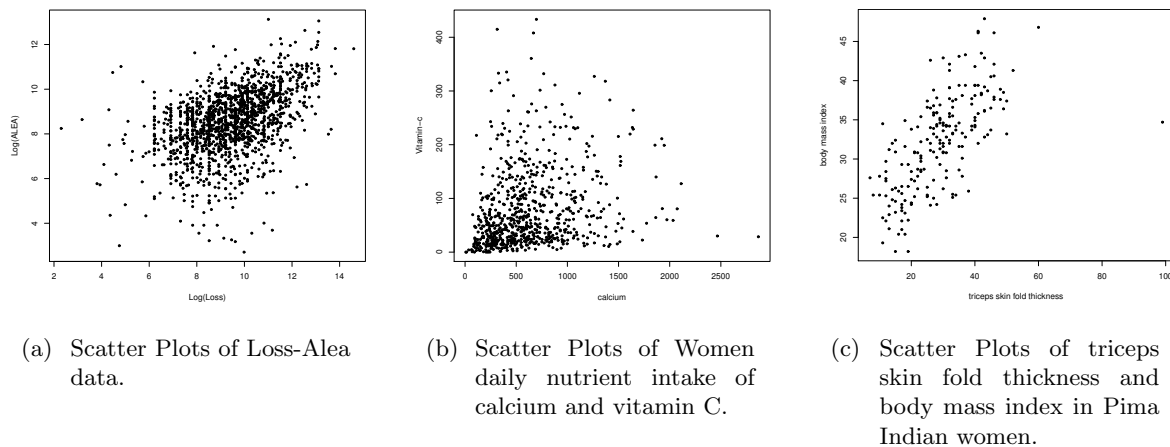
---

## 5. APPLICATIONS

---

To demonstrate the performance of new dependence coefficients estimation in previous sections, the Gumbel, Clayton and Frank copula is fit to the following three real data sets:

- The first data set is comprised of 1500 general liability claims randomly chosen from late settlement lags (Frees *et al.* [12]) and was provided by Insurance Services Office, Inc. Each claim consists of an indemnity payment (the loss) and an allocated loss adjustment expense (ALAE). The data is available in the R package “copula”. For simplicity, 34 censored data have not been used.
- According to the manual of R’s package “lcopula”, the nutrient data frame consists of quintuples consisting of four-day measurements for the intake of calcium, iron, protein, vitamin A and C for the women aged from 25 to 50 in the United States as part of the “Continuing Survey of Food Intakes of Individuals” program. The processed data has 737 measurements from a cohort study of the United States Department of Agriculture (USDA) and is available online at the University of Pennsylvania repository. The main concern is to estimate the dependence coefficients of Women’s daily nutrient intake of calcium and vitamin C.
- A population of women who were at least 21 years old, of Pima Indian heritage, and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria by using R’s package of “MASS”. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. The training set “Pima.tr” contains a randomly selected set of 200 subjects. An application is illustrated for determining dependence coefficients of Triceps skinfold thickness and body mass index in Pima Indian women.



**Figure 4:** Scatter plots of real data sets.

Figure 4 shows the scatter plots of the three data sets. When Figure 4 is examined, the dependence structure between involved random variables is obvious. In order to assess the goodness-of-fit results, the Cramér von Mises (*CvM*) statistic is used:

$$(5.1) \quad CvM = n \int_0^1 \left( \widehat{K}_n(t) - K_{\widehat{\theta}}(t) \right)^2 dK_{\widehat{\theta}}(t),$$

where  $\widehat{K}_n$  is the empirical Kendall distribution function as a non-parametric estimator of  $K(t)$ . The dependence parameter  $\theta$  is estimated by means of the Pseudo-likelihood method. The statistic is evaluated by the relevant  $p$ -value obtained by running 10.000 Monte Carlo samples as the method is described in Berg [2] and Genest *et al.* [14]. All goodness-of-fit results and parametric estimation of dependence coefficients are presented in Table 7 while Table 8 provides the estimation results of  $\tau$ ,  $\lambda_L$  and  $\lambda_U$  based on Bézier curve for three data sets.

**Table 7:** Goodness-of-fit results based on  $K(t)$  for three reel data sets.

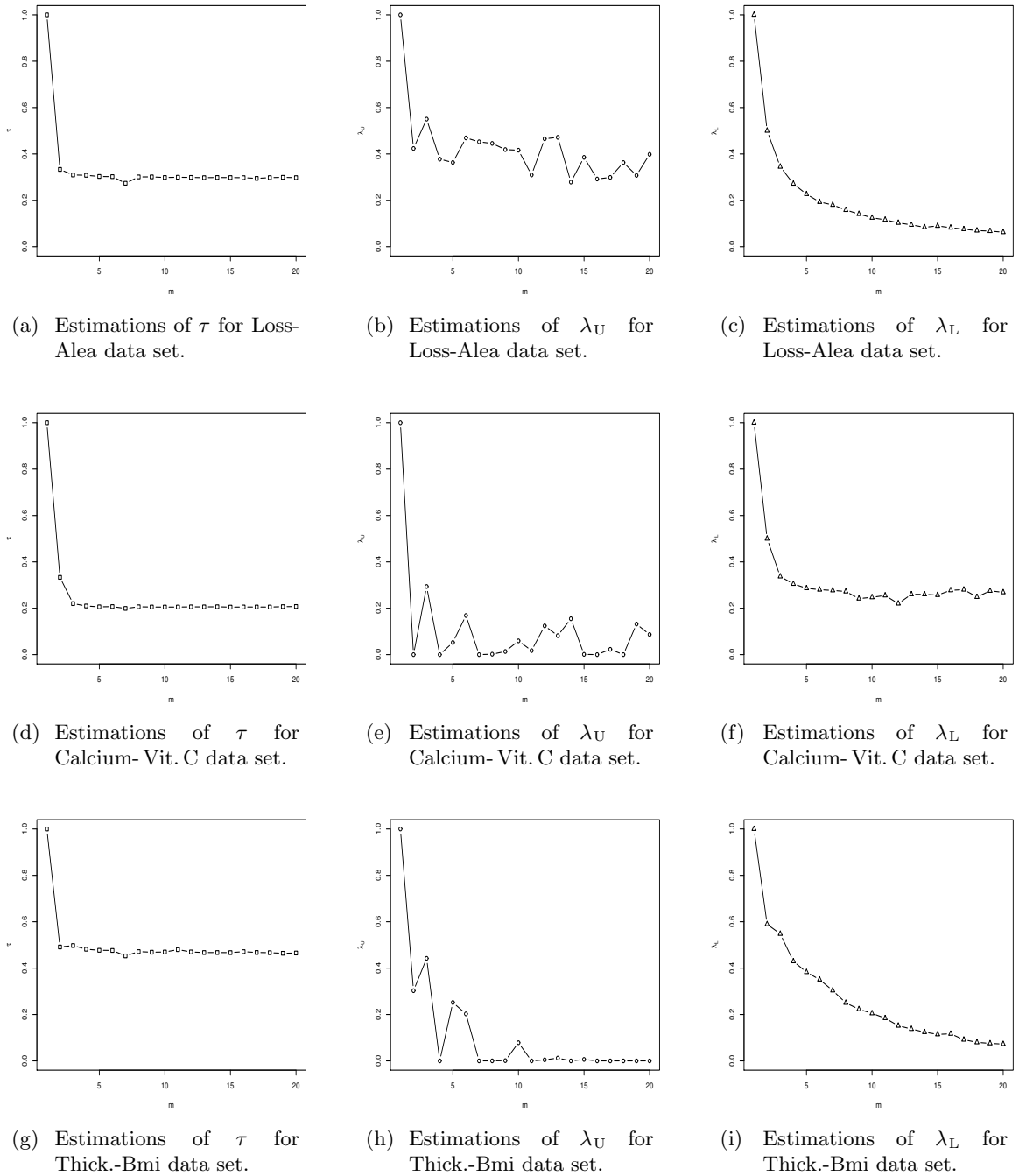
Data	Copula	Parameter	$\widehat{\tau}$	$\widehat{\lambda}_L$	$\widehat{\lambda}_U$	<i>CvM</i>	$p$ -value
Loss-Alea	Gumbel	1.4607	0.3154	0	0.3927	0.0414	0.8291
	Frank	3.0942	0.3154	0	0	0.2293	0.0292
	Clayton	0.9214	0.3154	0.4713	0	1.4181	0.0000
Calcium- Vit. C	Gumbel	1.2665	0.2104	0	0.2714	0.5627	0.0000
	Frank	1.9651	0.2104	0	0	0.3546	0.0011
	Clayton	0.5330	0.2104	0.2724	0	0.0505	0.6073
Thick.-Bmi	Gumbel	2.0933	0.5222	0	0.6074	0.1393	0.0221
	Frank	6.1568	0.5222	0	0	0.0711	0.2252
	Clayton	2.1866	0.5222	0.7283	0	0.2343	0.0014

**Table 8:** The estimation of  $\tau$ ,  $\lambda_U$  and  $\lambda_L$  for three reel data sets.

Data	Est. Meth.	$m = 5$	$m = 10$	$m = 15$	$m = 20$
Loss-Alea	$\widehat{\tau}^m$	0.3030	0.2984	0.2981	0.2979
	$\widehat{\lambda}_U^m$	0.3631	0.4161	0.3852	0.3982
	$\widehat{\lambda}_L^m$	0.2267	0.1248	0.0897	0.0626
Calcium- Vit. C	$\widehat{\tau}^m$	0.2061	0.2051	0.2049	0.2075
	$\widehat{\lambda}_U^m$	0.0523	0.0593	0.0116	0.0865
	$\widehat{\lambda}_L^m$	0.2863	0.2472	0.2568	0.2687
Thick.-Bmi	$\widehat{\tau}^m$	0.4769	0.4694	0.4668	0.4651
	$\widehat{\lambda}_U^m$	0.2517	0.0783	0.0062	0.0001
	$\widehat{\lambda}_L^m$	0.3826	0.2049	0.1143	0.0724

The results in Table 7 represent that Gumbel copula is a good choice for variables Loss-Alea with a  $p$ -value of 0.8291. It is concluded from Table 8 that as the degree of polynomial increases, estimation of  $\lambda_U$  and  $\lambda_L$  approach to the parametric estimate of dependence coefficients of Gumbel copula for Insurance data. For Calcium and Vitamin-C data, Clayton copula fits the data well with  $p$ -value of 0.6073. For the estimation of  $\lambda_L$ ,  $\widehat{\lambda}_L^{20}$  is closure to the parametric estimate of  $\lambda_L = 0.2714$ . Also, it is obtained that the estimation of  $\lambda_U$  approaches

to the parametric estimate of  $\lambda_U = 0$  as polynomial degree increases. For the triceps skinfold thickness and body mass index in Pima Indian women, Frank copula provides the best fit with  $p$ -value of 0.2252 from a statistical point of view. Tables 7 and 8 indicate that the estimation of  $\lambda_U$  and  $\lambda_L$  approaches to the parametric estimate of  $\lambda_U = 0$  and  $\lambda_L = 0$ . In addition, Figure 5 shows the estimations of dependence coefficients for three real data sets depending on the polynomial degree  $m = 1, \dots, 20$ . It can be concluded that, as the polynomial degree increases the estimation of dependence coefficients gets closure to the real values.



**Figure 5:** Estimations of dependence coefficients of data sets for degree  $m = 1, 2, \dots, 20$ .

---

## 6. CONCLUSION

---

In this study, a method of estimating the dependence coefficients of bivariate Archimedean family of copula is proposed. The Kendall's tau, lower tail dependence and upper tail dependence are estimated by using the Bézier curve. The new estimator of the dependence coefficients are flexible by the polynomial degree of  $m$ . A Monte Carlo simulation study is performed to measure the performance of the proposed estimation method for  $\tau$ ,  $\lambda_U$ , and  $\lambda_L$ . The performance according to the different levels of dependence size is investigated as well. The simulation results show that the new estimator of  $\tau$ ,  $\lambda_U$ , and  $\lambda_L$  presented a good performance. Besides, the new estimators of  $\tau$ ,  $\lambda_U$ , and  $\lambda_L$  indicated a satisfactory performance for the three data sets examined.

---

## REFERENCES

---

- [1] BARBE, P.; GENEST, C.; GHOUDI, K. and RÉMILLARD, B. (1996). On Kendall's process, *Journal of Multivariate Analysis*, **58**, 197–229.
- [2] BERG, D. (2009). Copula goodness-of-fit testing: an overview and power comparison, *The European Journal of Finance*, **15**, 675–701.
- [3] CAILLAULTY, C. and GUÉGAN, D. (2007). Empirical Estimation of Tail Dependence Using Copulas. Application to Asian Markets, *Quantitative Finance*, **5**, 489–501.
- [4] CLAYTON, D.G. (1978). Model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika*, **65**, 141–151.
- [5] DIMITROVA, D.; KAISHEV, K. and PENEV, I. (2008). GeD spline estimation of multivariate Archimedean copulas, *Computational Statistics and Data Analysis*, **52**, 3570–3582.
- [6] DUNCAN, M. (2005). *Applied Geometry for Computer Graphics and CAD*, Springer-Verlag, London.
- [7] EMBRECHTS, P.; MCNEIL, A. and STRAUMANN, D. (2002). *Correlation and dependence in risk management: properties and pitfalls*. In “Risk Management: Value at Risk and Beyond?” (M.A.H. Dempster, Ed.), Cambridge, 176–223.
- [8] ERDOĞAN, M.S.; DIŞIBÜYÜK, Ç. and ORUÇ, Ö.E. (2019). An alternative distribution function estimation method using rational Bernstein polynomials, *Journal of Computational and Applied Mathematics*, **353**, 232–242.
- [9] FERREIRA, H. and FERREIRA, M. (2012). Tail dependence between order statistics, *Journal of Multivariate Analysis*, **105**(1), 176–192.
- [10] FERREIRA, M. (2013). Nonparametric estimation of the tail-dependence coefficient, *REVSTAT – Statistical Journal*, **11**(1), 1–16.
- [11] FRANK, M.J. (1979). On the simultaneous associativity of  $F(x, y)$  and  $x + y - F(x, y)$ , *Aequationes Mathematicae*, **19**, 194–226.
- [12] FREES, E. and VALDEZ, E. (1998). Understanding relationships using copulas, *North American Actuarial Journal*, **2**, 1–25.
- [13] GENEST, C. and MACKAY, R.J. (1986). Copules Archimédiennes et familles de lois bidimensionnelles dont les marges sont données, *The Canadian Journal of Statistics*, **14**, 145–149.

- [14] GENEST, C.; QUESSY, J.F. and RÉMILLARD, B. (2006). Goodness-of-fit procedures for copula models based on the probability integral transform, *Scandinavian Journal of Statistics*, **33**, 337–366.
- [15] GENEST, C. and RIVEST, L. (1993). Statistical inference procedures for bivariate Archimedean copulas, *Journal of the American Statistical Association*, **88**(423), 1034–1043.
- [16] GOEGEBEUR, Y. and GUILLOU, A. (2012). Asymptotically unbiased estimation of the coefficient of tail dependence, *Scandinavian Journal of Statistics*, **40**(1), 174–189.
- [17] GUMBEL, E.J. (1960). Distributions des valeurs extrêmes en plusieurs dimensions, *Publications de l'Institut de Statistique de l'Université de Paris*, **19**(9), 171–173.
- [18] KOLLO, T.; PETTERE, G. and VALGE, M. (2017). Tail dependence of skew  $t$ -copulas, *Communications in Statistics – Simulation and Computation*, **46**(2), 1024–1034.
- [19] MICHIELS, F.; KOCH, I. and DE SCHEPPER, A. (2011). A new method for the construction of bivariate Archimedean copulas based on the  $\lambda$  function, *Communications in Statistics – Theory and Methods*, **40**(15), 2670–2679.
- [20] NELSEN, R.B. (2006). *An Introduction to Copulas*, Springer-Verlag, New York (NY).
- [21] SCHMIDT, R. and STADTMÜLLER, U. (2006). Nonparametric estimation of tail dependence, *Scandinavian Journal of Statistics*, **33**, 307–335.
- [22] SUSAM, S.O. and UCER HUDAVERDI, B. (2018). Testing independence for Archimedean copula based on Bernstein estimate of Kendall distribution function, *Journal of Statistical Computation and Simulation*, **88**(13), 2589–2599.
- [23] SUSAM, S.O. and UCER HUDAVERDI, B. (2020). A goodness-of-fit test based on Bézier curve estimation of Kendall distribution, *Journal of Statistical Computation and Simulation*, **90**(13), 1194–1215.
- [24] SUSAM, S.O. and UCER HUDAVERDI, B. (2022). On construction of Bernstein–Bézier type bivariate Archimedean copula, *REVSTAT – Statistical Journal*, **20**(3), 337–351.

---

---

## Estimation in Weibull Distribution Under Progressively Type-I Hybrid Censored Data

---

---

Authors: YASIN ASAR  

– Department of Mathematics and Computer Sciences, Necmettin Erbakan University,  
Turkey  
yasinasar@hotmail.com , yasar@erbakan.edu.tr

REZA ARABI BELAGHI 

– Department of Statistics, University of Tabriz,  
Iran  
rezaarabi11@gmail.com

Received: October 2019

Revised: January 2021

Accepted: January 2021

Abstract:

- In this article, we consider the estimation of unknown parameters of Weibull distribution when the lifetime data are observed in the presence of progressively type-I hybrid censoring scheme. The Newton–Raphson algorithm, Expectation–Maximization (EM) algorithm and Stochastic EM algorithm are utilized to derive the maximum likelihood estimates for the unknown parameters. Moreover, Bayesian estimators using Tierney–Kadane Method and Markov Chain Monte Carlo method are obtained under three different loss functions, namely, squared error loss, linear-exponential and generalized entropy loss functions. Also, the shrinkage pre-test estimators are derived. An extensive Monte Carlo simulation experiment is conducted under different schemes so that the performances of the listed estimators are compared using mean squared error, confidence interval length and coverage probabilities. Asymptotic normality and MCMC samples are used to obtain the confidence intervals and highest posterior density intervals respectively. Further, a real data example is presented to illustrate the methods. Finally, some conclusive remarks are presented.

Keywords:

- *Bayesian estimation; EM algorithm; SEM algorithm; Tierney–Kadane’s approximation; progressively type-I hybrid censoring; Weibull distribution.*

AMS Subject Classification:

- 62F10, 62N01, 62N05.

---

## 1. INTRODUCTION

---

Censored data occurs commonly in reliability and survival analysis. There are mainly two censoring schemes which are type-I censoring where the life-testing experiment stops at a predetermined time, say  $T$  and type-II censoring, where the life-testing experiment stops when predetermined number of failures, say  $m$ , are observed. Epstein [19] proposed the hybrid censoring scheme which is the mixture of type-I and type-II censoring schemes. The hybrid censoring scheme has become quite popular in the reliability and life-testing experiments so far. For example, see the papers of Chen and Bhattacharya [13], Childs *et al.* [15], Kundu and Joarder [26], Balakrishnan and Kundu [10]. It is worth mentioning that the book of Balakrishnan and Cramer [8] discussed the topics of progressive censoring and progressive hybrid censoring in detail as separate chapters. In these schemes, it is allowed to remove the units only at the terminal points of the experiments. However, Kundu and Joarder [26] introduced another scheme which is called the type-I progressively hybrid censoring scheme (type-I PHCS) such that it allows removals of units during the test time. For more information on progressive censoring, we refer to Balakrishnan and Aggarwala [7], Balakrishnan [6] and Balakrishnan and Cramer [8]. Type-I PHCS can be viewed as a mixture of type-I progressive censoring and hybrid censoring as follows: Assume that there are  $n$  identical units in a lifetime experiment with the progressive censoring scheme  $(R_1, R_2, \dots, R_m)$ ,  $1 \leq m \leq n$  and the lifetime experiment ends at a predetermined time  $T \in (0, \infty)$  and  $n, m, R_i$ 's are all fixed non-negative integers. At the time of first failure, say  $X_{1:m:n}$ ,  $R_1$  units randomly removed from the remaining  $n - 1$  units. Similarly, when the second failure occurs at the time  $X_{2:m:n}$ ,  $R_2$  units are removed from the remaining  $n - R_1 - 2$  units. This process continues up to the end of experiment which occurs at the time  $\min(X_{m:m:n}, T)$ . Therefore, if the  $m$ -th failure occurs before time  $T$ , the experiment ends at the time  $X_{m:m:n}$  and all the remaining units  $R_m = n - \sum_{i=1}^{m-1} R_i - m$  are removed. However, if the experiment ends at time  $T$  with only  $J$  failures,  $0 \leq J < m$ , then all the remaining units  $R_J^* = n - \sum_{i=1}^J R_i - J$  are removed and the test ends at time  $T$ . Therefore, under type-I PHCS we have the following two cases:

- Case I:  $\{X_{1:m:n}, X_{2:m:n}, \dots, X_{m:m:n}\}$  if  $X_{m:m:n} \leq T$ .
- Case II:  $\{X_{1:m:n}, X_{2:m:n}, \dots, X_{J:m:n}\}$  if  $X_{J:m:n} < T < X_{J+1:m:n}$ .

Due to the fact that the lifetime distributions of many experimental units can be modeled by a two-parameter Weibull distribution which is one of the most commonly used model in reliability and lifetime data analysis, we consider the Weibull distribution in this paper. The probability distribution function (PDF) and cumulative distribution function (CDF) of two parameter Weibull distribution are given as follows:

$$(1.1) \quad f(x; \alpha, \beta) = \alpha \beta x^{\alpha-1} \exp\{-\beta x^\alpha\},$$

$$(1.2) \quad F(x; \alpha, \beta) = 1 - \exp\{-\beta x^\alpha\},$$

where  $\alpha > 0$  is the shape parameter and  $\beta > 0$  is the scale parameter.

Ng *et al.* [34] used the estimation method, along with Fisher information matrix, in the context of optimal progressive censoring schemes for the Weibull distribution. Banerjee and Kundu [12] considered the statistical inference on Weibull parameters when the data are type-II hybrid censored, maximum likelihood estimation (MLE), approximate MLE and Bayes



estimation techniques were studied by the authors. Balakrishnan and Kateri [9] proposed an alternative approach based on a graphical method, which also shows the existence and uniqueness of the MLEs. Lin *et al.* [30] studied the MLEs and the approximate MLEs (AMLEs) of the parameters of Weibull distribution under adaptive type-II progressive hybrid censoring. Huang and Wu [20] discussed the maximum likelihood estimation and Bayesian estimation of Weibull parameters under progressively type-II censoring scheme. Lin *et al.* [28] investigated the maximum likelihood estimation and Bayesian estimation for a two-parameter Weibull distribution based on adaptive type-I progressively hybrid censored data which was introduced by Lin and Huang [29]. Jia *et al.* [21] studied the exact inference on Weibull parameters under multiple type-I censoring. Mokhtari *et al.* [32] discussed the approximate and Bayesian inferential procedures for the progressively type-II hybrid censored data from the Weibull distribution. However, this type of censoring is identical to what we called as type-I progressive hybrid censored data. This paper will be different from [32] in three directions. Firstly, we introduce a new approach for inference about the Weibull distribution based on expectation-maximization (EM) and stochastic expectation-maximization (SEM) methods. We will show that both EM and SEM will result to have better estimates in the sense of having smaller biases and mean square errors. Secondly, we will derive the shrinkage estimators based on the ML estimates resulting to have higher deficiencies. Finally, in the Bayesian approach, different loss functions such as squared error loss (SEL), linear-exponential (LINEX), and general entropy loss (GEL) will be applied with both informative and non-informative priors.

The rest of the paper is organized as follows: In Section 2, MLE of the parameters are introduced by using Newton–Raphson (NR) algorithm, EM algorithm and SEM algorithm, also the Fisher information matrix is obtained. In Section 3, Bayes estimation for the parameters of Weibull distribution under the assumption of independent priors using different loss functions such as SEL, LINEX and GEL loss functions. Moreover, Tierney and Kadane [44] (T–K) approximations under these loss functions are also computed and Markov-Chain Monte Carlo (MCMC) method is also presented to estimate the parameters. In Section 4, a shrinkage pre-test estimation method is discussed. Extensive Monte Carlo simulations are conducted and results are discussed in Section 5. A real data example is presented in Section 6 to illustrate the findings of the study. Finally, some conclusive remarks are given in Section 7.

---

## 2. MAXIMUM LIKELIHOOD ESTIMATION

---

Let  $\mathbf{X} = (X_{1:m:n}, \dots, X_{r:m:n})$  represents the type-I progressively hybrid censored sample of size  $r$  from a sample of size  $n$  drawn from a population with probability distribution given in Equation (1.1). Throughout this paper, we will denote  $X_{i:m:n}$  by  $X_{(i)}$ ,  $i = 1, 2, \dots, r$ . Then the likelihood function of  $(\alpha, \beta)$  given the observed data  $\mathbf{x}$  can be written as

$$(2.1) \quad L(\alpha, \beta | \mathbf{x}) \propto \prod_{i=1}^r f(x_{(i)}; \alpha, \beta) [1 - F(x_{(i)}; \alpha, \beta)]^{R_i} [1 - F(\mathcal{C}; \alpha, \beta)]^{R_T},$$

where  $r = m$ ,  $\mathcal{C} = x_{(m)}$ ,  $R_T = 0$  in Case I, and  $r = d$ ,  $\mathcal{C} = T$ ,  $R_T = n - d - \sum_{i=1}^d R_i$  in Case II.

Based on the observed data, the log-likelihood function can be expressed as

$$\begin{aligned}
 l(\alpha, \beta | \mathbf{x}) &= \ln L(\alpha, \beta | \mathbf{x}) \\
 (2.2) \quad &= r \ln(\alpha\beta) + (\alpha - 1) \sum_{i=1}^r \ln(x_{(i)}) - \beta \sum_{i=1}^r \left\{ x_{(i)}^\alpha (1 + R_i) \right\} - \beta C^\alpha R_T.
 \end{aligned}$$

Taking the derivatives of Equation (2.2) with respect to  $\alpha$  and  $\beta$  and equating them to zero, one can obtain the following likelihood equations for  $\alpha$  and  $\beta$  respectively:

$$(2.3) \quad \frac{\partial l(\alpha, \beta | \mathbf{x})}{\partial \alpha} = \frac{r}{\alpha} + \sum_{i=1}^r \ln(x_{(i)}) - \beta \sum_{i=1}^r \left\{ (1 + R_i) x_{(i)}^\alpha \ln(x_{(i)}) \right\} - \beta C^\alpha \ln(C) R_T = 0,$$

$$(2.4) \quad \frac{\partial l(\alpha, \beta | \mathbf{x})}{\partial \beta} = \frac{r}{\beta} - \sum_{i=1}^r \left\{ x_{(i)}^\alpha (1 + R_i) \right\} - C^\alpha R_T = 0.$$

Solving Equation (2.4) yields the MLE of  $\beta$  which is given by

$$(2.5) \quad \hat{\beta} = \frac{r}{C^{\hat{\alpha}} R_T + \sum_{i=1}^r \left\{ x_{(i)}^{\hat{\alpha}} (1 + R_i) \right\}}.$$

Now, substituting Equation (2.5) into (2.3), the MLE of  $\alpha$  can be obtained by solving the following nonlinear equation:

$$\frac{r}{\hat{\alpha}} + \frac{r \left[ \sum_{i=1}^r \left\{ (1 + R_i) x_{(i)}^{\hat{\alpha}} \ln(x_{(i)}) \right\} + R_T C^{\hat{\alpha}} \ln(C) \right]}{R_T C^{\hat{\alpha}} + \sum_{i=1}^r \left\{ x_{(i)}^{\hat{\alpha}} (1 + R_i) \right\}} = 0.$$

The second partial derivatives of the log-likelihood equation are obtained as follows:

$$(2.6) \quad \frac{\partial^2 l(\alpha, \beta | \mathbf{x})}{\partial \alpha^2} = -\frac{r}{\alpha^2} - \beta \sum_{i=1}^r \left\{ (1 + R_i) x_{(i)}^\alpha \ln(x_{(i)})^2 \right\} - \beta C^\alpha \ln(C)^2 R_T,$$

$$(2.7) \quad \frac{\partial^2 l(\alpha, \beta | \mathbf{x})}{\partial \alpha \partial \beta} = -\sum_{i=1}^r \left\{ (1 + R_i) x_{(i)}^\alpha \ln(x_{(i)}) \right\} - C^\alpha \ln(C) R_T,$$

$$(2.8) \quad \frac{\partial^2 l(\alpha, \beta | \mathbf{x})}{\partial \beta^2} = \frac{-r}{\beta^2}.$$

Now, using Equations (2.6)–(2.8), the Fisher’s information matrix  $\mathbf{I}(\alpha, \beta)$  can be formed by

$$(2.9) \quad \mathbf{I}(\alpha, \beta) = E \begin{bmatrix} -\frac{\partial^2 l(\alpha, \beta | \mathbf{x})}{\partial \alpha^2} & -\frac{\partial^2 l(\alpha, \beta | \mathbf{x})}{\partial \alpha \partial \beta} \\ -\frac{\partial^2 l(\alpha, \beta | \mathbf{x})}{\partial \alpha \partial \beta} & -\frac{\partial^2 l(\alpha, \beta | \mathbf{x})}{\partial \beta^2} \end{bmatrix}.$$

It is well-known that (see [27]) the distribution of MLEs  $(\hat{\alpha}, \hat{\beta})$  is a bivariate normal distribution with

$$N\left((\alpha, \beta), \mathbf{I}^{-1}(\alpha, \beta)\right),$$

where  $\mathbf{I}^{-1}(\alpha, \beta)$  is the covariance matrix. Moreover, one can approximate the covariance matrix evaluated at  $(\hat{\alpha}, \hat{\beta})$  by the following observed information matrix:

$$(2.10) \quad \mathbf{I}(\hat{\alpha}, \hat{\beta}) = \begin{bmatrix} -\frac{\partial^2 l(\alpha, \beta | \mathbf{x})}{\partial \alpha^2} & -\frac{\partial^2 l(\alpha, \beta | \mathbf{x})}{\partial \alpha \partial \beta} \\ -\frac{\partial^2 l(\alpha, \beta | \mathbf{x})}{\partial \alpha \partial \beta} & -\frac{\partial^2 l(\alpha, \beta | \mathbf{x})}{\partial \beta^2} \end{bmatrix}_{(\hat{\alpha}, \hat{\beta})}.$$

**2.1. Expectation-Maximization algorithm**

The EM algorithm proposed by Dempster *et al.* [16] can be used to obtain the MLEs of the parameters  $\alpha$  and  $\beta$ . It is known that the EM algorithm converges more reliably than NR. Since type-I PHCS can be considered as an incomplete data problem (see [33]), it is possible to apply EM algorithm to obtain the MLEs of the parameters. Now, let us denote the incomplete (censored) data by  $\mathbf{Z} = (Z_1, Z_2, \dots, Z_r)$  where  $Z_j = (Z_{j1}, Z_{j2}, \dots, Z_{jR_j})$ ,  $j = 1, 2, \dots, r$ , such that  $Z_j$  denotes the lifetimes of censored units at the time of  $x_{(j)}$ . Similarly, let  $Z_T$  denotes the lifetimes of censored units at the time of  $T$ . Now, combining both the observed and censored data, one can obtain the complete data which is given by  $\mathbf{W} = (\mathbf{X}, \mathbf{Z})$ . The corresponding likelihood equation of the complete data can be obtained as follows:

$$(2.11) \quad L_W(\alpha, \beta | \mathbf{x}) = \prod_{i=1}^r \left\{ f(x_{(i)}; \alpha, \beta) \prod_{j=1}^{R_i} f(z_{ij}; \alpha, \beta) \right\} \prod_{j=1}^{R_T} f(z_{Tj}; \alpha, \beta).$$

Therefore, the log-likelihood equation can be easily obtained by taking the natural logarithm of Equation (2.11) as follows:

$$(2.12) \quad \begin{aligned} l_W(\alpha, \beta | \mathbf{x}) &= \ln(L_W(\alpha, \beta | \mathbf{x})) \\ &= \sum_{i=1}^r \ln(\alpha \beta x_{(i)}^{\alpha-1} \exp\{-\beta x_{(i)}^\alpha\}) \\ &\quad + \sum_{i=1}^r \sum_{j=1}^{R_i} \ln(\alpha \beta z_{ij}^{\alpha-1} \exp\{-\beta z_{ij}^\alpha\}) + \sum_{j=1}^{R_T} \ln(\alpha \beta z_{Tj}^{\alpha-1} \exp\{-\beta z_{Tj}^\alpha\}) \\ &= n \ln \alpha + n \ln \beta + (\alpha - 1) \sum_{i=1}^r \ln(x_{(i)}) - \beta \sum_{i=1}^r x_{(i)}^\alpha + (\alpha - 1) \sum_{i=1}^r \sum_{j=1}^{R_i} \ln(z_{ij}) \\ &\quad - \beta \sum_{i=1}^r \sum_{j=1}^{R_i} z_{ij}^\alpha + (\alpha - 1) \sum_{j=1, r \neq m}^{R_T} \ln(z_{Tj}) - \beta \sum_{j=1, r \neq m}^{R_T} z_{Tj}^\alpha. \end{aligned}$$

Note that the last two terms of Equation (2.12), should be considered only for the Case II. Based on the complete sample, the MLEs of the parameters  $\alpha$  and  $\beta$  can be obtained by taking the derivatives of (2.12) with respect to  $\alpha$  and  $\beta$  respectively and equating them to zero as follows:

$$(2.13) \quad \begin{aligned} \frac{\partial l_W(\alpha, \beta | \mathbf{x})}{\partial \alpha} &= \frac{n}{\alpha} + \sum_{i=1}^r \ln(x_{(i)}) - \beta \sum_{i=1}^r x_{(i)}^\alpha \ln(x_{(i)}) + \sum_{i=1}^r \sum_{j=1}^{R_i} \ln(z_{ij}) \\ &\quad - \beta \sum_{i=1}^r \sum_{j=1}^{R_i} z_{ij}^\alpha \ln(z_{ij}) + \sum_{j=1, r \neq m}^{R_T} \ln(z_{Tj}) - \beta \sum_{j=1, r \neq m}^{R_T} z_{Tj}^\alpha \ln(z_{Tj}) = 0, \end{aligned}$$

$$(2.14) \quad \frac{\partial l_W(\alpha, \beta | \mathbf{x})}{\partial \beta} = \frac{n}{\beta} - \sum_{i=1}^r x_{(i)}^\alpha - \sum_{i=1}^r \sum_{j=1}^{R_i} z_{ij}^\alpha - \sum_{j=1, r \neq m}^{R_T} z_{Tj}^\alpha = 0.$$

Now, the conditional expectation of the log-likelihood equation of the complete data given the observations should be computed in the E-step of the algorithm. However, the following

conditional expectations are necessary to be computed:

$$\begin{aligned}
 E\left(\frac{\partial l_W(\alpha, \beta | \mathbf{x})}{\partial \alpha} \mid x_{(i)}, T\right) &= \frac{n}{\alpha} + \sum_{i=1}^r \ln(x_{(i)}) - \beta \sum_{i=1}^r x_{(i)}^\alpha \ln(x_{(i)}) \\
 &\quad + \sum_{i=1}^r \sum_{j=1}^{R_i} E\left[\ln(Z_{ij}) (1 - \beta Z_{ij}^\alpha) \mid Z_{ij} > x_{(i)}\right] \\
 &\quad + \sum_{j=1, r \neq m}^{R_T} E\left[\ln(Z_{Tj}) (1 - \beta Z_{Tj}^\alpha) \mid Z_{Tj} > T\right],
 \end{aligned}
 \tag{2.15}$$

$$\begin{aligned}
 E\left(\frac{\partial l_W(\beta, \beta | \mathbf{x})}{\partial \beta} \mid x_{(i)}, T\right) &= \frac{n}{\beta} - \sum_{i=1}^r x_{(i)}^\alpha - \sum_{i=1}^r \sum_{j=1}^{R_i} E\left[Z_{ij}^\alpha \mid Z_{ij} > x_{(i)}\right] \\
 &\quad - \sum_{j=1, r \neq m}^{R_T} E\left[Z_{Tj}^\alpha \mid Z_{Tj} > T\right].
 \end{aligned}
 \tag{2.16}$$

In order to compute the expectations given above, making use of the theorem proved in [33], the conditional probability function of the censored data given the observed data can be obtained as follows:

$$f(z_i | \mathcal{C}^*, \alpha, \beta) = \frac{f(z_i, \alpha, \beta)}{1 - F(\mathcal{C}^*, \alpha, \beta)}, \quad Z_i > \mathcal{C}^*,
 \tag{2.17}$$

such that  $\mathcal{C}^* = x_{(i)}$  for  $i = 1, 2, \dots, r$  and  $\mathcal{C}^* = T$  for  $i = T$ . Thus, the following expectations can be obtained:

$$\begin{aligned}
 \mathcal{E}_1(\mathcal{C}^*, \alpha, \beta) &= E\left[Z^\alpha \mid Z > \mathcal{C}^*\right] = \frac{1}{1 - F(\mathcal{C}^*, \alpha, \beta)} \int_{\mathcal{C}^*}^{\infty} t^\alpha f(t) dt \\
 &= \frac{e^{-\beta \mathcal{C}^{*\alpha}}}{1 - F(\mathcal{C}^*, \alpha, \beta)} \frac{(1 + \beta \mathcal{C}^{*\alpha})}{\beta},
 \end{aligned}
 \tag{2.18}$$

$$\begin{aligned}
 \mathcal{E}_2(\mathcal{C}^*, \alpha, \beta) &= E\left(\ln(Z) (1 - \beta Z^\alpha) \mid Z > \mathcal{C}^*\right) \\
 &= \frac{1}{1 - F(\mathcal{C}^*, \alpha, \beta)} \int_{\mathcal{C}^*}^{\infty} \ln(t) (1 - \beta t^\alpha) f(t) dt.
 \end{aligned}
 \tag{2.19}$$

Since it is hard to obtain a closed form solution to Equation (2.19), the integral is approximated via Monte Carlo integration method in the simulation. After updating the missing data with the expectations above in the E-step, the log-likelihood function is maximized in the M-step at the current state, say  $\hat{\alpha}_k$  and  $\hat{\beta}_k$  being the estimators of  $\alpha$  and  $\beta$  and the following updating equations are computed:

$$\begin{aligned}
 \hat{\alpha}_{k+1} &= n \left\{ - \sum_{i=1}^r \ln(x_{(i)}) + \hat{\beta}_{k+1} \sum_{i=1}^r x_{(i)}^{\hat{\alpha}_k} \ln(x_{(i)}) - \sum_{i=1}^r R_i \mathcal{E}_2(x_{(i)}, \hat{\alpha}_k, \hat{\beta}_{k+1}) \right. \\
 &\quad \left. - R_T \mathcal{E}_2(T, \hat{\alpha}_k, \hat{\beta}_{k+1}) \right\}^{-1},
 \end{aligned}
 \tag{2.20}$$

$$\hat{\beta}_{k+1} = n \left\{ \sum_{i=1}^r x_{(i)}^{\hat{\alpha}_k} + \sum_{i=1}^r R_i \mathcal{E}_1(x_{(i)}, \hat{\alpha}_k, \hat{\beta}_k) + R_T \mathcal{E}_1(T, \hat{\alpha}_k, \hat{\beta}_k) \right\}^{-1}.
 \tag{2.21}$$

The EM estimates of  $(\alpha, \beta)$  can be computed by an iterative procedure using Equation (2.21) and the iterations can be terminated when  $|\hat{\alpha}_{k+1} - \alpha_k| + |\hat{\beta}_{k+1} - \beta_k| < \epsilon$  where  $\epsilon > 0$  is a small real number.

---

## 2.2. Stochastic Expectation-Maximization algorithm

---

The computations in the E-step of EM algorithm is complex. Therefore, Wei and Tanner [46] proposed a Monte Carlo version of EM algorithm. However, the M-step of this algorithm may take so much time. Diebolt and Celeux [18] introduced a stochastic-EM (SEM) algorithm by considering simulated values from the conditional distribution. Asl *et al.* [4] used this algorithm successfully. In the SEM algorithm, firstly, one needs to generate  $R_i$  number of samples of  $z_{ij}$  where  $i = 1, 2, \dots, r$  and  $j = 1, 2, \dots, R_i$  using the following conditional CDF:

$$(2.22) \quad F(z_{ij}; \alpha, \beta | z_{ij} > x_{(i)}) = \frac{F(z_{ij}; \alpha, \beta) - F(x_{(i)}; \alpha, \beta)}{1 - F(x_{(i)}; \alpha, \beta)}, \quad z_{ij} > x_{(i)}.$$

Now, using Equations (2.13) and (2.14), the estimators of  $(\alpha, \beta)$  at the  $k + 1$  step of the algorithm can be obtained as follows:

$$(2.23) \quad \hat{\alpha}_{k+1} = n \left[ - \sum_{i=1}^r \ln(x_{(i)}) + \hat{\beta}_{k+1} \sum_{i=1}^r x_{(i)}^{\hat{\alpha}_k} \ln(x_{(i)}) - \sum_{i=1}^r \sum_{j=1}^{R_i} \ln(z_{ij}) \left( 1 - \hat{\beta}_{k+1} z_{ij}^{\hat{\alpha}_k} \right) - \sum_{j=1, r \neq m}^{R_T} \ln(z_{Tj}) \left( 1 - \hat{\beta}_{k+1} z_{Tj}^{\hat{\alpha}_k} \right) \right]^{-1},$$

$$(2.24) \quad \hat{\beta}_{k+1} = n \left[ \sum_{i=1}^r x_{(i)}^{\hat{\alpha}_k} + \sum_{i=1}^r \sum_{j=1}^{R_i} z_{ij}^{\hat{\alpha}_k} + \sum_{j=1, r \neq m}^{R_T} z_{Tj}^{\hat{\alpha}_k} \right]^{-1}.$$

Similarly, the iterations can be terminated when  $|\hat{\alpha}_{k+1} - \alpha_k| + |\hat{\beta}_{k+1} - \beta_k| < \epsilon$  where  $\epsilon > 0$  is a small real number.

---

## 2.3. Fisher information matrix

---

In this subsection, by making use of the idea of missing information principle proposed by Louis [31], we can obtain the observed Fisher information matrix. Louis [31] suggested the following relation:

$$(2.25) \quad \mathbf{I}_X(\psi) = \mathbf{I}_W(\psi) - \mathbf{I}_{W|X}(\psi),$$

where  $\psi = (\alpha, \beta)'$ ,  $\mathbf{I}_X(\psi)$ ,  $\mathbf{I}_W(\psi)$  and  $\mathbf{I}_{W|X}(\psi)$  are the observed, complete and missing information matrices respectively. Now, the complete information matrix of a complete data set following the Weibull distribution can be obtained as

$$(2.26) \quad \mathbf{I}_W(\psi) = -E \left( \frac{\partial^2 \ln \mathcal{L}}{\partial \psi^2} \right) = E \left[ \begin{array}{cc} \frac{n}{\alpha^2} + \beta \sum_{i=1}^n x_i^\alpha & \sum_{i=1}^n x_i^\alpha \ln x_i \\ \sum_{i=1}^n x_i^\alpha \ln x_i & \frac{n}{\beta^2} \end{array} \right] = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix},$$

where

$$\begin{aligned} b_{11} &= \frac{n}{\alpha^2} + n\alpha\beta^2 \int_0^\infty \frac{x^{2\alpha-1} \ln(x)}{\exp(\beta x^\alpha)} dx, \\ b_{12} &= b_{21} = n\alpha\beta \int_0^\infty \frac{x^{2\alpha-1} \ln(x)}{\exp(\beta x^\alpha)} dx, \\ b_{22} &= \frac{n}{\beta^2}, \end{aligned}$$

and  $\ln \mathcal{L}(\psi) = n \ln \alpha + n \ln \beta + (\alpha - 1) \sum_{i=1}^n x_i + \beta \sum_{i=1}^n x_i^\alpha$  is the corresponding log-likelihood equation. Moreover, the missing information matrix  $\mathbf{I}_{W|X}(\psi)$  is given by

$$(2.27) \quad \mathbf{I}_{W|X}(\psi) = \sum_{i=1}^r R_i \mathbf{I}_{W|X}^{(i)}(\psi) + R_T \mathbf{I}_{W|X}^*(\psi),$$

where  $\mathbf{I}_{W|X}^{(i)}(\psi)$  and  $\mathbf{I}_{W|X}^*(\psi)$  are the information matrices of a single observation from a truncated Weibull distribution from left at  $x_{(i)}$  and  $T$  respectively, such that

$$\mathbf{I}_{W|X}^{(i)}(\psi) = -E \left( \frac{\partial^2 \ln \mathcal{L}}{\partial \psi^2} \ln \left\{ f(z_{ij}; \psi \mid z_{ij} > x_{(i)}) \right\} \right).$$

Now to calculate the missing information matrix  $\mathbf{I}_{W|X}^{(i)}(\psi)$ , the conditional distribution given in Equation (2.17) is used to obtain the following

$$L_f = \ln \left( f(z_{ij} \mid z_{ij} > x_{(i)}) \right) = \ln(\alpha) + \ln(\beta) + (\alpha - 1) \ln(z_{ij}) - \beta z_{ij}^\alpha + \beta x_{(i)}^\alpha.$$

The second partial derivatives of  $L_f$  are obtained as follows:

$$\begin{aligned} \frac{\partial^2 L_f}{\partial \alpha^2} &= -\frac{1}{\alpha^2} - \beta z_{ij}^\alpha \ln(z_{ij})^2 + \beta x_{(i)}^\alpha \ln(x_{(i)})^2, \\ \frac{\partial^2 L_f}{\partial \alpha \partial \beta} &= -z_{ij}^\alpha \ln(z_{ij}) + x_{(i)}^\alpha \ln(x_{(i)}), \\ \frac{\partial^2 L_f}{\partial \beta^2} &= -\frac{1}{\beta^2}. \end{aligned}$$

Now, in order to obtain the information matrices, the negative expected values of the quantities above are computed respectively as follows:

$$\begin{aligned} E \left( -\frac{\partial^2 L_f}{\partial \alpha^2} \right) &= \frac{1}{\alpha^2} + \beta \mathcal{E}_4(x_{(i)}, \alpha, \beta) - \beta x_{(i)}^\alpha \ln(x_{(i)})^2, \\ E \left( -\frac{\partial^2 L_f}{\partial \alpha \partial \beta} \right) &= \mathcal{E}_3(x_{(i)}, \alpha, \beta) - x_{(i)}^\alpha \ln(x_{(i)}), \\ E \left( -\frac{\partial^2 L_f}{\partial \beta^2} \right) &= \frac{1}{\beta^2}, \end{aligned}$$

where

$$\begin{aligned} \mathcal{E}_3(\mathcal{C}^*, \alpha, \beta) &= E(Z^\alpha \ln(Z) \mid Z > \mathcal{C}^*) = \frac{1}{1 - F(\mathcal{C}^*, \alpha, \beta)} \int_{\mathcal{C}^*}^\infty t^\alpha \ln(t) f(t) dt, \\ \mathcal{E}_4(\mathcal{C}^*, \alpha, \beta) &= E(Z^\alpha \ln(Z)^2 \mid Z > \mathcal{C}^*) = \frac{1}{1 - F(\mathcal{C}^*, \alpha, \beta)} \int_{\mathcal{C}^*}^\infty t^\alpha \ln(t)^2 f(t) dt. \end{aligned}$$

Using similar arguments, the information matrix  $\mathbf{I}_{W|X}^*(\psi)$  can also be computed easily. Then, using (2.25)–(2.26), the asymptotic variance-covariance matrix of  $\hat{\psi}$  can be computed by inverting the observed information matrix  $\mathbf{I}_X(\hat{\psi})$ . Note that  $\hat{\psi}$  is computed using the NR estimates.

---

### 3. BAYESIAN ESTIMATION

---

In this section, following Kundu [25], we consider the Bayesian estimation for the parameters of the Weibull distribution under the assumption that the random variables  $\alpha$  and  $\beta$  have independent gamma priors such that  $\alpha \sim \text{Gamma}(a, b)$  and  $\beta \sim \text{Gamma}(c, d)$ . Therefore, the joint prior density of  $\alpha$  and  $\beta$  can be written as

$$\pi(\alpha, \beta) \propto \alpha^{a-1} \beta^{c-1} \exp\{-(b\alpha + d\beta)\}, \quad a, b, c, d > 0.$$

Now, the posterior distribution of  $\alpha$  and  $\beta$  can be obtained as follows:

$$\begin{aligned} \pi(\alpha, \beta | \mathbf{x}) &= \frac{L(\alpha, \beta | \mathbf{x}) \pi(\alpha, \beta)}{\int_0^\infty \int_0^\infty L(\alpha, \beta | \mathbf{x}) \pi(\alpha, \beta) d\alpha d\beta} \\ (3.1) \quad &= \frac{\left(\prod_{i=1}^r x_{(i)}^{\alpha-1}\right) \beta^{c+r-1} \alpha^{a+r-1}}{\Gamma(c+r) \Psi(a, c, \mathbf{x})} \exp\left\{d - b\alpha + \sum_{i=1}^r (1 + R_i) x_{(i)}^\alpha + C^\alpha R_T\right\}, \end{aligned}$$

where

$$\Psi(a, c, \mathbf{x}) = \int_0^\infty \frac{\alpha^{a+r-1} \exp\{-b\alpha\} \left(\prod_{i=1}^r x_{(i)}^{\alpha-1}\right)}{\left[d + \sum_{i=1}^r (1 + R_i) x_{(i)}^\alpha + C^\alpha R_T\right]^{a+c+r}} d\alpha.$$

In this paper, three different loss functions are considered. One of them is the most commonly used squared error loss function (SEL) which is defined as follows:

$$L_S(\hat{t}(\psi), t(\psi)) = (\hat{t}(\psi) - t(\psi))^2,$$

where  $\hat{t}(\psi)$  is an estimator of  $t(\psi)$ . SEL is a symmetric loss function which gives equal weights to both underestimation and overestimation. However, in certain situation overestimation and underestimation may have serious consequences ([37]). In such cases using SEL may not be appropriate. One remedy is to use linear-exponential (LINEX) loss function. LINEX is an asymmetric loss function introduced by Varian [45] as follows:

$$L_L(\hat{t}(\psi), t(\psi)) = e^{\nu(\hat{t}(\psi)-t(\psi))} - \nu(\hat{t}(\psi) - t(\psi)) - 1, \quad \nu \neq 0.$$

The LINEX loss function is a convex function whose shape is determined by the value of  $\nu$ . The negative (positive) value of  $\nu$  gives more weight to overestimation (underestimation) and its magnitude reflects the degree of asymmetry. It is seen that, for  $\nu = 1$ , the function is quite asymmetric with overestimation being costlier than underestimation. If  $\nu < 0$ , it rises almost exponentially when the estimation error  $\hat{t}(\psi) - t(\psi) < 0$  and almost linearly if  $\hat{t}(\psi) - t(\psi) > 0$ . For small values of  $|\nu|$ , the LINEX loss function is almost symmetric and not far from squared error loss function.

Under the SEL function, the Bayes estimators of  $\alpha$  and  $\beta$  which are the expected values of the corresponding posterior distributions are computed respectively as follows:

$$(3.2) \quad \hat{\alpha}_S = E(\pi(\alpha | \mathbf{x})) = \frac{\Psi(a+1, c-1, \mathbf{x})}{\Psi(a, c, \mathbf{x})}$$

and

$$(3.3) \quad \hat{\beta}_S = E(\pi(\beta | \mathbf{x})) = (a + c + r) \frac{\Psi(a, c+1, \mathbf{x})}{\Psi(a, c, \mathbf{x})}.$$

Since the Bayes estimators given above includes the complicated integral function  $\Psi(a, c+1, \mathbf{x})$  we also consider using the Bayes estimate of  $t(\psi)$  under the LINEX loss function is given by

$$\hat{t}_L(\psi) = -\frac{1}{\nu} \ln \left[ E_t(e^{-\nu t(\psi)} | \mathbf{x}) \right] = -\frac{1}{\nu} \ln \left[ \int_0^\infty \int_0^\infty e^{-\nu t(\psi)} \pi(\alpha, \beta | \mathbf{x}) d\alpha d\beta \right].$$

Another asymmetric loss function that gained more attention is the general entropy loss (GEL) function given by

$$L_{GEL}(\hat{t}(\psi), t(\psi)) = \left( \frac{\hat{t}(\psi)}{t(\psi)} \right)^\kappa - \kappa \ln \left( \frac{\hat{t}(\psi)}{t(\psi)} \right) - 1, \quad \kappa \neq 0,$$

where  $\kappa$  is the shape parameter showing the departure from symmetry. When  $\kappa > 0$ , the overestimation is considered to be more serious than underestimation and for  $\kappa < 0$  vice versa. The Bayes estimator under GEL function is given by

$$\hat{t}_{GEL}(\psi) = \left[ E_t(t(\psi)^{-\kappa} | \mathbf{x}) \right]^{-1/\kappa} = \left[ \int_0^\infty \int_0^\infty t(\psi)^{-\kappa} \pi(\alpha, \beta | \mathbf{x}) d\alpha d\beta \right]^{-1/\kappa}.$$

### 3.1. Tierney–Kadane approximation

In this subsection, the approximation method of Tierney and Kadane [44] is used to obtain the approximate Bayes estimators under SEL, LINEX and GEL loss functions. Now, we consider the following functions:

$$(3.4) \quad \Delta(\alpha, \beta) = \frac{1}{n} \ln [L(\alpha, \beta | \mathbf{x}) \pi(\alpha, \beta)],$$

$$(3.5) \quad \Delta^*(\alpha, \beta) = \frac{1}{n} \ln [L(\alpha, \beta | \mathbf{x}) \pi(\alpha, \beta) t(\psi)].$$

Now assume that  $(\tilde{\alpha}_\Delta, \tilde{\beta}_\Delta)$  and  $(\tilde{\alpha}_{\Delta^*}, \tilde{\beta}_{\Delta^*})$  respectively maximize the functions  $\Delta(\alpha, \beta)$  and  $\Delta^*(\alpha, \beta)$ . Then the approximation method of Tierney and Kadane [44] is given by

$$\tilde{t}_{SEL}(\alpha, \beta) = \sqrt{\frac{|\Sigma^*|}{|\Sigma|}} \exp \left[ n \left( \Delta_1^*(\tilde{\alpha}_{\Delta^*}, \tilde{\beta}_{\Delta^*}) - \Delta(\tilde{\alpha}_\Delta, \tilde{\beta}_\Delta) \right) \right],$$

where  $|\Sigma|$  and  $|\Sigma^*|$  are the negative of inverses the second derivative matrices of  $\Delta(\alpha, \beta)$  and  $\Delta_1^*(\alpha, \beta)$  respectively obtained at  $(\tilde{\alpha}_\Delta, \tilde{\beta}_\Delta)$  and  $(\tilde{\alpha}_{\Delta^*}, \tilde{\beta}_{\Delta^*})$ . The function  $\Delta(\alpha, \beta)$  can be easily obtained by using the Equation (3.4) as follows:

$$(3.6) \quad \Delta(\alpha, \beta) = \frac{1}{n} \left[ \ln(M) + (\alpha - 1) \sum_{i=1}^r \ln(x_{(i)}) - \beta \left( d + b\alpha + \sum_{i=1}^r (1 + R_i) x_{(i)}^\alpha + C^\alpha R_T \right) + (a + c + r - 1) \ln(\beta) + (a + r - 1) \ln(\alpha) \right],$$



where  $M = \frac{d^c b^a}{\Gamma(c)\Gamma(a)}$ . Now, differentiating Equation (3.6) with respect to  $\alpha$  and  $\beta$  solving for these parameters, one gets the following equations:

$$\begin{aligned} \tilde{\alpha}_\Delta &= (a + r - 1) \left[ \beta \left( b + \sum_{i=1}^r (1 + R_i) x_{(i)}^\alpha + C^\alpha R_T \right) - \sum_{i=1}^r \ln(x_{(i)}) \right]^{-1}, \\ \tilde{\beta}_\Delta &= (a + c + r - 1) \left[ \sum_{i=1}^r (1 + R_i) x_{(i)}^\alpha + C^\alpha R_T + d + b\alpha \right]^{-1}. \end{aligned}$$

Since it is easy to obtain the second derivatives and the related Hessian matrices, we skip this part. Thus under the SEL function, the approximate Bayes estimators are computed by

$$\begin{aligned} \tilde{\alpha}_{\text{SEL}} &= \sqrt{\frac{|\Sigma^*|}{|\Sigma|}} \exp \left[ n \left( \Delta_{1\alpha}^*(\tilde{\alpha}_{\Delta^*}, \tilde{\beta}_{\Delta^*}) - \Delta(\tilde{\alpha}_\Delta, \tilde{\beta}_\Delta) \right) \right], \\ \tilde{\beta}_{\text{SEL}} &= \sqrt{\frac{|\Sigma^*|}{|\Sigma|}} \exp \left[ n \left( \Delta_{1\beta}^*(\tilde{\alpha}_{\Delta^*}, \tilde{\beta}_{\Delta^*}) - \Delta(\tilde{\alpha}_\Delta, \tilde{\beta}_\Delta) \right) \right], \end{aligned}$$

where  $\Delta_{1\alpha}^*(\alpha, \beta) = \Delta(\alpha, \beta) + \frac{1}{n} \ln(\alpha)$  for  $t(\alpha, \beta) = \alpha$  and  $\Delta_{1\beta}^*(\alpha, \beta) = \Delta(\alpha, \beta) + \frac{1}{n} \ln(\beta)$  for  $t(\alpha, \beta) = \beta$ .

One can also compute the Bayes estimators under the LINEX loss and get

$$\tilde{t}_{\text{LINEX}}(\alpha, \beta) = \sqrt{\frac{|\Sigma^*|}{|\Sigma|}} \exp \left[ n \left\{ \Delta_2^*(\tilde{\alpha}_{\Delta^*}, \tilde{\beta}_{\Delta^*}) - \Delta(\tilde{\alpha}_\Delta, \tilde{\beta}_\Delta) \right\} \right].$$

Letting  $t(\alpha, \beta) = e^{-\nu\alpha}$ , one gets  $\Delta_{2\alpha}^*(\alpha, \beta) = \Delta(\alpha, \beta) - \frac{1}{n} \nu\alpha$  and letting  $t(\alpha, \beta) = e^{-\nu\beta}$ ,  $\Delta_{2\beta}^*(\alpha, \beta) = \Delta(\alpha, \beta) - \frac{1}{n} \nu\beta$ . Thus, approximate Bayes estimators under LINEX function are computed as

$$\begin{aligned} \tilde{\alpha}_{\text{LINEX}} &= -\frac{1}{\nu} \ln \left( \sqrt{\frac{|\Sigma^*|}{|\Sigma|}} \exp \left[ n \left( \Delta_{2\alpha}^*(\tilde{\alpha}_{\Delta^*}, \tilde{\beta}_{\Delta^*}) - \Delta(\tilde{\alpha}_\Delta, \tilde{\beta}_\Delta) \right) \right] \right), \\ \tilde{\beta}_{\text{LINEX}} &= -\frac{1}{\nu} \ln \left( \sqrt{\frac{|\Sigma^*|}{|\Sigma|}} \exp \left[ n \left( \Delta_{2\beta}^*(\tilde{\alpha}_{\Delta^*}, \tilde{\beta}_{\Delta^*}) - \Delta(\tilde{\alpha}_\Delta, \tilde{\beta}_\Delta) \right) \right] \right). \end{aligned}$$

Finally, letting  $t(\alpha, \beta) = \alpha^{-\kappa}$ , one gets  $\Delta_{3\alpha}^*(\alpha, \beta) = \Delta(\alpha, \beta) - \frac{\kappa}{n} \ln(\alpha)$  and letting  $t(\alpha, \beta) = \beta^{-\kappa}$ ,  $\Delta_{3\beta}^*(\alpha, \beta) = \Delta(\alpha, \beta) - \frac{\kappa}{n} \ln(\beta)$ . Thus, approximate Bayes estimators under GEL function are obtained by

$$\begin{aligned} \tilde{\alpha}_{\text{GEL}} &= \left( \sqrt{\frac{|\Sigma^*|}{|\Sigma|}} \exp \left[ n \left( \Delta_{3\alpha}^*(\tilde{\alpha}_{\Delta^*}, \tilde{\beta}_{\Delta^*}) - \Delta(\tilde{\alpha}_\Delta, \tilde{\beta}_\Delta) \right) \right] \right)^{-1/\kappa}, \\ \tilde{\beta}_{\text{GEL}} &= \left( \sqrt{\frac{|\Sigma^*|}{|\Sigma|}} \exp \left[ n \left( \Delta_{3\beta}^*(\tilde{\alpha}_{\Delta^*}, \tilde{\beta}_{\Delta^*}) - \Delta(\tilde{\alpha}_\Delta, \tilde{\beta}_\Delta) \right) \right] \right)^{-1/\kappa}. \end{aligned}$$

---

### 3.2. MCMC method

---

Metropolis–Hastings (MH) algorithm, a method for generating random samples from the posterior distribution using a proposal density, is considered in this subsection. A sym-

metric proposal density of type  $q(\theta'|\theta) = q(\theta|\theta')$  may be considered generally, where  $\theta$  is the parameter vector of the distribution considered. Following Dey *et al.* [17], we consider a bivariate normal distribution as the proposal density such that  $q(\theta'|\theta) = N(\theta|\mathbf{V}_\theta)$  where  $\mathbf{V}_\theta$  is the covariance matrix and  $\theta = (\alpha, \beta)$ . Although, the bivariate normal distribution may generate negative observations, the domain of both shape and scale parameters of Weibull distribution is positive. Therefore, the following steps of MH algorithm is used to generate MCMC sample from the posterior density given by (3.1):

- (1) Set the initial parameter values as  $\theta = \theta_0$ .
- (2) For  $j = 1, 2, \dots, N$ , repeat the following steps:
  - (i) Set  $\theta = \theta_{j-1}$ ;
  - (ii) Generate new parameters  $\lambda$  from bivariate normal  $N_2(\ln(\theta), \mathbf{V}_\theta)$ ;
  - (iii) Compute  $\theta_{\text{new}} = \exp(\lambda)$ ;
  - (iv) Calculate  $\gamma = \min\left(1, \frac{\pi(\theta_{\text{new}}|\mathbf{x})\theta_{\text{new}}}{\pi(\theta|\mathbf{x})\theta}\right)$ ;
  - (v) Set  $\theta_j = \theta_{\text{new}}$  with probability  $\lambda$ , otherwise  $\theta_j = \theta$ .

After generating the MCMC sample, some of the initial samples, say  $N_0$ , can be discarded as burn-in process and the estimations can be computed via the remaining ones ( $M = N - N_0$ ) under SEL, LINEX and GEL loss functions as follows:

$$\begin{aligned} \hat{t}_{\text{SEL}}(\psi) &= \frac{1}{M} \sum_{i=1}^M t(\psi_i), \\ \hat{t}_{\text{LINEX}}(\psi) &= -\frac{1}{\nu} \ln\left(\frac{1}{M} \sum_{i=1}^M \exp(-\nu t(\psi_i))\right), \\ \hat{t}_{\text{GEL}}(\psi) &= \left(\frac{1}{M} \sum_{i=1}^M (t(\psi_i)^{-\kappa})\right)^{-1/\kappa}. \end{aligned}$$

The main advantage of MCMC method over Tierney–Kadane method is that the MCMC samples can also be used to compute highest posterior density (HPD) intervals. Chen and Shao [14] proposed a method to compute the HPD intervals using MCMC samples. This method has been used in the literature extensively. Now, consider the posterior density  $\pi(\theta|\mathbf{x})$ . Assume that the  $p$ -th quantile of the distribution is given by  $\theta^{(p)} = \inf\{\theta: \Pi(\theta|\mathbf{x}) \geq p; 0 < p < 1\}$  where  $\Pi(\theta|\mathbf{x})$  denotes the posterior distribution function of  $\theta$ . Now, for a given  $\theta^*$ , a simulation consistent estimator of  $\Pi(\theta^*|\mathbf{x})$  can be computed as

$$\Pi(\theta^*|\mathbf{x}) = \frac{1}{M} \sum_{i=1}^M I(\theta \leq \theta^*),$$

where  $I(\theta \leq \theta^*)$  is an indicator function. Then, the estimate of  $\Pi(\theta^*|\mathbf{x})$  is given as

$$\hat{\Pi}(\theta^*|\mathbf{x}) = \begin{cases} 0 & \text{if } \theta^* < \theta_{(N_0)}, \\ \sum_{j=N_0}^i \gamma_j & \text{if } \theta_{(i)} < \theta^* < \theta_{(i+1)}, \\ 1 & \text{if } \theta_{(M)}, \end{cases}$$

where  $\gamma_j = 1/M$  and  $\theta_{(j)}$  is the  $j$ -th ordered value of  $\theta_j$ .  $\theta^{(p)}$  can be approximated by the following:

$$\theta^{(p)} = \begin{cases} \theta_{(N_0)} & \text{if } p = 0, \\ \theta_{(j)} & \text{if } \sum_{j=N_0}^{i-1} \gamma_j < p < \sum_{j=N_0}^i \gamma_j. \end{cases}$$

Now, one can construct the  $100(1-p)\%$  confidence intervals where  $0 < p < 1$  as  $(\hat{\theta}^{j/s}, \hat{\theta}^{(j+[(1-p)s])/s})$ ,  $j = 1, 2, \dots, s - [(1-p)s]$  such that  $[v]$  denotes the greatest integer less than or equal to  $v$ . At the end, the HPD credible interval of  $\theta$  is the one having the shortest length.

---

#### 4. SHRINKAGE ESTIMATION

---

In the problem of statistical inference there may be some non-sample prior information that practitioner may have from previous experiences or knowledge Saleh [39]. For example, medical experts may know the average time of that a vaccine may take to relief a pain according their medical knowledge. This non-sample Prior information on the parameters in a statistical model generally leads to an improved inference procedure in problems of statistical inference. Restricted models arise from the incorporation of the known prior information in the model in the form of a constraint. The estimators obtained from restricted (unrestricted) model is known as the restricted (unrestricted) estimators. The results of an analysis of the restricted and unrestricted models can be weighted against loss of efficiency and validity of the constraints in deciding a choice between these two extreme inference methods, when a full confidence may not be in the prior information (see [2]).

Bancroft [11] was the first to consider a pre-test procedure when there is doubt that the prior information is not certain (uncertain prior information). After the pioneering study [11], pre-test estimators has gained much attention. Thompson [43] defined an efficient shrinkage estimator. Following [43], shrinkage estimation of the Weibull parameters has been discussed by a number of authors, including [41], [35], [36] and [42]. We also refer to the following book and papers among others: [22], [40], [39], [23].

Now suppose that there is an uncertain prior information in the form of  $\theta = \theta_0$  where  $\theta$  is the parameter of a distribution of interest. Our aim is to estimate  $\theta$  using a pre-test estimation strategy and this prior information. Therefore, we consider the following hypothesis to check the validity of this information:

$$\begin{aligned} H_0: \theta &= \theta_0, \\ H_1: \theta &\neq \theta_0. \end{aligned}$$

It is known that under  $H_0$ , the asymptotic distribution of  $\sqrt{D}(\hat{\theta} - \theta_0)$  is normal with  $N(0, \sigma_{\hat{\theta}}^2)$  and the related test statistics can be defined as follows:

$$W_D = \left( \frac{\sqrt{D}(\hat{\theta} - \theta_0)}{\sigma_{\hat{\theta}}} \right)^2.$$

One can reject the null hypothesis when  $W_D > \chi_1^2(\lambda)$  based on the distribution of  $W_D$  where  $\lambda$  can be treated as the degree of trust in the prior information about the parameter such that  $\theta = \theta_0$ , see [39] and [1]. Thus, the shrinkage pre-test estimator (SPT) can be defined as:

$$\hat{\theta}_{\text{SPT}} = \lambda\theta_0 + (1 - \lambda)\hat{\theta}I(W_D < \chi_1^2(\lambda)),$$

where  $I(A)$  is the indicator of the set  $A$ .

---

## 5. MONTE CARLO SIMULATION EXPERIMENTS

---

In this section, we conduct a simulation study to illustrate the performance of the different estimation techniques discussed in this paper by considering  $(n, m) = (30, 15)$ , different values of predetermined time  $T = 1.0, 2.0$ , and the real values of the parameters are chosen as  $\alpha = 0.5$  and  $\beta = 1.5$  in all cases. The following three schemes are considered in the simulation:

- Scheme 1:  $R = (0^{m-1}, n - m)$ ;
- Scheme 2:  $R = (n - m, 0^{m-1})$ ;
- Scheme 3:  $R = (2^5, 0^{m-6}, n - m - 10)$ .

It is noted that Scheme 1 is the type-II censoring such that  $n - m$  units are removed from the experiment at the time of the  $m$ -th failure; in Scheme 2,  $n - m$  units are removed at the time of the first failure. However, in Scheme 3, a progressive type-II censoring scheme allowing different numbers of censoring within the experiment is considered. The progressive type-II censored data from Weibull distribution is generated using algorithm proposed by Balakrishnan and Aggarwala [7]. The maximum likelihood estimators of  $\alpha$  and  $\beta$  are obtained using NR, EM and SEM algorithms. In computing the Bayes estimates, two different priors are used such as the non-informative priors as  $a = b = c = d = 0$  and the informative priors where we assume that we have past samples from Weibull( $\alpha, \beta$ ) distribution, say  $K$  samples and their corresponding MLEs as  $(\hat{\alpha}_j, \hat{\beta}_j)$ ,  $j = 1, 2, \dots, K$ . Now, equating the sample means and variances of these values to the means and variances of gamma priors respectively and solving the equations for  $K = 1000$ , and  $n = 30$  being the sample size of past samples, we obtain the following informative prior values,  $a = 43.77$ ,  $b = 83.45$ ,  $c = 24.24$ ,  $d = 15.47$ .

Bayes estimates are computed under SEL, LINEX, GEL loss functions. Notice that for the LINEX loss function, we considered two values of  $\nu$  as  $\nu = -0.5, 0.5$  giving more weight to underestimation and overestimation respectively. Similarly, two choices of  $\kappa$  such as  $\kappa = -0.5, 0.5$  are taken into account under GEL function. Moreover, 6000 MCMC samples are generated and MCMC estimations are computed under the listed loss function and respective parameter values. The first 1000 MCMC samples are considered as a burn-in sample so that the average values and MSEs are computed via the remaining 5000 samples for each replicate in the simulation.

For the shrinkage estimators, the test statistic  $W_D$  is calculated and then shrinkage pre-test (SPT) estimators are obtained. The distribution of the test statistic  $W_D$  is computed under the null hypothesis, that is,  $H_0: \theta = \theta_0$ . Moreover, we take  $\lambda = 0.5$  giving equal weight to both restricted and unrestricted estimators and the type one test error is set to 0.05 in testing the hypothesis, prior values of the parameters are taken as  $\alpha_0 = 0.7$ ,  $\beta_0 = 1.7$  for practical purposes. The MLE shrinkage pre-test estimators are obtained using NR algorithm and also the Bayes estimator with T-K method under different loss functions.

Totally, 5000 repetitions are carried out and average values (Avg), mean squared errors (MSE), confidence/credible interval lengths (IL) and coverage probabilities (CP) are obtained for the purpose of comparison. MSEs of the estimators are computed as follows:

$$\text{MSE}(\hat{\theta}) = \frac{1}{5000} \sum_{i=1}^{5000} (\hat{\theta}_i - \theta)^2,$$

where  $\hat{\theta}_i$  is NR, EM, SEM, SPT estimators and Bayes estimators under SEL loss function in the  $i$ -th replication. However, the MSEs of Bayes estimators under LINEX and GEL loss functions are computed respectively by

$$\begin{aligned} \text{MSE}_{\text{LINEX}}(\hat{\theta}) &= \frac{1}{5000} \sum_{i=1}^{5000} \left( e^{\nu(\hat{\theta}_i - \theta)} - \nu(\hat{\theta}_i - \theta) - 1 \right), \\ \text{MSE}_{\text{GEL}}(\hat{\theta}) &= \frac{1}{5000} \sum_{i=1}^{5000} \left( \left( \frac{\hat{\theta}_i}{\theta} \right)^\kappa - \kappa \ln \left( \frac{\hat{\theta}_i}{\theta} \right) - 1 \right). \end{aligned}$$

All of the computations are performed using the R Statistical Program [38]. All the results are presented in Tables 1–5.

Based on Table 1, we can conclude that EM and SEM estimates are quiet preferable to the NR method for all schemes and  $T$ s. Both MSEs and Avgs for EM and SEM estimates are the close to each other and they are smaller than those of NR method. We also observe that as  $m$  increase, the values of MSEs and Avgs decrease, generally.

The results of Bayes estimates based on TK and MCMC methods are reported in Tables 2–3. From these tables, it is evident that all the Bayes estimates based on informative priors have very small MSEs compared to the MLEs. We also see that the Bayes estimates based on informative priors are better than those that are based on non-informative priors in all schemes and  $(T, n, m)$ s. However, EM and SEM estimates are better than non-informative Bayes estimates based on SEL in terms of MSE and Avg. So we can conclude that Bayes estimates even with non informative priors are preferable to the NR, for all schemes and  $T$ s. When we compare MSEs of T–K and MCMC methods, we observed that they are generally close to each other. However, T–K is better in some of the cases and vice versa in some others. However, the MCMC has the advantage of construction of the credible intervals. Thus, we can say that MCMC is preferable since it gives more information.

The performances of SPT estimators are given in Table 4. According to Table 4, we can say that SPT estimators based on informative T–K method have better performance than SPT based on NR methods in the sense of both MSE and Avg, generally. Moreover, SPT with T–K method based on GEL function seems to have the least MSE values among others. SPT estimator based on NR method has smaller MSE values than NR estimator when we consider the parameter  $\beta$ , and both methods have closer MSE values for the parameter  $\alpha$ .

Finally, the confidence intervals and coverage probabilities are summarized in Table 5. It is observed that when we use non-informative priors the estimated CPs are smaller than the nominal CPs. Moreover, the expected ILs of non-informative methods are less than that of NR method. However, the estimated CPs of NR are slightly more than the non-informative method. Further, we observe that the CIs based on informative priors are better than the ones based on the non-informative priors and the once based on NR, in terms of having smaller ILs but higher CPs.

**Table 1:** Average values (Avg) and the corresponding MSEs of the estimators NR, EM and SEM.

T	R		NR		EM		SEM	
			$\alpha$	$\beta$	$\alpha$	$\beta$	$\alpha$	$\beta$
1	1	Avg	0.5559	1.8433	0.5276	1.6415	0.5352	1.6669
		MSE	0.0224	0.6931	0.0112	0.2207	0.0139	0.2788
	2	Avg	0.5279	1.6480	0.5239	1.5958	0.5294	1.6065
MSE		0.0158	0.3385	0.0135	0.1772	0.0141	0.1946	
2	1	Avg	0.5559	1.8433	0.5276	1.6416	0.5353	1.6670
		MSE	0.0224	0.6930	0.0112	0.2206	0.0139	0.2788
	2	Avg	0.5280	1.6412	0.5233	1.5947	0.5287	1.6020
MSE		0.0137	0.3045	0.0124	0.1723	0.0129	0.1869	
3	1	Avg	0.5435	1.7540	0.5315	1.6490	0.5330	1.6485
		MSE	0.0175	0.5108	0.0127	0.2552	0.0131	0.2647
	2	Avg	0.5476	1.7676	0.5339	1.6578	0.5353	1.6567
MSE		0.0172	0.5001	0.0126	0.2494	0.0130	0.2593	

**Table 2:** Average values (Avg) and the corresponding MSEs of the Bayes estimators with T-K approximation.

T	R		SEL		LINEX				GEL			
			$\alpha$	$\beta$	$\nu = -0.5$		$\nu = 0.5$		$\kappa = -0.5$		$\kappa = 0.5$	
					$\alpha$	$\beta$	$\alpha$	$\beta$	$\alpha$	$\beta$	$\alpha$	$\beta$
Informative Priors												
1	1	Avg	0.5210	1.5773	0.5220	1.5946	0.5200	1.5600	0.5192	1.5665	0.5155	1.5449
		MSE	0.0018	0.0257	0.0002	0.0036	0.0002	0.0029	0.0008	0.0012	0.0007	0.0011
	2	Avg	0.5199	1.5632	0.5209	1.5807	0.5189	1.5460	0.5180	1.5522	0.5141	1.5302
MSE		0.0018	0.0252	0.0002	0.0034	0.0002	0.0029	0.0008	0.0012	0.0007	0.0012	
2	1	Avg	0.5206	1.5702	0.5215	1.5869	0.5196	1.5537	0.5188	1.5598	0.5151	1.5389
		MSE	0.0019	0.0273	0.0002	0.0037	0.0002	0.0031	0.0008	0.0013	0.0008	0.0012
	2	Avg	0.5210	1.5773	0.5220	1.5946	0.5200	1.5600	0.5192	1.5665	0.5155	1.5449
MSE		0.0018	0.0257	0.0002	0.0036	0.0002	0.0029	0.0008	0.0012	0.0007	0.0011	
3	1	Avg	0.5193	1.5605	0.5203	1.5772	0.5183	1.5442	0.5175	1.5501	0.5137	1.5291
		MSE	0.0018	0.0268	0.0002	0.0036	0.0002	0.0031	0.0008	0.0013	0.0008	0.0013
	2	Avg	0.5210	1.5719	0.5220	1.5885	0.5200	1.5554	0.5192	1.5615	0.5156	1.5408
MSE		0.0019	0.0271	0.0002	0.0037	0.0002	0.0031	0.0008	0.0013	0.0008	0.0012	
Non-Informative Priors												
1	1	Avg	0.5519	1.8793	0.5441	1.9056	0.5353	1.6135	0.5560	1.9979	0.5397	1.8118
		MSE	0.0211	0.7592	0.0022	0.0746	0.0022	0.0329	0.0083	0.0243	0.0086	0.0250
	2	Avg	0.5298	1.6345	0.5322	1.7031	0.5248	1.5557	0.5234	1.5990	0.5100	1.5181
MSE		0.0159	0.3310	0.0020	0.0425	0.0019	0.0355	0.0065	0.0140	0.0067	0.0144	
2	1	Avg	0.5411	1.7500	0.5408	1.8062	0.5341	1.6223	0.5384	1.7550	0.5262	1.6597
		MSE	0.0170	0.5159	0.0020	0.0582	0.0020	0.0402	0.0065	0.0166	0.0067	0.0172
	2	Avg	0.5519	1.8793	0.5441	1.9057	0.5353	1.6136	0.5560	1.9979	0.5397	1.8118
MSE		0.0211	0.7591	0.0022	0.0746	0.0021	0.0329	0.0083	0.0243	0.0086	0.0250	
3	1	Avg	0.5290	1.6203	0.5315	1.6773	0.5254	1.5573	0.5237	1.5920	0.5125	1.5253
		MSE	0.0137	0.2934	0.0017	0.0366	0.0017	0.0322	0.0056	0.0118	0.0057	0.0122
	2	Avg	0.5453	1.7632	0.5451	1.8196	0.5384	1.6361	0.5427	1.7685	0.5307	1.6741
MSE		0.0167	0.5052	0.0020	0.0568	0.0020	0.0389	0.0062	0.0152	0.0064	0.0159	

**Table 3:** Average values (Avg) and the corresponding MSEs of the Bayes estimators with MCMC method.

T	R		SEL		LINEX				GEL			
					$\nu = -0.5$		$\nu = 0.5$		$\kappa = -0.5$		$\kappa = 0.5$	
			$\alpha$	$\beta$	$\alpha$	$\beta$	$\alpha$	$\beta$	$\alpha$	$\beta$	$\alpha$	$\beta$
Informative Priors												
1	1	Avg	0.5210	1.5770	0.5220	1.5944	0.5200	1.5601	0.5192	1.5663	0.5155	1.5448
		MSE	0.0018	0.0262	0.0002	0.0039	0.0002	0.0030	0.0008	0.0015	0.0008	0.0012
	2	Avg	0.5199	1.5631	0.5209	1.5806	0.5188	1.5462	0.5179	1.5522	0.5140	1.5304
MSE		0.0018	0.0253	0.0002	0.0037	0.0002	0.0029	0.0009	0.0014	0.0008	0.0012	
3	Avg	0.5206	1.5703	0.5216	1.5871	0.5197	1.5540	0.5188	1.5599	0.5151	1.5391	
	MSE	0.0019	0.0277	0.0002	0.0041	0.0002	0.0032	0.0009	0.0016	0.0008	0.0013	
2	1	Avg	0.5210	1.5770	0.5220	1.5944	0.5200	1.5601	0.5192	1.5663	0.5155	1.5449
		MSE	0.0018	0.0262	0.0002	0.0039	0.0002	0.0030	0.0008	0.0015	0.0008	0.0012
	2	Avg	0.5193	1.5604	0.5203	1.5770	0.5183	1.5442	0.5174	1.5500	0.5137	1.5292
MSE		0.0018	0.0271	0.0002	0.0039	0.0002	0.0031	0.0009	0.0015	0.0008	0.0013	
3	Avg	0.5210	1.5720	0.5220	1.5887	0.5201	1.5558	0.5192	1.5617	0.5156	1.5410	
	MSE	0.0019	0.0275	0.0002	0.0040	0.0002	0.0031	0.0009	0.0016	0.0009	0.0013	
Non-Informative Priors												
1	1	Avg	0.5411	1.7748	0.5455	1.9792	0.5368	1.6503	0.5335	1.7117	0.5180	1.5932
		MSE	0.0176	0.4644	0.0024	0.1791	0.0022	0.0412	0.0073	0.0273	0.0068	0.0121
	2	Avg	0.5286	1.6289	0.5323	1.7081	0.5249	1.5607	0.5219	1.5890	0.5086	1.5091
MSE		0.0158	0.3208	0.0021	0.0665	0.0020	0.0370	0.0069	0.0159	0.0066	0.0127	
3	Avg	0.5380	1.7158	0.5414	1.8220	0.5346	1.6347	0.5320	1.6738	0.5199	1.5910	
	MSE	0.0161	0.4175	0.0022	0.1139	0.0020	0.0445	0.0067	0.0201	0.0063	0.0127	
2	1	Avg	0.5411	1.7748	0.5455	1.9793	0.5368	1.6504	0.5335	1.7117	0.5180	1.5933
		MSE	0.0176	0.4643	0.0024	0.1791	0.0022	0.0412	0.0073	0.0273	0.0068	0.0120
	2	Avg	0.5280	1.6166	0.5311	1.6804	0.5249	1.5604	0.5224	1.5839	0.5112	1.5181
MSE		0.0136	0.2856	0.0018	0.0582	0.0017	0.0338	0.0059	0.0137	0.0057	0.0110	
3	Avg	0.5422	1.7293	0.5455	1.8353	0.5389	1.6483	0.5363	1.6876	0.5245	1.6057	
	MSE	0.0159	0.4065	0.0021	0.1128	0.0020	0.0432	0.0065	0.0192	0.0061	0.0117	

**Table 4:** Average values (Avg) and the corresponding MSEs of the SPT estimators.

T	R		NR		SEL		LINEX		GEL	
			$\alpha$	$\beta$	$\alpha$	$\beta$	$\alpha$	$\beta$	$\alpha$	$\beta$
1	1	Avg	0.5911	1.8251	0.5879	1.6367	0.5870	1.6278	0.5828	1.6198
		MSE	0.0263	0.7304	0.0113	0.0249	0.0111	0.0225	0.0106	0.0208
	2	Avg	0.5576	1.6367	0.5771	1.6301	0.5759	1.6212	0.5708	1.6131
MSE		0.0190	0.1464	0.0104	0.0233	0.0102	0.0210	0.0097	0.0193	
3	Avg	0.5708	1.7106	0.5749	1.6329	0.5738	1.6244	0.5690	1.6166	
	MSE	0.0205	0.2600	0.0102	0.0247	0.0101	0.0224	0.0096	0.0208	
2	1	Avg	0.5911	1.8252	0.5879	1.6367	0.5870	1.6279	0.5828	1.6198
		MSE	0.0263	0.7304	0.0113	0.0249	0.0111	0.0225	0.0106	0.0208
	2	Avg	0.5556	1.6352	0.5698	1.6274	0.5686	1.6189	0.5631	1.6108
MSE		0.0173	0.1323	0.0097	0.0239	0.0095	0.0218	0.0090	0.0203	
3	Avg	0.5750	1.7253	0.5750	1.6335	0.5738	1.6249	0.5689	1.6170	
	MSE	0.0204	0.2508	0.0102	0.0249	0.0101	0.0227	0.0096	0.0211	

**Table 5:** Confidence intervals and coverage probabilities of NR and MCMC methods. (U: upper, L: lower, IL: interval length, CP: coverage probability.)

$T$	$R$		NR				MCMC: Informative				MCMC: Non-Informative			
			L	U	IL	CP	L	U	IL	CP	L	U	IL	CP
1	1	$\alpha$	0.2952	0.8166	0.5215	95.54	0.4063	0.6483	0.2420	99.90	0.3172	0.8193	0.5021	92.90
		$\beta$	0.4102	3.2764	2.8661	97.12	1.1124	2.1299	1.0175	99.74	0.7989	3.4874	2.6885	92.80
	2	$\alpha$	0.2972	0.7585	0.4614	95.02	0.4027	0.6526	0.2498	99.88	0.3232	0.7854	0.4622	94.64
		$\beta$	0.6301	2.6658	2.0357	95.78	1.0944	2.1159	1.0215	99.98	0.7981	2.7963	1.9982	94.52
	3	$\alpha$	0.3203	0.7668	0.4466	94.58	0.4067	0.6481	0.2414	99.90	0.3371	0.7796	0.4424	93.50
		$\beta$	0.6540	2.8541	2.2001	96.38	1.1111	2.1114	1.0003	99.96	0.8604	2.9774	2.1170	92.98
2	1	$\alpha$	0.2952	0.8167	0.5215	95.54	0.4063	0.6483	0.2420	99.90	0.3172	0.8193	0.5021	92.90
		$\beta$	0.4103	3.2764	2.8661	97.12	1.1124	2.1299	1.0175	99.74	0.7990	3.4875	2.6885	92.80
	2	$\alpha$	0.3161	0.7400	0.4239	94.92	0.4044	0.6491	0.2447	99.66	0.3375	0.7609	0.4235	94.30
		$\beta$	0.7199	2.5624	1.8426	95.68	1.1020	2.0980	0.9959	99.88	0.8484	2.6496	1.8013	93.82
	3	$\alpha$	0.3258	0.7695	0.4436	94.86	0.4074	0.6481	0.2407	99.86	0.3424	0.7820	0.4396	93.72
		$\beta$	0.6702	2.8649	2.1947	97.20	1.1140	2.1110	0.9970	99.86	0.8740	2.9859	2.1119	93.46

## 6. REAL DATA EXAMPLE

We consider a data set reported by [5] representing the strength measured in GigaPascal (GPa) for single carbon fibres, and impregnated 1000-carbon fibre tows. Single fibres were tested under tension at gauge lengths of 10 mm. This data was analyzed by [3] considering a hybrid censoring scheme for the Weibull distribution. Following [3], we analyze this data set using two-parameter Weibull distribution after subtracting 1.75. The authors recorded that the validity of the Weibull model based on the Kolmogorov–Smirnov (K–S) test is full-filled, namely,  $K-S = 0.072$  and  $p\text{-value} = 0.885$ .

To compute the Bayes estimates, since we have no prior information about the unknown parameters, we assume the non-informative priors by setting  $a = b = c = d = 0$ . Taking  $m = 40$  and  $T = 2$ , we use the following schemes:

- Scheme 1:  $R = (0^{39}, 23)$ ;
- Scheme 2:  $R = (23, 0^{39})$ ;
- Scheme 3:  $R = (2, 0^{10}, 2^3, 0^{10}, 2^3, 0^{10}, 3^3)$ .

We have produced 60000 MCMC samples and the first 10000 of them are considered as the burn-in sample. We have provided the histograms of the samples for each parameter in Figures 1–2 and also some diagnostics showing the efficiency of the MCMC algorithm in Figures 3–5. The acceptance rate after the burn-in sample is close to 0.36 and it is stable. Therefore, it can be said that the MCMC algorithm works well.

In SPT estimates, since we don't have any prior information about parameters, we use the Bayes estimates as an estimated prior information. Then we substitute them in the SPT formulae as  $\hat{\theta}_{\text{SPT}} = \lambda\theta_0 + (1-\lambda)\hat{\theta}_{\text{Bayes}} I(W_D < \chi_1^2(\lambda))$  by setting  $\lambda = 0.5$  and  $\alpha = 0.05$ .



All the estimation methods considered in this paper are applied to this data and the estimated parameter values are reported in Table 6. We observe that the estimated values of  $\alpha$  and  $\beta$  based on all the methods are closer to each other. Further, it can be seen that the Bayes estimates based on the two different methods are quite closer to each other which also show the stability of the MCMC algorithm. Moreover, asymptotic confidence intervals of NR method and HPD intervals of MCMC method are given in Table 7. According to this table, we can say that NR confidence intervals are mostly wider than the ones obtained via MCMC. This situation is also coincident with the simulation results.

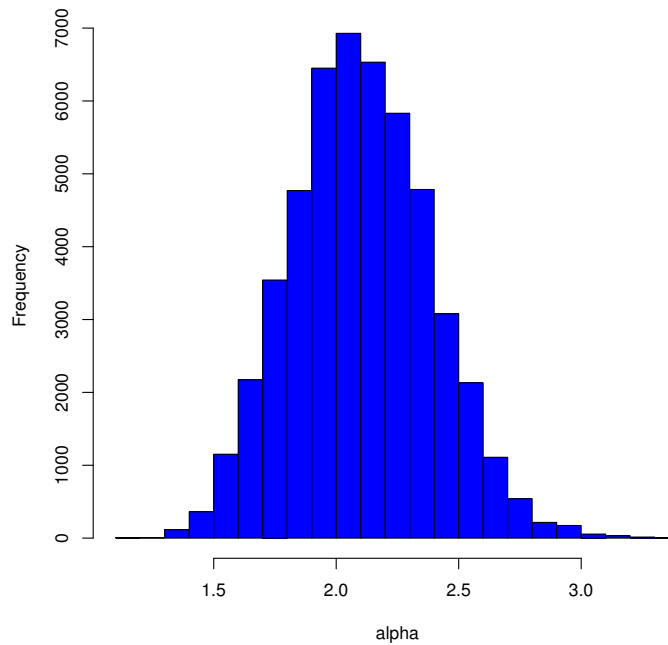


Figure 1: Histogram of the MCMC samples of the parameter  $\alpha$ .

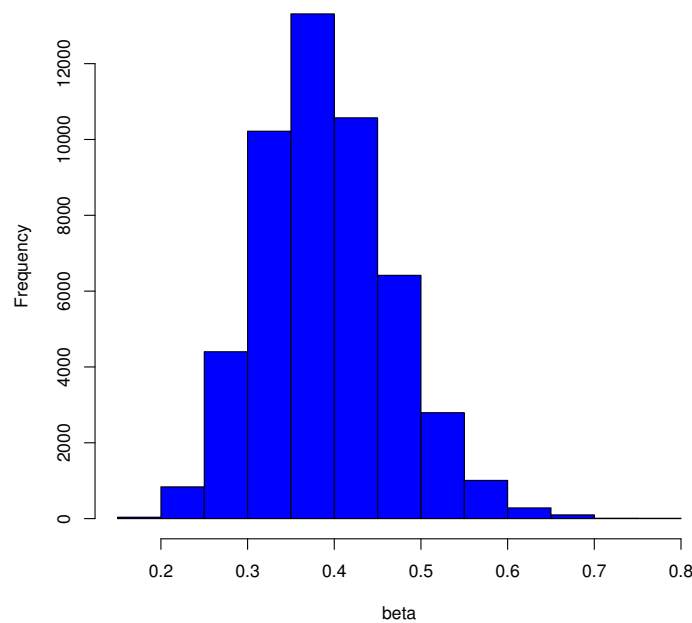
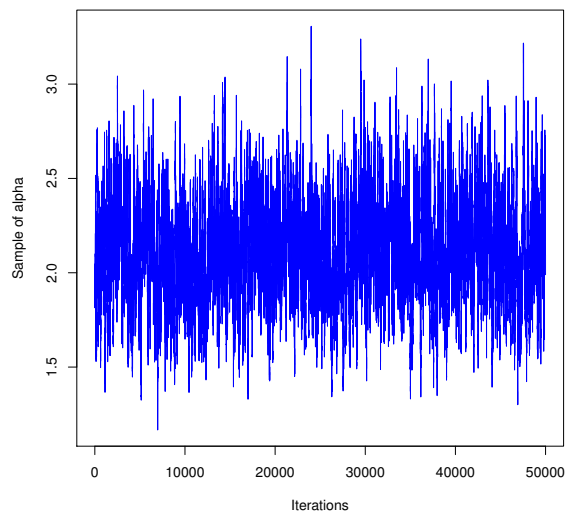
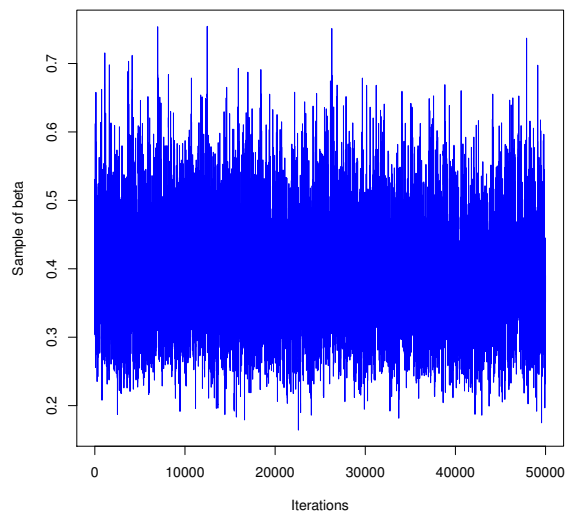


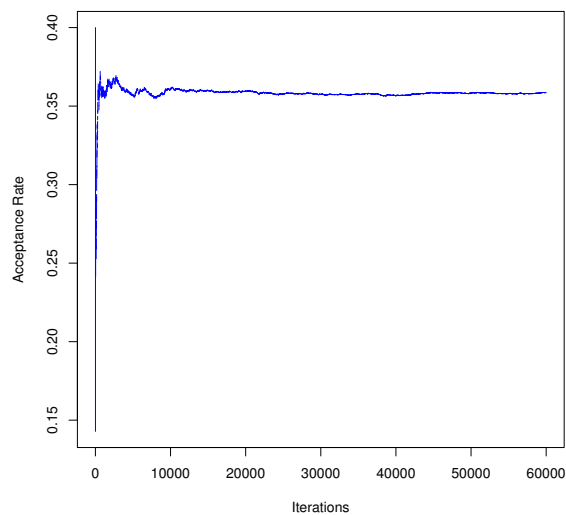
Figure 2: Histogram of the MCMC samples of the parameter  $\beta$ .



**Figure 3:** MCMC samples of the parameter  $\alpha$  vs iterations.



**Figure 4:** MCMC samples of the parameter  $\beta$  vs iterations.



**Figure 5:** Acceptance rate of MCMC samples.

**Table 6:** Estimation values of listed methods for Carbon Fibre data.

	Scheme 1		Scheme 2		Scheme 3	
	$\alpha$	$\beta$	$\alpha$	$\beta$	$\alpha$	$\beta$
MLE Method						
NR	2.2542	0.3980	2.3058	0.3918	2.1169	0.3884
EM	2.2641	0.3975	2.2952	0.3986	2.1128	0.3908
SEM	2.2515	0.3981	2.3046	0.3922	2.1304	0.3892
Tierney–Kadane Method						
SEL	2.2505	0.3991	2.3041	0.3942	2.1104	0.3899
LINEX ( $\nu = -0.5$ )	2.2764	0.4005	2.3310	0.3963	2.1303	0.3914
LINEX ( $\nu = 0.5$ )	2.2260	0.3978	2.2793	0.3924	2.0915	0.3885
GEL ( $\kappa = -0.5$ )	2.2393	0.3958	2.2929	0.3893	2.1012	0.3863
GEL ( $\kappa = 0.5$ )	2.2169	0.3890	2.2704	0.3793	2.0827	0.3792
MCMC Method						
SEL	2.2496	0.3980	2.3042	0.3933	2.1028	0.3915
LINEX ( $\nu = -0.5$ )	2.2735	0.3994	2.3288	0.3953	2.1232	0.3929
LINEX ( $\nu = 0.5$ )	2.2261	0.3967	2.2802	0.3914	2.0828	0.3900
GEL ( $\kappa = -0.5$ )	2.2390	0.3947	2.2937	0.3885	2.0932	0.3878
GEL ( $\kappa = 0.5$ )	2.2176	0.3880	2.2725	0.3788	2.0739	0.3805
Shrinkage Method						
NR	2.2524	0.3985	2.3049	0.3930	2.1137	0.3892
SEL	2.2505	0.3991	2.3041	0.3942	2.1104	0.3899
LINEX ( $\nu = 0.5$ )	2.2634	0.3998	2.3175	0.3953	2.1204	0.3906
GEL ( $\kappa = 0.5$ )	2.2449	0.3974	2.2985	0.3917	2.1058	0.3881

**Table 7:** Confident intervals and interval lengths of NR and MCMC methods for Carbon Fibre data (U: upper, L: lower, IL: interval length).

Scheme	Method	$\alpha$			$\beta$		
		L	U	IL	L	U	IL
1	NR	1.6321	2.8764	1.2443	0.2539	0.5420	0.2880
	MCMC	1.6668	2.8725	1.2057	0.2682	0.5540	0.2857
2	NR	1.6740	2.9376	1.2636	0.2175	0.5660	0.3485
	MCMC	1.7418	2.9404	1.1986	0.2408	0.5834	0.3426
3	NR	1.5703	2.6636	1.0933	0.2417	0.5351	0.2933
	MCMC	1.5744	2.6737	1.0993	0.2584	0.5558	0.2974

---

## 7. CONCLUSIVE REMARKS

---

In this paper, we discussed the estimation of parameters of Weibull distribution under type-I progressively hybrid censoring scheme using both classical and Bayesian strategies. Namely, MLE is obtained using NR, EM and SEM algorithms and Bayesian estimators are computed via T–K approximation and MCMC method under SEL, LINEX and GEL loss functions. We have also proposed the shrinkage preliminary test estimators based on NR and T–K with informative priors using equal weights on the prior information and the sample information.

A real data application and extensive Monte Carlo simulations have been considered to compare the estimators in terms of MSE and Avg and also we compared the lengths of CIs and CPs. According to the results, EM algorithm beats the other ML estimates. However, we observed that both the T–K and MCMC methods perform quite closely. Finally, we found out that shrinkage preliminary test estimates have satisfactory performances in the presence of having proper prior information.

---

## ACKNOWLEDGMENTS

---

This paper was written while Dr. Yasin Asar visited McMaster University and he was supported by The Scientific and Technological Research Council of Turkey (TUBITAK), BIDEB-2219 Postdoctoral Research Program, Project No.: 1059B191700537.

---

## REFERENCES

---

- [1] AHMED, S.E. (2014). *Penalty, Shrinkage And Pretest Strategies: Variable Selection and Estimation*, Springer, New York.
- [2] AHMED, S.E. and SALEH, A.M.E. (1990). Estimation strategies for the intercept vector in a simple linear multivariate normal regression model, *Computational Statistics & Data Analysis*, **10**(3), 193–206.
- [3] ASGHARZADEH, A.; VALIOLLAHI, R. and KUNDU, D. (1990). Prediction for future failures in Weibull distribution under hybrid censoring, *Journal of Statistical Computation and Simulation*, **85**(4), 824–838.
- [4] ASL, M.N.; BELAGHI, R.A. and BEVRANI, H. (2018). Classical and Bayesian inferential approaches using Lomax model under progressively type-I hybrid censoring, *Journal of Computational and Applied Mathematics*, **343**, 397–412.
- [5] BADER, M.G. and PRIEST, A.M. (1982). *Statistical aspects of fibre and bundle strength in hybrid composites*. In “Progress in Science and Engineering of Composites”, ICCM-IV, Tokyo, 1129–1136.
- [6] BALAKRISHNAN, N. (2007). Progressive censoring methodology: an appraisal, *Test*, **16**(2), 211–259.
- [7] BALAKRISHNAN, N. and AGGARWALA, R. (2000). *Progressive Censoring: Theory, Methods, and Applications*, Springer, New York.
- [8] BALAKRISHNAN, N. and CRAMER, E. (2014). *The Art Of Progressive Censoring: Applications To Reliability And Quality. Statistics For Industry And Technology*, Birkhäuser, New York.
- [9] BALAKRISHNAN, N. and KATERI, M. (2008). On the maximum likelihood estimation of parameters of Weibull distribution based on complete and censored data, *Statistics & Probability Letters*, **78**(17), 2971–2975.
- [10] BALAKRISHNAN, N. and KUNDU, D. (2013). Hybrid censoring: models, inferential results and applications, *Computational Statistics & Data Analysis*, **57**(1), 166–209.
- [11] BANCROFT, T.A. (1944). On biases in estimation due to the use of preliminary tests of significance, *The Annals of Mathematical Statistics*, **15**(2), 190–204.

- [12] BANERJEE, A. and KUNDU, D. (2008). Inference based on type-II hybrid censored data from a Weibull distribution, *IEEE Transactions on Reliability*, **57**(2), 369–378.
- [13] CHEN, S. and BHATTACHARYA, G.K. (1988). Exact confidence bounds for an exponential parameter under hybrid censoring, *Communications in Statistics – Theory and Methods*, **17**, 1857–1870.
- [14] CHEN, M.H. and SHAO, Q.M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals, *Journal of Computational and Graphical Statistics*, **8**(1), 69–92.
- [15] CHILDS, A.; CHANDRASEKAR, B. and BALAKRISHNAN, N. (2008). *Exact likelihood inference for an exponential parameter under progressive hybrid censoring schemes*. In “Statistical Models and Methods for Biomedical and Technical Systems” (F. Vonta, M. Nikulin, N. Limnios, and C. Huber-Carol, Eds.), Birkhäuser Boston, 319–330.
- [16] DEMPSTER, A.P.; LAIRD, N.M. and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1–22.
- [17] DEY, S.; SINGH, S.; TRIPATHI, Y.M. and ASGHARZADEH, A. (2016). Estimation and prediction for a progressively censored generalized inverted exponential distribution, *Statistical Methodology*, **32**, 185–202.
- [18] DIEBOLT, J. and CELEUX, G. (1993). Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions, *Stochastic Models*, **9**(4), 599–613.
- [19] EPSTEIN, B. (1954). Truncated life tests in the exponential case, *The Annals of Mathematical Statistics*, **25**(3), 555–564.
- [20] HUANG, S.R. and WU, S.J. (2012). Bayesian estimation and prediction for Weibull model with progressive censoring, *Journal of Statistical Computation and Simulation*, **82**(11), 1607–1620.
- [21] JIA, X.; NADARAJAH, S. and GUO, B. (2018). Exact inference on Weibull parameters with multiply type-I censored data, *IEEE Transactions on Reliability*, **67**(2), 432–445.
- [22] JUDGE, G.G. and BOCK, M.E. (1978). *The Statistical Implications of Pre-test and Stein-rule Estimators in Econometrics*, North-Holland, Amsterdam.
- [23] KIBRIA, B.M.G. and SALEH, A.M.E. (2010). Preliminary test estimation of the parameters of exponential and Pareto distributions for censored samples, *Statistical Papers*, **51**(4), 757–773.
- [24] KIM, C.; JUNG, J. and CHUNG, Y. (2011). Bayesian estimation for the exponentiated Weibull model under type-II progressive censoring, *Statistical Papers*, **52**(1), 53–70.
- [25] KUNDU, D. (2008). Bayesian inference and life testing plan for the Weibull distribution in presence of progressive censoring, *Technometrics*, **50**(2), 144–154.
- [26] KUNDU, D. and JOARDER, A. (2011). Analysis of type-II progressively hybrid censored data, *Computational Statistics Data Analysis*, **50**(10), 2509–2528.
- [27] LAWLESS, J.F. (2011). *Statistical Models And Methods For Lifetime Data*, Vol. 362, John Wiley & Sons.
- [28] LIN, C.T.; CHOU, C.C. and HUANG, Y.L. (2012). Inference for the Weibull distribution with progressive hybrid censoring, *Computational Statistics & Data Analysis*, **56**(3), 451–467.
- [29] LIN, C.T. and HUANG, Y.L. (2012). On progressive hybrid censored exponential distribution, *Journal of Statistical Computation and Simulation*, **82**(5), 689–709.
- [30] LIN, C.T.; NG, H.K.T. and CHAN, P.S. (2009). Statistical inference of type-II progressively hybrid censored data with Weibull lifetimes, *Communications in Statistics – Theory and Methods*, **38**(10), 1710–1729.
- [31] LOUIS, T.A. (1982). Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)*, **44**(2), 226–233.

- [32] MOKHTARI, E.B.; RAD, A.H. and YOUSEFZADEH, F. (2011). Inference for Weibull distribution based on progressively type-II hybrid censored data, *Journal of Statistical Planning and Inference*, **141**(8), 2824–2838.
- [33] NG, H.K.T.; CHAN, P.S. and BALAKRISHNAN, N. (2002). Estimation of parameters from progressively censored data using EM algorithm, *Computational Statistics & Data Analysis*, **39**(4), 371–386.
- [34] NG, H.K.T.; CHAN, P.S. and BALAKRISHNAN, N. (2004). Optimal progressive censoring plans for the Weibull distribution, *Technometrics*, **46**(4), 470–481.
- [35] PANDEY, M. (1983). Shrunk estimators of Weibull shape parameter in censored samples, *IEEE Transactions on Reliability*, **32**(2), 200–203.
- [36] PANDEY, M. and SINGH U.S. (1983). Shrunk estimators of Weibull shape parameter from type-II censored samples, *IEEE Transactions on Reliability*, **42**(1), 81–86.
- [37] PARSIAN, A. and KIRMANI, S.N.U.A. (2002). *Estimation under LINEX loss function*. In “Handbook of Applied Econometrics and Statistical Inference”, Vol. 165, 53–76.
- [38] R CORE TEAM (2020). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- [39] SALEH, A.M.E. (2006). *Theory of Preliminary Test and Stein-type Estimations with Applications*, Wiley, New York.
- [40] SALEH, A.M.E. and KIBRIA, B.M.G. (1993). Performance of some new preliminary test ridge regression estimators and their properties, *Communications in Statistics – Theory and Methods*, **22**(10), 2747–2764.
- [41] SINGH, J. and BHATKULIKAR, S.G. (1978). Shrunk estimation in Weibull distribution, *Sankhya: The Indian Journal of Statistics, Series B*, 382–393.
- [42] SINGH, H.P. and SHUKLA, S.K. (1993). Estimation in the two-parameter Weibull distribution with prior information, *IAPQR Transactions*, **25**(2), 107–118.
- [43] THOMPSON, J.R. (1968). Some shrinkage techniques for estimating the mean, *Journal of the American Statistical Association*, **63**(321), 113–122.
- [44] TIERNEY, L. and KADANE, J.B. (1986). Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association*, **81**(393), 82–86.
- [45] VARIAN, H.R. (1975). *A Bayesian approach to real estate assessment*. In “Studies in Bayesian Econometric and Statistics” (in honor of Leonard J. Savage), 195–208.
- [46] WEI, G.C. and TANNER, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms, *Journal of the American Statistical Association*, **85**(411), 699–704.

---

---

## The Destructive Zero-Inflated Power Series Cure Rate Models for Carcinogenesis Studies

---

---

Authors: RODRIGO R. PESCIM  

– Department of Statistics, Londrina State University,  
Londrina, PR, Brazil  
[rrpescim@uel.br](mailto:rrpescim@uel.br)

ADRIANO K. SUZUKI 

– Department of Applied Mathematics and Statistics, University of São Paulo,  
São Carlos, SP, Brazil  
[suzuki@icmc.usp.br](mailto:suzuki@icmc.usp.br)

GAUSS M. CORDEIRO 

– Department of Statistics, Federal University of Pernambuco,  
Recife, PE, Brazil  
[gausscordeiro@gmail.com](mailto:gausscordeiro@gmail.com)

EDWIN M. M. ORTEGA 

– Department of Exact Science, University of São Paulo,  
Piracicaba, SP, Brazil  
[edwin@usp.br](mailto:edwin@usp.br)

Received: January 2020

Revised: January 2021

Accepted: January 2021

### Abstract:

- In this paper, we propose a new flexible survival model called the destructive zero-inflated power series cure rate model. This new model describes a realistic interpretation for the biological mechanism of the occurrence of the event of interest in studies of carcinogenesis in the presence of the competing latent causes. The maximum likelihood method is used for estimating the model parameters. For different sample sizes, various scenarios are simulated to evaluate the precision of the estimates. The usefulness of the new cure rate survival model is illustrated by means of a cutaneous melanoma data set.


### Keywords:

- *cure rate model; destructive cure model; simulation study; survival analysis; Weibull distribution; zero inflated power series distribution.*

### AMS Subject Classification:

- 62N02, 62F99.

---

 Corresponding author.

---

## 1. INTRODUCTION

---

Cancer is the name given to a set of more than 100 diseases that have in common the disordered growth of cells, which invade tissues and organs. Dividing rapidly, these cells tend to be very aggressive and uncontrollable, determining the formation (carcinogenesis process) of malignant tumors, which can spread to other regions of the body. The carcinogenesis process (cancer formation), in general, occurs slowly and may take several years for a cancer cell proliferate and give rise to a visible tumor. That process goes through several stages (initiation of a tumor, promotion and progression) before reaching the tumor. Statistics show that cancer is one of the most important public health concern around the world and for this reason it is crucial to estimate its prevalence, incidence, and mortality/survival rates [17]. An overview of descriptive cancer data on this disease is a first step to appreciate control measures and preventive interventions in a global context of progressive cancer burden [21].

Cutaneous malignant melanoma, a type of skin cancer that originates in melanocytes (cells that produce melanin, a substance that determines skin color), is a tumor whose incidence is increasing dramatically in persons with light-colored skin in all parts of the world. As with other cancers, there are several causes of malignant melanoma formation such as environmental (imminent exposure to ultraviolet radiation), genetic and immunological factors. In most studies, the incidence doubles every 6 to 10 years. In years of potential life loss, melanoma is second to adult leukemia, as it affects younger individuals, causing a major public health problem [1]. According to World Health Organization, about 132,000 new cases of cutaneous melanoma are diagnosed worldwide each year. In particular, the American Cancer Society estimated that there will be 96,000 new cases of cutaneous melanoma in the United States and 7,000 deaths from this disease in 2020. In addition, approximately 57,000 new cases of invasive cutaneous melanoma will occur in men and 39,000 new cases in women in 2020. On the other hand, there has been a great improvement in the survival of patients with cutaneous melanoma, mainly due to the early detection of the tumor, in recent years. In general, it is taken that “cured” is related to survival beyond 5 years for patients with melanoma. This may be due to earlier diagnosis, when tumors are still at a thinner depth, as well as improved treatment and surgical techniques [8].

Survival models with a cure fraction for cutaneous melanoma data have played an important role in survival analysis in recent years. These types of survival models cover situations in which there are persons not susceptible to the occurrence of the event of interest. Consequently, a fraction (or proportion) of these individuals are not expected to experience the event of interest, that is, these individuals are considered not susceptible or “cured” in the survival analysis context. The proportion of cured individuals is denoted by the cure fraction. Cure rate models have the main purpose to include in their formulation the possibility of estimating the cure rate and they have been widely studied by several authors and used for modeling time-to-event data for various types of cancers, including breast cancer, non-Hodgkins lymphoma, leukemia, prostate cancer and melanoma.

The most popular type of cure rate models are the mixture (or Berkson–Gage) cure model [2] and the promotion time cure rate model ([27] and [7]). While the Berkson–Gage cure model is based on the assumption that only one cause is responsible for the occurrence of the event of interest, that is, the unknown number of causes of the event of interest is



assumed to be a Bernoulli random variable, in the promotion time cure model the number of causes follows a Poisson distribution. In a biological context, the occurrence of the event of interest might be due to one of many competing causes [14], with the number of causes and the distribution of survival times associated with each cause [11] being unknown which leads to a latent competing causes structure. In this sense, the event of interest can be the death of a patient or a tumor recurrence, which can happen because of unknown competing causes [21]. These latent competing causes can be assigned to metastasis-component tumor cells left active after an initial treatment. A metastasis-component tumor cell is a tumor cell having the potential of metastasizing [27]. The statistical literature on distributions which accommodate different numbers of latent competitors have as the main works in the books by [19] and [16] as well as the review paper by [26] and the papers of [10], [29], [6], [24] and [4] can be mentioned as key references.

More recently, [23] extended the works of [27] and [7] by considering a cure rate model (also known as a destructive weighted Poisson cure rate model) to deal with the assumption that each initiated cell (competing cause) becomes cancerous with probability one. They argue that this development is a much more realistic alternative to the cure rate model in explaining the biological mechanism underlying the occurrence of the event in presence of a cure fraction. This is because the proposed cure rate survival model presumes that the original number of lesions, or altered cells are not repaired or eliminated after some intensive treatment, and this group (which is represented by a variable) of unrepaired cells (or latent factors) are potentially competing to give rise to a tumor, or risk of failure. Figure 1 represents the destructive model in a diagram form.

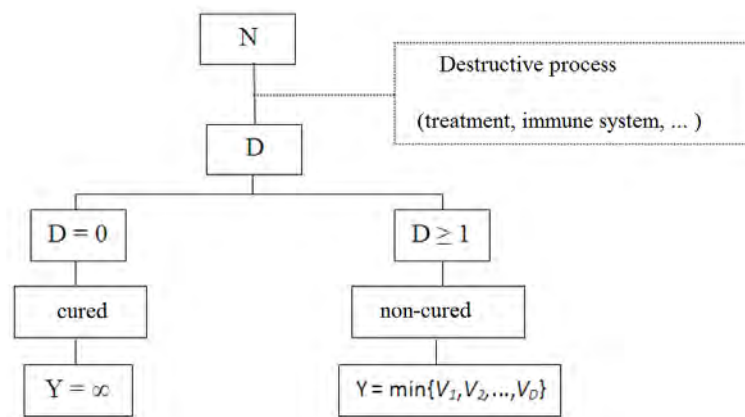


Figure 1: Representation of the proposed destructive model in a diagram form.

However, there is an amount (or proportion) of cells that have not been initiated (normal cells), which includes repaired cells, that are not being explained properly by those cure rate models that consider the number of initiated cells related to the occurrence of a tumor being a random variable that follows the power series family of distributions which has as special cases the Poisson, Bernoulli, geometric, negative binomial, etc.

In a biological context, it is noted that there is a much larger number of cells that are not initiated (normal cells) than cells that are initiated (and consequently become malignant

cells), which leads to an “excess of not initiated cells” (or “excess number of zero counts”) in relation to cells which are lesioned.

In this sense, the excess of zeros (not initiated cells) can be explained in terms of zero-inflated models as follows:

1. First, there exist an amount of not initiated cells (zeros) which have never experienced any type of alterations or lesions (structural zeros).
2. On the other hand, there exist an amount of not initiated cells which have experienced alterations (or lesions), but those cells were repaired (sampling zeros).

Therefore, it is maybe desirable to construct tractable statistical models that can adequately incorporate a biological mechanism for the initiation process of carcinogenesis, and this is the main motivation for the present research work.

Here, we introduce a new cure rate survival model which extends the works of [23] and [4] by incorporating a structure to estimate the proportion of not initiated cells (those one that have never been altered/lesioned and those one that have been repaired). To create such structure, we use the concept of zero-inflated models by considering an extension of the discrete power series distributions by including an additional parameter  $\pi$ . Its interpretation is related to the proportion of repaired cells by means a repair system of the body. In this approach, we assume that the number of initiated cells follow the zero-inflated power series (ZIPS) [15] distribution, which is a suitable choice for modelling data sets that possesses excess of zeros and overdispersion. Furthermore, it provides a realistic interpretation related to the biological mechanism of the occurrence of the event of interest. Also, it includes a process of destruction of tumor cells after an initial treatment ([23], [3] and [22]).

The rest of the paper is outlined as follows. In Section 2, we formulate the new cure rate model. Some special models are reported in Section 3. Inference based on maximum-likelihood (ML) is discussed in Section 4. In Section 5, we perform a simulation study to verify the precision of the estimates of the model parameters. An application to a real data set on cutaneous melanoma is addressed in Section 6. Finally, Section 7 provides some concluding remarks.

---

## 2. MODEL FORMULATION

---

Let  $N$  be an unobservable (latent) random variable which follows the zero-inflated power series (ZIPS) distribution, denoting the initial number of initiated cells related to the occurrence (or recurrence) of a tumor for an individual in a population, with probability mass function (pmf)

$$(2.1) \quad P[N=n] = \begin{cases} \pi + (1-\pi) \frac{a_0}{g(\theta)}, & \text{for } n = 0, \\ (1-\pi) \frac{a_n \theta^n}{g(\theta)}, & \text{for } n = 1, 2, 3, \dots, \end{cases}$$

where  $0 < \pi < 1$ ,  $a_n > 0$  ( $a_n$  depends only on  $n$ ) and  $g(\theta) = \sum_{n=0}^{\infty} a_n \theta^n$  is a positive, finite and differentiable function.

Here, the parameter  $\pi$  is interpreted as the proportion of cells that have never experienced alterations (or modifications) in their genes, while the interpretation for the quantity  $(1 - \pi)$  refers to the proportion of cells which have been repaired from a body repair mechanism.

Also, we note that if  $\pi = 0$ , the ZIPS distribution reduces to the power series (PS) distribution proposed by [20]. Some important well-known discrete distributions belong to this family of distributions. For example, if  $g(\theta) = (1 + \theta)^m$  and  $m$  is positive integer, Equation (2.1) becomes the zero-inflated binomial (ZIBin) distribution. If  $g(\theta) = \exp(\theta)$ , it defines the zero-inflated Poisson (ZIP) distribution. Further, if  $g(\theta) = (1 + \theta)^{-\phi}$ ,  $\phi > 0$  and  $0 < \theta < 1$ , the zero-inflated negative binomial (ZINB) distribution is obtained from Equation (2.1), among others.

For the ZIPS random variable  $N$ , the probability generating function (pgf) is

$$(2.2) \quad \mathbb{A}_N(z) = \pi + (1 - \pi) \frac{g(\theta z)}{g(\theta)}, \quad \text{for } 0 \leq z \leq 1,$$

where the ratio  $g(\theta z)/g(\theta)$  is the pgf of the PS distribution. For more details, see [20].

The first consequence of a prolonged treatment (destructive process) is the possible formation of precancerous lesions into the genome of the cells. These cells are denoted as malignant cells. Given  $N = n$ , let  $X_j$ ,  $j = 1, 2, \dots, n$  be independent random variables (independent of  $N$ ) following a Bernoulli distribution with success probability  $p$  indicating presence of the  $j$ -th lesion. The pgf of the Bernoulli random variable  $X_j$  can be expressed as

$$(2.3) \quad \mathbb{A}_{X_j}(z) = 1 - p(1 - z), \quad \text{for } 0 \leq z \leq 1.$$

The variable  $D$ , representing the total number of malignant cells among the  $N$  initial cells (competing causes) which are not eliminated by the treatment is defined as

$$(2.4) \quad D = \begin{cases} X_1 + X_2 + \dots + X_N, & \text{if } N > 0, \\ 0, & \text{if } N = 0, \end{cases}$$

where  $D \leq N$ . The idea involved in (2.4) was suggested by [28] considering that the initial  $N$  cells are primary initiated malignant cells, where  $X_j$  in (2.4) represents the number of living malignant cells that are descendants of the  $j$ -th initiated malignant cell during some time interval. In this case,  $D$  denotes the total number of living malignant cells at some specific time. The time to event for the  $j$ -th competing cause is represented by  $V_j$ ,  $j = 1, \dots, D$ . Conditional on  $D$ , the  $V_j$ 's are assumed iid with cumulative distribution function  $F(t)$  and survival function  $S(t) = 1 - F(t)$ . Also, we note that the total number of malignant cells  $D$  and the time  $V_j$  are not observable.

As pointed out by [3], in the competing causes scenario, the number of unrepaired lesions  $D$  in (2.4) and the time  $V$  taken to transform these lesions into a detectable tumor are both not observable (latent variables). In this context, we denote  $V$  a progression time. Thus, the observed time to the event of interest (the patient's death) is defined by the following random variable

$$(2.5) \quad Y = \min\{V_1, \dots, V_D\}$$

for  $D \geq 1$ , and  $Y = \infty$  if  $D = 0$ , which leads to a proportion  $p_0$  of the population which is called the cured fraction.

Under this setup, [23] showed that the survival function for the population of the random variable  $Y$  in (2.5) has the form

$$S_{\text{pop}}(y) = P[Y \geq y] = \mathbb{A}_D(S(y)) = \sum_{d=0}^{\infty} P[D=d] \{S(y)\}^d = \mathbb{A}_N(\mathbb{A}_{X_j}(S(y))),$$

where  $S(\cdot)$  is the survival function for non-cured population and  $\mathbb{A}_D(\cdot)$  is the pgf for the variable  $D$ . Combining (2.2) and (2.3), the survival function of the observable lifetime of the event of interest can be expressed as

$$(2.6) \quad S_{\text{pop}}(y) = \pi + (1 - \pi) \frac{g(\theta[1 - pF(y)])}{g(\theta)},$$

where  $F(y) = 1 - S(y)$ . Hereafter, Equation (2.6) is referred to as the destructive zero inflated power series (DZIPS) cure rate model. This model includes two important special cases: For  $\pi = 0$ , it reduces to the destructive power series (DPS) cure rate model and if  $\pi = 0$  in addition  $p = 1$ , it gives the power series (PS) cure rate model ([4]).

From model (2.6), the proportion  $p_0$  of cured individuals in the population is

$$p_0 = \lim_{y \rightarrow \infty} S_{\text{pop}}(y) = \pi + (1 - \pi) \frac{g(\theta(1 - p))}{g(\theta)}.$$

The density function associated with (2.6) can be expressed as

$$(2.7) \quad f_{\text{pop}}(y) = -\frac{dS_{\text{pop}}(y)}{dy} = -\left[ (1 - \pi) \frac{g'(\theta[1 - pF(y)])}{g(\theta)} \right],$$

where  $g'(\cdot) = dg(\cdot)/d\theta$ ,  $f(y) = dF(y)/dy$  denotes the proper density function of the time  $V$  to the event in (2.6). Note that the function  $f_{\text{pop}}(y)$  is a proper function, whereas  $S_{\text{pop}}(y)$  is not a proper survival function.

### 3. SPECIAL CASES OF THE DZIPS CURE MODEL

In this section, we present some specific models that arise from the ZIPS model formulation. Here, we consider situations where  $N$  is a random variable which follows the zero-inflated Poisson, zero-inflated binomial, zero-inflated negative binomial, and zero-inflated geometric distributions.

#### 3.1. The destructive zero-inflated Poisson (DZIP) cure model

If we consider  $a_n = \frac{1}{n!}$  and  $g(\theta) = \exp(\theta)$  in (2.1), the number of initiated cells  $N$  follows a ZIP distribution with  $\theta > 0$  and  $\pi \in (0, 1)$  and pmf

$$(3.1) \quad P_{\text{ZIP}}[N = n] = \begin{cases} \pi + (1 - \pi) e^{-\theta}, & \text{for } n = 0, \\ (1 - \pi) \frac{e^{-\theta} \theta^n}{n!}, & \text{for } n = 1, 2, 3, \dots \end{cases}$$

The corresponding survival function of the DZIP cure model is

$$(3.2) \quad S_{\text{pop}}(y) = \pi + (1 - \pi) e^{-\theta p F(y)}.$$

The cure rate is  $p_0 = \pi + (1 - \pi) e^{-\theta p}$ , and the corresponding density function takes the form

$$(3.3) \quad f_{\text{pop}}(y) = (1 - \pi) \theta p f(y) e^{-\theta p F(y)}.$$

There are some important special cases in (3.2). For  $\pi = 0$ , it follows the destructive Poisson cure model defined by Rodrigues *et al.* [23]. We introduce the zero-inflated Poisson cure model for  $p = 1$ , whereas for  $\pi = 0$  in addition  $p = 1$ , it follows the promotion time cure model studied by [27] and [7].

### 3.2. The destructive zero-inflated binomial (DZIBin) cure model

If we have  $a_n = \binom{m}{n}$  and  $g(\theta) = (1 + \theta)^m$  in (2.1), the number of initiated cells  $N$  follows a ZIBin distribution with parameters  $\frac{\theta}{1+\theta}$ ,  $\pi \in (0, 1)$  ( $m$  is a positive integer) and pmf

$$P_{\text{ZIBin}}[N = n] = \begin{cases} \pi + (1 - \pi) \left(\frac{1}{1 + \theta}\right)^m, & \text{for } n = 0, \\ (1 - \pi) \binom{m}{n} \left(\frac{\theta}{1 + \theta}\right)^n \left(\frac{1}{1 + \theta}\right)^{m-n}, & \text{for } n = 1, 2, 3, \dots \end{cases}$$

The survival function of the DZIBin cure model has the form

$$(3.4) \quad S_{\text{pop}}(y) = \pi + (1 - \pi) \left[1 - \frac{\theta p F(y)}{1 + \theta}\right]^m.$$

Here, the cure fraction is given by  $p_0 = \pi + (1 - \pi) \left[1 - \frac{\theta p}{1 + \theta}\right]^m$ . So, the density function of the DZIBin cure model can be expressed as

$$(3.5) \quad f_{\text{pop}}(y) = (1 - \pi) \frac{m \theta p f(y)}{1 + \theta} \left[1 - \frac{\theta p F(y)}{1 + \theta}\right]^{m-1}.$$

The DZIBin cure model in (3.4) with  $\pi = 0$  in addition to  $p = m = 1$  coincides with the mixture (Berkson–Gage) cure model pioneered by [2].

### 3.3. The destructive zero-inflated negative binomial (DZINB) cure model

If we consider  $a_n = \frac{\Gamma(\phi^{-1} + n)}{n! \Gamma(\phi^{-1})}$ ,  $g(\theta) = (1 - \theta)^{-1/\phi}$  and  $\theta = \frac{\eta \phi}{1 + \eta \phi}$  in (2.1), the number of initiated cells  $N$  follows a ZINB distribution with  $\eta > 0$ ,  $\phi \geq -1$ ,  $\eta \phi > 0$  and  $\pi \in (0, 1)$ , with pmf

$$P_{\text{ZINB}}[N = n] = \begin{cases} \pi + (1 - \pi) (1 + \eta \phi)^{-1/\phi}, & \text{for } n = 0, \\ (1 - \pi) \frac{\Gamma(\phi^{-1} + n)}{n! \Gamma(\phi^{-1})} \left(\frac{\eta \phi}{1 + \eta \phi}\right)^n (1 + \eta \phi)^{-1/\phi}, & \text{for } n = 1, 2, 3, \dots \end{cases}$$

where  $\Gamma(\cdot)$  denotes the gamma function.

The survival function of the DZINB cure model has the form

$$(3.6) \quad S_{\text{pop}}(y) = \pi + (1 - \pi) [1 + \eta \phi p F(y)]^{-1/\phi},$$

the cure fraction is  $p_0 = \pi + (1 - \pi) [1 + \eta \phi p]^{-1/\phi}$ , and the associated density function becomes

$$(3.7) \quad f_{\text{pop}}(y) = (1 - \pi) \eta p f(y) [1 + \eta \phi p F(y)]^{-(1/\phi)-1}.$$

The DZINB cure model in (3.6) with  $\pi = 0$  reduces to the destructive negative binomial model [4], whereas the negative binomial cure rate model [5] is a special case of (3.6) when  $\pi = 0$  and  $p = 1$ .

### 3.4. The destructive zero-inflated geometric (DZIG) cure model

Moreover, the destructive zero-inflated geometric (DZIG) cure rate model with parameter  $\theta = \eta/(1 + \eta)$  is one more important special case of (3.6) when  $\phi = 1$  leading to

$$(3.8) \quad S_{\text{pop}}(y) = \pi + (1 - \pi) [1 + \eta p F(y)]^{-1},$$

the cure fraction is  $p_0 = \pi + (1 - \pi) [1 + \eta p]^{-1}$  and the density function reduces to

$$(3.9) \quad f_{\text{pop}}(y) = (1 - \pi) \eta p f(y) [1 + \eta p F(y)]^{-2}.$$

## 4. INFERENCE AND ESTIMATION

Here, we consider the situation when the time to event of interest is not completely observed and is subject to right censoring. Let  $C_i$  denote the censoring time. We observe  $t_i = \min\{Y_i, C_i\}$  and  $\delta_i = 1$  if  $Y_i$  is the observed time to the event defined before and  $\delta_i = 0$  if it is right censored, for  $i = 1, \dots, n$ . Let  $\gamma$  represent the parameter vector of the distribution for the unobserved lifetime in (2.5). Here, we note that the DZIPS cure rate models in Section 3 are unidentifiable according to [18]. So, to overcome this problem, we propose to relate the model parameters  $p$  and  $\theta$  (or  $\eta$ ) to covariates  $\mathbf{x}_{i1} = (x_{i11}, x_{i12}, \dots, x_{i1p_1})^\top$  and  $\mathbf{x}_{i2} = (x_{i21}, x_{i22}, \dots, x_{i2p_2})^\top$ , respectively, without common elements and  $\mathbf{x}_{i2}$  without a column of intercepts. Here, the systematic components are

$$(4.1) \quad \log\left(\frac{p_i}{1 - p_i}\right) = \mathbf{x}_{i1}^\top \boldsymbol{\beta}_1 \quad \text{and} \quad \log(\theta_i) = \mathbf{x}_{i2}^\top \boldsymbol{\beta}_2,$$

where  $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \dots, \beta_{1p_1})^\top$  and  $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \dots, \beta_{2p_2})^\top$  represent the associated parameter vectors. A critical issue is the selection of covariates to be included in the link functions in (4.1). More precisely, given a link function and a set potential covariates, the problem is to find and fit the “best” model under a “selected” subset of covariates [3]. In fact, to choose which explanatory variables will be connected to the parameters  $p_i$  and  $\theta_i$  is not an easy task because it depends on several factors such as the type of cancer, the covariates available in the study, patient history, etc. It is always important to work together with the medical team to take any kind of decision. Moreover, for readers interested in this discussion, we suggest [12] and [9].

From  $n$  pairs of times and censoring indicators  $(y_1, \delta_1), \dots, (y_n, \delta_n)$ , the observed full likelihood function under non-informative censoring can be expressed as

$$(4.2) \quad L(\boldsymbol{\nu}, \mathbf{D}) \propto \prod_{i=1}^n \{f_{\text{pop}}(t_i; \boldsymbol{\nu})\}^{\delta_i} \{S_{\text{pop}}(t_i; \boldsymbol{\nu})\}^{1-\delta_i},$$

where  $\boldsymbol{\nu} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\gamma}^\top)^\top$ ,  $\mathbf{D} = (\mathbf{t}, \boldsymbol{\delta}, \mathbf{x}_1, \mathbf{x}_2)$ ,  $\mathbf{t} = (t_1, \dots, t_n)$ ,  $\mathbf{x}_1 = (\mathbf{x}_{11}, \dots, \mathbf{x}_{n1})$ ,  $\mathbf{x}_2 = (\mathbf{x}_{21}, \dots, \mathbf{x}_{n2})$ , and  $f_{\text{pop}}(\cdot; \boldsymbol{\nu})$  and  $S_{\text{pop}}(\cdot; \boldsymbol{\nu})$  are defined in Equations (2.7) and (2.6), respectively.

Next, we assume a Weibull distribution for the observed lifetime in (2.5) with cdf and pdf (for  $z > 0$ )

$$F(z; \boldsymbol{\gamma}) = 1 - \exp(-z^{\gamma_1} e^{\gamma_2}) \quad \text{and} \quad f(z; \boldsymbol{\gamma}) = \gamma_1 z^{\gamma_1-1} \exp(\gamma_2 - z^{\gamma_1} e^{\gamma_2}),$$

respectively,  $\boldsymbol{\gamma}^\top = (\gamma_1, \gamma_2)^\top$ ,  $\gamma_1 > 0$  and  $\gamma_2 > 0$ . The choice of the Weibull distribution is due to the fact that this lifetime distribution is a very popular model and it has been extensively used over the past decades for modeling data in reliability, engineering and biological studies. Also, the pdf and cdf of the Weibull distribution have closed-forms which provide simple expressions for its survival and hazard functions.

The ML estimation of the parameter vector  $\boldsymbol{\nu}$  can be implemented by numerical maximization of the log-likelihood function  $\ell(\boldsymbol{\nu}, \mathbf{D}) = \log L(\boldsymbol{\nu}, \mathbf{D})$  using R software. Further, confidence intervals and hypothesis tests can be based on the large sample normal distribution of the maximum likelihood estimator (MLE) with the variance-covariance matrix given by the inverse of the Fisher information. More specifically, under conditions that are fulfilled for the parameter vector  $\boldsymbol{\nu}$  in the interior of the parameter space but not on the boundary, the asymptotic distribution of  $\sqrt{n}(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu})$  is multivariate normal  $N_{p_1+p_2+2}(0, K(\boldsymbol{\nu})^{-1})$ , where  $K(\boldsymbol{\nu})$  is the information matrix. The asymptotic covariance matrix  $K(\boldsymbol{\nu})^{-1}$  of  $\hat{\boldsymbol{\nu}}$  can be approximated by the inverse of the  $(p_1 + p_2 + 2) \times (p_1 + p_2 + 2)$  observed information matrix  $-\ddot{\mathbf{L}}(\boldsymbol{\nu}, \mathbf{D})$ . The elements of the observed information matrix  $-\ddot{\mathbf{L}}(\boldsymbol{\nu}, \mathbf{D})$  are calculated numerically. The approximate multivariate normal distribution  $N_{p_1+p_2+2}(0, -\ddot{\mathbf{L}}(\boldsymbol{\nu}, \mathbf{D})^{-1})$  for  $\hat{\boldsymbol{\nu}}$  can be used in the classical way to construct approximate confidence regions for some parameters in  $\boldsymbol{\nu}$ . Also, we can use the likelihood ratio (LR) statistic for comparing some special models with the DZIPS regression model.

---

## 5. SIMULATION STUDY

---

In this section, we conduct a simulation study in order to evaluate some properties of the MLEs. For each individual  $i$  ( $i = 1, \dots, n$ ), the number of competing risks of the event of interest  $N$  is generated from the ZIP and ZINB distributions given in (3.1) and (3.6), respectively. We assume covariates  $x_{i11}$  and  $x_{i21}$  generated from a Bernoulli distribution with parameter 0.5 and exponential distribution with parameter one, respectively. Also, we consider the systematic components

$$(5.1) \quad \log\left(\frac{p_i}{1-p_i}\right) = \beta_{10} + \beta_{11} x_{i11} \quad \text{and} \quad \log(\psi_i) = \beta_{21} x_{i21},$$

where  $\psi_i$  is the parameter  $\theta_i$  and  $\eta_i$  in the DZIP and DZINB cure rate models, respectively.

We simulate from the DZIP cure fraction distribution with parameters  $\pi = 0.25$ ,  $\gamma_1 = 2$ ,  $\gamma_2 = -0.5$ ,  $\beta_{10} = -1$ ,  $\beta_{11} = 0.5$  and  $\beta_{21} = 1.25$ ; and from the DZINB distribution under two setups: the first assuming  $\pi = 0.25$ ,  $\gamma_1 = 2$ ,  $\gamma_2 = -0.5$ ,  $\phi = 1$ ,  $\beta_{10} = -1.5$ ,  $\beta_{11} = 0.75$  and  $\beta_{21} = 1.5$  (DZIG distribution), and the second with  $\pi = 0.25$ ,  $\gamma_1 = 2$ ,  $\gamma_2 = -0.5$ ,  $\phi = 1.5$ ,  $\beta_{10} = -0.5$ ,  $\beta_{11} = 1.5$  and  $\beta_{21} = 1.25$  (DZINB distribution). The censoring times are sampled from the uniform distribution in the  $(0, \tau)$  interval, where  $\tau$  controls the censoring proportion of the uncured population. Here, the proportions of censored observations are approximately 62%, 68% and 70%, respectively. The results are obtained from 1,000 Monte Carlo simulations where, in each replication, a random sample of size  $n = 50, 100, 250, 500$  and 750 is drawn.

**Table 1:** Summaries of the performance of the DZIP cure model.

Sample size ( $n$ )	Parameter	Summaries of parameters			
		Mean	Bias	MSE	CP
50	$\pi$	0.2161	-0.0339	0.0265	0.930
	$\gamma_1$	2.2849	0.2849	0.2951	0.933
	$\gamma_2$	-0.5761	-0.0761	0.2439	0.930
	$\beta_{10}$	-0.8371	0.1629	5.7489	0.960
	$\beta_{11}$	1.3094	0.8094	19.3054	0.973
	$\beta_{21}$	1.4400	0.1900	0.1870	0.924
100	$\pi$	0.2302	-0.0198	0.0124	0.952
	$\gamma_1$	2.1194	0.1194	0.0970	0.931
	$\gamma_2$	-0.5096	-0.0096	0.0885	0.946
	$\beta_{10}$	-0.9897	0.0103	1.4746	0.957
	$\beta_{11}$	0.6473	0.1473	3.4111	0.966
	$\beta_{21}$	1.3148	0.0648	0.0516	0.954
250	$\pi$	0.2424	-0.0076	0.0053	0.924
	$\gamma_1$	2.0399	0.0399	0.0277	0.950
	$\gamma_2$	-0.5099	-0.0099	0.0325	0.960
	$\beta_{10}$	-1.0200	-0.0200	0.1691	0.947
	$\beta_{11}$	0.5799	0.0799	0.5453	0.953
	$\beta_{21}$	1.2507	0.0007	0.0165	0.959
500	$\pi$	0.2473	-0.0027	0.0023	0.947
	$\gamma_1$	2.0165	0.0165	0.0144	0.927
	$\gamma_2$	-0.4921	0.0079	0.0168	0.938
	$\beta_{10}$	-1.0035	-0.0035	0.0734	0.949
	$\beta_{11}$	0.5008	0.0008	0.0616	0.962
	$\beta_{21}$	1.2326	-0.0174	0.0074	0.956
750	$\pi$	0.2494	-0.0006	0.0015	0.939
	$\gamma_1$	2.0127	0.0127	0.0082	0.951
	$\gamma_2$	-0.4945	0.0055	0.0106	0.950
	$\beta_{10}$	-1.0048	-0.0048	0.0489	0.940
	$\beta_{11}$	0.5068	0.0068	0.0407	0.958
	$\beta_{21}$	1.2329	-0.0171	0.0049	0.958



Tables 1, 2 and 3 display the averages of the MLEs (mean), bias, mean square errors (MSE) and coverage probabilities (CP) for nominal 95% of the DZIP, DZIG and DZINB cure models, respectively. We conclude from these results that (for all parameters) the MSEs of the MLEs decay toward zero when the sample size increases, as expected under standard asymptotic theory. In fact, the estimates tend to be closer to the true parameter values and the CPs converge to the nominal level when the sample size  $n$  increases.

**Table 2:** Summaries of the performance of the DZIG cure model.

Sample size ( $n$ )	Parameter	Summaries of parameters			
		Mean	Bias	MSE	CP
50	$\pi$	0.2222	-0.0278	0.0353	0.939
	$\gamma_1$	2.3558	0.3558	0.4846	0.946
	$\gamma_2$	-0.5849	-0.0849	0.4056	0.914
	$\beta_{10}$	-1.3522	0.1478	7.4278	0.968
	$\beta_{11}$	1.9488	1.1988	27.0050	0.976
	$\beta_{21}$	1.7590	0.2590	0.5628	0.906
100	$\pi$	0.2467	-0.0033	0.0214	0.943
	$\gamma_1$	2.1990	0.1990	0.1914	0.929
	$\gamma_2$	-0.5569	-0.0569	0.1624	0.933
	$\beta_{10}$	-1.4941	0.0059	1.4062	0.962
	$\beta_{11}$	1.4290	0.6790	9.6464	0.977
	$\beta_{21}$	1.6210	0.1210	0.1626	0.937
250	$\pi$	0.2491	-0.0009	0.0086	0.940
	$\gamma_1$	2.0800	0.0800	0.0503	0.943
	$\gamma_2$	-0.5223	-0.0223	0.0568	0.945
	$\beta_{10}$	-1.4777	0.0223	0.2496	0.965
	$\beta_{11}$	0.8922	0.1422	1.0476	0.959
	$\beta_{21}$	1.5239	0.0239	0.0490	0.945
500	$\pi$	0.2540	0.0040	0.0042	0.933
	$\gamma_1$	2.0324	0.0324	0.0210	0.951
	$\gamma_2$	-0.5020	-0.0020	0.0230	0.959
	$\beta_{10}$	-1.4628	0.0372	0.1143	0.963
	$\beta_{11}$	0.7960	0.0460	0.1245	0.962
	$\beta_{21}$	1.4891	-0.0109	0.0223	0.939
750	$\pi$	0.2556	0.0056	0.0033	0.921
	$\gamma_1$	2.0276	0.0276	0.0144	0.938
	$\gamma_2$	-0.5092	-0.0092	0.0183	0.932
	$\beta_{10}$	-1.4529	0.0471	0.0842	0.944
	$\beta_{11}$	0.7859	0.0359	0.0848	0.958
	$\beta_{21}$	1.4873	-0.0127	0.0153	0.938

**Table 3:** Summaries of the performance of the DZINB cure model.

Sample size ( $n$ )	Parameter	Summaries of parameters			
		Mean	Bias	MSE	CP
50	$\pi$	0.0976	-0.1524	0.0436	0.995
	$\gamma_1$	2.9864	0.9864	2.1222	0.969
	$\gamma_2$	-0.8296	-0.3296	0.8123	0.931
	$\phi$	2.9119	2.4119	14.9794	0.999
	$\beta_{10}$	0.1421	1.1421	23.6676	0.983
	$\beta_{11}$	3.3878	2.6378	67.0179	0.992
	$\beta_{21}$	2.1163	0.8663	2.0872	0.966
100	$\pi$	0.1392	-0.1108	0.0331	0.984
	$\gamma_1$	2.4785	0.4785	0.6232	0.970
	$\gamma_2$	-0.6812	-0.1812	0.3434	0.942
	$\phi$	1.9185	1.4185	6.3233	1.000
	$\beta_{10}$	-0.7544	0.2456	6.3957	0.966
	$\beta_{11}$	2.4230	1.6730	26.6289	0.984
	$\beta_{21}$	1.6889	0.4389	0.5671	0.976
250	$\pi$	0.2005	-0.0495	0.0155	0.975
	$\gamma_1$	2.1611	0.1611	0.1189	0.977
	$\gamma_2$	-0.5978	-0.0978	0.1137	0.969
	$\phi$	1.0213	0.5213	1.1649	1.000
	$\beta_{10}$	-0.9418	0.0582	0.6769	0.961
	$\beta_{11}$	1.2954	0.5454	5.1548	0.982
	$\beta_{21}$	1.3871	0.1371	0.0895	0.986
500	$\pi$	0.2362	-0.0138	0.0066	0.956
	$\gamma_1$	2.0704	0.0704	0.0408	0.970
	$\gamma_2$	-0.5519	-0.0519	0.0576	0.960
	$\phi$	0.6995	0.1995	0.4031	0.996
	$\beta_{10}$	-0.9677	0.0323	0.1758	0.962
	$\beta_{11}$	0.9161	0.1661	1.0521	0.975
	$\beta_{21}$	1.3053	0.0553	0.0342	0.977
750	$\pi$	0.2468	-0.0032	0.0037	0.953
	$\gamma_1$	2.0480	0.0480	0.0261	0.964
	$\gamma_2$	-0.5188	-0.0188	0.0346	0.960
	$\phi$	0.5932	0.0932	0.1934	0.978
	$\beta_{10}$	-0.9744	0.0256	0.1111	0.957
	$\beta_{11}$	0.8271	0.0771	0.2584	0.975
	$\beta_{21}$	1.2754	0.0254	0.0199	0.971

---

## 6. APPLICATION: CUTANEOUS MELANOMA DATA

---

In this section, we illustrate the usefulness of the DZIPS cure rate regression with an application to a real data set on cancer recurrence. The data are part of a study on cutaneous melanoma (a type of malignant cancer) extracted from [25] on 205 patients observed for the evaluation of postoperative in the period from 1962 to 1977. The cutaneous melanoma data contain information about the survival times of patients after surgery for malignant melanoma which were collected at Odense University Hospital [13].

In general, the standard treatment of cutaneous melanoma consists of broad excision of primary tumor or cicatrices at a distance of at least 5 cm down to the fascia, though not including this. On the face the tumor was removed at a distance of only 2 cm. Lymphonodectomy was only undertaken when lymph nodes were clinically suspected. The clinical data and follow-up were based on information from the case histories of the patients. For more details, see [13].

The observed survival time range, approximately from 0 to 15 years (with mean equal to 5.9 years), refers to the time until the patient’s death or the censoring time. There are 72% of censoring, corresponding to the patients which had died from other causes or were still alive at the end of the study. The following variables involved in the study for each patient are:  $y_i$ : observed time (in years),  $x_{i11}$ : tumor thickness (in mm, mean = 2.92 and standard deviation = 2.96) and  $x_{i21}$ : ulceration status (absent,  $n = 115$ ; present,  $n = 90$ ). As we mentioned earlier, the identifiability issue is avoided if the parameter  $p$  is linked only to tumor thickness, while the parameter  $\theta$  (or  $\eta$ ) is linked to the ulceration status in the DZIP, DZINB and DZIG regressions. The survival function for these cure rate regressions are:

- **DZIP survival function**

$$S(y_i | \mathbf{x}_i) = \pi + (1 - \pi) \exp\left\{-\theta_i p_i [1 - \exp(-y_i^{\gamma_1} e^{\gamma_2})]\right\},$$

where

$$p_i = \frac{\exp(\beta_{10} + \beta_{11} x_{i11})}{1 + \exp(\beta_{10} + \beta_{11} x_{i11})} \quad \text{and} \quad \theta_i = \exp(\beta_{20} + \beta_{21} x_{i21}).$$

- **DZINB survival function**

$$S(y_i | \mathbf{x}_i) = \pi + (1 - \pi) \left\{1 + \eta_i \phi p_i [1 - \exp(-y_i^{\gamma_1} e^{\gamma_2})]\right\}^{-1/\phi},$$

where

$$p_i = \frac{\exp(\beta_{10} + \beta_{11} x_{i11})}{1 + \exp(\beta_{10} + \beta_{11} x_{i11})} \quad \text{and} \quad \eta_i = \exp(\beta_{20} + \beta_{21} x_{i21}).$$

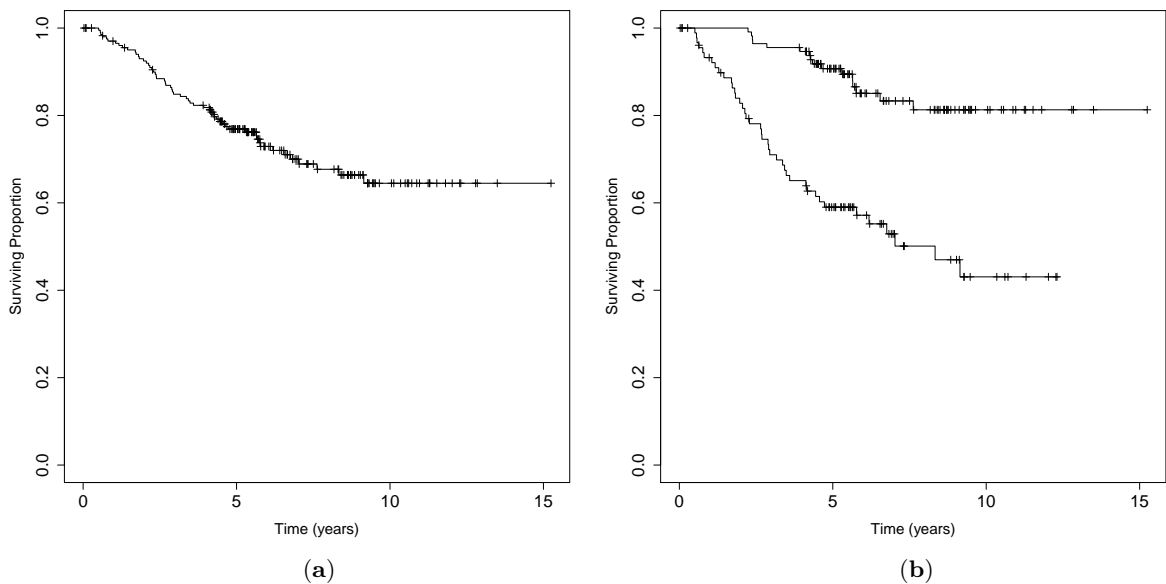
- **DZIG survival function**

$$S(y_i | \mathbf{x}_i) = \pi + (1 - \pi) \left\{1 + \eta_i p_i [1 - \exp(-y_i^{\gamma_1} e^{\gamma_2})]\right\}^{-1},$$

where

$$p_i = \frac{\exp(\beta_{10} + \beta_{11} x_{i11})}{1 + \exp(\beta_{10} + \beta_{11} x_{i11})} \quad \text{and} \quad \eta_i = \exp(\beta_{20} + \beta_{21} x_{i21}).$$

Figure 2(a) shows that the Kaplan–Meier survival function estimate confirms a plateau around 0.64 and this fact indicates the presence of a proportion of patients for whom the malignant melanoma will never occur again, and then, those patients can be considered as cured. Also, the empirical Kaplan–Meier curves stratified by ulceration status (upper: absent, lower: present) are displayed in Figure 2(b) and they reveal that the ulceration affects the lifetime of the patients with malignant melanoma.



**Figure 2:** (a) Kaplan–Meier curve for the cutaneous melanoma data. (b) Kaplan–Meier curves stratified by ulceration status (upper: present, lower: absent).

For model comparison, we fit the DZIP, DZINB and DZIG cure models described in Section 3 to the cutaneous melanoma data. The special cases of these models were also fitted to these data, i.e., the Poisson ( $\pi = 0$  and  $p = 1$ ), the negative binomial ( $\pi = 0$  and  $p = 1$ ) and the geometric ( $\pi = 0$ ,  $p = 1$  and  $\phi = 1$ ) models. We note that these special models belong to the PS cure models proposed by [4]. For these models, the destructive process is absent and consequently, the parameter  $\theta$  (or  $\eta$ ) is linked to both variables (ulceration status and tumor thickness). In order to compare the models, we use the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). The results of the DZIPS cure models and its sub-models are reported in Table 4.

**Table 4:** The values of  $\max \log L(\cdot)$ , the AIC and BIC statistics for the Destructive Zero-Inflated Poisson (DZIP), Destructive Zero-Inflated Negative Binomial (DZINB), Destructive Zero-Inflated Geometric (DZIG), Poisson, negative binomial and geometric cure models.

Survival Cure Rate Model	$\max \log L(\cdot)$	AIC	BIC
Destructive Zero-Inflated Poisson	-201.18	416.3	439.6
Destructive Zero-Inflated Negative Binomial	-198.95	413.9	440.5
Destructive Zero-Inflated Geometric	-199.93	<b>413.8</b>	<b>437.1</b>
Poisson	-207.83	425.6	442.2
Negative Binomial	-201.52	423.0	439.7
Geometric	-205.42	420.8	437.4

According to the criteria in Table 4, the DZIG cure rate regression is the best model and so, it is selected as our working model. For this regression, we estimate the unknown parameters via ML method. All computations are performed using the R software.

The survival function for the DZIG cure rate regression is

$$S(y_i; \hat{\pi}, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\beta}_1, \hat{\beta}_2) = \hat{\pi} + (1 - \hat{\pi}) \left\{ 1 + \hat{\eta}_i \hat{p}_i \left[ 1 - \exp(-\hat{y}_i^{\hat{\gamma}_1} e^{\hat{\gamma}_2}) \right] \right\}^{-1},$$

where

$$\hat{p}_i = \frac{\exp(\hat{\beta}_{10} + \hat{\beta}_{11} x_{i11})}{1 + \exp(\hat{\beta}_{10} + \hat{\beta}_{11} x_{i11})} \quad \text{and} \quad \hat{\eta}_i = \exp(\hat{\beta}_{20} + \hat{\beta}_{21} x_{i21}).$$

Here, for the cutaneous melanoma data set, the vectors  $\hat{\beta}_1$  and  $\hat{\beta}_2$  are

$$\hat{\beta}_1 = (\beta_{10}, \beta_{11})^\top \quad \text{and} \quad \hat{\beta}_2 = (\beta_{20}, \beta_{21})^\top.$$

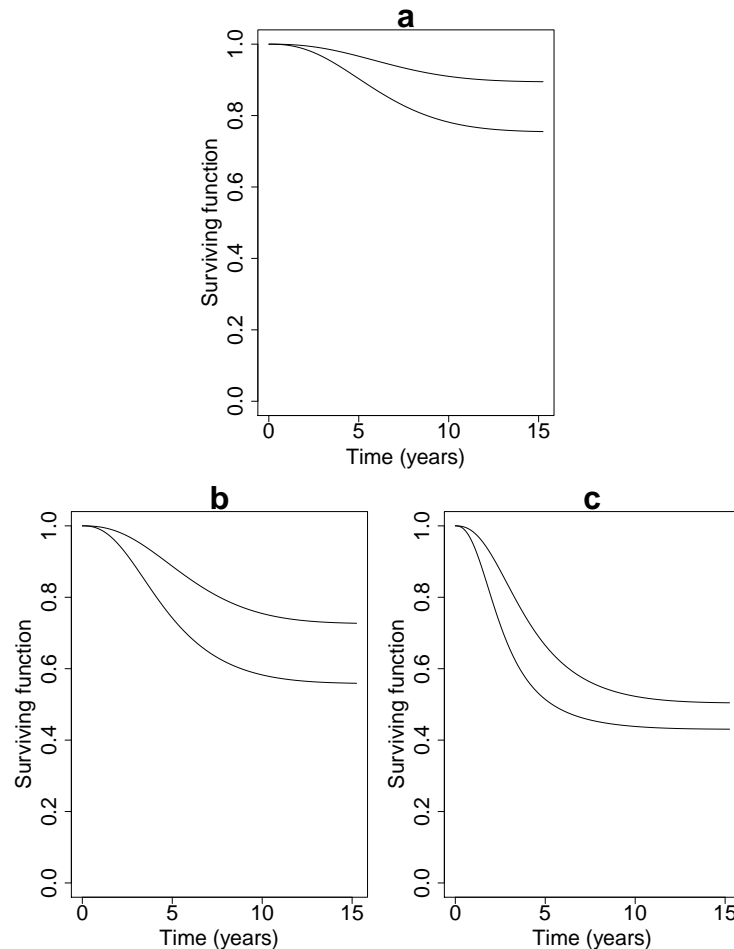
Table 5 gives the MLEs of the parameters, their standard errors and  $p$ -values from the fitted regression. We note from the fitted DZIG cure rate regression that ulceration status and tumor thickness are significant sloppy 1% and there is a significant difference for the presence or absent of ulceration status and also a difference related to the thickness of the tumor. Thus, those variables have influenced on the survival times of the patients. The estimate of the parameter  $\pi$  is 0.3895, and as mentioned earlier in Section 1, this indicates a proportion of those cells which never experience alterations/lesions. Consequently, the proportion of cells that were repaired by a repair system of the organism is  $(1 - \pi) = 0.6105$  (or 61.05%).

**Table 5:** Results from the fitted DZIG cure rate regression.

Parameter	Estimate	Standard Error	$p$ -value
$\gamma_1$	2.41	0.28	—
$\gamma_2$	-5.00	0.61	—
$\pi$	0.38	0.24	—
$\beta_{10}$	-4.41	0.93	< 0.001
$\beta_{11}$	0.86	0.26	0.001
$\beta_{20}$	2.59	0.88	0.003
$\beta_{21}$	3.76	0.74	< 0.001

Figure 3 displays the estimated survival function of the DZIG cure rate regression for patients with 0.320 mm, 1.940 mm and 4.254 mm tumor thickness, which correspond to the 5%, 50% and 80% tumor thickness quantiles. The survival rate decreases more rapidly for patients with thicker tumors in presence of ulceration. On the other hand, for patients with less thick tumor in presence of ulceration, the survival rate does not fall below 75% as shown in Figure 3(a).

Finally, we turn our attention to the role of the ulceration status and thickness tumor covariates on the estimation of the surviving fraction ( $p_0$ ). To estimate the proportion of cured individuals, we use Equation (4.1) and the MLEs of the parameters. So, for the DZIG cure regression, the estimated cure fraction  $\hat{p}_0 = \hat{\pi} + (1 - \hat{\pi}) [1 + \hat{\eta} \hat{p}]^{-1}$  is 0.6450. This result is confirmed in Figure 2(a). Also, we note that the cure rate decreases when tumor thickness size increases and it is smaller for patients with presence of ulceration.



**Figure 3:** Estimated survival function from the DZIG cure rate regression stratified by ulceration status (upper: absent, lower: present) for patients with tumor thickness equal to: (a) 0.320 mm, (b) 1.940 mm, and (c) 4.254 mm.

---

## 7. CONCLUDING REMARKS

---

In this paper, we propose the destructive zero-inflated power series (DZIPS) family of cure rate models by extending the works of [23] and [4]. The DZIPS models are very flexible and contain special models such as the zero-inflated binomial (ZIBin), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), zero-inflated geometric (ZIG) models, among others. The proposed model allows estimation of the cure fraction by incorporating a systematic component to estimate the proportion of not initiated cells (those one that have never been altered/lesioned and those one that have been repaired). Hence, this extended family of models is very flexible in many practical situations. An application to a real cutaneous melanoma data set demonstrates that it can be used quite effectively to provide better interpretation for the underlying biological mechanism, in addition to offering a better fit than the other commonly used cure rate models.

---

## ACKNOWLEDGMENTS

---

The authors are grateful to the editor and the referee for helpful comments and suggestions. We gratefully acknowledge grants from CNPq, Brazil.

---

## REFERENCES

---

- [1] BARRAL, A.M. (2001). *Immunological Studies in Malignant melanoma: importance of tnfr and the thioredoxin system*, Doctorate Thesis, Linkoping University, Linkoping, Sweden.
- [2] BERKSON, J. and GAGE, R.P. (1952). Survival curve for cancer patients following treatment, *Journal of the American Statistical Association*, **47**, 501–515.
- [3] BORGES, P.; RODRIGUES, J. and BALAKRISHNAN, N. (2012). Correlated destructive generalized power series cure rate models and associated inference with an application to a cutaneous melanoma data, *Computational Statistics and Data Analysis*, **56**, 1703–1713.
- [4] CANCHO, V.G.; BANDYOPADHYAY, D.; LOUZADA, F. and YIQI, B. (2013). The destructive negative binomial cure rate model with a latent activation scheme, *Statistical Methodology*, **13**, 48–68.
- [5] CANCHO, V.G.; RODRIGUES, J. and DE CASTRO, M. (2011). A flexible model for survival data with a cure rate: a Bayesian approach, *Journal of Applied Statistics*, **38**, 57–70.
- [6] DE CASTRO, M.; CANCHO, V.G. and RODRIGUES, J. (2009). A Bayesian long-term survival model parametrized in the cured fraction, *Biometrical Journal*, **51**, 443–455.
- [7] CHEN, M.H.; IBRAHIM, J.G. and SINHA, D. (1999). A new Bayesian model for survival data with a surviving fraction, *Journal of the American Statistical Association*, **94**, 909–919.
- [8] CLEGG, L.X.; FEUER, E.J.; MIDTHUNE, D.N.; FAY, M.P. and HANKEY, B.F. (2002). Impact of reporting delay and reporting error on cancer incidence rates and trends, *Journal of the National Cancer Institute*, **94**, 1537–1545.
- [9] COLLET, D. (1994). *Modelling Survival Data in Medical Research*, Chapman and Hall, London.
- [10] COONER, F.; BANERJEE, S.; CARLIN, B. and SINHA, D. (2007). Flexible cure rate modelling under latent activation schemes, *Journal of the American Statistical Association*, **102**, 560–572.
- [11] COX, D. and OAKES, D. (1984). *Analysis of Survival Data*, Chapman and Hall, London.
- [12] DRAPER, N.R. and SMITH, H. (1984). *Applied Regression Analysis*, John Wiley and Sons, New York.
- [13] DRZEWIECKI, K.T.; LADEFOGED, C. and CHRISTENSEN, H.E. (1980). Biopsy and prognosis for cutaneous malignant melanomas in clinical stage I, *Scandinavian Journal of Plastic and Reconstructive Surgery*, **14**, 141–144.
- [14] GORDON, N.H. (1990). Application of the theory of finite mixtures for the estimation of cure rates of treated cancer patients, *Statistics in Medicine*, **9**, 397–407.
- [15] GUPTA, P.L.; GUPTA, R.L. and TRIPATHI, R.C. (1995). Inflated modified power series distributions with applications, *Communications in Statistics – Theory and Methods*, **24**, 2355–2374.

- [16] IBRAHIM, J.G.; CHEN, M.H. and SINHA, D. (2001). *Bayesian Survival Analysis*, Springer-Verlag, New York.
- [17] INSTITUTO NACIONAL DO CÂNCER. *Tipos de câncer: melanoma cutâneo*, Rio de Janeiro: INCA, 2020. <https://www.inca.gov.br/tipos-de-cancer/cancer-de-pele-melanoma>
- [18] LI, C.S.; TAYLOR, J. and SY, J. (2001). Identifiability of cure models, *Statistics and Probability Letters*, **54**, 389–395.
- [19] MALLER, R.A. and ZHOU, X. (1996). *Survival Analysis with Long-Term Survivors*, John Wiley and Sons, New York.
- [20] NOACK, A. (1950). On a class of discrete random variables, *Annals of Mathematical Statistics*, **21**, 127–132.
- [21] ORTEGA, E.M.M.; CORDEIRO, G.M.; CAMPELO, A.K.; KATTAN, M.W. and CANCHO, V.G. (2015). A power series beta Weibull regression model for predicting breast carcinoma, *Statistics in Medicine*, **44**, 1366–1388.
- [22] PESCIM, R.R.; ORTEGA, E.M.M.; SUZUKI, A.K.; CANCHO, V.G. and CORDEIRO, G.M. (2019). A new destructive Poisson odd log-logistic generalized half-normal cure rate model, *Communication in Statistics – Theory and Methods*, **48**, 2113–2128.
- [23] RODRIGUES, J.; DE CASTRO, M.; BALAKRISHNAN, N. and CANCHO, V.G. (2011). Destructive weighted Poisson cure rate models, *Lifetime Data Analysis*, **17**, 333–346.
- [24] RODRIGUES, J.; DE CASTRO, M.; CANCHO, V.G. and LOUZADA-NETO, F. (2009). On the unification of the long-term survival models, *Statistics and Probability Letters*, **79**, 753–759.
- [25] SCHEIKE, T. (2009). Timereg Package. R Package Version 3.4.0. With contributions from T. Martinussen and J. Silver. R package version 1.1-6.
- [26] TSODIKOV, A.D.; IBRAHIM, J.G. and YAKOVLEV, A.Y. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models, *Journal of the American Statistical Association*, **98**, 1063–1078.
- [27] YAKOVLEV, A.Y. and TSODIKOV, A.D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*, World Scientific, Singapore.
- [28] YANG, G.L. and CHEN, C.W. (1991). A stochastic two-stage carcinogenesis model: a new approach to computing the probability of observing tumor in animal bioassays, *Mathematical Biosciences*, **104**, 247–258.
- [29] ZHAO, Y.; LEE, A.H.; YAU, K.K.W. and BURKE, V. (2009). A score test for assessing the cured proportion in the long-term survivor mixture model, *Statistics in Medicine*, **28**, 3454–3466.



---

---

## Single Index Regression Model for Functional Quasi-Associated Time Series Data

---

---

Authors: SALIM BOUZEBDA  

– Alliance Sorbonne Université, Université de Technologie de Compiègne,  
L.M.A.C., Compiègne, France  
[salim.bouzebda@utc.fr](mailto:salim.bouzebda@utc.fr)

ALI LAKSACI 

– Department of Mathematics, College of Science, King Khalid University,  
Abha, 61413, Saudi Arabia  
[alilak@yahoo.fr](mailto:alilak@yahoo.fr)

MUSTAPHA MOHAMMEDI

– Université Djillali Liabès,  
BP 89, 22000, Sidi Bel Abbès, Algérie  
[mustapha.mohammedi@gmail.com](mailto:mustapha.mohammedi@gmail.com)

Received: March 2020

Revised: February 2021

Accepted: February 2021

Abstract:

- The mixing condition is often considered to modeling the functional time series data. Alternatively, in this work we consider the problem of nonparametric estimation of the regression function in Single Functional Index Model (SFIM) under the quasi-association dependence condition. The main result of this work is the establishment of the asymptotic properties of the estimator, such as the almost complete convergence rates. Furthermore, the asymptotic normality of the constructed are obtained under some mild conditions. We finally discuss how to apply our result to construct the confidence intervals. Finally, the finite-sample performances of the model and the estimation method are illustrated using the analysis of simulated data.

Keywords:

- *single functional index model; functional Hilbert space; kernel regression estimation; mixing; weak dependence; quasi-associated variables; almost complete convergence; asymptotic normality.*

AMS Subject Classification:

- 62G05, 62G08, 62L12, 62G20.

---

## 1. INTRODUCTION

---

The statistical study of single index models have been investigated and developed by several authors from a practical and theoretical point of view. The case of a vector explanatory variable was studied by [19] and [20]. The single index models are very popular in the econometric community because it respond two important preoccupations. The first concerns dimension reduction since this type of model makes it possible to provide a solution to the problem of the curse of dimensionality, in the sense that pure nonparametric models are highly affected by dimensionality effects while semiparametric ideas are more appealing candidates. The second is related to the interpretability of the index  $\theta$  introduced in these models, for more details on refer to [8], [18] and [3] for an overview on methodological issues on FDA. Therefore, the single functional index model accumulate the advantages of single index model, and inherits the potential of the functional linear model in terms of applications. The interested reader, for the semiparametric and the nonparametric functional models, may refer [17], [24, 25], [27] and [7] for survey on the topics.

The modelization of functional data, has been developed intensively. The motivation of such statistical analysis is justified by the recent technological development of the measuring instruments that offers the opportunity to observe phenomena in an increasingly accurate way, but this accuracy obviously generates a large amount of data observed over a finer grid, which can be considered as observations varying over a continuum. The most theoretical results are obtained under independence condition. However, in practice, it is rarely that we have an independent identically distributed observations of functional nature. The functional time series presents the more realistic situation. Thus it is really crucial to study the functional statistical models when the usual independence condition on the statistical sample is relaxed. In this paper, we consider the problem of the nonparametric estimation of the regression function in single functional index model when the data are weakly dependant.

Usually the dependence structure is modelled with the strong mixing hypothesis, in this paper we focus in some more general correlation, that is the quasi-associated condition. The latter has been introduced for real valued random fields by [5], which generalizes the positively associated variables introduced in [13].

From practical point of view, this kind of data has great importance in practice, in particular, in reliability theory, mathematical physics and in percolation theory (see, for instance, [28]) for more discussion on the practical interest of these random variables. Moreover, from the theoretical point of view, the concept of quasi-association correlation can be viewed as a particular case of the weak dependence condition for real-valued stochastic processes introduced by [12] which allows treating the mixing condition and association correlation in a unified approach.

Noting that the single index model is a semi-parametric regression model, thus, it couples the advantages of both parametric and nonparametric regression models. Because of these advantages, it has received an increasing amount of attention in the nonparametric regression literature. Key references on this topic in multivariate statistic are [21] and [20] for previous results and [30] for more recent advances and references.

However, in the literature of functional statistic, the single functional index model is strictly limited in the case where the data is functional (a curve). The first result in this context, was given by [15]. They obtained the almost complete convergence of the regression function  $r(\cdot)$  in the independent and identically distributed (i.i.d.) case. The generalization of this result to the dependent case has been studied by [26]. [29] uses a Bayesian method to estimate the bandwidths in the kernel form error density and regression function, under an autoregressive error structure, and according to empirical studies, the author considered that the single functional index model gives improved estimation and prediction accuracies compared to any nonparametric functional regression considered. [27] have proposed a new automatic and location-adaptive procedure for estimating regression in a Functional Single-Index Model (FSIM) based on  $k$ -Nearest Neighbours ideas. Motivated by the analysis of imaging data, [23] proposed a novel functional varying-coefficient single-index model to carry out the regression analysis of functional response data on a set of covariates of interest. This method represents a new extension of varying-coefficient single-index models for scalar responses collected from cross-sectional and longitudinal studies. By simulation and real data analysis, the authors demonstrated the advantages of the proposed estimate. [31] have considered the problem of predicting the real-valued response variable using explanatory variables containing both multivariate random variable and random curve. The authors considered the functional partial linear single-index model in order to treat the multivariate random variable as linear part and the random curve as functional single-index part, respectively.

The concept of quasi-association for random variables taking its values in a Hilbert space has been investigated by [10], and obtained some limit theorems for this type of variables. More recently, [11] studied the asymptotic normality of regression function under quasi-associated data when the explanatory variable takes its values in a Hilbert space.

The main purpose of the present paper is to establish the asymptotic properties of the estimator  $\widehat{r}_\theta(\cdot)$ , when the variables are functional quasi-associated and in single index structure, such as the almost complete convergence rates. Furthermore, the asymptotic distribution is obtained under some mild conditions.

We point out that the mixing and the association concern two distinct classes of processes but not disjoint and offer two complementary approaches to study the dependence. Moreover, the functional quasi-associated data analysis has great importance in various domains such as the reliability theory or the statistical mechanics. Furthermore, it should be noted that the dependence condition considered here allow to avoid the widely used strong mixing condition which is very easy to verified in practice.

The rest of this work is organized as follows. In Section 2, we describe the single index regression model for functional data and in the quasi-associated framework, the next section is devoted to the introduction of the notation and hypotheses needed to state our main results. In Section 4, we will establish our main results of the almost complete convergence of the kernel estimators and the asymptotic normality under non restrictive conditions. In Section 5.2, we discuss the impact of our contribution in practice application of our results for the construction of the confidence interval. In Section 6 we perform a short simulation study to show that our proposed model works well for finite samples. To avoid interrupting the flow of the presentation, all mathematical developments are relegated to the Section 7.

---

## 2. MODEL AND ESTIMATOR

---

We start by giving a definition of quasi-association adapted to the functional framework. In the real valued random fields, [5] define the quasi-association dependence in the Definition 2.1 and it adapted to functional random variables in the Definition 2.2 given in [10] as follows.

**Definition 2.1.** A sequence  $(X_n)_{n \in \mathbb{N}}$  of r.v.'s is said to be quasi-associated, if for any disjoint subsets  $I$  and  $J$  of  $\mathbb{N}$  and all bounded Lipschitz functions  $f_1: \mathbb{R}^{|I|} \rightarrow \mathbb{R}$  and  $f_2: \mathbb{R}^{|J|} \rightarrow \mathbb{R}$  satisfying:

$$(2.1) \quad \left| \text{Cov} \left( f_1(X_i, i \in I), f_2(X_j, j \in J) \right) \right| \leq \text{Lip}(f_1) \text{Lip}(f_2) \sum_{i \in I} \sum_{j \in J} |\text{Cov}(X_i, X_j)|,$$

where  $|I|$  denotes cardinality of a finite set  $I$ , and the Lipschitz of a function  $f(\cdot)$  is defined by

$$\text{Lip}(f) = \sup_{x \neq y} \frac{|f(x) - f(y)|}{\|x - y\|_1}, \quad \text{with} \quad \|(x_1, \dots, x_k)\|_1 = \sum_{k=1}^n |x_k|.$$

**Definition 2.2.** A sequence  $(X_i)_{i \in \mathbb{N}}$  of r.v.'s taking values in a Hilbert space  $H$  is called quasi-associated relative to an orthonormal basis  $\{e_p: p \geq 1\}$  of  $H$ , if for any  $p \geq 1$ ,  $(\langle X_i, e_1 \rangle, \dots, \langle X_i, e_p \rangle)_{i \in \mathbb{N}}$  is a sequence of random vectors quasi-associates.

Now, we consider a sequence of quasi-associated random variables  $\{(X_i, Y_i)\}_{i \in \mathbb{N}}$  identically distributed as  $(X, Y)$ , which are valued in  $H \times \mathbb{R}$ , where  $H$  is a separable real Hilbert space with inner product  $\langle \cdot, \cdot \rangle$  and a orthonormal basis  $\{e_p: p \geq 1\}$ . We consider the semi-metric  $d_\theta(\cdot, \cdot)$  associated to the single-index  $\theta \in H$  defined by  $\forall u, v \in H$ :

$$d_\theta(u, v) := |\langle \theta, u - v \rangle|.$$

The purpose of this paper is to study the estimation of the nonparametric regression of  $Y$  given  $\langle \theta, X \rangle$  structure, denoted by

$$(2.2) \quad r(\langle \theta, X_i = x \rangle) = \mathbb{E}(Y | \langle \theta, X_i = x \rangle).$$

Such structure suppose that the explanation of  $Y$  from  $X$  is done through an fixed functional index  $\theta$  in  $\Theta$ . Now, we suppose that exists a  $\theta \in \Theta \subset H$  where the observations  $(X_i, Y_i)_{i=1, \dots, n}$  are related by the following relation:

$$(2.3) \quad Y_i = r(\langle \theta, X_i \rangle) + \varepsilon_i, \quad \forall i = 1, \dots, n,$$

where  $r(\cdot)$  is a real function, and for  $i = 1, \dots, n$ ,  $\varepsilon_i$  is a real random variable such that  $\mathbb{E}(\varepsilon_i | X_i) = 0$ . We consider that the single functional index model is identifiable, i.e., if the regression function is differentiable and if  $\langle \theta, e_1 \rangle = 1$ , where  $e_1$  is the first element of an orthonormal basis of  $H$ . Then, if  $r_1(\langle \theta_1, x \rangle) = r_2(\langle \theta_2, x \rangle)$  implies that  $r_1 \equiv r_2$  and  $\theta_1 \equiv \theta_2$ .

This hypothesis that we consider is demonstrated by [15] once we have the differentiability of the regression operator  $r(\cdot)$ . For more details on the problem of identifiability of the single functional index model, one can refer to the last reference. The kernel estimator  $\widehat{r}_\theta(\cdot)$  of regression operator  $r_\theta(\cdot) = r(\langle \theta, \cdot \rangle)$  is defined by

$$(2.4) \quad \widehat{r}_{\theta,n}(x) = \frac{\sum_{i=1}^n Y_i K_i(x)}{\sum_{i=1}^n K_i(x)}, \quad \text{for all } x \in H,$$

where  $K_i(x) := K\left(\frac{\langle \theta, x - X_i \rangle}{h_n}\right)$  is the kernel function and  $h_n$  is the bandwidth parameter decreases to zero as  $n$  goes to infinity.

---

### 3. ASSUMPTIONS AND NOTATION

---

In the sequel, we will denote by  $C$  and/or  $C'$  some strictly positive constants and by  $\lambda_r$  the covariance coefficient defined as:

$$\lambda_r := \sup_{s \geq r} \sum_{|i-j| \geq s} \lambda_{i,j},$$

where

$$\lambda_{i,j} = \sum_{k \geq 1} \sum_{l \geq 1} |\text{Cov}(X_i^k, X_j^l)| + \sum_{k \geq 1} |\text{Cov}(X_i^k, Y_j)| + \sum_{l \geq 1} |\text{Cov}(Y_i, X_j^l)| + |\text{Cov}(Y_i, Y_j)|,$$

with  $X_i^p := \langle X_i, e_p \rangle$ . In our analysis, we shall assume the following assumptions:

**(H<sub>1</sub>)** Let  $E_i(x) := \langle \theta, x - X_i \rangle$  so that  $E_i(x)$  is a real-valued random variable,

$$G_\theta(x, h_n) := \mathbb{P}(|E_i(x)| \leq h_n) > 0,$$

and  $G_\theta(x, \cdot)$  is differentiable at 0.

**(H<sub>2</sub>)** The random pair  $\{(X_i, Y_i), i \in \mathbb{N}\}$  is quasi-associated such that:

(i) The covariance coefficient satisfies

$$\lambda_k \leq C e^{-ak} \quad \text{for some } a > 0, \quad C > 0;$$

(ii) The process  $(X_i)_i$  satisfies

$$\max_{i \neq j} \left\{ \mathbb{P}\left(|E_i| \leq h_n, |E_j| \leq h_n\right) \right\} := \psi_\theta(x, h_n) > 0,$$

where  $\psi_\theta(x, \cdot)$  is differentiable at 0;

(iii) The response observations  $(Y_i)_i$  are such that, almost surely

$$\forall i \neq j \quad \mathbb{E}(|Y_i Y_j| | X_i, X_j) \leq C < \infty$$

$$\text{and} \quad \mathbb{E}(|Y|^p | X = x) \leq C < \infty \quad \text{for } p > 4.$$

(H<sub>3</sub>) For all  $u, v \in H$  we have

$$|r_\theta(u) - r_\theta(v)| \leq C |\langle \theta, u - v \rangle|^\beta, \quad \text{for certain } \beta > 0.$$

(H<sub>4</sub>) The kernel  $K(\cdot)$  is a Lipschitzian function on  $[0, 1]$  such that

$$C \mathbf{1}_{[0,1]}(t) < K(t) < C' \mathbf{1}_{[0,1]}(t).$$

(H<sub>5</sub>) There exists a sequence of positive real numbers  $\delta_n$  such that

$$\begin{cases} \delta_n^{p-2} \chi_\theta^{(p-4)/2p}(x, h_n) \rightarrow 0, \\ \sum_n n \delta_n^{-p} < \infty, \end{cases}$$

where  $\chi_\theta(x, h_n) = \max(\psi_\theta(x, h_n), G_\theta^2(x, h_n))$  and  $p$  is given in (H<sub>2</sub>).

---

### Some comments on the assumptions

---

All the assumptions are standard in this context of semiparametric functional data analysis. The concentration property of the explanatory variable in small balls under single index topological structure is defined in the assumption (H<sub>1</sub>). The quasi-association features of the underlying functional time series is explored through the condition (H<sub>2</sub>). It covers the three fundamental aspects of the considered process. The correlation's level of the data is quantified by the geometric form of the covariance coefficient  $\lambda_k$ , while the local dependency of the data is expressed by the function  $\psi_\theta(x, h_n)$  allowing to emphasize the functional component of the time series  $(X_i)_i$ . It should be noted that the conditional moments integrability in (H<sub>2</sub>)(iii) is usual in the regression data analysis. It was used by [16] for the nonparametric case and by [1] in the single functional index case. It is less restrictive than the exponential version assumed by [10]. Finally, let us mention that the hypothesis (H<sub>3</sub>) is used to control the regularity condition of the link function with respect the single index. This kind of assumption is needed to evaluate the bias in the asymptotic results of this paper, while the conditions (H<sub>4</sub>) and (H<sub>5</sub>) are classical technical assumptions in NFDA.

---

## 4. MAIN RESULTS

---

### 4.1. The almost consistency

---

Our aim is to establish the almost complete convergence (a.co.)<sup>1</sup> of  $\widehat{r}_\theta(x)$  to  $r_\theta(x)$ , and the main result is given by the following theorem.

**Theorem 4.1.** *Under the assumptions (H<sub>1</sub>)–(H<sub>5</sub>), we have, as  $n \rightarrow \infty$ ,*

$$(4.1) \quad \widehat{r}_{\theta,n}(x) - r_\theta(x) = O_{\text{a.co.}} \left( h_n^\beta + \sqrt{\frac{\chi_\theta^{1/2}(x, h_n) \log n}{n G_\theta^2(x, h_n)}} \right).$$

Let

$$(4.2) \quad \widehat{r}_{\theta,0}(x) := \frac{1}{n \mathbb{E} K_1(x)} \sum_{i=1}^n K_i(x) \quad \text{and} \quad \widehat{r}_{\theta,1}(x) := \frac{1}{n \mathbb{E} K_1(x)} \sum_{i=1}^n Y_i K_i(x).$$

Let us consider the following decomposition:

$$\begin{aligned} \widehat{r}_{\theta,n}(x) - r_\theta(x) &= \frac{\widehat{r}_{\theta,1}(x)}{\widehat{r}_{\theta,0}(x)} - r_\theta(x) \\ &= \frac{1}{\widehat{r}_{\theta,0}(x)} \left[ (\widehat{r}_{\theta,1}(x) - \mathbb{E}(\widehat{r}_{\theta,1})) - (r_\theta(x) - \mathbb{E}(\widehat{r}_{\theta,1})) \right] - \frac{r_\theta(x)}{\widehat{r}_{\theta,0}(x)} (\widehat{r}_{\theta,0} - 1) \\ &= \frac{1}{\widehat{r}_{\theta,0}(x)} \left[ (\widehat{r}_{\theta,1}(x) - \widehat{r}_{\theta,2}(x)) + (\widehat{r}_{\theta,2}(x) - \mathbb{E}(\widehat{r}_{\theta,2})) \right] \\ &\quad + \frac{1}{\widehat{r}_{\theta,0}(x)} \left[ (\mathbb{E}(\widehat{r}_{\theta,2}(x)) - \mathbb{E}(\widehat{r}_{\theta,1})) - (r_\theta(x) - \mathbb{E}(\widehat{r}_{\theta,1})) \right] - \frac{r_\theta(x)}{\widehat{r}_{\theta,0}(x)} (\widehat{r}_{\theta,0} - 1), \end{aligned}$$

where

$$(4.3) \quad \widehat{r}_{\theta,2}(x) := \frac{1}{n \mathbb{E} K_1(x)} \sum_{i=1}^n \widehat{Y}_i K_i(x).$$

The real variable  $Y$  response is not necessarily bounded. For this, we introduce the truncated random variable  $\widehat{Y}$ , defined by  $\widehat{Y}_i = Y_i \mathbb{1}_{\{|Y_i| \leq \delta_n\}}$ . The proof of the Theorem 4.1 is based on the following Lemmas.

---

<sup>1</sup>We say that the sequence  $(\Theta_n)_n$  converges a.co. to zero, if and only if

$$\forall \tau > 0, \quad \sum_{n \geq 1} \mathbb{P}(|\Theta_n| > \tau) < \infty.$$

Furthermore, we say that  $\Theta_n = O_{\text{a.co.}}(\theta_n)$  if there exists  $\tau_0 > 0$  such that

$$\sum_{n \geq 1} \mathbb{P}(|\Theta_n| > \tau_0 \theta_n) < \infty.$$

**Lemma 4.1** (See [22]). *Let  $X_1, \dots, X_n$  be real random variables such that  $\mathbb{E}X_i = 0$  and  $\mathbb{P}(|X_i| \leq M) = 1$ , for all  $i = 1, \dots, n$  and some  $M < \infty$ . Let*

$$\sigma_n^2 = \text{Var} \left( \sum_{i=1}^n X_i \right).$$

*Assume, furthermore, that there exist  $K < \infty$  and  $\beta > 0$  such that, for all  $u$ -tuplets  $(s_1, \dots, s_u)$  and all  $v$ -tuplets  $(t_1, \dots, t_v)$  with  $1 \leq s_1 \leq \dots \leq s_u \leq t_1 \leq \dots \leq t_v \leq n$ , the following inequality is fulfilled:*

$$\left| \text{Cov}(X_{s_1} \dots X_{s_u}, X_{t_1} \dots X_{t_v}) \right| \leq K^2 M^{u+v-2} v e^{-\beta(t_1-s_u)}.$$

Then,

$$\mathbb{P} \left( \left| \sum_{i=1}^n X_i \right| \geq t \right) \leq \exp \left\{ - \frac{t^2/2}{A_n + B_n^{\frac{1}{3}} t^{\frac{5}{3}}} \right\},$$

for  $A_n \leq \sigma_n^2$  and

$$B_n = \left( \frac{16 n K^2}{9 A_n (1 - e^{-\beta})} \vee 1 \right) \left( \frac{2(K \vee M)}{1 - e^{-\beta}} \right).$$

**Lemma 4.2.** *Under the assumptions (H<sub>1</sub>)–(H<sub>5</sub>), we have, as  $n \rightarrow \infty$ ,*

$$(4.4) \quad \left| \widehat{r}_{\theta,2}(x) - \mathbb{E}(\widehat{r}_{\theta,2}) \right| = O_{\text{a.co.}} \left( \sqrt{\frac{\chi_{\theta}^{1/2}(x, h_n) \log n}{n G_{\theta}^2(x, h_n)}} \right).$$

**Lemma 4.3.** *Under the assumptions (H<sub>1</sub>), (H<sub>2</sub>)(i,ii)–(H<sub>5</sub>), we have, as  $n \rightarrow \infty$ ,*

$$(4.5) \quad \left| \widehat{r}_{\theta,0}(x) - 1 \right| = O_{\text{a.co.}} \left( \sqrt{\frac{\chi_{\theta}^{1/2}(x, h_n) \log n}{n G_{\theta}^2(x, h_n)}} \right).$$

**Lemma 4.4.** *Under the assumptions of Lemma 4.3, we have, as  $n \rightarrow \infty$ ,*

$$(4.6) \quad \exists \eta > 0 \quad \text{such that} \quad \sum_{i=1}^n \mathbb{P} \left( \left| \widehat{r}_{\theta,0}(x) \right| < \eta \right) < \infty.$$

**Lemma 4.5.** *Under the assumptions (H<sub>1</sub>), (H<sub>4</sub>)–(H<sub>5</sub>), we have, as  $n \rightarrow \infty$ ,*

$$(4.7) \quad \left| r_{\theta}(x) - \mathbb{E}(\widehat{r}_{\theta,1}) \right| = O(h_n^{\beta}).$$

**Lemma 4.6.** *Under the assumptions (H<sub>1</sub>), (H<sub>3</sub>)–(H<sub>5</sub>), we have, as  $n \rightarrow \infty$ ,*

$$(4.8) \quad \left| \mathbb{E}(\widehat{r}_{\theta,2}) - \mathbb{E}(\widehat{r}_{\theta,1}) \right| = O \left( \sqrt{\frac{\chi_{\theta}^{1/2}(x, h_n) \log n}{n G_{\theta}^2(x, h_n)}} \right).$$

**Lemma 4.7.** *Under the assumptions (H<sub>1</sub>), (H<sub>2</sub>)(iii)–(H<sub>5</sub>), we have, as  $n \rightarrow \infty$ ,*

$$(4.9) \quad \left| \widehat{r}_{\theta,1}(x) - \widehat{r}_{\theta,2}(x) \right| = O_{\text{a.co.}} \left( \sqrt{\frac{\chi_{\theta}^{1/2}(x, h_n) \log n}{n G_{\theta}^2(x, h_n)}} \right).$$



---

**4.2. The asymptotic normality**

---

Now, we study the asymptotic normality of  $\widehat{r}_\theta(x)$ . To do that, we assume that the function

$$\varphi_\theta(x) := \mathbb{E}(Y_1^2 \mid \langle \theta, X_1 = z \rangle), \quad z \in H,$$

exists and is uniformly continuous in some neighborhood of  $z$ . Moreover, we modify slightly the assumptions  $(\mathbf{H}_1)$ ,  $(\mathbf{H}_4)$  and  $(\mathbf{H}'_5)$  is required:

$(\mathbf{H}'_1)$  The concentration property  $(\mathbf{H}_1)$  holds. Moreover, there exists a function  $\beta_x(\cdot)$  such that

$$\forall s \in [0, 1], \quad \lim_{h_n \rightarrow 0} G_\theta(x, s h_n) / G_\theta(x, h_n) = \beta_x(s).$$

$(\mathbf{H}'_4)$  The kernel  $K(\cdot)$  satisfies  $(\mathbf{H}_3)$  and is a differentiable function on  $]0, 1[$  with derivative  $K'(\cdot)$  such that  $-\infty < C < K'(\cdot) < C' < 0$ .

$(\mathbf{H}'_5)$  There exists a sequence of positive real numbers  $\gamma_n$  such that

$$\begin{cases} \gamma_n \chi_\theta(x, h_n) \rightarrow 0, \\ n^{3/2} \chi_\theta^{p/p-2}(x, h_n) \rightarrow 0. \end{cases}$$

**Theorem 4.2.** Under the assumptions  $(\mathbf{H}'_1)$ – $(\mathbf{H}_2)$ ,  $(\mathbf{H}_3)$ ,  $(\mathbf{H}'_4)$ ,  $(\mathbf{H}'_5)$  and if

$$n h_n^{2\beta} G_\theta(x, h_n) \rightarrow 0,$$

we have, for all  $x \in \mathcal{A}$ ,

$$(4.10) \quad \sqrt{n G_\theta(x, h_n)} (\widehat{r}_{\theta,n}(x) - r_\theta(x)) \xrightarrow{\mathcal{D}} N(0, \sigma_\theta^2(x)), \quad \text{as } n \rightarrow \infty,$$

where

$$\sigma_\theta^2(x) = \frac{\beta_2 (\varphi_\theta(x) - r_\theta^2(x))}{\beta_1^2},$$

with

$$\beta_j = - \int_0^1 (K^j)'(s) \beta_x(s) ds, \quad \text{for } j = 1, 2,$$

and

$$\mathcal{A} = \left\{ x \in H : \sigma_\theta^2(x) \neq 0 \right\}.$$

We can use the same decomposition as in the proof of Theorem 4.1, where  $\delta_n$  is replaced by  $\gamma_n$  in  $\widehat{r}_{\theta,2}(x)$ . Observe that the consistency of  $\widehat{r}_{\theta,0}$  to 1 is shown in Lemma 4.3 and, under the consideration  $n h_n^{2\beta} G_\theta(x, h_n) \rightarrow 0$ , we get

$$\sqrt{n G_\theta(x, h_n)} (r_\theta(x) - \mathbb{E}(\widehat{r}_{\theta,1})) \rightarrow 0.$$

---

<sup>2</sup>  $\xrightarrow{\mathcal{D}}$  denotes the convergence in distribution.

Moreover, by straightforward modification of the proofs of Lemmas 4.7 and 4.6, we obtain, under  $(H'_5)$ ,

$$\sqrt{n G_\theta(x, h_n)} |\hat{r}_{\theta,1}(x) - \hat{r}_{\theta,2}(x)| \longrightarrow 0, \quad \text{in probability,}$$

and

$$\sqrt{n G_\theta(x, h_n)} (\mathbb{E}(\hat{r}_{\theta,2}) - \mathbb{E}(\hat{r}_{\theta,1})) \longrightarrow 0.$$

So, all it remains to show is the following intermediate lemma.

**Lemma 4.8.** *Under the hypotheses of Theorem 4.2, we have, as  $n \rightarrow \infty$ ,*

$$(4.11) \quad \sqrt{n G_\theta(x, h_n)} \left( \hat{r}_{\theta,2}(x) - r_\theta(x) \hat{r}_{\theta,0}(x) - \mathbb{E}(\hat{r}_{\theta,2}(x) - r_\theta(x) \hat{r}_{\theta,0}(x)) \right) \xrightarrow{\mathcal{D}} N(0, \sigma_\theta^2(x)).$$

---

## 5. DISCUSSION AND APPLICATIONS

---

### 5.1. On the weak functional time series data analysis

---

The functional time series data analysis is one of the most important subject in functional data analysis (FDA). It is motivated by the rarity of the independent identically distributed observations functional observations in practice. The functional time series presents the more realistic situation. At this stage, the most of the existing studies on functional dependent data are developed under mixing assumption, namely, strong mixing framework. However, in this contribution, we investigate functional semiparametric regression under weak dependency condition of the quasi-associated correlation. From theoretical point of view this consideration allows to increase the scope of application of the proposed functional model. Indeed, it is well known that the mixing conditions are very hard to check and there exists lot of usual process fail to verify the mixing assumption. [4] have listed a numerous process, we quote, for instance, Bernoulli shifts class, Markov processes driven by discrete innovations and the AR(1) process with  $\rho < 1/2$  and Bernoulli innovation among others. Thus we can say that the important feature of our study is to analyse the functional time series data without the mixing assumption. In addition we point out that our study generalize also the classical association (negative or positive). Thus the quasi-associated functional time series data is sufficiently weak to cover a large class of weak functional time series data. Finally, let us precise that our theoretical development explore the dependence structure of the data through the convergence rate. The latter contains the additional  $\chi_\theta(x, h_n)$  that is control the local dependency of the data. It is clear that this dependency condition impact significantly the convergence rate of the estimator compared to the independent situation. Of course the independent case is more fast than the dependent one.

---

## 5.2. Application to the confidence intervals

---

The purpose of a confidence interval is to supplement the functional estimate at a point with information about the uncertainty in this estimate. It is a direct application of the Central Limit Theorem (CLT). In order to provide a confidence interval for the regression function in single functional model, we need first to propose a consistent estimator of the variance  $\sigma_\theta^2(x)$ . A natural consistent estimator of this variance is obtained by estimating the parameters involved in this quantity such as  $(\beta_j)_{j=1,2}$  and  $\varphi_\theta(\cdot)$ . A natural estimator of  $\beta_j$  is

$$(5.1) \quad \widehat{\beta}_j = \frac{1}{n G_\theta(x, h)} \sum_{i=1}^n K^j \left( \frac{\langle \theta, x - X_i \rangle}{h_n} \right), \quad j = 1, 2,$$

while the Nadaraya–Waston type estimator  $\varphi(\cdot)$  is

$$(5.2) \quad \widehat{\varphi}_n(x) = \frac{\sum_{i=1}^n Y_i^2 K \left( \frac{\langle \theta, x - X_i \rangle}{h_n} \right)}{\sum_{i=1}^n K \left( \frac{\langle \theta, x - X_i \rangle}{h_n} \right)}.$$

Consequently, by combining the Equations (5.1), (5.2) with the definition of  $\widehat{r}_{\theta,n}(x)$  consistent estimator of  $\sigma_\theta^2(x)$  denoted by  $\widehat{\sigma}_\theta^2(x)$ , it follows that the asymptotic confidence band at asymptotic level  $1 - \alpha$  for  $r_\theta(x)$  is

$$(5.3) \quad \widehat{r}_{\theta,n}(x) \pm \mathcal{U}_{1-\frac{\alpha}{2}} \left( \frac{\widehat{\sigma}_\theta^2(x)}{n G_\theta(x, h)} \right)^{\frac{1}{2}}.$$

Let us note the  $\left( \frac{\widehat{\sigma}_\theta^2(x)}{n G_\theta(x, h)} \right)$  is easy to compute and does not require the estimation of  $G_\theta(x, h)$ . The latter will be removed by a simple manipulation.

---

## 5.3. On the applicability of the SFIM

---

From theoretical point of view, it is well known that the single index model is one of the most important additive models used to improve the convergence rate of the nonparametric approach. This model keeps this feature in functional statistics. However, the applicability of this model in practice requires an additional works that is the determination of the functional index  $\theta$  and the smoothing parameter  $h$  which are often unknown in practice. This issue has been widely addressed in the nonfunctional case, but, remains not fully explored in the functional statistics. The readers interested by this topics can refer to [29] and the references therein (for recent advances in this topic). Thus, the estimation of the functional index and/or the bandwidth  $h_n$  in the quasi-associated functional time series case is an important prospect of the present contribution. As preliminary step, we present in this paragraph some selector rules compatible with our context of the functional time series data analysis. The first one is the Least Squares Cross-Validation (LSCV) rule, defined by

$$(5.4) \quad (\widehat{\theta}, \widehat{h}) = \underset{\substack{h_n \in H_n \\ \theta \in \Theta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (Y_i - \widehat{r}_{\theta,n}^i(X_i))^2,$$

where  $\widehat{r}_{\theta,n}^{-i}$  is the leave-one-out estimator of  $\widehat{r}_{\theta,n}$ . This kind of cross-validation is widely used in the nonparametric prediction problems to select the bandwidth parameter in the kernel smoothing. It was popularized in semi-parametric functional data analysis by [1]. The second one is the Maximum Likelihood Cross-Validation (MLCV) rule, expressed by

$$(5.5) \quad (\widehat{\theta}, \widehat{h}) = \underset{\substack{h_n \in H_n \\ \theta \in \Theta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \log \widehat{f}(Y_i | \widehat{r}_{\theta,n}^{-i}(X_i)),$$

where  $\widehat{f}(\cdot | \cdot)$  is the estimator of the conditional density of  $Y$  given  $\langle \theta, X \rangle$ . This criterion can be viewed as generalization of the rule (5.4) when the conditional distribution is Gaussian. Of course in practice we must optimize these rule over finite subset  $\Theta$  of index. Similarly to [1], we propose to select the optimal index from the following subset:

$$\Theta = \Theta_n = \left\{ \theta \in H, \theta = \sum_{i=1}^k c_i e_i, \|\theta\| = 1, \text{ and } \exists j \in [1, k] \text{ such that } \langle \theta, e_j \rangle > 0 \right\},$$

where  $(e_i)_{i=1, \dots, k}$  is finite basis functions of the Hilbert subspace spanned by the covariates  $(X_i)_i$  and  $(c_i)_i$  some real calibrated constants allowing to insure the identifiability of the model. The common way is to choose the  $(c_i)_i$  with calibration from the subset  $\{-1, 0, 1\}$ . Finally let us point both rules (5.4) and (5.5) we can take  $H_n$  as the subsets of the  $p$ -quantiles of the vector distance  $D = D_{ij} = \|X_i - X_j\|$ .

---

## 6. A SIMULATION STUDY

---

This section is devoted to some simulation experiments allowing to highlight the finite sample performance of the proposed SFIM-regression in different situations. This empirical study has two main purposes: The first one is to show the easy implantation of the SFIM in practice and the second one is to control the effect of the principal settings of the study (such as, the dependence's level, the type of the functional index, the smoothing degree of the link functions and the nature of the conditional distribution) in the efficiency of this functional model. For these objects, we simulate a functional time series data using the following SFIM equation:

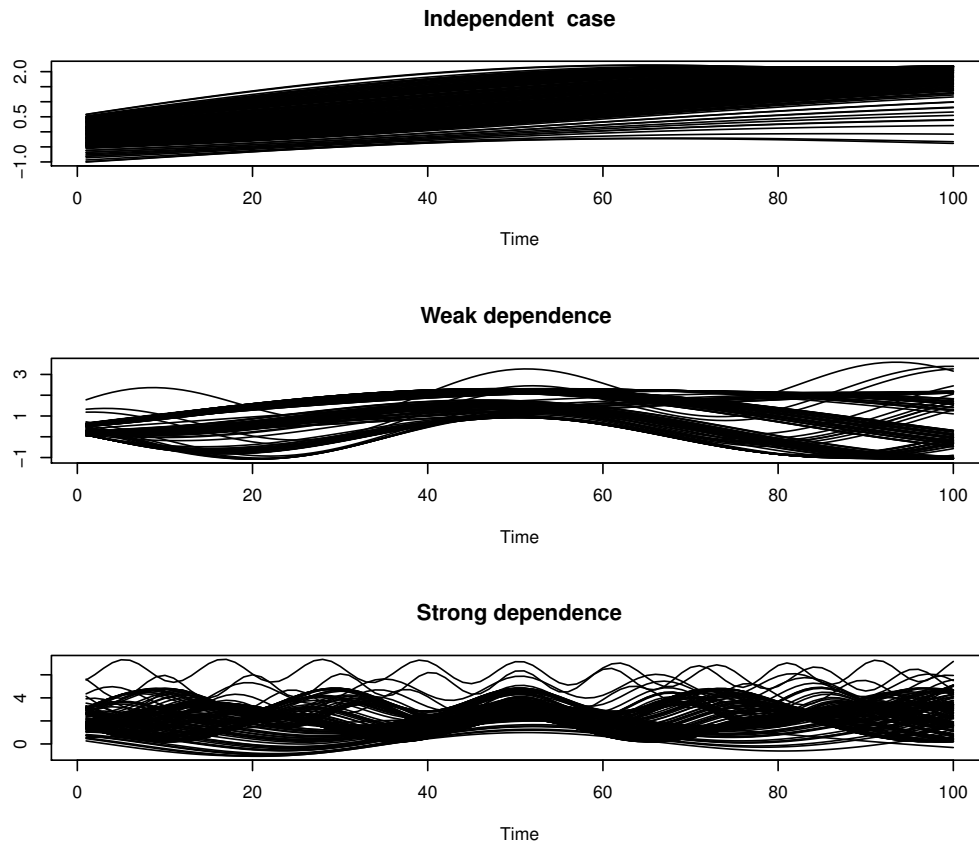
$$(6.1) \quad Y_i = r(\langle \theta, X_i \rangle) + \epsilon_i \quad \text{for } i = 1, \dots, n = 150,$$

where the  $\epsilon_i$ 's are generated independently according to a normal distribution  $\mathcal{N}(0, 1)$ . The functional regressors are generated by the following formula:

$$X_i(t) = \cos(W_i t) + \sin(W_i t) + .2(W_i t), \quad t \in [-\pi, +\pi],$$

and  $W_i$  is selected random variable. Three levels of dependency are considered that are independent, quasi-associated (weak-dependency) and  $\alpha$ -mixing (strong dependency). For the independent case, we take  $(W_i)_i$  as sample of  $\mathcal{N}(0, 1)$ . The quasi-associated case is carried out by generating the process  $(W_i)_i$  as non-strong mixing autoregressive of order 1. It obtained by taking the coefficient of the autoregressive  $\rho = 0.1$  and the innovation random variable as Binom(10, 0.25). It is shown in [6] that this kind of process fails to satisfy the  $\alpha$ -mixing assumption. However, this process is quasi-associated because it can be treated as linear process with positive coefficients. Concerning the strong dependency, we drown  $W$  from an autoregressive of order 1 with  $\rho = 0.75$  and the  $\chi_2(4)$  as innovation random variable.

The strong mixing property of this kind of process has been proved by [2]. The following Figure 1 shows the shape of  $n = 150$  curves  $X_i$ 's for three situations (independent, quasi-associated and strong dependency). The curves are discretized in the same grid formed by 100 points  $[-\pi, \pi]$ .

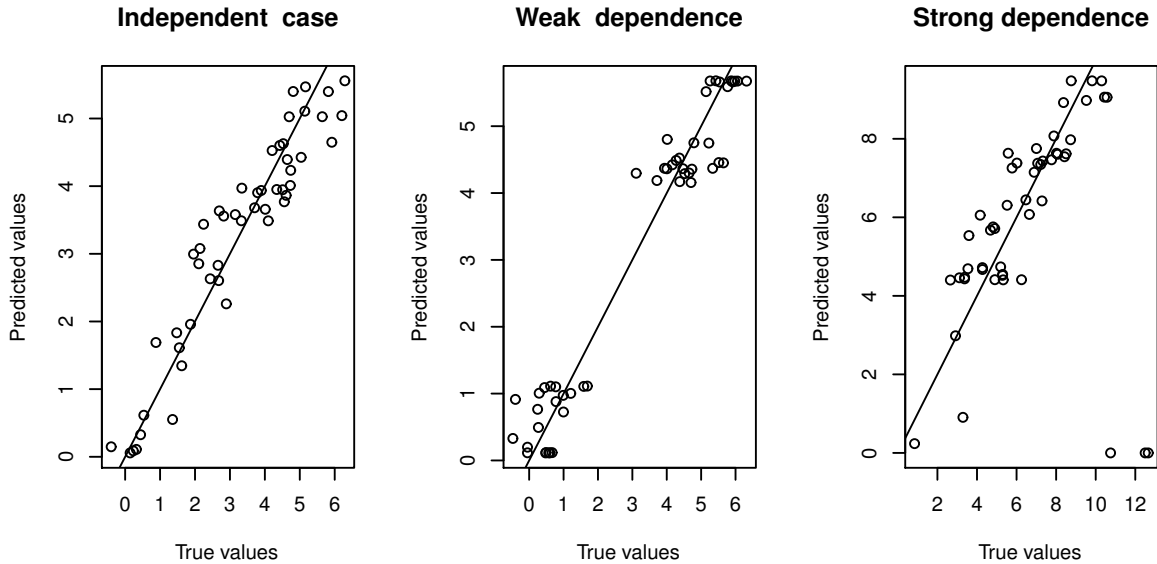


**Figure 1:** The shape of the regressors in the three cases.

In the first illustration, we control the effect of the degree of dependency on the prediction's quality using the single functional index regression. For this goal, we generate the scalar response  $Y_i$  by taking  $r_1(x) = 3 \log(1 + x^2)$  as link function and  $\theta_1 = e_1$  is the first element of the Karhunen–Loève basis functions. Explicitly  $\theta_1$  is the eigenfunction associated to the first eigenvalue of the covariance operator of the process  $(X_i)_i$ . It is eligible functionals index because it belongs in the same Hilbert subspace of the functional variable and is an element of  $\Theta_n$  (see the previous section).

Undoubtedly, the easy implementation of any statistical approach in practice is closely linked to the flexibility of the choice of parameters involved in this approach. At this stage the bandwidth parameter  $h_n$  and the functional index  $\theta$  are the principal parameters of the estimator. In this first illustration, we use the least squares cross-validation rule (5.4) described in the previous section to determine  $\theta$ . The mentioned rule is optimized over  $\Theta_n$  associated to the Karhunen–Loève basis functions (for  $k = 5$ ). For sake of brevity, we use the default smoothing parameter  $h_n$  of R-package `fda.usc` and quadratic kernel on  $(0, 1)$ .

The obtained results are given in the Figure 2. The latter gives a global overview on the behaviours of SFIM-predictor with respect the dependence's level. In this figure we plot the true values  $(Y_i)_i$  versus the predicted values for the three situations (independent, quasi-associated and strong dependency).



**Figure 2:** The SFIM-prediction results.

The results are not surprising. The SFIM-predictor has a satisfactory degree of performance. However, its behavior is strongly affected by the correlation of the data. The quality of prediction decreases with the degree of the dependency. The performance of the prediction procedures is tested by comparing the Mean Square Prediction Error defined by:

$$\text{MSPE} = \frac{1}{150} \sum_{i=1}^{150} \left( Y_i - \widehat{r}_{\widehat{\theta}_1, n}(X_i) \right)^2, \quad \widehat{\theta}_1 \text{ being the optimizer of (5.4).}$$

For this first illustration, we have obtained 0.23 for the independent against 0.92 for the quasi-associated and 1.78 for the strong mixing case.

Now, in order to give comprehensive empirical analysis for this semi-parametric model, we examine, in this second illustration, the impact of the other characteristics (the type of the functional index, the smoothing degree of the link functions and the nature of the conditional distribution) on the SFIM-prediction. More precisely, we compare two link functions (smooth and unsmooth (discontinuous in some points)), two functional indexes (eligible and ineligible) and two conditional distributions (Gaussian and non-Gaussian). This comparison will be carried out for the three previous dependence situations (independent, quasi-associated and strong mixing). We keep the data of the first illustration as perfect situation of the SFIM-prediction (eligible index, smooth link function and Gaussian conditional distribution).

Now, for the other situations, we follow the same algorithm of the first illustration to generate the output observations  $(Y_i)_i$ . To do that, we simulate with an arbitrary functional index expressed by the normalised function

$$\theta_2(t) = 0.15 t \sin(t)$$

and the link function

$$r_2(x) = r_1(x) \mathbb{1}_{[0,.5]} - r_1^2(x) \mathbb{1}_{[-1,-.5]}.$$

The last factor of SFIM-prediction is the conditional distribution of  $Y$  given  $X$ . The latter is explicitly given by the distribution of  $\epsilon_i$  shifted by  $r(\langle \theta, x \rangle)$ . For this second illustration, we generate the white noise  $\epsilon_i$  from normal mixture distribution  $(0.75) \mathcal{N}(0, 1) + 0.25 \mathcal{N}(.5, 2)$ . To quantify the impact of the conditional distribution on the SFIM-prediction we compare the two selector rules of the functional index (5.4) and (5.5). Of course both rules coincide when the conditional distribution is Gaussian. Finally, we point out that we have used the same kernel and the same bandwidth as in the first illustration and the conditional distribution in the rule (5.5) is computed by the routine `npcdist` in the R-package `np`. The results on this comparison study are presented in Table 1. It contains the MSPE for the six scenarios mentioned before.

**Table 1:** Comparison of the MSPE errors of the SFIM-prediction.

Dependency case	Conditional distribution	SFIM		CV-rule	
		Index	Function	LSCV	MLCV
Independent	Gaussian	Eligible	Smooth	0.23	0.24
		Ineligible	Smooth	0.71	0.76
		Eligible	Discontinuous	0.57	0.64
		Ineligible	Discontinuous	1.23	1.36
	Normal Mixture	Eligible	Smooth	0.41	0.33
		Ineligible	Smooth	0.93	0.67
		Eligible	Discontinuous	0.79	0.62
		Ineligible	Discontinuous	1.56	0.95
Quasi-associated	Gaussian	Eligible	Smooth	0.92	0.97
		Ineligible	Smooth	1.62	1.71
		Eligible	Discontinuous	1.27	1.29
		Ineligible	Discontinuous	2.09	2.14
	Normal Mixture	Eligible	Smooth	1.18	0.97
		Ineligible	Smooth	1.54	1.18
		Eligible	Discontinuous	1.41	1.07
		Ineligible	Discontinuous	2.14	1.92
Strong mixing	Gaussian	Eligible	Smooth	1.78	1.88
		Ineligible	Smooth	2.23	2.34
		Eligible	Discontinuous	2.17	2.25
		Ineligible	Discontinuous	2.57	2.59
	Normal Mixture	Eligible	Smooth	1.93	1.57
		Ineligible	Smooth	2.37	2.05
		Eligible	Discontinuous	2.18	1.93
		Ineligible	Discontinuous	2.68	2.15

The simulation results of Table 1 show that the prediction is strongly affected by the different features of the data (dependence degree) as well as the model (the smoothing property of the link function). This statement incorporates the theoretical result that relates the convergence rate of the estimator to the correlation of the data and the regularity assumption of the model. In addition the choice of the functional index impact also the performance of the prediction by the SFIM. In particular the two rules (5.4) and (5.5) are equivalent when the conditional distribution is Gaussian while the selector criterion (5.5) is more adequate for the mixture case. Overall, both criterion give a satisfactory level of accuracy even in the critical situation when the index is illegible and the link function is discontinuous.

---

## 7. PROOFS OF THE INTERMEDIATE RESULTS

---

This section is devoted to the proofs of our results. The previously presented notation continues to be used in the following.

---

### Proof of Lemma 4.2

---

The proof of this lemma is based on inequality given in Lemma 4.1 on the variables

$$\widehat{\Delta}_i(x) := \frac{1}{n \mathbb{E}(K_1(x))} [\widehat{Z}_i - \mathbb{E}(\widehat{Z}_i)], \quad i = 1, \dots, n,$$

where  $\widehat{Z}_i = \widehat{Y}_i K_i(x)$ , and we have

$$\begin{aligned} \mathbb{E}(\widehat{\Delta}_i) &= 0, \\ \|\widehat{\Delta}_i\|_\infty &\leq \frac{2 \delta_n}{n G_\theta(x, h_n)} \|K\|_\infty, \\ \text{Lip}(\widehat{\Delta}_i) &\leq 2 \text{Lip}(K) \frac{\delta_n}{n G_\theta(x, h_n) h_n}, \\ \widehat{r}_{\theta,2}(x) - \mathbb{E}(\widehat{r}_{\theta,2}(x)) &= \sum_{i=1}^n \widehat{\Delta}_i. \end{aligned}$$

We start by evaluating the covariance term  $\text{Cov}(\widehat{\Delta}_{s_1}, \dots, \widehat{\Delta}_{s_u}, \widehat{\Delta}_{t_1}, \dots, \widehat{\Delta}_{t_v})$ , for all  $(s_1, \dots, s_u) \in \mathbb{N}^u$  and  $(t_1, \dots, t_v) \in \mathbb{N}^v$  with  $1 \leq s_1 \leq \dots \leq s_u \leq t_1 \leq \dots \leq t_v \leq n$ . If  $m = t_1 - s_u = 0$ , using the fact that, for all  $p > 0$ ,

$$\mathbb{E}(K_1^p(x)) = O(G_\theta(x, h_n)),$$

and under the second part of (H<sub>2</sub>)(iii), we readily obtain

$$\begin{aligned} \left| \text{Cov}(\widehat{\Delta}_{s_1} \dots \widehat{\Delta}_{s_u}, \widehat{\Delta}_{t_1} \dots \widehat{\Delta}_{t_v}) \right| &\leq \left( \frac{1}{n \mathbb{E}(K_1(x))} \right)^{u+v} \mathbb{E}(|\widehat{Z}_{s_1} \dots \widehat{Z}_{s_u}^2 \dots \widehat{Z}_{t_v}|) \\ &\leq \left( \frac{C \delta_n \|K\|_\infty}{n G_\theta(x, h_n)} \right)^{u+v} \mathbb{E}(Y_{s_u}^2 K_{s_u}^2) \\ &\leq \left( \frac{C \delta_n}{n G_\theta(x, h_n)} \right)^{u+v} G_\theta(x, h_n). \end{aligned}$$



If  $m = t_1 - s_u > 0$ , by quasi-association of the sequence  $(\widehat{Z}_n)$ , we infer that

$$\begin{aligned}
 \left| \text{Cov}(\widehat{\Delta}_{s_1} \dots \widehat{\Delta}_{s_u}, \widehat{\Delta}_{t_1} \dots \widehat{\Delta}_{t_v}) \right| &\leq 4 \left( \frac{\delta_n \text{Lip}(K)}{n G_\theta(x, h_n) h_n} \right)^2 \left( \frac{2 \delta_n \|K\|_\infty}{n G_\theta(x, h_n)} \right)^{u+v-2} \sum_{i=1}^u \sum_{j=1}^v \lambda_{s_i, t_j} \\
 &\leq C^{u+v} \left( \frac{\text{Lip}(K)}{h_n} \right)^2 \left( \frac{\delta_n}{n G_\theta(x, h_n)} \right)^{u+v} (u \wedge v) \lambda_{t_1 - s_u} \\
 (7.1) \qquad &\leq C^{u+v} \left( \frac{\text{Lip}(K)}{h_n} \right)^2 \left( \frac{\delta_n}{n G_\theta(x, h_n)} \right)^{u+v} v e^{-am}.
 \end{aligned}$$

On the other hand, making use of the first part of the condition  $(H_2)$ (iii) we may write

$$\begin{aligned}
 \left| \text{Cov}(\widehat{\Delta}_{s_1} \dots \widehat{\Delta}_{s_u}, \widehat{\Delta}_{t_1} \dots \widehat{\Delta}_{t_v}) \right| &\leq \left( \frac{C \delta_n \|K\|_\infty}{n G_\theta(x, h_n)} \right)^{u+v-2} \left| \text{Cov}(\widehat{\Delta}_{s_u}, \widehat{\Delta}_{t_1}) \right| \\
 &\leq \left( \frac{C \delta_n \|K\|_\infty}{n G_\theta(x, h_n)} \right)^{u+v-2} \left( \left| \mathbb{E}(\widehat{\Delta}_{s_u} \widehat{\Delta}_{t_1}) \right| + \mathbb{E}|\widehat{\Delta}_{s_u}| \mathbb{E}|\widehat{\Delta}_{t_1}| \right) \\
 &\leq \left( \frac{C \delta_n \|K\|_\infty}{n G_\theta(x, h_n)} \right)^{u+v-2} \left( \frac{C}{n G_\theta(x, h_n)} \right)^2 \delta_n^2 \chi_\theta(x, h_n).
 \end{aligned}$$

It follows that

$$(7.2) \qquad \left| \text{Cov}(\widehat{\Delta}_{s_1}, \dots, \widehat{\Delta}_{s_u}, \widehat{\Delta}_{t_1}, \dots, \widehat{\Delta}_{t_v}) \right| \leq C^{u+v} \left( \frac{\delta_n}{n G_\theta(x, h_n)} \right)^{u+v} \chi_\theta(x, h_n).$$

Moreover, by multiplying a  $\tau$ -power of (7.1) and  $(1 - \tau)$ -power of (7.2) for some  $\frac{1}{4} < \tau < \frac{1}{2}$ , we obtain an upper-bound of the covariance as follows for  $1 \leq s_1 \leq \dots \leq s_u \leq t_1 \leq \dots \leq t_v \leq n$ :

$$\left| \text{Cov}(\widehat{\Delta}_{s_1} \dots \widehat{\Delta}_{s_u}, \widehat{\Delta}_{t_1} \dots \widehat{\Delta}_{t_v}) \right| \leq C^{u+v} \left( \frac{\delta_n}{n G_\theta(x, h_n)} \right)^{u+v} \left( \frac{\text{Lip}(K)}{h_n} \right)^{2\tau} \left( \sqrt{\chi_\theta(x, h_n)} \right)^{2(1-\tau)} v e^{-a\tau m}.$$

So, by  $(H_5)$ , we have

$$\left| \text{Cov}(\widehat{\Delta}_{s_1} \dots \widehat{\Delta}_{s_u}, \widehat{\Delta}_{t_1} \dots \widehat{\Delta}_{t_v}) \right| \leq \left( \frac{C \delta_n}{n G_\theta(x, h_n)} \right)^{u+v-2} \left( \frac{C \delta_n}{n G_\theta(x, h_n)} \right)^2 \sqrt{\chi_\theta(x, h_n)} v e^{-a\tau m},$$

where

$$M_n = \frac{C \delta_n}{n G_\theta(x, h_n)} \quad \text{and} \quad K_n = \frac{C \chi_\theta^{1/4}(x, h_n) \delta_n}{n G_\theta(x, h_n)}.$$

It remains to calculate  $\text{Var}\left(\sum_{i=1}^n \widehat{\Delta}_i\right)$ :

$$\begin{aligned}
 \text{Var}\left(\sum_{i=1}^n \widehat{\Delta}_i\right) &= \left( \frac{1}{n \mathbb{E}(K_1(x))} \right)^2 \sum_i \sum_j \text{Cov}(\widehat{Z}_i, \widehat{Z}_j) \\
 &= \left( \frac{1}{n \mathbb{E}(K_1(x))} \right)^2 \left[ n \text{Var}(\widehat{Z}_1) + \sum_i \sum_{j \neq i} \text{Cov}(\widehat{Z}_i, \widehat{Z}_j) \right] \\
 &= \left( \frac{1}{n \mathbb{E}(K_1(x))} \right)^2 \left[ n T_1 + T_{ij} \right].
 \end{aligned}$$

Now, under the assumption (H<sub>5</sub>), we obtain for the first term:

$$\begin{aligned} T_1 &= \text{Var}(\widehat{Z}_1) = \mathbb{E}(\widehat{Y}_1^2 K_1^2(x)) - \left(\mathbb{E}(\widehat{Y}_1 K_1(x))\right)^2 \\ &\leq \mathbb{E}(Y_1^2 K_1^2(x)) \\ &\leq \mathbb{E}\left(K_1^2(x) \mathbb{E}(Y_1^2|X)\right) \\ &\leq C \mathbb{E}(K_1^2(x)). \end{aligned}$$

For all  $j \geq 1$ , we have

$$(7.3) \quad \mathbb{E}(K_1^j(x)) = O(G_\theta(x, h_n)),$$

and

$$T_1 = \text{Var}(\widehat{Z}_1) = O(\chi_\theta^{1/2}(x, h_n)).$$

We readily obtain that

$$(7.4) \quad \frac{1}{n \left(\mathbb{E}(K_1(x))\right)^2} T_1 \leq \frac{C \chi_\theta^{1/2}(x, h_n)}{n G_\theta^2(x, h_n)}.$$

For the second term, we have the following decomposition

$$T_{ij} = \sum_i \sum_{0 < |i-j| \leq u_n} \text{Cov}(\widehat{Z}_i, \widehat{Z}_j) + \sum_i \sum_{|i-j| > u_n} \text{Cov}(\widehat{Z}_i, \widehat{Z}_j) = J_1 + J_2,$$

where  $(u_n)$  is a sequence of positive integer and

$$\lim_{n \rightarrow \infty} u_n = \infty.$$

Now, under the assumptions (H<sub>2</sub>), we have

$$\begin{aligned} |J_1| &= \sum_i \sum_{0 < |i-j| \leq u_n} |\text{Cov}(\widehat{Z}_i, \widehat{Z}_j)| \leq n u_n \left[ \max_{i \neq j} \left| \mathbb{E}(K_i(x) K_j(x)) \right| + \left(\mathbb{E}(K_1(x))\right)^2 \right] \\ &\leq C n u_n \chi_\theta(x, h_n). \end{aligned}$$

Making use of the condition (H<sub>2</sub>)(i), we infer that

$$\begin{aligned} |J_2| &= \sum_i \sum_{|i-j| > u_n} |\text{Cov}(\widehat{Z}_i, \widehat{Z}_j)| \leq C \delta_n^2 \left(\frac{\text{Lip}(K)}{h_n}\right)^2 \sum_i \sum_{|i-j| > u_n} \lambda_{i,j} \\ &\leq C n \delta_n^2 h_n^{-2} e^{-a u_n}. \end{aligned}$$

This implies that

$$|T_{ij}| \leq \sum_{i=1}^n \sum_{i \neq j} |\text{Cov}(\widehat{Z}_i, \widehat{Z}_j)| \leq C \left( n u_n \chi_\theta(x, h_n) + n \delta_n^2 h_n^{-2} e^{-a u_n} \right).$$

Next, taking

$$u_n = \frac{1}{a} \log \left( \frac{\delta_n^2 a}{h_n^2 \chi_\theta(x, h_n)} \right).$$

Observe that (H<sub>5</sub>) insure that

$$\sqrt{\chi_\theta(x, h_n) \log(\delta_n)} \rightarrow 0,$$

which allows to write that

$$(7.5) \quad T_{ij} = o\left(n \chi_\theta^{1/2}(x, h_n)\right) \rightarrow 0.$$

It follows that

$$\text{Var}\left(\sum_{i=1}^n \widehat{\Delta}_i\right) = O\left(\frac{\chi_\theta^{1/2}(x, h_n)}{n G_\theta^2(x, h_n)}\right).$$

The conditions of Lemma 4.1 are verified for

$$\begin{aligned} K_n &= \frac{C \chi_\theta^{1/4}(x, h_n) \delta_n}{n G_\theta(x, h_n)}, & M_n &= \frac{C \delta_n}{n G_\theta(x, h_n)}, \\ A_n &= \frac{\chi_\theta^{1/2}(x, h_n)}{n G_\theta^2(x, h_n)}, \\ B_n &= \left(\frac{16 n K^2}{9 A_n (1 - e^{-\beta})} \vee 1\right) \left(\frac{2(K \vee M)}{1 - e^{-\beta}}\right) = \frac{\delta_n}{n G_\theta(x, h_n)}. \end{aligned}$$

So, we apply the inequality in [22] to the random variables  $\widehat{\Delta}_i$  to infer that

$$\begin{aligned} \mathbb{P}\left(\left|\widehat{r}_{\theta,2}(x) - \mathbb{E}(\widehat{r}_{\theta,2}(x))\right| > \varepsilon \sqrt{\frac{\chi_\theta^{1/2}(x, h_n) \log n}{n G_\theta^2(x, h_n)}}\right) &= \mathbb{P}\left(\left|\sum_{i=1}^n \widehat{\Delta}_i\right| > \varepsilon \sqrt{\frac{\chi_\theta^{1/2}(x, h_n) \log n}{n G_\theta^2(x, h_n)}}\right) \\ &\leq \exp\left(\frac{-\varepsilon^2 \chi_\theta^{1/2}(x, h_n) \log n}{n G_\theta^2(x, h_n) L_\theta(n)}\right), \end{aligned}$$

where

$$L_\theta(n) = \left(\frac{\chi_\theta^{1/2}(x, h_n)}{n G_\theta^2(x, h_n)} + \left(\frac{\delta_n}{n G_\theta^2(x, h_n)}\right)^{\frac{1}{3}} \left(\frac{\chi_\theta^{1/2}(x, h_n) \log n}{n G_\theta^2(x, h_n)}\right)^{\frac{5}{6}}\right).$$

Then we finally obtain that

$$\begin{aligned} \mathbb{P}\left(\left|\widehat{r}_{\theta,2}(x) - \mathbb{E}(\widehat{r}_{\theta,2})\right| > \varepsilon \sqrt{\frac{\chi_\theta^{1/2}(x, h_n) \log n}{n G_\theta^2(x, h_n)}}\right) &\leq \exp\left(\frac{-\varepsilon^2 \log n}{C + \left(\delta_n^2 \chi_\theta^{-1/2}(x, h_n) \log^5 n\right)^{\frac{1}{6}}}\right) \\ &\leq C_1 \exp(-\varepsilon^2 \log(n)). \end{aligned}$$

The proof is achieved by a suitable choice of  $\varepsilon$ . □

---

### Proof of Lemma 4.3

---

The proof of this lemma is similar to the proof of the previous Lemma 4.2. Since  $\widehat{Y}_i = 1$ , it suffices to replace  $\widehat{\Delta}_i$  by

$$\widetilde{\Delta}_i = \frac{1}{n \mathbb{E}(K_1(x))} \left[K_i(x) - \mathbb{E}(K_i(x))\right], \quad i = 1, \dots, n.$$

Thus we obtain, under  $(H_1)$ – $(H_5)$ ,

$$\begin{aligned} \mathbb{P}\left(\left|\widehat{r}_{\theta,0}(x) - 1\right| > \varepsilon \sqrt{\frac{\log n}{n G_{\theta}(x, h_n)}}\right) &= \mathbb{P}\left(\left|\sum_{i=1}^n \widetilde{\Delta}_i\right| > \varepsilon \sqrt{\frac{\chi_{\theta}^{1/2}(x, h_n) \log n}{n G_{\theta}^2(x, h_n)}}\right) \\ &\leq C'_1 \exp(-\varepsilon^2 \log(n)). \end{aligned}$$

Thus the proof is complete. □

**Proof of Lemma 4.4**

Notice that we have

$$\left\{|\widehat{r}_{\theta,0}(x)| \leq \frac{1}{2}\right\} \subset \left\{|\widehat{r}_{\theta,0}(x) - 1| > \frac{1}{2}\right\},$$

that implies that

$$\mathbb{P}\left(|\widehat{r}_{\theta,0}(x)| \leq \frac{1}{2}\right) \leq \mathbb{P}\left(|\widehat{r}_{\theta,0}(x) - 1| > \frac{1}{2}\right).$$

Under the hypothesis  $(H_1)$ – $(H_5)$  and by applying Lemma 4.3, we deduce that

$$\sum_n \mathbb{P}\left(|\widehat{r}_{\theta,0}(x)| \leq \frac{1}{2}\right) \leq \sum_n \mathbb{P}\left(|\widehat{r}_{\theta,0}(x) - 1| > \frac{1}{2}\right) < \infty.$$

Then, for  $\eta = \frac{1}{2}$ , we have  $\sum_n \mathbb{P}\left(|\widehat{r}_{\theta,0}(x)| \leq \eta\right) < \infty$ . Thus the proof is complete. □

**Proof of Lemma 4.5**

One can easily see that we have

$$\begin{aligned} \left|r_{\theta}(x) - \mathbb{E}(\widehat{r}_{\theta,1}(x))\right| &= \left|r_{\theta}(x) - \mathbb{E}\left(\frac{1}{n \mathbb{E}(K_1(x))} \sum_{i=1}^n Y_i K_i(x)\right)\right| \\ &= \frac{1}{\mathbb{E}(K_1(x))} \left[|r_{\theta}(x) \mathbb{E}(K_1(x)) - \mathbb{E}(Y_1 K_1(x))|\right] \\ &= \frac{1}{\mathbb{E}(K_1(x))} \mathbb{E}\left[ (|r_{\theta}(x) - r_{\theta}(X_1)|) K_1(x) \right] \leq C h_n^{\beta}. \end{aligned}$$

This readily implies that we have

$$r_{\theta}(x) - \mathbb{E}(\widehat{r}_{\theta,1}) = O(h_n^{\beta}).$$

Thus the proof is complete. □

---

**Proof of Lemma 4.6**

---

We first observe that we have

$$\begin{aligned} |\mathbb{E}(\widehat{r}_{\theta,2}) - \mathbb{E}(\widehat{r}_{\theta,1})| &= \frac{1}{n \mathbb{E} K_1(x)} \left| \mathbb{E} \left( \sum_{i=1}^n Y_i \mathbf{1}_{\{|Y_i| > \delta_n\}} K_i(x) \right) \right| \\ &\leq \mathbb{E} \left( |Y_1| \mathbf{1}_{|Y_1| > \delta_n} K_1(x) \right) \left( \mathbb{E}(K_1(x)) \right)^{-1}. \end{aligned}$$

The Hölder’s inequality allows to write that, for  $\alpha = \frac{p}{2}$  and  $\beta$  such that  $\frac{1}{\alpha} + \frac{1}{\beta} = 1$ ,

$$\begin{aligned} \left| \mathbb{E}(\widehat{r}_{\theta,2}(x)) - \mathbb{E}(\widehat{r}_{\theta,1}(x)) \right| &\leq \frac{1}{G_\theta(x, h_n)} \mathbb{E}^{1/\alpha} [ |Y^\alpha| \mathbf{1}_{\{|Y| \geq \delta_n\}} ] \mathbb{E}^{1/\beta} [ K_1^\beta ] \\ &\leq \frac{1}{G_\theta(x, h_n)} \delta_n^{-1} \mathbb{E}^{1/\alpha} [ |Y^p| ] G_\theta^{1/\beta}(x, h_n) \\ &\leq C \delta_n^{-1} G_\theta^{(1-\beta)/\beta}(x, h_n). \end{aligned}$$

Hence, we obtain from (H<sub>5</sub>) that

$$\left| \mathbb{E}(\widehat{r}_{\theta,2}(x)) - \mathbb{E}(\widehat{r}_{\theta,1}(x)) \right| = o \left( \sqrt{\frac{\chi_\theta^{1/2}(x, h_n) \log n}{n G_\theta^2(x, h_n)}} \right).$$

Thus the proof is complete. □

---

**Proof of Lemma 4.7**

---

By (H<sub>5</sub>) and we apply the Markov’s inequality to show that,  $\forall \epsilon > 0$ ,

$$\begin{aligned} \mathbb{P} \left( |\widehat{r}_{\theta,1}(x) - \widehat{r}_{\theta,2}(x)| > \epsilon \right) &= \mathbb{P} \left( \frac{1}{n \mathbb{E} K_1(x)} \left| \sum_{i=1}^n Y_i \mathbf{1}_{\{|Y_i| > \delta_n\}} K_i(x) \right| > \epsilon \right) \\ &\leq n \mathbb{P}(|Y_1| > \delta_n) \leq n \delta_n^{-p} \mathbb{E}(|Y|^p) \leq C n \delta_n^{-p}. \end{aligned}$$

Since

$$\sum_{n \geq 1} n \delta_n^{-p} < \infty,$$

then there exists  $\epsilon_0 > 0$ , such that

$$(7.6) \quad \sum_{n \geq 1} \mathbb{P} \left( |\widehat{r}_{\theta,1}(x) - \widehat{r}_{\theta,2}(x)| > \epsilon_0 \sqrt{\frac{\chi_\theta^{1/2}(x, h_n) \log n}{n G_\theta^2(x, h_n)}} \right) < \infty,$$

which completes the proof of the lemma. □

---

**Proof of Lemma 4.8**


---

Let us introduce the following sum  $S_n = \sum_{i=1}^n L_{ni}$ , where

$$L_{ni} = \frac{\sqrt{G_\theta(x, h_n)}}{\sqrt{n} \mathbb{E}(K_1(x))} \left( (\widehat{Y}_i - r_\theta(x)) K_i(x) - \mathbb{E} \left( (\widehat{Y}_i - r_\theta(x)) K_i(x) \right) \right).$$

Therefore

$$S_n = \sqrt{n G_\theta(x, h_n)} \left( (\widehat{r}_{\theta,2}(x) - r_\theta(x)) \widehat{r}_{\theta,0}(x) - \mathbb{E}(\widehat{r}_{\theta,2}(x) - r_\theta(x)) \widehat{r}_{\theta,0}(x) \right).$$

Thus, our claimed result is now

$$(7.7) \quad S_n \rightarrow \mathcal{N}(0, \sigma_\theta^2(x)).$$

To do that, we use the basic technique of [9] for which we split  $S_n$  into

$$S_n = T_n + T'_n + \zeta_k,$$

with

$$T_n = \sum_{j=1}^k \eta_j \quad \text{and} \quad T'_n = \sum_{j=1}^k \xi_j,$$

where

$$\eta_j := \sum_{i \in I_j} L_{ni}, \quad \xi_j := \sum_{i \in J_j} L_{ni}, \quad \zeta_k := \sum_{i=k(p+q)+1}^n L_{ni},$$

with

$$I_j = \left\{ (j-1)(p+q) + 1, \dots, (j-1)(p+q) + p \right\},$$

$$J_j = \left\{ (j-1)(p+q) + p + 1, \dots, j(p+q) \right\},$$

and  $p = p_n$ ,  $q = q_n$  two sequences of natural numbers tending to  $\infty$ , such that

$$p = O(G_\theta^{-1}(x, h_n)), \quad q = o(p) \quad \text{and} \quad k = \left\lfloor \frac{n}{p+q} \right\rfloor,$$

where  $\lfloor \cdot \rfloor$  stands for the integer part. Firstly, observe that we have  $\frac{kq}{n} \rightarrow 0$ , and  $\frac{kp}{n} \rightarrow 1$ ,  $\frac{q}{n} \rightarrow 0$ , which imply that  $\frac{p}{n} \rightarrow 0$  as  $n \rightarrow \infty$ . Now, our asymptotic normality results are a consequence of the following statements:

$$(7.8) \quad \mathbb{E}(T'_n)^2 + \mathbb{E}(\zeta_k)^2 \rightarrow 0$$

and

$$(7.9) \quad T_n \rightarrow \mathcal{N}(0, \sigma_\theta^2(x)).$$

For (7.8), we write

$$\mathbb{E}(T'_n)^2 = k \text{Var}(\xi_1) + 2 \sum_{1 \leq i < j \leq k} |\text{Cov}(\xi_i, \xi_j)|$$

and

$$\text{Var}(\xi_1) \leq q \text{Var}(L_{n1}) + 2 \sum_{1 \leq i < j \leq q} |\text{Cov}(L_{ni}, L_{nj})|.$$

Similarly to (7.4), we infer that

$$\text{Var}(L_{n1}) = O(n^{-1}),$$

which implies that

$$kq \text{Var}(L_{n1}) = O\left(\frac{kq}{n}\right) \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

On the other hand, we use the same arguments as those used in (7.5) to conclude that

$$(7.10) \quad k \sum_{1 \leq i < j \leq q} |\text{Cov}(L_{ni}, L_{nj})| = o\left(\frac{kq}{n}\right) \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

Thus, the limit of the first term of  $\mathbb{E}(T'_n)^2$  is equal to 0. Next, by using stationarity, we can write

$$\begin{aligned} \sum_{1 \leq i < j \leq k} |\text{Cov}(\xi_i, \xi_j)| &= \sum_{l=1}^{k-1} (k-l) |\text{Cov}(\xi_1, \xi_{l+1})| \\ &\leq k \sum_{l=1}^{k-1} |\text{Cov}(\xi_1, \xi_{l+1})| \\ &\leq k \sum_{l=1}^{k-1} \sum_{(i,j) \in J_1 \times J_{l+1}} \text{Cov}(L_{ni}, L_{nj}). \end{aligned}$$

It is clear that, for all  $(i, j) \in J_1 \times J_j$ , we have  $|i - j| \geq p + 1 > p$ , and then

$$\begin{aligned} \sum_{1 \leq i < j \leq k} |\text{Cov}(\xi_i, \xi_j)| &\leq k \frac{C \gamma_n^2}{n h_n^2 G_\theta(x, h_n)} \sum_{i=1}^p \sum_{\substack{j=2p+q+1, \\ |i-j| > p}}^{k(p+q)} \lambda_{i,j} \\ &\leq \frac{C k p \gamma_n^2}{n h_n^2 G_\theta(x, h_n)} \lambda_p \\ &\leq \frac{C \gamma_n^2}{G_\theta^3(x, h_n)} e^{-ap} \rightarrow 0. \end{aligned}$$

Finally, we get

$$\mathbb{E}(T'_1)^2 \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Since  $(n - k(p + q)) \leq p$ , we have by the same manner

$$\begin{aligned} \mathbb{E}(\zeta_k)^2 &\leq (n - k(p + q)) \text{Var}(L_{n1}) + 2 \sum_{1 \leq i < j \leq n} |\text{Cov}(L_{ni}, L_{nj})| \\ &\leq p \text{Var}(L_{n1}) + 2 \sum_{1 \leq i < j \leq n} |\text{Cov}(L_{ni}, L_{nj})| \\ &\leq \frac{C p}{n} + o(1). \end{aligned}$$

Hence,

$$\mathbb{E}(\zeta_k)^2 \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

So, it remains to proof the asymptotic normality (7.9). The proof is standard. Indeed, it is based in the following assertions

$$(7.11) \quad \left| \mathbb{E} \left( e^{it \sum_{j=1}^k \eta_j} \right) - \prod_{j=1}^k \mathbb{E} (e^{it \eta_j}) \right| \rightarrow 0,$$

and

$$(7.12) \quad k \operatorname{Var}(\eta_1) \rightarrow \sigma_\theta^2(x), \quad k \mathbb{E}(\eta_1^2 \mathbb{1}_{\{\eta_1 > \epsilon \sigma_\theta(x)\}}) \rightarrow 0.$$

To prove (7.11), notice that

$$(7.13) \quad \begin{aligned} & \left| \mathbb{E} \left( e^{it \sum_{j=1}^k \eta_j} \right) - \prod_{j=1}^k \mathbb{E} (e^{it \eta_j}) \right| \\ & \leq \left| \mathbb{E} \left( e^{it \sum_{j=1}^k \eta_j} \right) - \mathbb{E} \left( e^{it \sum_{j=1}^{k-1} \eta_j} \right) \mathbb{E} (e^{it \eta_k}) \right| + \left| \mathbb{E} \left( e^{it \sum_{j=1}^{k-1} \eta_j} \right) - \prod_{j=1}^{k-1} \mathbb{E} (e^{it \eta_j}) \right| \\ & = \left| \operatorname{Cov} \left( e^{it \sum_{j=1}^{k-1} \eta_j}, e^{it \eta_k} \right) \right| + \left| \mathbb{E} \left( e^{it \sum_{j=1}^{k-1} \eta_j} \right) - \prod_{j=1}^{k-1} \mathbb{E} (e^{it \eta_j}) \right| \end{aligned}$$

and, successively, we have

$$(7.14) \quad \begin{aligned} & \left| \mathbb{E} \left( e^{it \sum_{j=1}^k \eta_j} \right) - \prod_{j=1}^k \mathbb{E} (e^{it \eta_j}) \right| \\ & \leq \left| \operatorname{Cov} \left( e^{it \sum_{j=1}^{k-1} \eta_j}, e^{it \eta_k} \right) \right| + \left| \operatorname{Cov} \left( e^{it \sum_{j=1}^{k-2} \eta_j}, e^{it \eta_{k-1}} \right) \right| + \dots + \left| \operatorname{Cov} (e^{it \eta_2}, e^{it \eta_1}) \right|. \end{aligned}$$

The use of the quasi-associated propriety permits to write that

$$\left| \operatorname{Cov} (e^{it \eta_2}, e^{it \eta_1}) \right| \leq \frac{C t^2 \gamma_n^2}{n G_\theta^3(x, h_n)} \sum_{i \in I_1} \sum_{j \in I_2} \lambda_{i,j}.$$

Applying this inequality to each term on the right-hand side of (7.14) in order to obtain

$$\begin{aligned} & \left| \mathbb{E} \left( e^{it \sum_{j=1}^k \eta_j} \right) - \prod_{j=1}^k \mathbb{E} (e^{it \eta_j}) \right| \\ & \leq \frac{C t^2 \gamma_n^2}{n G_\theta^3(x, h_n)} \left[ \sum_{i \in I_1} \sum_{j \in I_2} \lambda_{i,j} + \sum_{i \in I_1 \cup I_2} \sum_{j \in I_3} \lambda_{i,j} + \dots + \sum_{i \in I_1 \cup \dots \cup I_{k-1}} \sum_{j \in I_k} \lambda_{i,j} \right]. \end{aligned}$$

Observe that for every  $2 \leq l \leq k-1$ ,  $(i, j) \in I_l \times I_{l+1}$ , we have  $|i-j| \geq q+1 > q$ , then

$$\sum_{i \in I_1 \cup \dots \cup I_{l-1}} \sum_{j \in I_l} \lambda_{i,j} \leq p \lambda_q.$$

Therefore, inequality (7.13) becomes

$$\left| \mathbb{E} \left( e^{it \sum_{j=1}^k \eta_j} \right) - \prod_{j=1}^k \mathbb{E} (e^{it \eta_j}) \right| \leq \frac{C t^2 \gamma_n^2}{n G_\theta^3(x, h_n)} k p \lambda_q \leq \frac{C t^2 \gamma_n^2}{n G_\theta^3(x, h_n)} k p e^{-aq} \rightarrow 0.$$



Concerning (7.12), we use the same arguments as in to conclude that

$$\lim_{n \rightarrow \infty} k \operatorname{Var}(\eta_1) = \lim_{n \rightarrow \infty} k p \operatorname{Var}(L_{n1}).$$

On the other hand

$$\operatorname{Var}(L_{n1}) = \frac{G_\theta(x, h_n)}{n \mathbb{E}^2(K_1(x))} \operatorname{Var}\left(\left(\widehat{Y}_1 - r_\theta(x)\right) K_1(x)\right).$$

It can be written as

$$\begin{aligned} \operatorname{Var}(L_{n1}) &= \frac{G_\theta(x, h_n)}{n \mathbb{E}^2(K_1(x))} \left\{ \mathbb{E}\left(K_1^2(x) (Y_1 - r_\theta(x))^2\right) - \mathbb{E}\left[K_1^2(x) (Y_1 - r_\theta(x))^2 \mathbb{1}_{|Y_1| > \gamma_n}\right] \right\} \\ &\quad - \frac{G_\theta(x, h_n)}{n \mathbb{E}^2(K_1(x))} \left( \mathbb{E}\left(K_1(x) (Y_1 - r_\theta(x)) \mathbb{1}_{|Y_1| < \gamma_n}\right) \right)^2. \end{aligned}$$

By combining the same ideas used in the proof of Lemma 4.6 to those used by [14], we show that

$$(7.15) \quad \operatorname{Var}(L_{n1}) = \frac{\sigma_\theta^2(x)}{n} + o\left(\frac{1}{n}\right).$$

Therefore,

$$k \operatorname{Var}(\eta_1) = \frac{k p \sigma_\theta^2(x)}{n} + o\left(\frac{k p}{n}\right) \rightarrow \sigma_\theta^2(x).$$

For the second part of (7.12), we use the fact that

$$|\eta_1| \leq C p |L_{n1}| \leq \frac{C \gamma_n p}{\sqrt{n G_\theta(x, h_n)}},$$

and Tchebychev inequality to get

$$\begin{aligned} k \mathbb{E}(\eta_1^2 \mathbb{1}_{\{|\eta_1| > \epsilon \sigma_\theta(x)\}}) &\leq \frac{C \gamma_n^2 p^2 k}{n G_\theta(x, h_n)} \mathbb{P}(|\eta_1| > \epsilon \sigma_\theta(x)) \\ &\leq \frac{C \gamma_n^2 p^2 k}{n G_\theta(x, h_n)} \frac{\operatorname{Var}(\eta_1)}{\epsilon^2 \sigma_\theta^2(x)} = O\left(\frac{\gamma_n^2 p^2}{n G_\theta(x, h_n)}\right), \end{aligned}$$

which completes the proof. □

---

## ACKNOWLEDGMENTS

---

The authors are indebted to the Editor-in-Chief, the Associate Editor and the referee for their very valuable comments and suggestions which led to a considerable improvement of the manuscript.

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through the Research Groups Program under grant number R.G.P. 2/82/42.

---

**REFERENCES**


---

- [1] AIT-SAÏDI, A.; FERRATY, F.; KASSA, R. and VIEU, P. (2008). Cross-validated estimations in the single-functional index model, *Statistics*, **42**(6), 475–494.
- [2] ANDREWS, D. (1983). First order autoregressive processes and strong mixing, Cowles Foundation Discussion Papers 664, Cowles Foundation for Research in Economics, Yale University.
- [3] ANEIROS, G.; CAO, R.; FRAIMAN, R.; GENEST, C. and VIEU, P. (2019). Recent advances in functional data analysis and high-dimensional statistics, *J. Multivariate Anal.*, **170**, 3–9.
- [4] ANGO NZE, P.; BÜHLMANN, P. and DOUKHAN, P. (2002). Weak dependence beyond mixing and asymptotics for nonparametric regression, *Ann. Statist.*, **30**(2), 397–430.
- [5] BULINSKI, A. and SUQUET, C. (2001). Normal approximation for quasi-associated random fields, *Statistics and Probability Letters*, **54**, 215–226.
- [6] CHERNICK, M.R. (1981). A limit theorem for the maximum of autoregressive processes with uniform marginal distributions, *Ann. Probab.*, **9**(1), 145–149.
- [7] CHOWDHURY, J. and CHAUDHURI, P. (2019). Nonparametric depth and quantile regression for functional data, *Bernoulli*, **25**(1), 395–423.
- [8] CUEVAS, A. (2014). A partial overview of the theory of statistics with functional data, *J. Statist. Plann. Inference*, **147**, 1–23.
- [9] DOOB, J.L. (1953). *Stochastic Processes*, John Wiley & Sons, Inc., New York; Chapman & Hall, Limited, London.
- [10] DOUGE, L. (2010). Théorèmes limites pour des variables quasi-associées Hilbertiennes, *Ann. I.S.U.P.*, **54**(1-2), 51–60.
- [11] DOUGE, L. (2018). Nonparametric regression estimation for quasi-associated Hilbertian processes.
- [12] DOUKHAN, P. and LOUHICHI, S. (1999). A new weak dependence condition and applications to moment inequalities, *Stochastic Processes and their Applications*, **84**(2), 313–342.
- [13] ESARY, J.D.; PROSCHAN, F. and WALKUP, D.W. (1967). Association of random variables, with applications, *Ann. Math. Statist.*, **38**, 1466–1474.
- [14] FERRATY, F.; MAS, A. and VIEU, P. (2007). Nonparametric regression on functional data: inference and practical aspects, *Aust. N. Z. J. Stat.*, **49**(3), 267–286.
- [15] FERRATY, F.; PEUCH, A. and VIEU, P. (2003). Modèle à indice fonctionnel simple, *C. R. Math. Acad. Sci. Paris*, **336**(12), 1025–1028.
- [16] FERRATY, F. and VIEU, P. (2006). *Nonparametric Functional Data Analysis*, Springer Series in Statistics. Springer, New York. Theory and practice.
- [17] GEENENS, G. (2011). *A nonparametric functional method for signature recognition*. In “Recent Advances in Functional Data Analysis and Related Topics”, *Contrib. Statist.*, pages 141–147. Physica-Verlag/Springer, Heidelberg.
- [18] GOIA, A. and VIEU, P. (2016). An introduction to recent advances in high/infinite dimensional statistics [Editorial], *J. Multivariate Anal.*, **146**, 1–6.
- [19] HÄRDLE, W.; HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models, *Ann. Statist.*, **21**(1), 157–178.
- [20] HRISTACHE, M.; JUDITSKY, A. and SPOKOINY, V. (2001). Direct estimation of the index coefficient in a single-index model, *Ann. Statist.*, **29**(3), 595–623.
- [21] ICHIMURA, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models, *J. Econometrics*, **58**(1-2), 71–120.

- [22] KALLABIS, R.S. and NEUMANN, M.H. (2006). An exponential inequality under weak dependence, *Bernoulli*, **12**(2), 333–350.
- [23] LI, J.; HUANG, C. and ZHU, H. (2017). A functional varying-coefficient single-index model for functional response data, *J. Amer. Statist. Assoc.*, **112**(519), 1169–1181.
- [24] LING, N. and VIEU, P. (2018). Nonparametric modelling for functional data: selected survey and tracks for future, *Statistics*, **52**(4), 934–949.
- [25] LING, N. and VIEU, P. (2021). On semiparametric regression in functional data analysis, *Wiley Interdiscip. Rev. Comput. Stat.*, **13**(2), e1538, 15.
- [26] MASRY, E. (2005). Nonparametric regression estimation for dependent functional data: asymptotic normality, *Stochastic Process. Appl.*, **115**(1), 155–177.
- [27] NOVO, S.; ANEIRO, G. and VIEU, P. (2019). Automatic and location-adaptive estimation in functional single-index regression, *J. Nonparametr. Stat.*, **31**(2), 364–392.
- [28] RICHARD E. BARLOW, F.P. (1975). *Statistical theory of reliability and life testing: probability models*. In “International Series in Decision Processes Series in Quantitative Methods for Decision Making”, Holt, Rinehart and Winston, 1975.
- [29] SHANG, H.L. (2020). Estimation of a functional single index model with dependent errors and unknown error density, *Comm. Statist. Simulation Comput.*, **49**(12), 3111–3133.
- [30] STRZALKOWSKA-KOMINIAK, E. and CAO, R. (2013). Maximum likelihood estimation for conditional distribution single-index models under censoring, *J. Multivariate Anal.*, **114**, 74–98.
- [31] WANG, G.; FENG, X.N. and CHEN, M. (2016). Functional partial linear single-index model, *Scand. J. Stat.*, **43**(1), 261–274.



---

---

## On Analyzing Non-Monotone Failure Data

---


---

Authors: MUHAMMAD MANSOOR  

– Department of Statistics, Govt. Sadiq Egerton College,  
Bahawalpur, Pakistan  
[mansoor.abbasi143@gmail.com](mailto:mansoor.abbasi143@gmail.com)

M.H. TAHIR 

– Department of Statistics, The Islamia University of Bahawalpur,  
Bahawalpur, Pakistan  
[mtahir.stat@gmail.com](mailto:mtahir.stat@gmail.com)

GAUSS M. CORDEIRO 

– Department of Statistics, Federal University of Pernambuco,  
Recife, PE, Brazil  
[gauss@de.ufpe.br](mailto:gauss@de.ufpe.br)

EDWIN M.M. ORTEGA 

– Department of Exact Science, ESALQ, University of São Paulo,  
Piracicaba, SP, Brazil  
[edwin@usp.br](mailto:edwin@usp.br)

AYMAN ALZAATREH 

– Department of Mathematics and Statistics, American University of Sharjah,  
Sharjah, United Arab Emirates  
[aalzaatreh@aus.edu](mailto:aalzaatreh@aus.edu)

Received: December 2019

Revised: January 2021

Accepted: March 2021

Abstract:

- A new two-parameter distribution is defined for modeling non-monotone lifetime data. It is constructed based on the logistic-G family and the exponential distribution. Its hazard rate properties are different than those of the well-known distributions. Some of its statistical properties are presented. An extended regression based on the logarithm of the random variable of the introduced distribution is defined. The new regression can provide better fits than other special regressions for analyzing real data. The performance of the maximum likelihood estimates is investigated from a simulation study. Three lifetime data sets are used to prove empirically the usefulness of the new models.


Keywords:

- *bathtub failure rate; exponential distribution; logistic distribution; maximum likelihood estimation; regression model.*

AMS Subject Classification:

- 60E05, 62N05, 62F10.

---

 Corresponding author.

---

## 1. INTRODUCTION

---

Recently, many distributions have been defined for modeling lifetime data. The Weibull distribution has survival and hazard rate functions in closed-forms; see Murthy *et al.* [14]. Gupta and Kundu [8] introduced the exponentiated exponential (EE) distribution as an alternative to the gamma and Weibull distributions. It has many properties similar to those of the gamma and Weibull with closed-form survival and hazard rate functions; see Gupta and Kundu [9]. The hazard rate functions (hrfs) of the gamma, Weibull and EE distributions can not be upside-down bathtub and bathtub shapes but only monotonically increasing, monotonically decreasing or constant shapes.

Taking into account these points, we define a new two-parameter alternative to the above distributions to overcome the above-mentioned drawback. Further, it is common in practical situations to use an appropriate regression based on an asymmetric distribution for censored data and survival time data. Recently, various papers have been published on that subject such as those by Lanjoni *et al.* [10], Cordeiro *et al.* [5], among others. Another objective of this work is to propose a location-scale regression based on the logistic-exponential distribution named the log-logistic exponential regression. It is a new regression that can be applied to data sets with the presence of censored data.

The paper is outlined as follows. In Section 2, the new *logistic-G* (LG) family is introduced and some of its structural properties are studied. A special model of the LG family called the *logistic-exponential* (LE) distribution is presented in Section 3. Some of its mathematical properties are addressed in Section 4. The parameters of the LE distribution are estimated by maximum likelihood (ML) in Section 5. Further, a Monte Carlo simulation study is conducted to assess the performance of the ML method. An extended regression model is proposed and studied in Section 6. In Section 7, the usefulness of the new models is shown empirically by means of three real data sets. Finally, Section 8 offers some concluding remarks.

---

## 2. THE NEW LG FAMILY

---

Alzaatreh *et al.* [2] defined the *T-X family* of distributions as follows. Let  $r(t)$  be the probability density function (pdf) of a random variable (rv)  $T \in [a, b]$  for  $-\infty \leq a < b < \infty$  and let  $W(\cdot): [0, 1] \rightarrow \mathbb{R}$  be an adequate link function. The cumulative distribution function (cdf) of the *T-X family* is

$$F(x; \boldsymbol{\xi}) = \int_a^{W[G(x; \boldsymbol{\xi})]} r(t) dt,$$

where  $\boldsymbol{\xi}$  is the parameter vector of  $G$ .

Based on the above definition, if the function  $W[G(x; \boldsymbol{\xi})]$  is monotonically non-increasing with  $W(0) \rightarrow b$  and  $W(1) \rightarrow a$ , one can redefine the *T-X family* cdf as

$$(2.1) \quad F(x; \boldsymbol{\xi}) = 1 - \int_a^{W[G(x; \boldsymbol{\xi})]} r(t) dt.$$

Let  $T$  be a logistic rv with pdf  $r(t) = \alpha e^{-\alpha t}(1 + e^{-\alpha t})^{-2}$  and support in  $\mathbb{R}$ , where  $\alpha > 0$ . By setting  $W[G(x; \boldsymbol{\xi})] = \log\{-\log[G(x; \boldsymbol{\xi})]\}$ , a monotonically non-increasing function in  $G(x; \boldsymbol{\xi})$ , the cdf of the LG family follows from (2.1):

$$(2.2) \quad F(x; \alpha, \boldsymbol{\xi}) = 1 - \left[1 + \left\{-\log[G(x; \boldsymbol{\xi})]\right\}^{-\alpha}\right]^{-1}, \quad x \in \mathbb{R}.$$

If  $g(x; \boldsymbol{\xi}) = dG(x; \boldsymbol{\xi})/dx$ , the associated pdf to (2.2) is

$$(2.3) \quad f(x; \alpha, \boldsymbol{\xi}) = \frac{\alpha g(x; \boldsymbol{\xi}) \left\{-\log[G(x; \boldsymbol{\xi})]\right\}^{-\alpha-1}}{G(x; \boldsymbol{\xi}) \left[1 + \left\{-\log[G(x; \boldsymbol{\xi})]\right\}^{-\alpha}\right]^2}.$$

The dependence on the baseline vector  $\boldsymbol{\xi}$  and  $\alpha$  is omitted and then  $G(x) = G(x; \boldsymbol{\xi})$  and  $f(x) = f(x; \alpha, \boldsymbol{\xi})$ . Hereafter, a rv with pdf (2.3) is denoted by  $X \sim \text{LG}(\alpha, \boldsymbol{\xi})$ .

The hrf of  $X$  has the form

$$(2.4) \quad h(x) = \frac{\alpha g(x) \left\{-\log[G(x)]\right\}^{-\alpha-1}}{G(x) \left[1 + \left\{-\log[G(x)]\right\}^{-\alpha}\right]}.$$

The quantile function (qf) of  $X$  follows by inverting  $F(x) = u$  in (2.2):

$$(2.5) \quad Q(u) = Q_G(e^{-v}),$$

where  $Q_G(v) = G^{-1}(v)$  is the parent qf and  $v = [(1-u)/u]^{1/\alpha}$ . Then, the solution of the nonlinear equation  $X = Q(U)$  has density (2.3) if  $U$  has a uniform  $U(0, 1)$  distribution.

Equation (2.5) gives a simple interpretation for the LG family. If  $T$  has a logistic density  $r(t)$  with shape parameter  $\alpha$ , the LG family is obtained from the qf of the  $G$  distribution by  $X = Q_G(e^{-e^T})$ .

**Proposition 2.1.** *Let  $c = \inf\{x: G(x) > 0\}$ . The asymptotics of Equations (2.2), (2.3) and (2.4) when  $x \rightarrow c$  are:*

$$\begin{aligned} F(x) &\sim \left\{-\log[G(x)]\right\}^{-\alpha}, \\ f(x) &\sim \frac{\alpha g(x)}{G(x)} \left\{-\log[G(x)]\right\}^{-\alpha-1}, \\ h(x) &\sim \frac{\alpha g(x)}{G(x)} \left\{-\log[G(x)]\right\}^{-\alpha-1}. \end{aligned}$$

**Proposition 2.2.** *The asymptotics of Equations (2.2), (2.3) and (2.4) when  $x \rightarrow \infty$  are given by*

$$1 - F(x) \sim \bar{G}(x)^\alpha, \quad f(x) \sim \alpha g(x) \bar{G}(x)^{\alpha-1} \quad \text{and} \quad h(x) \sim \frac{\alpha g(x)}{\bar{G}(x)}.$$

**Theorem 2.1.** *The Shannon’s entropy of the LG family takes the form*

$$(2.6) \quad \eta_X = E \left[ \log \left\{ g \left[ G^{-1} \left( e^{-e^T} \right) \right] \right\} \right] - B \left( 1 - \frac{1}{\alpha}, 1 + \frac{1}{\alpha} \right) - \log \alpha + 2,$$

where  $B(\cdot, \cdot)$  is the beta function.

**Proof:** Alzaatreh *et al.* [2] obtained the Shannon entropy of the T-X family, where  $W[G(x)] = -\log[1-G(x)]$ . One can use their same technique to obtain this entropy for the LG family in (2.2) when  $W[G(x)] = \log\{-\log[G(x)]\}$  as

$$(2.7) \quad \eta_X = \mathbb{E} \left[ \log \left\{ g \left[ G^{-1} \left( e^{-e^T} \right) \right] \right\} \right] - \mathbb{E}(e^T) + \mu_T + \eta_T,$$

where  $\mu_T$  and  $\eta_T$  are the mean and Shannon entropy of the rv  $T$ , respectively. If  $T$  has the logistic distribution, (2.6) follows easily from (2.7). □

### 2.1. Linear representation

We can rewrite Equation (2.2) as

$$(2.8) \quad F(x) = \frac{\left\{ -\log[G(x)] \right\}^{-\alpha}}{1 + \left\{ -\log[G(x)] \right\}^{-\alpha}}.$$

The power series  $\left\{ -\log[G(x)] \right\}^{-\alpha} = \sum_{k=0}^{\infty} p_k [1-G(x)]^k$  holds, where  $p_0 = 1, p_1 = -\alpha/2, p_2 = (3\alpha^2 - 5\alpha)/24, p_3 = (-\alpha^3 + 5\alpha^2 - 6\alpha)/48$ , etc. The radius of convergence of this series is infinite for  $0 < G(x) < 1$  and then it converges for all real numbers  $x$  with great rapidity.

Then, we can express Equation (2.8) as a ratio of two convergent power series of  $G(x)$ :

$$F(x) = \frac{\sum_{k=0}^{\infty} p_k [1-G(x)]^k}{\sum_{k=0}^{\infty} q_k [1-G(x)]^k} = \sum_{k=0}^{\infty} b_k [1-G(x)]^k.$$

Here,  $q_0 = 1 + p_0, b_0 = p_0/q_0$  and, for  $k \geq 1, q_k = p_k$  and

$$b_k = \frac{1}{q_0} \left( p_k - \frac{1}{q_0} \sum_{r=1}^k q_r b_{k-r} \right).$$

Further,  $F(x)$  can be rewritten as

$$F(x) = \sum_{k=0}^{\infty} b_k [1-G(x)]^k = \sum_{j=0}^{\infty} \sum_{k=j}^{\infty} (-1)^j b_k \binom{k}{j} G(x)^j$$



and then

$$(2.9) \quad F(x) = \sum_{j=0}^{\infty} d_j G(x)^j,$$

where  $d_j = \sum_{k=j}^{\infty} (-1)^j b_k \binom{k}{j}$  and  $G(x)^j$  denotes the exponentiated-G (“exp-G” for short) cdf with power parameter  $j$ .

Hence, the density of  $X$  has a linear representation in terms of exp-G densities, namely

$$(2.10) \quad f(x) = \sum_{j=0}^{\infty} d_{j+1} h_{j+1}(x),$$

where  $h_{j+1}(x) = (j + 1) g(x) G(x)^j$  is the exp-G density with power parameter  $j + 1$ . Some exp-G properties are addressed in more than 50 papers cited by Tahir and Nadarajah [19].

Clearly, some mathematical properties of the LG family can be derived from Equation (2.10) and those exp-G properties.

---

## 2.2. Moments

---

Let  $Y_{j+1}$  be a rv having density  $h_{j+1}(x)$ . The  $n$ -th moment of  $X$  follows from (2.10) as

$$(2.11) \quad \mathbb{E}(X^n) = \sum_{j=0}^{\infty} d_{j+1} \mathbb{E}(Y_{j+1}^n) = \sum_{j=0}^{\infty} (j + 1) d_{j+1} \tau_{n,j},$$

where  $\tau_{n,j} = \int_{-\infty}^{\infty} x^n G(x)^j g(x) dx = \int_0^1 Q_G(u)^n u^j du$ . Cordeiro and Nadarajah [4] determined the quantity  $\tau_{n,j}$  for the normal, beta, gamma and Weibull distributions. Their developments can be used to other distributions.

The  $n$ -th incomplete moment of  $X$ , say  $m_n(y) = \int_0^y x^n f(x) dx$ , is given by

$$(2.12) \quad \begin{aligned} m_n(y) &= \sum_{j=0}^{\infty} d_{j+1} \int_0^y x^n h_{j+1}(x) dx \\ &= \sum_{j=0}^{\infty} (j + 1) d_{j+1} \int_0^{G(y)} Q_G(u)^n u^j du. \end{aligned}$$

The main application of the first incomplete moment  $m_1(y)$  refers to the deviations from the mean and median and the Bonferroni and Lorenz curves of  $X$ . A further important application is related to the mean residual life (MRL) of  $X$ , i.e. the function measuring the remaining life expectancy at age  $t$ , given by  $\nu(t) = [1 - m_1(t)] / [1 - F(t)] - t$ . This function is like the density and generating functions: for a distribution with a finite mean, it completely determines the distribution. The use of the MRL is a helpful tool in model building.

---

### 2.3. Generating function

---

The moment generating function (mgf)  $M(t) = \mathbb{E}(e^{tX})$  of  $X$  can be determined from (2.10) as

$$(2.13) \quad M(t) = \sum_{j=0}^{\infty} d_{j+1} M_{j+1}(t) = \sum_{i=0}^{\infty} (j+1) d_{j+1} \rho(t, j),$$

where  $M_{j+1}(t)$  is the mgf of  $Y_{j+1}$  and  $\rho(t, j) = \int_0^1 \exp[t Q_G(u)] u^j du$ .

Hence,  $M(t)$  can be determined from the exp-G generating function. The characteristic function of  $X$  is simply  $M(-\mathbf{i}t)$ , where  $\mathbf{i} = \sqrt{-1}$ , and it always exists, even when the generating function does not.

---

## 3. THE LE DISTRIBUTION

---

Consider the baseline exponential with cdf  $G(x) = 1 - e^{-\lambda x}$ . The cdf of the LE distribution can be determined from (2.2) as

$$(3.1) \quad F(x) = F(x; \alpha, \lambda) = 1 - \left[ 1 + \left\{ -\log(1 - e^{-\lambda x}) \right\}^{-\alpha} \right]^{-1}.$$

Hereafter, let  $X \sim \text{LE}(\alpha, \lambda)$  have the cdf (3.1). The pdf of  $X$  is

$$(3.2) \quad f(x) = \frac{\alpha \lambda \left\{ -\log(1 - e^{-\lambda x}) \right\}^{-\alpha-1}}{(e^{\lambda x} - 1) \left[ 1 + \left\{ -\log(1 - e^{-\lambda x}) \right\}^{-\alpha} \right]^2}.$$

The hrf of  $X$  becomes

$$(3.3) \quad h(x) = \frac{\alpha \lambda \left\{ -\log(1 - e^{-\lambda x}) \right\}^{-\alpha-1}}{(e^{\lambda x} - 1) \left[ 1 + \left\{ -\log(1 - e^{-\lambda x}) \right\}^{-\alpha} \right]}.$$

Equation (3.1) has two parameters  $\alpha$  and  $\lambda$  such as the gamma, log-normal, Weibull and EE distributions. The LE model has closed-form survival and hazard functions like the Weibull and EE distributions.

Figures 1 and 2 display some plots of the density and hrf of  $X$  for selected values of  $\alpha$  when  $\lambda = 1$ . Figure 1 shows that the LE density is a right-skewed distribution. The plots in Figure 2 indicate that the hrf of  $X$  can have decreasing failure rate (DFR), bathtub (BT) and decreasing-increasing-decreasing (DID) shapes. The limiting behavior of this hrf is  $\lim_{x \rightarrow \infty} h(x) = \alpha$  and  $\lim_{x \rightarrow 0} h(x) = \infty$ , and it always approaches  $\alpha$  when  $X$  goes to infinity.

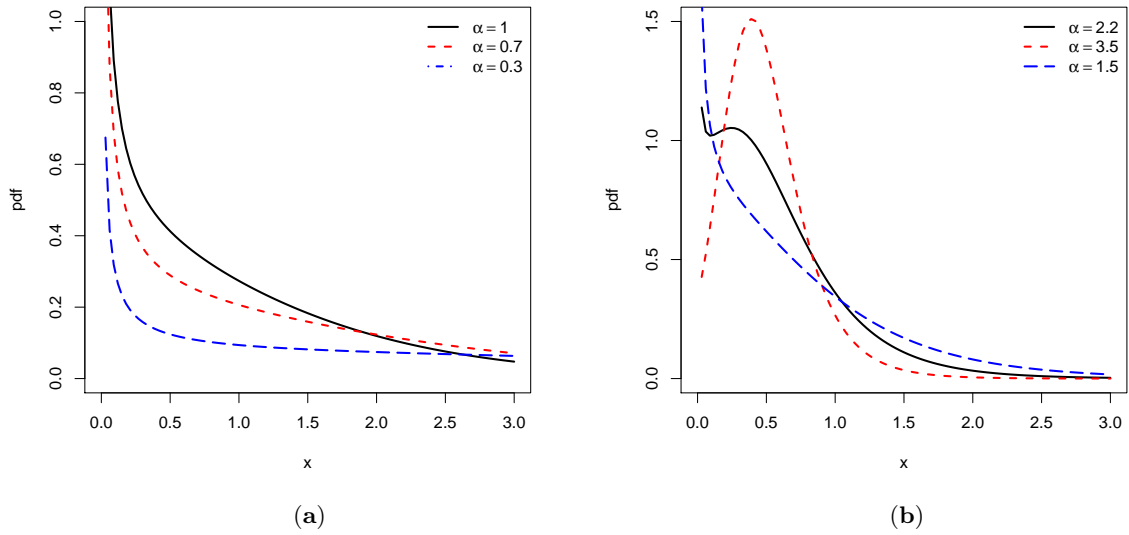


Figure 1: Plots of the LE density varying  $\alpha$  with  $\lambda = 1$ .

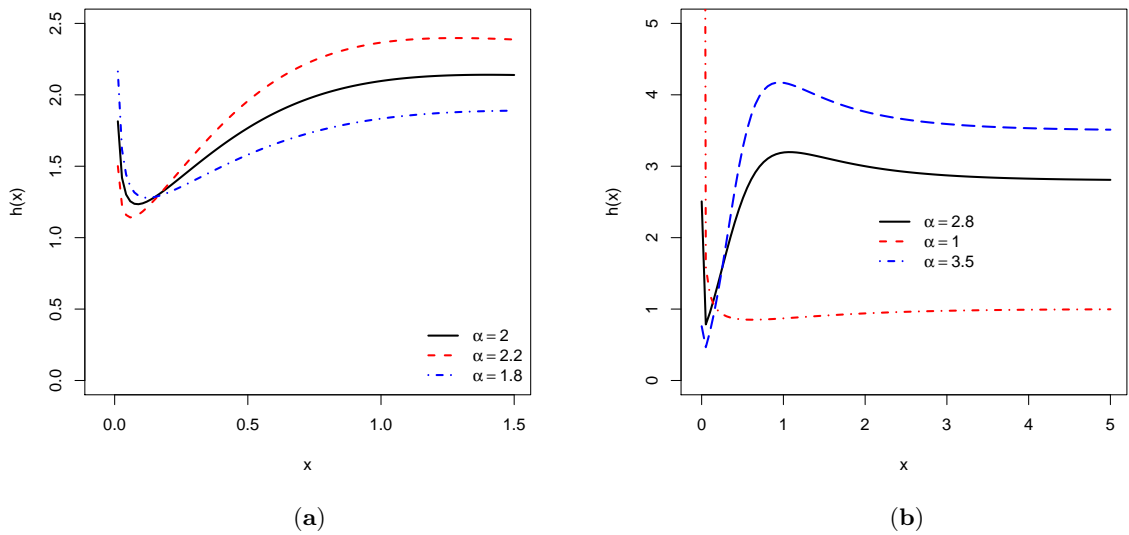


Figure 2: Plots of the LE hrf varying  $\alpha$  for  $\lambda = 1$ .

---

#### 4. PROPERTIES OF LE DISTRIBUTION

---

In this section, we obtain some properties of the LE distribution.

---

##### 4.1. Asymptotics and shapes

---

**Proposition 4.1.** *The asymptotics of the cdf, pdf and hrf of  $X$  when  $x \rightarrow 0$  are:*

$$\begin{aligned} F(x) &\sim 1 - \left\{ 1 + [-\log(\lambda x)]^{-\alpha} \right\}^{-1}, \\ f(x) &\sim \frac{\alpha}{x} [-\log(\lambda x)]^{-\alpha-1} \left\{ 1 + [-\log(\lambda x)]^{-\alpha} \right\}^{-2}, \\ h(x) &\sim \frac{\alpha}{x} [-\log(\lambda x)]^{-\alpha-1} \left\{ 1 + [-\log(\lambda x)]^{-\alpha} \right\}^{-1}. \end{aligned}$$

**Proposition 4.2.** *The asymptotics of the cdf, pdf and hrf of  $X$  when  $x \rightarrow \infty$  are*

$$1 - F(x) \sim e^{-\alpha\lambda x}, \quad f(x) \sim \alpha\lambda e^{-\alpha\lambda x} \quad \text{and} \quad h(x) \sim \alpha\lambda.$$

---

##### 4.2. Transformation

---

If  $Y$  has the logistic distribution with parameter  $\alpha$ , then  $X = -\lambda^{-1} \log(1 - e^{-e^Y})$  follows the LE( $\alpha, \lambda$ ) model.

---

##### 4.3. Mode

---

**Lemma 4.1.** *The modes of the LE density are the solutions of  $k(x) = 0$ , where*

$$k(x) = -\lambda - \frac{\lambda}{e^{\lambda x} - 1} \left[ 1 - \frac{\alpha + 1}{\left\{ -\log(1 - e^{-\lambda x}) \right\}} + \frac{2\alpha \left\{ -\log(1 - e^{-\lambda x}) \right\}^{-\alpha-1}}{1 + \left\{ -\log(1 - e^{-\lambda x}) \right\}^{-1}} \right].$$

---

##### 4.4. Quantile function

---

The qf of  $X$  is  $Q(u) = -\lambda^{-1} \log(1 - e^{-v})$ ,  $u \in (0, 1)$ , where  $v = [(1 - u)/u]^{1/\alpha}$ .

---

#### 4.5. Shannon entropy

---

**Theorem 4.1.** *The Shannon entropy of  $X$  is*

$$(4.1) \quad \eta_X = \frac{\lambda}{\lambda-1} - B\left(1 - \frac{1}{\alpha}, 1 + \frac{1}{\alpha}\right) - \log \alpha + 2.$$

**Proof:** For the LE distribution, the result holds:

$$\mathbb{E}\left[\log\left\{g\left[G^{-1}\left(e^{-e^T}\right)\right]\right\}\right] = \mathbb{E}(e^T) = \frac{\lambda}{\lambda-1}.$$

Equation (4.1) follows by substituting the above result in (2.6). □

---

#### 4.6. Moments and generating function

---

The LE density comes from (2.10) as

$$f(x) = \sum_{j=0}^{\infty} d_{j+1} (j+1) \lambda e^{-\lambda x} (1 - e^{-\lambda x})^j.$$

The moments of  $X$  follow from the EE distribution and (2.11):

$$(4.2) \quad \mu'_n = \mathbb{E}(X^n) = n! \sum_{j,l=0}^{\infty} \frac{(-1)^l (j+1) d_{j+1} A(j,l)}{\lambda^{j+1} (l+1)^{n+1}},$$

where  $A(j,l) = j(j-1)\cdots(j-l)/l!$ .

The skewness and kurtosis of  $X$  for some values of  $\alpha$  by taking  $\lambda = 1$  are displayed in Figure 3. The distribution of  $X$  is right-skewed. For fixed  $\lambda$ , the skewness is a decreasing function of  $\alpha$ , whereas the kurtosis decreases steadily towards asymptotic limits when  $\alpha$  increases.

The  $n$ -th incomplete moment of  $X$  is obtained from (2.12):

$$(4.3) \quad m_n(y) = \lambda^{-n} \sum_{j=0}^{\infty} (j+1) d_{j+1} A_n^*(j+1),$$

where

$$A_n^*(j+1) = \sum_{p=0}^{\infty} \frac{(-1)^p}{(p+1)^{r+1}} \binom{j}{p} \gamma\left(n+1, (p+1)\lambda y\right), \quad n = 1, 2, \dots,$$

and  $\gamma(p, x) = \int_0^x w^{p-1} e^{-w} dw$  (for  $p > 0$ ) is the incomplete gamma function.

The mgf of  $X$  follows from (2.13) as

$$(4.4) \quad M(t) = \Gamma\left(1 - \frac{t}{\lambda}\right) \sum_{j=0}^{\infty} \frac{(j+1)! d_{j+1}}{\Gamma\left(j+2 - \frac{t}{\lambda}\right)}.$$

Equations (4.2), (4.3) and (4.4) are the main results of this section.

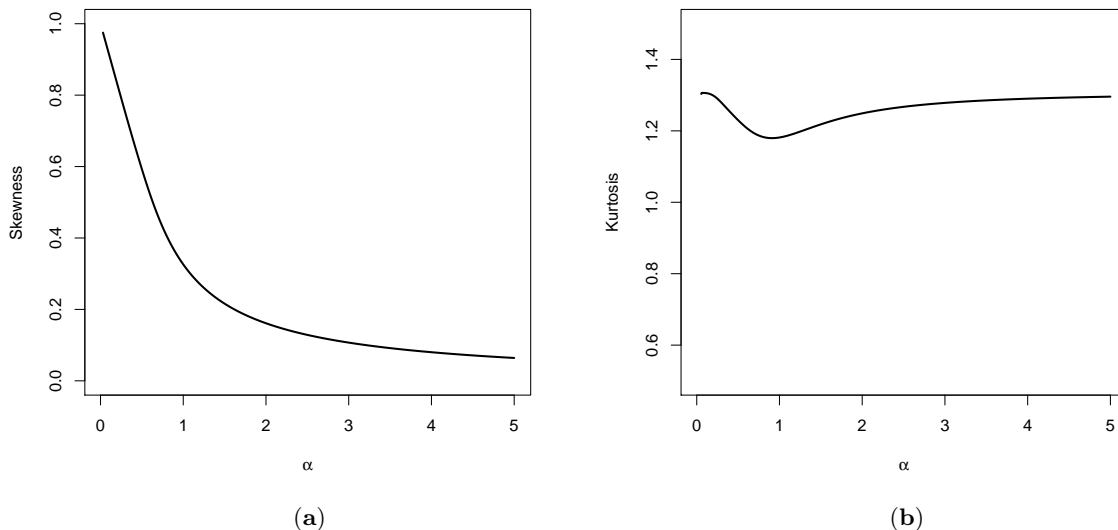


Figure 3: (a) Skewness and (b) Kurtosis plots of  $X$  for  $\lambda = 1$ .

---

#### 4.7. Order statistics

---

Order statistics make their appearance in many areas of statistical theory and practice. Suppose  $X_1, \dots, X_n$  is a random sample from the LE distribution. Let  $X_{i:n}$  denote the  $i$ -th order statistic. The pdf of  $X_{i:n}$  can be expressed as

$$f_{i:n}(x) = K \sum_{j=0}^{n-i} (-1)^j \binom{n-i}{j} f(x) F(x)^{j+i-1},$$

where  $K = 1/B(i, n - i + 1)$ .

Gradshteyn and Ryzhik [7] provided a power series raised to a positive integer  $n$ :

$$(4.5) \quad \left( \sum_{i=0}^{\infty} a_i u^i \right)^n = \sum_{i=0}^{\infty} b_{n,i} u^i,$$

where the coefficients  $b_{n,i}$  (for  $i = 1, 2, \dots$ ) satisfy the recurrence equation (with  $b_{n,0} = a_0^n$ )

$$b_{n,i} = (i a_0)^{-1} \sum_{m=1}^i [m(n+1) - i] a_m b_{n,i-m}.$$

The density function of  $X_{i:n}$  can be reduced to

$$(4.6) \quad f_{i:n}(x) = \sum_{r,k=0}^{\infty} m_{r,k} \pi_{EE}(x; \lambda, r+k+1),$$

where  $\pi_{EE}(x; \lambda, r+k+1)$  (for  $r, k \geq 0$ ) denotes the EE density function with parameters  $\lambda$  and  $r+k+1$ , and

$$m_{r,k} = \frac{n! (r+1) (i-1)! d_{r+1}}{(r+k+1)} \sum_{j=0}^{n-i} \frac{(-1)^j f_{j+i-1,k}}{(n-i-j)! j!}.$$

Here,  $d_r$  is defined in (2.9) and the quantities  $f_{j+i-1,k}$  follow recursively from (for  $k \geq 1$ )

$$f_{j+i-1,k} = (k d_0)^{-1} \sum_{m=1}^k [m(j+i) - k] d_m f_{j+i-1,k-m},$$

and  $f_{j+i-1,0} = d_0^{j+i-1}$ .

Equation (4.6) shows that the pdf of the LE order statistics is a double linear combination of EE densities. Therefore, several mathematical quantities of these order statistics can be derived from this result.

---

## 5. ESTIMATION

---

The maximum likelihood estimates (MLEs) enjoy desirable properties for constructing confidence intervals. We consider the estimation of the unknown parameters of the LE distribution by the maximum likelihood method. Further works could be addressed using different methods to estimate the LE parameters such as moments, least squares, weighted least squares, bootstrap, Jackknife, Cramér–von-Mises, Anderson–Darling, Bayesian, among others, and compare the estimators from these methods.

Let  $x_1, \dots, x_n$  be  $n$  observed values from the LE distribution given in Equation (3.2) with vector of parameters  $\Theta = (\alpha, \lambda)^T$ . The log-likelihood  $\ell = \ell(\Theta)$  for  $\Theta$  is

$$\begin{aligned} \ell = & n \log(\alpha \lambda) - \sum_{i=1}^n \log(e^{\lambda x_i} - 1) - (\alpha + 1) \sum_{i=1}^n \log\{-\log(1 - e^{-\lambda x_i})\} \\ (5.1) \quad & - 2 \sum_{i=1}^n \log\left[1 + \{-\log(1 - e^{-\lambda x_i})\}^{-\alpha}\right]. \end{aligned}$$

Equation (5.1) can be maximized either directly by using well-known platforms such as R (optim function), SAS (PROC NLMIXED) and Ox program (MaxBFGS subroutine).

---

### 5.1. Simulation results

---

We examine the accuracy of the MLEs of the parameters of the LE distribution using Monte Carlo simulations. The simulation analysis is carried out by generating 5,000 samples for some sample sizes and parameter combinations. Table 1 gives the average biases (Biases) of the MLEs, mean square errors (MSEs), coverage probabilities (CPs) and average widths (AWs) of 95% confidence intervals for  $\alpha$  and  $\lambda$ . These results indicate that the MLEs are accurate. The biases, MSEs and AWs of  $X$  are small for large samples. Further, the CPs are quite close to the 95% nominal levels. So, we conclude that the MLEs can be used for estimating and constructing confidence intervals for the model parameters.

**Table 1:** Simulation results.

Parameter	$n$	$\alpha = 0.3, \lambda = 1$				$\alpha = 0.8, \lambda = 1$			
		Bias	MSE	CP	AW	Bias	MSE	CP	AW
$\alpha$	25	-0.006	0.003	0.92	0.210	-0.091	0.029	0.92	0.599
	50	-0.004	0.002	0.93	0.149	-0.070	0.016	0.96	0.437
	75	-0.004	0.001	0.94	0.121	-0.073	0.013	0.96	0.354
	100	-0.002	0.001	0.95	0.106	-0.067	0.011	0.95	0.309
$\lambda$	25	0.013	0.018	0.93	0.510	-0.039	0.083	0.95	0.988
	50	0.011	0.009	0.95	0.360	-0.035	0.044	0.96	0.707
	75	0.006	0.006	0.96	0.292	-0.059	0.028	0.95	0.559
	100	0.005	0.004	0.95	0.254	-0.053	0.021	0.95	0.488
Parameter	$n$	$\alpha = 1.5, \lambda = 1$				$\alpha = 3, \lambda = 1$			
		Bias	MSE	CP	AW	Bias	MSE	CP	AW
$\alpha$	25	0.175	0.149	0.96	1.362	0.156	0.400	0.94	2.260
	50	0.111	0.067	0.95	0.932	0.084	0.188	0.94	1.565
	75	0.097	0.046	0.96	0.758	0.048	0.104	0.95	1.267
	100	0.090	0.036	0.95	0.653	0.044	0.084	0.96	1.095
$\lambda$	25	0.002	0.068	0.93	0.968	0.020	0.025	0.92	0.572
	50	-0.007	0.034	0.95	0.677	0.011	0.012	0.93	0.401
	75	-0.013	0.021	0.96	0.551	0.007	0.007	0.96	0.326
	100	-0.027	0.016	0.96	0.465	0.002	0.005	0.95	0.280

---

## 6. THE LOG-LOGISTIC EXPONENTIAL REGRESSION WITH CENSORED DATA

---

If  $X$  follows the LE distribution (3.2),  $Y = \log(X)$  will have the *log-logistic exponential* (LLE) distribution. The density function of  $Y$  (for  $y \in \mathbb{R}$ ), parameterized in terms of  $\lambda = e^{-\mu}$ , takes the form

$$(6.1) \quad f(y) = \frac{\alpha \exp[(y - \mu) - \exp(y - \mu)] \left[ -\log\left\{1 - \exp[-\exp(y - \mu)]\right\} \right]^{-\alpha-1}}{\left\{1 - \exp[-\exp(y - \mu)]\right\} \left\{1 + \left[ -\log\left\{1 - \exp[\exp(y - \mu)]\right\} \right]^{-\alpha} \right\}^2},$$

where  $\mu \in \mathbb{R}$  is a location parameter and  $\alpha$  is a positive shape parameter.

We refer to Equation (6.1) as the LLE distribution, say  $Y \sim \text{LLE}(\alpha, \mu)$ . Thus,

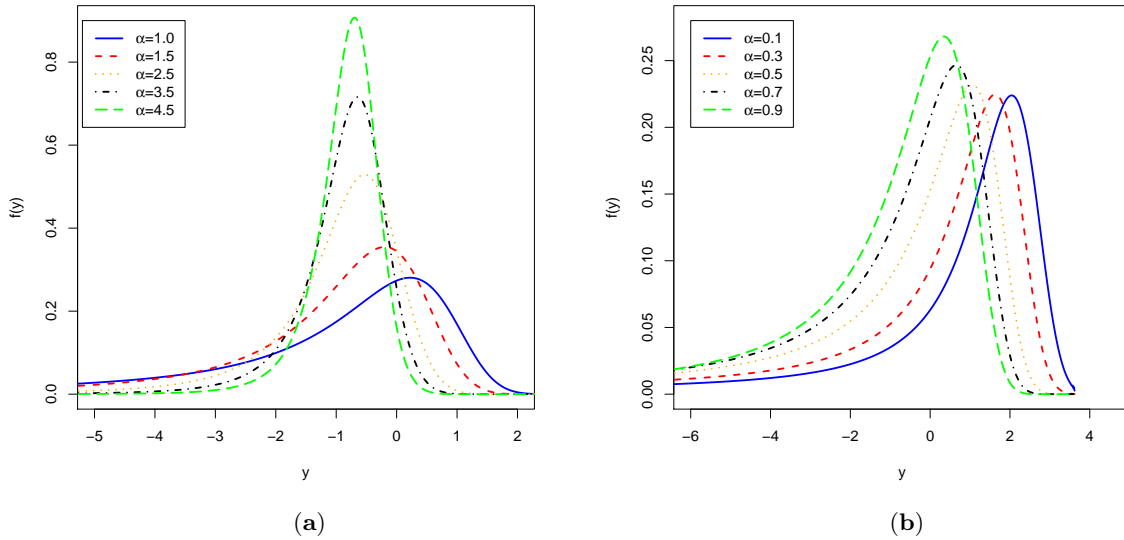
$$\text{if } X \sim \text{LE}(\alpha, \lambda) \text{ then } Y = \log(X) \sim \text{LLE}(\alpha, \mu).$$

Some shapes of the density function of  $Y$  are given in Figure 4.

The survival function of  $Y$  is

$$(6.2) \quad S(y) = \frac{1}{1 + \left[ -\log\left\{1 - \exp[-\exp(y - \mu)]\right\} \right]^{-\alpha}}.$$





**Figure 4:** The LLE density function. (a) For different values of  $\alpha > 1$  with  $\mu = 0$ . (b) For different values of  $\alpha < 1$  with  $\mu = 0$ .

The density function of  $Z = (Y - \mu)$  is

$$(6.3) \quad \pi(z; \alpha) = \frac{\alpha \exp[z - \exp(z)] \left[ -\log\{1 - \exp[-\exp(z)]\} \right]^{-\alpha-1}}{\left\{ 1 - \exp[-\exp(z)] \right\} \left\{ 1 + \left[ -\log\{1 - \exp[-\exp(z)]\} \right]^{-\alpha} \right\}^2}, \quad z \in \mathbb{R}.$$

Based on the LLE density, we propose the location-scale linear regression

$$(6.4) \quad y_i = \mathbf{v}_i^\top \boldsymbol{\beta} + z_i, \quad i = 1, \dots, n,$$

where the random error  $z_i$  has density function (6.3),  $\mathbf{v}_i^\top = (v_{i1}, \dots, v_{ip})$  is the vector of explanatory variables,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$  and  $\alpha$  are unknown parameters. The parameter  $\mu_i = \mathbf{v}_i^\top \boldsymbol{\beta}$  is the location of  $y_i$ . The location parameter vector  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^\top$  is represented by a linear model  $\boldsymbol{\mu} = \mathbf{V}\boldsymbol{\beta}$ , where  $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_n)^\top$  is a known model matrix. Equation (6.4) is referred to as the LLE regression for censored data and opens new possibilities for fitting several types of data. It is an extension of the log-exponential regression for censored data.

Consider a sample  $(y_1, \mathbf{v}_1), \dots, (y_n, \mathbf{v}_n)$  of  $n$  independent observations, where each random response is defined by  $y_i = \min\{\log(X_i), \log(D_i)\}$  assuming that the observed lifetimes and censoring times are independent. Let  $F$  and  $D$  be the sets of individuals for which  $y_i$  is the log-lifetime or log-censoring, respectively.

The log-likelihood function for the vector of parameters  $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^\top)^\top$  from regression (6.4) is

$$(6.5) \quad \begin{aligned} l(\boldsymbol{\theta}) = & r \log(\alpha) + \sum_{i \in F} z_i - \sum_{i \in F} \exp(z_i) - (\alpha + 1) \sum_{i \in F} \log \left[ -\log\{1 - \exp[-\exp(z_i)]\} \right] \\ & - \sum_{i \in F} \log \left\{ 1 - \exp[-\exp(z_i)] \right\} - 2 \sum_{i \in F} \log \left\{ 1 + \left[ -\log\{1 - \exp[-\exp(z_i)]\} \right]^{-\alpha} \right\} \\ & - \sum_{i \in D} \log \left\{ 1 + \left[ -\log\{1 - \exp[-\exp(z_i)]\} \right]^{-\alpha} \right\}, \end{aligned}$$

where  $z_i = (y_i - \mathbf{v}_i^\top \boldsymbol{\beta})$ , and  $r$  is the number of uncensored observations (failures). The MLE  $\hat{\boldsymbol{\theta}}$  of the vector of unknown parameters can be determined by maximizing the log-likelihood (6.5) using the subroutine `NLMixed` in SAS.

The `NLMixed` procedure of SAS has been exhaustively used to estimate the parameters for several distributions. Further, Molenberghs *et al.* [13] adopted this procedure to obtain the estimates in generalized linear models for repeated measures with normal and conjugate random effects, whereas Vangeneugden *et al.* [20] used it to calculate the estimates of extended random-effects models for repeated and overdispersed counts.

The estimated survival function for  $y_i$  ( $\hat{z}_i = y_i - \mathbf{v}_i^\top \hat{\boldsymbol{\beta}}$ ) is

$$(6.6) \quad S(y_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \frac{1}{1 + \left[ -\log \left\{ 1 - \exp \left[ -\exp(y_i - \mathbf{v}_i^\top \hat{\boldsymbol{\beta}}) \right] \right\} \right]^{-\hat{\boldsymbol{\alpha}}}}.$$

We can adopt likelihood ratio (LR) statistics in the usual way for comparing some special models with the LLE regression.

---

## 7. EMPIRICAL ILLUSTRATIONS WITH LIFETIME DATA

---

We now prove empirically that the LE distribution is a good alternative to the gamma, log-normal, Weibull, EE, Nadarajah–Haghighi (NH) introduced by Nadarajah and Haghighi [16], power Lindley (PL) defined by Ghitney *et al.* [6], exponentiated Lindley (EL) studied by Nadarajah *et al.* [15], Birnbaum–Saunders (BS) and inverse Gaussian (IG) distributions. For model comparison, we adopt the Anderson–Darling ( $A^*$ ), Cramér–von Mises ( $W^*$ ) and Kolmogorov–Smirnov (K-S) measures. The cdfs of the EE, NH, PL, EL, BS and pdf of the IG distributions (for  $x > 0$ ) are, respectively,

$$\begin{aligned} F_{EE}(x; \alpha, \lambda) &= (1 - e^{-\lambda x})^\alpha, & \alpha, \lambda > 0, \\ F_{NH}(x; \alpha, \lambda) &= 1 - e^{1 - (1 + \lambda x)^\alpha}, & \alpha, \lambda > 0, \\ F_{PL}(x; \beta, \theta) &= 1 - \left( \frac{1 + \theta + \theta x^\beta}{1 + \theta} \right) e^{-\theta x^\beta}, & \beta, \theta > 0, \\ F_{EL}(x; \alpha, \theta) &= \left[ 1 - \left( \frac{1 + \theta + \theta x}{1 + \theta} \right) e^{-\theta x} \right]^\alpha, & \alpha, \theta > 0, \\ F_{BS}(x; \alpha, \beta) &= \Phi \left[ \frac{1}{\alpha} \left\{ \left( \frac{x}{\beta} \right)^{1/2} - \left( \frac{\beta}{x} \right)^{1/2} \right\} \right], & \alpha, \beta > 0, \\ F_{IG}(x; \mu, \lambda) &= \sqrt{\frac{\lambda}{2\pi x^3}} \exp \left[ -\lambda(x - \mu)^2 / (2x\mu^2) \right], & \mu, \lambda > 0. \end{aligned}$$

---

### 7.1. Application 1: Failure of electrical appliances in life test

---

The data set taken from Lawless [11] represents the 1000 cycles to failure for a group of 60 electrical appliances in a life test. These data were also analyzed by Chesneau *et al.* [3] and Mazucheli *et al.* [12]. Some descriptive statistics for these data are:  $n = 60$ ,  $\bar{x} = 2.19297$ ,

$s = 1.920062$ , skewness = 1.2614 and kurtosis = 2.23207. The histogram displayed in Figure 5(a) and the skewness indicates that the distribution is right-skewed. The TTT plot (Aarset [1]) is given in Figure 5(b). It is first convex and then concave, which suggests a bathtub failure rate. So, the LE distribution could in principle be appropriate for modeling the current data.

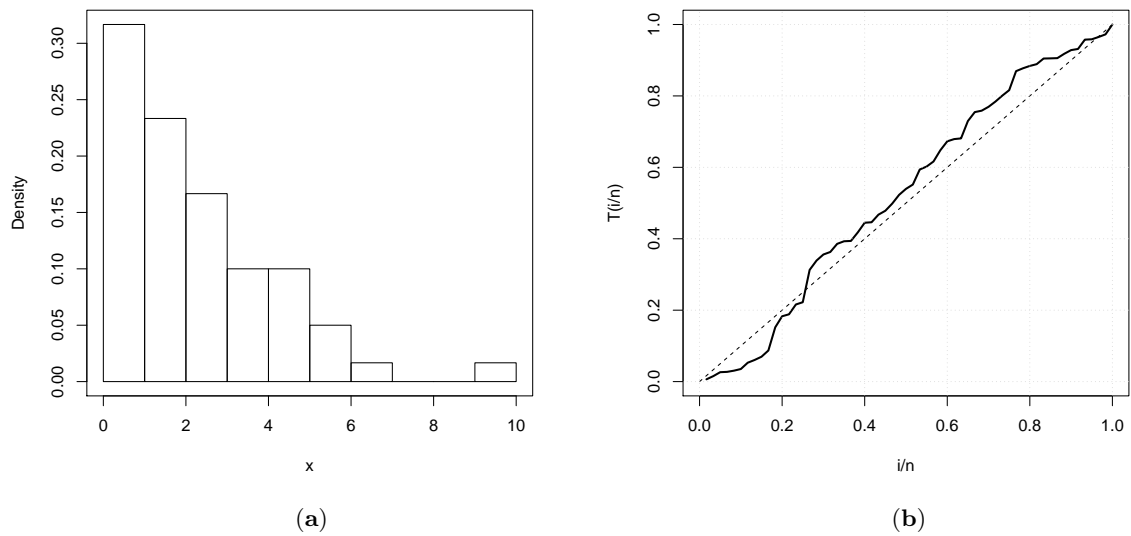


Figure 5: (a) Histogram. (b) TTT plot for failure data.

Table 2: Estimated quantities and goodness-of-fit measures for failure data.

Distribution	Estimates		A*	W*	K-S	K-S p-value
LE( $\alpha, \lambda$ )	1.9798 (0.2555)	0.2625 (0.0357)	0.3258	0.0374	0.0547	0.9491
Gamma( $\alpha, \theta$ )	0.9307 (0.1486)	2.3562 (0.4909)	0.7184	0.1042	0.0897	0.6860
Weibull( $c, \lambda$ )	1.0008 (0.1066)	0.4555 (0.0814)	0.7154	0.1036	0.0777	0.8342
Log-normal( $\mu, \sigma$ )	0.1597 (0.1858)	1.4392 (0.1313)	2.5241	0.4291	0.1653	0.0666
NH( $\alpha, \lambda$ )	1.6133 (0.8016)	0.2274 (0.1575)	0.4574	0.0615	0.0914	0.6632
EE( $\alpha, \lambda$ )	0.9159 (0.1502)	0.4311 (0.0735)	0.7103	0.1028	0.0921	0.6543
PL( $\beta, \theta$ )	0.8883 (0.0891)	0.8042 (0.1031)	0.6467	0.0766	0.0766	0.8155
EL( $\alpha, \theta$ )	0.7522 (0.1274)	0.6203 (0.0873)	0.4615	0.0644	0.0698	0.8522
IG( $\mu, \lambda$ )	2.1929 (0.7513)	0.3113 (0.1104)	4.6132	0.8576	0.30548	0.0000
BS( $\alpha, \beta$ )	1.9391 (0.1824)	0.6483 (0.1111)	2.4479	0.4343	0.3719	0.0000

Table 2 provides the MLEs of the parameters and the values of  $A^*$ ,  $W^*$  and K-S statistics and associated  $p$ -value for each fitted model. We can conclude that the LE distribution provides the best fit and has the ability to fit right-skewed data with BT failure rate. We also provide QQ-plots for all fitted models in Figure 6. Clearly, the new model provides the closest fit to the data.

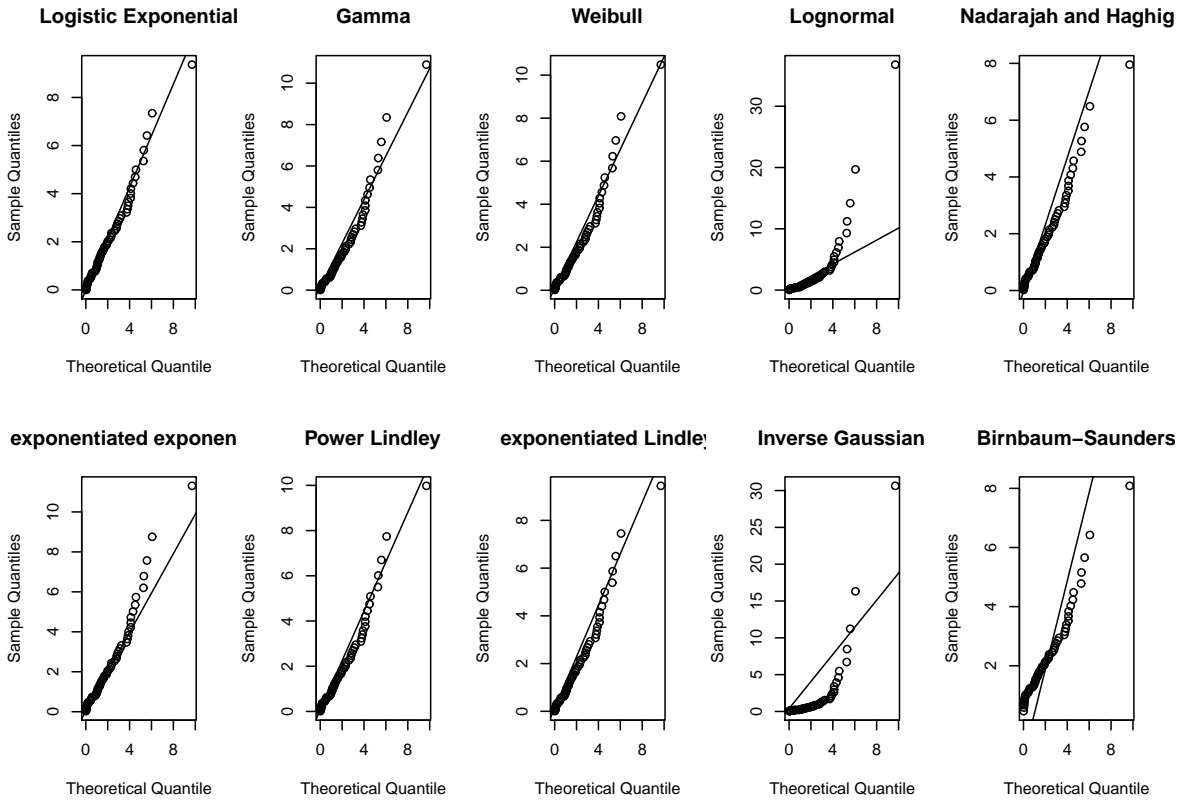


Figure 6: QQ-plots for failure data.

---

## 7.2. Application 2: Lung cancer patients data

---

This data is also taken from a study reported by Lawless [11]. These data represents 21 advanced lung cancer patients who were randomly assigned the chemotherapy treatments termed as standard. Survival times  $t$ , measured from the start of treatment for each patient. The main objective was to compare the effects of two chemotherapy treatments in prolonging survival time. The basic statistics for these data are:  $n = 21$ ,  $\bar{x} = 101.7619$ ,  $s = 110.8147$ , skewness = 1.29047 and kurtosis = 1.00438. The histogram displayed in Figure 7(a) and the skewness indicate that the distribution is right-skewed. The TTT plot of these data shown in Figure 7(b) indicates a decreasing failure rate.

The measures reported in Table 3 indicate that the LE model provides the most accurate fit to the data. Further, the QQ-plots for all fitted models in Figure 8 also suggest the same conclusion.

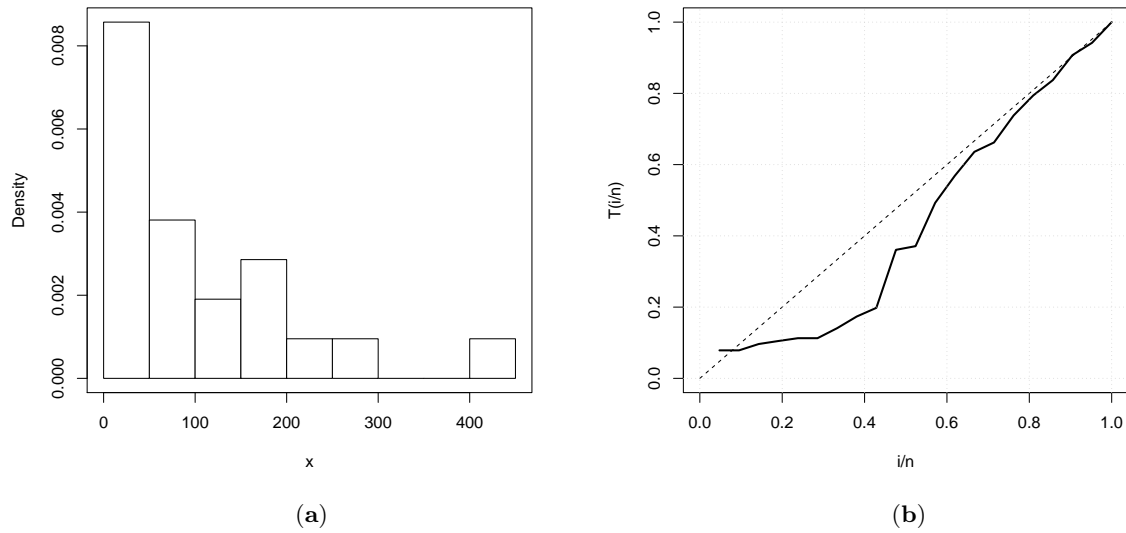


Figure 7: (a) Histogram. (b) TTT plot for cancer data.

Table 3: Estimated quantities and goodness-of-fit measures for cancer data.

Distribution	Estimates		A*	W*	K-S	K-S <i>p</i> -value
LE( $\alpha, \lambda$ )	0.8417 (0.1641)	0.0108 (0.0028)	0.5871	0.0872	0.1574	0.8755
Gamma( $\alpha, \theta$ )	1.2889 (0.2607)	57.7242 (10.7798)	0.6114	0.0912	0.1970	0.3887
Weibull( $c, \lambda$ )	0.8757 (0.1462)	0.0185 (0.0142)	0.6120	0.0922	0.1616	0.4425
Log-normal( $\mu, \sigma$ )	3.9144 (0.2832)	1.2982 (0.2003)	0.7087	0.1130	0.1503	0.2299
NH( $\alpha, \lambda$ )	0.6437 (0.2855)	0.0217 (0.0192)	0.6364	0.0975	0.15307	0.2088
EE( $\alpha, \lambda$ )	0.8301 (0.2288)	0.0087 (0.0025)	0.6056	0.0905	0.1701	0.3776
PL( $\beta, \theta$ )	0.6293 (0.1253)	0.1195 (0.0023)	0.6343	0.0965	0.1595	0.5590
EL( $\alpha, \theta$ )	0.4820 (0.1274)	0.0274 (0.0873)	0.6309	0.0928	0.3064	0.0386
IG( $\mu, \lambda$ )	101.0077 (38.6442)	32.1416 (9.9192)	0.6999	0.1152	0.4718	0.0150
BS( $\alpha, \beta$ )	1239.8960 (503.2201)	1880.1910 (642.4328)	1.0941	0.1844	0.5005	0.0000

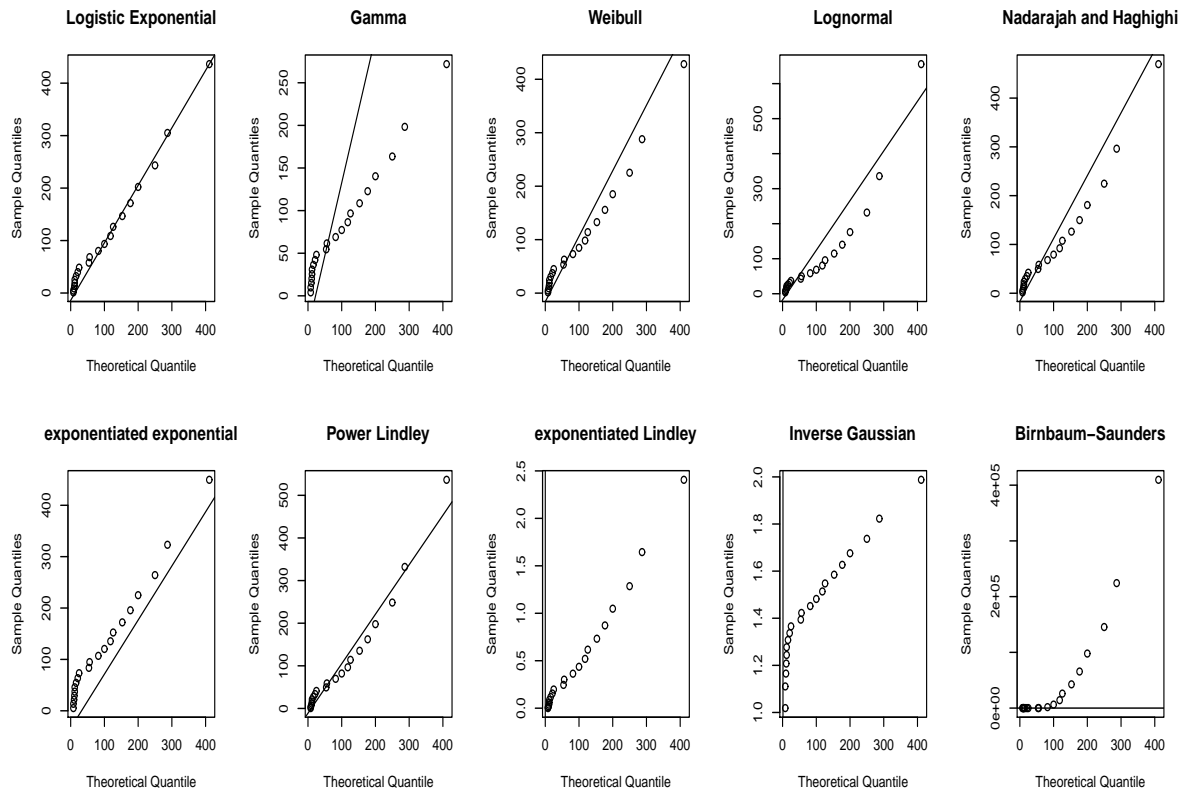


Figure 8: QQ-plots for cancer data.

---

### 7.3. Application 3: Entomology data

---

In this application, we take a data set from a study carried out at the Department of Entomology of the Luiz de Queiroz School of Agriculture, University of São Paulo. Such study aims to assess the longevity of the Mediterranean fruit fly (*ceratitis capitata*), which is considered a pest in agriculture. Instead of using an insecticide, Silva *et al.* [18] conducted a study using small portions of food containing substances extracted from a tree called *Azadirachta indica* which is best known internationally by the name “neem”. The experiment was completely randomized with 11 treatments, consisting of different extracts of the neem tree at concentrations of 39, 225, and 888 ppm, where the response variable is the lifetime of the adult flies in days after exposure to the treatments. From the results of the experiment, these 11 treatments are allocated into two groups, namely:

**Group 1:** Control 1 (deionized water); Control 2 (acetone –5%); aqueous extract of seeds (AES) (39 ppm); AES (225 ppm); AES (888 ppm); methanol extract of leaves (MEL) (225 ppm); MEL (888 ppm); and dichloromethane extract of branches (DMB) (39 ppm) 425.

**Group 2:** MEL (39 ppm); DMB (225 ppm); and DMB (888 ppm).

Lanjoni *et al.* [10] analyzed these data by fitting the log-Burr XII geometric type I (LBXIIGI) and log-Burr XII geometric type II (LBXIIGII) models. Recently, these data were also analyzed by Cordeiro *et al.* [5] and Zubair *et al.* [21] using the generalized Weibull-logistic regression and log-power-Cauchy negative-binomial regressions, respectively. Following the same procedure from these surveys, we compare the proposed model with these regressions in this application.

The response variable in the experiment is the lifetime of the adult lies in days after exposure to the treatments. The total sample size is  $n = 72$ . So, the variables used in this study are:

- $y_i$ : log-lifetime of ceratitis capitata adults in days;
- $\delta_i$ : censoring indicator;
- $v_{i1}$ : sex of the larvae;
- $v_{i2}$ : group (0 = group 1, 1 = group 2),  $i = 1, \dots, 74$ .

Lanjoni *et al.* [10] introduced two lifetime distributions by compounding the Burr XII and geometric distributions, and also defined two extended regressions based on the logarithms of these distributions. Let  $F$  and  $D$  be the sets of individuals for which  $y_i$  is the log-lifetime or log-censoring, respectively. We adopt the classical log-Weibull (LW) regression as an example to illustrate that the LE regression can provides better fits. In this case, the total log-likelihood function for the parameters  $\boldsymbol{\theta} = (\sigma, \boldsymbol{\beta}^\top)^\top$  is

$$l(\boldsymbol{\theta}) = r^* \log\left(\frac{1}{\sigma}\right) + \sum_{i \in F} z_i - \sum_{i \in F} \exp(z_i) - \sum_{i \in D} \exp(z_i),$$

where  $z_i = (y_i - \mathbf{v}_i^\top \boldsymbol{\beta}) / \sigma$ .

Next, we present results by fitting the regression (for  $i = 1, \dots, 172$ )

$$y_i = \beta_0 + \beta_1 v_{i1} + \beta_2 v_{i2} + \sigma z_i,$$

where  $y_i$  can follow the LLE, LBXIIGII and LBXIIGI distributions. For some fitted regressions, Table 4 lists the MLEs (and the corresponding standard errors in parentheses) of the parameters and the values of the following statistics: Akaike information criterion (AIC), Bayesian Information Criterion (BIC) and Consistent Akaike Information Criterion (CAIC). The computations are performed using the `NLMixed` subroutine in SAS. These results indicate that the LLE regression model with censored data could be chosen as the best regression. So, this regression is really competitive to the log-Weibull regression.

The MLEs of the parameters and their standard errors are listed in Table 4. Note that the covariate ( $v_2$ ) is significant at the 1% level, whereas the other covariate is not significant at the usual significance level.

**Table 4:** Estimated quantities,  $p$ -values in  $[\cdot]$  and goodness-of-fit measures from some regressions fitted to entomology data.

Regression	$\alpha$	$\beta_0$	$\beta_1$	$\beta_2$	AIC	CAIC	BIC
LLE	3.9444 (0.2774)	3.8567 (0.0607) [<0.0001]	0.0581 (0.0791) [0.4636]	-0.3474 (0.0882) [<0.0001]	334.5	334.7	347.1
LE		3.1724 (0.1198) [<0.0001]	0.1369 (0.1569) [0.3843]	-0.4430 (0.1766) [0.0130]	423.3	423.5	432.8
	$\sigma$	$\beta_0$	$\beta_1$	$\beta_2$			
LW	0.5151 (0.03256)	3.2435 (0.06309) [<0.0001]	0.1358 (0.08111) [0.0960]	-0.4158 (0.09124) [<0.0001]	344.3	344.6	356.9
	$\sigma$	$k$	$p$	$\beta_0$	$\beta_1$	$\beta_2$	
LBXIIGI	0.4877 (0.0596)	9.3993 (8.5578)	1E-8 (1E-9)	4.3085 (1.1316) [0.0002]	0.1104 (0.0962) [0.2525]	-0.4014 (0.0978) [<0.0001]	348.1 348.6 367.0
LBXIIGII	0.9107 (0.3379)	6.0541 (4.8403)	0.9798 (0.0247)	3.1649 (0.8847) [0.0005]	0.0354 (0.0803) [0.6605]	-0.3252 (0.0876) [0.0003]	335.7 336.2 354.6
LBXII	0.4877 (0.0597)	9.4002 (8.6141)	0	4.3085 (1.1349) [0.0002]	0.1104 (0.0963) [0.2528]	-0.4014 (0.0978) [<0.001]	346.1 346.4 361.8

Finally, we turn to a simplified model retaining only  $v_2$  as an explanatory variable

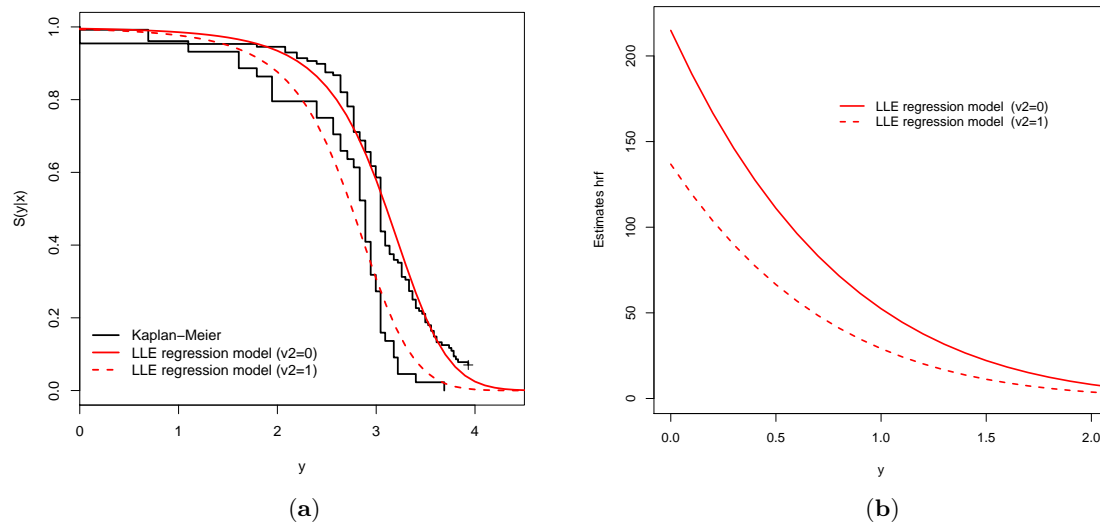
$$y_i = \beta_0 + \beta_2 v_{i2} + \sigma z_i .$$

The MLEs for the LLE regression model fitted to the data are given in Table 5. In order to assess if the model is appropriate, Figure 9(a) displays the plots of the empirical survival function and the estimated survival function from the fitted LLE regression. The plots of its hrfs in Figure 9(b) reveal decreasing shapes. There is a significant difference between the levels of the covariable  $v_2$ . In fact, this regression provides a good fit to these data.

**Table 5:** MLEs of the parameters from the fitted LLE regression model to the entomology data.

Model	$\alpha$	$\beta_0$	$\beta_2$
LLE	3.9489 (0.2777)	3.8845 (0.0475) [<0.001]	-0.3486 (0.0878) [0.0001]





**Figure 9:** Entomology data: (a) Estimated LLE survival function and empirical survival.  
(b) Estimated hrf.

---

## 8. CONCLUDING REMARKS

---

The Weibull, gamma and exponentiated-exponential distributions have two parameters and they are used quite often in survival analysis. These distributions can have increasing or unimodal probability density functions, and monotone hazard functions. However, none of which can have non-monotone hazard rate function shape. In many practical situations, one might observe non-monotone hazard rate functions, and clearly in those cases, none of these distribution functions can be used. The proposed LE distribution can have decreasing or unimodal density function shapes. It is also interesting to note that the hazard rate function possesses three different shapes: decreasing failure rate, bathtub and decreasing-increasing-decreasing.

Moreover, the LE distribution has only two parameters which makes estimating the parameters not very difficult. It may be mentioned that not too many two-parameter distributions can have non-monotone hazard function shape. Therefore, the proposed distribution will be quite useful. Furthermore, its survival and hazard rate functions have closed-form representations. Accordingly, this model can readily be utilized to analyze censored data sets. We also propose a new regression model that can be useful to model real data sets. The importance of the new models is proved empirically by means of three real data sets.

---

## ACKNOWLEDGMENTS

---

The authors are grateful for the comments and suggestions by the referees and the Editor. Their comments and suggestions have greatly improved the paper.

---

## REFERENCES

---

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

- [2] ALZAATREH, A.; LEE, C. and FAMOYE, F. (2013). A new method for generating families of continuous distributions, *Metron*, **71**, 63–79.
- [3] CHESNEAU, C.; BAKOUCH, H.S. and HUSSAIN, T. (2019). A new class of probability distributions via cosine and sine functions with applications, *Communications in Statistics – Simulation and Computation*, **48**, 2287–2300.
- [4] CORDEIRO, G.M. and NADARAJAH, S. (2011). Closed-form expressions for moments of a class of beta generalized distributions, *Brazilian Journal of Probability and Statistics*, **25**, 14–33.
- [5] CORDEIRO, G.M.; ORTEGA, E.M.M. and RAMIRES, T.H. (2015). A new generalized Weibull family of distributions: mathematical properties and applications, *Journal of Statistical Distributions and Applications*, **2**, 1–25.
- [6] GHITANY, M.E.; AL-MUTAIRI, D.K.; BALAKRISHNAN, N. and AL-ENEZI, L.J. (2013). Power Lindley distribution and associated inference, *Computational Statistics & Data Analysis*, **64**, 20–33.
- [7] GRADSHTEYN, I.S. and RYZHIK, I.M. (2000). *Tables of Integrals, Series and Products*, Academic Press, New York.
- [8] GUPTA, R.D. and KUNDU, D. (1999). Generalized exponential distributions, *Australian and Newzealand Journal of Statistics*, **41**, 173–188.
- [9] GUPTA, R.D. and KUNDU, D. (2007). Generalized exponential distribution: existing results and some recent developments, *Journal of Statistical Planning and Inference*, **137**, 3537–3547.
- [10] LANJONI, E.M.; ORTEGA, E.M.M. and CORDEIRO, G.M. (2016). Extended Burr XII regression models: theory and applications, *Journal of Agricultural, Biological and Environmental Statistics*, **21**, 203–224.
- [11] LAWLESS, J.F. (2003). *Statistical Models and Methods for Lifetime Data*, Wiley, New York.
- [12] MAZUCHELI, J.; GHITANY, M.E. and LOUZADA, F. (2014). Estimation methods for a two-parameter bathtub-shaped lifetime distribution, *Australian Journal of Basic and Applied Sciences*, **10**, 189–198.
- [13] MOLENBERGHS, G.; VERBEKE, G. and DEMÉTRIO, C.G.B. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects, *Statistical Science*, **25**, 325–347.
- [14] MURTHY, D.; XIE, M. and JIANG, R. (2004). *Weibull Models*, Wiley, New York.
- [15] NADARAJAH, S.; BAKOUCH, H. and TAHMASBI, R. (2011). A generalized Lindley distribution, *Sankhya*, **73**, 331–359.
- [16] NADARAJAH, S. and HAGHIGHI, F. (2011). An extension of the exponential distribution, *Statistics*, **45**, 543–558.
- [17] SILVA, G.O.; ORTEGA, E.M.M.; CANCHO, V.G. and BARRETO, M.L. (2008). Log-Burr XII regression models with censored data, *Computational Statistics & Data Analysis*, **52**, 3820–3842.
- [18] SILVA, M.A.; BEZERRA-SILVA, G.C.D.; VENDRAMIM, J.D. and MASTRANGELO, T. (2013). Sublethal effect of neem extract on Mediterranean fruit fly adults, *Revista Brasileira de Fruticultura*, **35**, 93–101.
- [19] TAHIR, M.H. and NADARAJAH, S. (2015). Parameter induction in continuous univariate distributions: well-established G families, *Anais da Academia Brasileira de Ciências*, **87**, 539–568.
- [20] VANGENEUGDEN, J.; DE MAZIERE, P.A.; VAN HULLE, M.M.; JAEGGLI, T.; VAN GOOL, L. and VOGELS, R. (2011). Distinct mechanisms for coding of visual actions in macaque temporal cortex, *Journal of Neuroscience*, **31**, 385–401.
- [21] ZUBAIR, M.; TAHIR, M.H.; CORDEIRO, G.M.; ALZAATREH, A. and ORTEGA, E.M.M. (2018). The power-Cauchy negative-binomial: properties and regression, *Journal of Statistical Distributions and Applications*, **5**, 1–17.

---

---

## Likelihood-Based Finite Sample Inference for Synthetic Data from Pareto Model

---

---

Authors: NUTAN MISHRA  

– Department of Mathematics and Statistics,  
University of South Alabama,  
Mobile, AL 36688, USA  
[nmishra@southalabama.edu](mailto:nmishra@southalabama.edu)

SANDIP BARUI 

– Quantitative Methods and Operations Management Area,  
Indian Institute of Management Kozhikode,  
Kozhikode, Kerala 673570, India  
[sandipbarui@iimk.ac.in](mailto:sandipbarui@iimk.ac.in)

Received: November 2019

Revised: March 2021

Accepted: March 2021

### Abstract:

- Statistical agencies often publish microdata or synthetic data to protect confidentiality of survey respondents. This is more prevalent in case of income data. In this paper, we develop likelihood-based finite sample inferential methods for a singly imputed synthetic data using plug-in sampling and posterior predictive sampling techniques under Pareto distribution, a well known income distribution. The estimators are constructed based on sufficient statistics and the estimation methods possess desirable properties. For example, the estimators are unbiased and confidence intervals developed are exact. An extensive simulation study is carried out to analyze the performance of the proposed methods.

### Keywords:

- *maximum likelihood estimation; plug-in sampling; posterior predictive sampling; synthetic data; exact confidence interval; pivotal quantity.*

### AMS Subject Classification:

- 62H10, 62H15, 62H12, 62J05, 62F10, 62E15, 62E10, 62E17, 62D99.

---

## 1. INTRODUCTION

---

Collecting and publishing public data (e.g., tax data) relevant to national interests, or to human race in a broader sense, thereby establishing transparency in government policies and aiding socio-economic development have been primary objectives of many statistical organizations. However, they are also responsible for protecting survey respondents' confidentiality since that leads to greater rates and accuracy in responses. The aforementioned issues have collectively led to the origin of synthetic data where sensitive values are treated like missing values and are imputed based on the underlying data distribution. Mere elimination of the key identifiers, e.g., name, address, unique government identification number, age, etc. may not be sufficient to provide full protection to respondent's identity, and hence, additional steps should be taken to this end. Synthetic data are generated in a way that the privacy and confidentiality of general public are not compromised, however, keeping underlying structure of the stochastic model that generated the data, intact. The inferences drawn from synthetic data are expected to reveal similar characteristics as the ones obtained actual data, hence, any decisions or actions based on synthetic data remain valid. Some well known techniques in this front involve cell suppression (method of withholding values of the primary risky cells and secondary nearly-risky cells by some random mechanism; [9], [7]), data swapping (a perturbation method of creating pairs with similar attributes and interchanging sensitive values between them; [5]), top coding/bottom coding (replacing confidential values of an attribute with the maximum or minimum or some other threshold values), random noise perturbation (method of contaminating data with random noises following some known distribution and applying statistical methods to estimate the true values ignoring the noises; [10], [11], [12]) and multiple imputation (replacing sensitive values with some aggregated measure obtained from multiple imputed values by utilizing the underlying stochastic nature of the data; [31], [35]) have been implemented widely for statistical disclosure control.

In this context, application of noise perturbed data and synthetic data have gained recognition only in the recent years. Under these techniques, random errors or noises are generated from a well known probability distribution and applied on quantitative data that need to be masked, either additively or multiplicatively. Many inherent characteristics and principal features of noise-perturbed data obtained from the actual microdata in order to protect privacy were studied by [10], [11], [12], [19], [22], [33], to name a few. Recently, [14] in their paper, developed a likelihood based inferential method under the assumption of multiplicative noise where data is obtained from a parametric model.

One of the early works to implement synthetic data for statistical disclosure control was accomplished by [30] where synthetic data is generated with a concept similar to multiple imputation [31]. Multiple imputation provides a framework in which each datum is replaced by a vector of  $m$  values sampled from a known probability distribution. In [30], the author suggested that multiple-imputation technique results in synthetic data that do not resemble any actual sampling unit while preserving inherent properties of the underlying distribution and confidentiality of the respondents. Detailed parametric and non-parametric inferential methods of analyses based on synthetic data were examined by [25].

An illustration on multiply imputed fully synthetic public use microdata with respect to inferences on various descriptive and analytic estimands, and degree of protection of confidentiality, was carried out by [27]. Modified adaptations on multiple imputation based framework in context of missing data, data confidentiality and measurement error was discussed in [29].

A likelihood-based finite sample inference was studied by [15] for a synthetic data obtained from an exponential distribution. Similar studies were carried out by [16] and [17] where synthetic data are generated from a normal distribution using posterior predictive sampling and plug-in sampling methods. Further discussions and developments in synthetic data methodology could be found in [28], [26] and [13].

Following the line of work similar to [15], in this paper we develop a likelihood-based inferential procedure for synthetic data using plug-in sampling and posterior predictive sampling where the true population is a two-parameter Pareto distribution. Define  $x = (x_1, \dots, x_n)^\top$  as the original microdata with a probability density function (pdf) given by  $f_\theta(x)$  where  $\theta$  is the parameter characterizing the underlying population. To illustrate the mechanism of plug-in sampling, let  $\hat{\theta} = \hat{\theta}(x)$  be a point estimate of  $\theta$ . Then, for a positive integer  $m$ , a synthetic data is given by  $Y = (y_1, \dots, y_m)$  where  $y_i = (y_{i1}, \dots, y_{in})^\top$ ;  $i = 1, \dots, m$  is a random sample generated from  $f_{\hat{\theta}}(\cdot)$ . On the other hand, posterior predictive sampling method assumes an appropriate prior distribution  $\pi(\theta)$  of  $\theta$ .  $\theta^*$  is chosen randomly from the posterior distribution  $\pi(\theta|x)$  of  $\theta$  given  $x$ . A synthetic data is given by  $Y = (y_1, \dots, y_m)$  where  $y_i = (y_{i1}, \dots, y_{in})^\top$ ;  $i = 1, \dots, m$  is a random sample generated from  $f_{\theta_i^*}(\cdot)$  where  $\theta_i^*$  is the value of  $\theta$  obtained by sampling from  $\pi(\theta|x)$  at  $i$ -th draw.

As discussed by [28], [26] and [13], for multiple imputed data sets, one may develop inference based on a scalar parameter  $Q = Q(\theta)$ . Let  $\eta = \eta(x)$  and  $\nu = \nu(x)$  be point estimator of  $Q(\theta)$  and estimator of variance of  $\eta$ , respectively. An estimator of  $Q$  obtained from the synthetic data  $Y$  is given by

$$(1.1) \quad \bar{\eta}_m = \frac{1}{m} \sum_{i=1}^m \eta_i$$

and an estimator of variance of  $\bar{\eta}_m$  is given by

$$(1.2) \quad V_m = \frac{1}{m(m-1)} \sum_{i=1}^m (\eta_i - \bar{\eta}_m)^2 + \frac{1}{m} \sum_{i=1}^m \nu_i,$$

where  $\eta_i = \eta(y_i)$  and  $\nu_i = \nu(y_i)$  for  $i = 1, \dots, m$ . For the upper  $\gamma/2$ -th quantile  $t_{\gamma/2, \nu}$  for a  $t$ -distribution with degrees of freedom

$$\nu = (m-1) \left[ 1 + \frac{(m-1) \sum_{i=1}^m \nu_i}{\sum_{i=1}^m (\eta_i - \bar{\eta}_m)^2} \right]^2,$$

an approximate interval estimate of  $Q(\theta)$  can be evaluated using  $(\bar{\eta}_m \pm t_{\gamma/2, \nu} \sqrt{V_m})$ .

Income data are often published by the statistical agencies as aggregates to ensure confidentiality at the cost of huge information loss. In order to circumvent this problem, these agencies use microdata in form of individual income data published synthetically. Again, Internal Revenue Service (IRS) releases tax return records of chosen individuals by masking their key identifiers because these are important source of information for policy makers, academicians or non-profit research organizations to analyze the influences of variation of tax policies on revenues or burden of tax on different social strata [3]. It is widely known that individual income can be well-modeled by Pareto distribution ([8]; [21]; [20]; [1]). The pdf of a random variable  $X$  following a Pareto distribution is given by

$$(1.3) \quad f_\theta(x) = \frac{\psi C^\psi}{x^{\psi+1}},$$

where  $x > C$ ,  $C$  is a scale parameter that denotes minimum threshold value for  $x$ ,  $\psi > 0$  is a shape parameter, and  $\theta = (C, \psi)^\top$ . In economics,  $\psi$  is known as the Pareto index [34] which is a measure related to breadth of the income distribution.

Though synthetic or imputed data are widely used to mask income related information of individuals [6], inferential procedures for a synthetic data generated from a Pareto model have not been studied yet, to the best of our knowledge. Therefore, in this paper, we study and develop inferential methods based on likelihood function for a model-based singly imputed synthetic data using plug-in and posterior predictive sampling methods when the original data is obtained from a Pareto distribution. The formulation and derivations of the inferential methodologies are mathematically more intensive, complex and challenging in comparison to the exponential [15] or normal [17] distributions, owing to the dependency between the scale parameter  $C$  and the Pareto random variable. In particular, for posterior predictive sampling, expressions for the estimators are either implicit or their derivations are intractable. However, the estimators that could be derived are sufficient for the concerned parameter and mostly exact in nature, except few which are build based on asymptotic normality of the ML estimators. Moreover, as argued by [16], developing inferential methods based on synthetic data requires generation of  $m$  random samples of size  $n$  with  $m > 1$ . However, situations arise when  $m$  may not be greater than one due to stricter privacy policies or to avoid high disclosure risks [15], and only a single synthetic version of the original data is available for study. Thus, a major motivation of this work is to establish valid inferential results based on a single synthetic data by properly utilizing the underlying model structure.

The rest of the paper is arranged as follows. In Section 2, discussion on methodology to estimate the parameters is provided. Section 3 deals with a simulation study which is carried out to validate the performance of our proposed method of estimation. Interpretation of the results of the simulation study are also discussed. Finally, concluding remarks are made in Section 4.

---

## 2. METHODOLOGY FOR DRAWING LIKELIHOOD BASED INFERENCE

---

Let  $X = (X_1, \dots, X_n)^\top$  represent the original data of size  $n$  where  $X_1, \dots, X_n$  are independent and identically distributed (iid) according to Pareto distribution with a pdf given in (1.3). The maximum likelihood (ML) estimators of  $C$  and  $\psi$  are, respectively, given by  $\hat{C} = X_{(1)} = \min\{X_1, \dots, X_n\}$  and  $\hat{\psi} = n \left[ \sum_{i=1}^n \log \left( \frac{X_i}{X_{(1)}} \right) \right]^{-1}$ . Note that the sampling distribution of  $\hat{C}$  is Pareto with scale parameter  $C$  and shape parameter  $n\psi$ . On the other hand,  $\hat{\psi}$  follows Inverse-Gamma (IG) distribution with parameters  $n$  and  $n\psi$  when  $C$  is known, and IG distribution with parameters  $n-1$  and  $n\psi$  when  $C$  is unknown [32]. Moreover,  $\hat{C}$  and  $\hat{\psi}$  are stochastically independent ([20]; [32]). Furthermore,  $\hat{C} = X_{(1)}$  is sufficient for  $C$  when  $\psi$  is known,  $\left( \prod_{i=1}^n X_i \right)^{1/n} = C e^{1/\hat{\psi}}$  is sufficient for  $\psi$  when  $C$  is known, and  $\hat{\theta} = \left( X_{(1)}, \sum_{i=1}^n \log \left( \frac{X_i}{X_{(1)}} \right) \right)^\top = (\hat{C}, n/\hat{\psi})^\top$  is jointly sufficient for  $\theta = (C, \psi)^\top$  when both  $C$  and  $\psi$  are unknown ([20]). Finally,  $\hat{C}$  and  $\hat{\psi}$  are both individually complete whereas  $(\hat{C}, \hat{\psi})^\top$  is jointly complete [32]. With this background, the following results are developed for synthetic data based on plug-in sampling.

---

## 2.1. Plug-in sampling

---

Let  $Y = (Y_1, Y_2, \dots, Y_N)^\top$  be a synthetic data of size  $N$  obtained by generating a random sample from a Pareto distribution with parameters  $\hat{C}$  and  $\hat{\psi}$ . For  $m$  multiply imputed synthetic data sets,  $N$  is generally taken as  $nm$ . However, our interest lies in the case where  $m = 1$  to incorporate stricter confidentiality as mentioned earlier. Hence, assuming the value of  $n$  known,  $N$  is considered to be equal to  $n$ . Once the synthetic data  $Y = (Y_1, Y_2, \dots, Y_n)^\top$  is obtained, our objective is to provide inference on  $\theta = (C, \psi)^\top$  based on  $Y$ . In the following subsections, we describe methodologies to draw inference on  $\theta$  under three scenarios, viz., inference on  $C$  when  $\psi$  is known, inference on  $\psi$  when  $C$  is known, and inference on  $\theta$  when both  $C$  and  $\psi$  are unknown.

---

### 2.1.1. Inference on $\psi$ when $C$ is known

---

Under this scenario,  $Y$  is generated from Pareto distribution with the value of  $C$  known. Let us define  $A = C^{-n} \prod_{i=1}^n Y_i$ .

**Theorem 2.1.** For  $i = 1, \dots, n$ ,  $y_i > C > 0$ ,  $\psi > 0$  and  $C$  known, the pdf of  $Y$  is given by

$$(2.1) \quad g_\psi(y) = \frac{2(\psi n)^n}{A C^n \Gamma(n)} \text{BesselK}\left(0, 2\sqrt{n\psi \log A}\right),$$

where  $\text{BesselK}(\cdot, \cdot)$  is the modified-Bessel function of second kind defined as

$$(2.2) \quad \text{BesselK}(n, z) = \sqrt{\frac{\pi}{2z}} \frac{e^{-z}}{(n - \frac{1}{2})!} \int_0^\infty e^{-t} t^{n-1/2} \left(1 - \frac{t}{2z}\right)^{n-1/2} dt,$$

for  $n \in \mathbb{R}$  and  $z \in \mathbb{C}$ .

**Proof:** For  $y_i > C > 0$ ,  $i = 1, \dots, n$ ,  $\psi > 0$  and known  $C$ , the conditional pdf of  $Y$  given  $\hat{\psi}$  is given by

$$g_1(y|\hat{\psi}) = \hat{\psi}^n C^{n\hat{\psi}} \left(\prod y_i\right)^{-\hat{\psi}-1}$$

and the conditional pdf of  $\hat{\psi}$  given  $\psi$  is given by

$$g_2(\hat{\psi}|\psi) = \frac{\psi^n n^n}{\Gamma(n)} \hat{\psi}^{-n-1} \exp(-\psi n/\hat{\psi}).$$

Thus,

$$\begin{aligned} g_\psi(y) &= g_1(y|\hat{\psi}) \times g_2(\hat{\psi}|\psi) \\ &= \frac{\psi^n n^n}{\Gamma(n)} \int_0^\infty C^{n\hat{\psi}} \left(\prod y_i\right)^{-\hat{\psi}-1} \exp(-n\psi/\hat{\psi}) \hat{\psi}^{-1} d\hat{\psi} \\ &= \frac{\psi^n n^n}{\Gamma(n) C^n} \int_0^\infty \hat{\psi}^{-1} A^{-\hat{\psi}-1} \exp(-n\psi/\hat{\psi}) d\hat{\psi}. \end{aligned} \quad \square$$

Many well known distributions can be expressed in the form of Bessel function. This special function, namely, modified Bessel function of second kind expressed in (2.2) can be computed for specified values of its argument using `Mathematica` version 12.2 [36].

*Uniformly minimum variance unbiased estimator and exact confidence interval for  $\psi$*

As discussed in ([20]),  $A$  is sufficient for  $\psi$  and complete. Let us define

$$(2.3) \quad \tilde{\psi} = \frac{n}{\sum_{i=1}^n \log(Y_i/C)} = n [\log(A)]^{-1}.$$

$\tilde{\psi}$  is also sufficient for  $\psi$  and complete. Hence,

$$(2.4) \quad E\{\tilde{\psi}\} = E\{E\{\tilde{\psi}|\hat{\psi}\}\} = E\left\{\frac{n\hat{\psi}}{n-1}\right\} = \frac{n^2}{(n-1)^2} \psi.$$

An unbiased estimator of  $\psi$  is  $\psi_u = \frac{(n-1)^2}{n^2} \tilde{\psi}$ .  $\psi_u$  is also a sufficient and complete statistic. Further, *Lehmann Scheffé theorem* ([4, Chapter 6]), implies that  $\psi_u$  is the uniformly minimum variance unbiased estimator (UMVUE) of  $\psi$  ([4, Chapter 7]). The variance of  $\psi_u$  is given by

$$(2.5) \quad \begin{aligned} V(\psi_u) &= V(E\{\psi_u|\hat{\psi}\}) + E\{V(\psi_u|\hat{\psi})\} \\ &= \left(\frac{n-1}{n}\right)^4 \left[ V(E\{\tilde{\psi}|\hat{\psi}\}) + E\{V(\tilde{\psi}|\hat{\psi})\} \right] \\ &= \left(\frac{n-1}{n}\right)^4 \left[ V\left(\frac{n\hat{\psi}}{n-1}\right) + E\left\{\frac{n^2\hat{\psi}^2}{(n-1)^2(n-2)}\right\} \right] \\ &= \left\{\frac{2n-3}{(n-2)^2}\right\} \psi^2. \end{aligned}$$

An estimate  $\widehat{V(\psi_u)}$  of  $V(\psi_u)$  is obtained using (2.5) by replacing  $\psi$  with  $\tilde{\psi}$ .

To find an exact CI for  $\psi$ , we construct a pivotal quantity based on the sufficient statistic  $\tilde{\psi}$ . Recall that  $\tilde{\psi}$  follows IG distribution with parameters  $(n, n\hat{\psi})$  when  $C$  is known. Then the conditional pdf of  $\tilde{\psi}$  is

$$(2.6) \quad g_2(\tilde{\psi}|\hat{\psi}) = \frac{\hat{\psi}^n n^n}{\Gamma(n)} \tilde{\psi}^{-n-1} \exp\left(-\frac{\hat{\psi}n}{\tilde{\psi}}\right).$$

Again, the conditional pdf of  $\hat{\psi}$  given  $\psi$  is given by

$$(2.7) \quad g_2(\hat{\psi}|\psi) = \frac{\psi^n n^n}{\Gamma(n)} \hat{\psi}^{-n-1} \exp\left(-\frac{\psi n}{\hat{\psi}}\right).$$

Combining (2.6) and (2.7), we obtain

$$(2.8) \quad h_\psi(\tilde{\psi}) = \frac{\psi^n n^{2n}}{[\Gamma(n)]^2} \int_0^\infty \hat{\psi}^{-1} \tilde{\psi}^{-n-1} \exp\left(-n\left[\frac{\psi}{\hat{\psi}} + \frac{\hat{\psi}}{\tilde{\psi}}\right]\right) d\hat{\psi}.$$



Taking substitution  $\psi_a = \frac{\tilde{\psi}}{\psi}$ , (2.8) can be written as

$$(2.9) \quad h_{\psi}(\tilde{\psi}) = \frac{\psi^n n^{2n}}{[\Gamma(n)]^2} \tilde{\psi}^{-n-1} \int_0^{\infty} \psi_a^{-1} \exp\left(-n\left[\frac{1}{\psi_a} + \frac{\psi \psi_a}{\tilde{\psi}}\right]\right) d\psi_a.$$

Further, considering a transformation of variable  $\tilde{\psi} \rightarrow W$  where  $W = \frac{\tilde{\psi}}{\psi}$  we obtain

$$(2.10) \quad h(w) = \frac{n^{2n}}{[\Gamma(n)]^2} w^{-n-1} \int_0^{\infty} \psi_a^{-1} \exp\left(-n\left[\frac{1}{\psi_a} + \frac{\psi_a}{w}\right]\right) d\psi_a,$$

which is independent of  $\psi$ . Hence,  $W = \frac{\tilde{\psi}}{\psi} = \frac{n(\log A)^{-1}}{\psi}$  is a pivot for  $\psi$ . For a given level of significance  $\gamma \in (0, 1)$ , we may obtain  $\kappa_2 > \kappa_1 > 0$  such that

$$(2.11) \quad \int_{\kappa_1}^{\kappa_2} h(w) dw = 1 - \gamma.$$

Therefore, an exact  $(1 - \gamma)$  100% CI for  $\psi$  is given by

$$(2.12) \quad \left(\frac{n(\log A)^{-1}}{\kappa_2}, \frac{n(\log A)^{-1}}{\kappa_1}\right).$$

$\kappa_1$  and  $\kappa_2$  are chosen such that the CI in (2.12) has the shortest length. For achieving that, we define the length of the CI in (2.12) as

$$L_{\psi} = n(\log A)^{-1} \left[\frac{1}{\kappa_1} - \frac{1}{\kappa_2}\right] = \tilde{\psi} \left[\frac{1}{\kappa_1} - \frac{1}{\kappa_2}\right].$$

The objective is to find  $\kappa_1$  and  $\kappa_2$  such that the expected value of  $L_{\psi}$  is minimum subject to (2.11). Applying Lagrangian multiplier technique, the Lagrangian function  $L_{\psi}(\kappa_1, \kappa_2, \lambda)$  is obtained as

$$(2.13) \quad L_{\psi}(\kappa_1, \kappa_2, \lambda) = \left[\frac{1}{\kappa_1} - \frac{1}{\kappa_2}\right] + \lambda \left(H_W(\kappa_2) - H_W(\kappa_1) - (1 - \gamma)\right),$$

where  $\lambda$  is the Lagrangian multiplier and  $H_W(w) = \int_0^w h(u) du$ . On taking partial derivatives of  $L_{\psi}(\kappa_1, \kappa_2, \lambda)$  in (2.13) with respect to  $\kappa_1$ ,  $\kappa_2$  and  $\lambda$  we solve (2.14) for  $\kappa_1$  and  $\kappa_2$  where

$$(2.14) \quad \begin{aligned} \kappa_1^2 h(\kappa_1) - \kappa_2^2 h(\kappa_2) &= 0, \\ H_W(\kappa_2) - H_W(\kappa_1) - (1 - \gamma) &= 0. \end{aligned}$$

---

### Maximum likelihood estimator and asymptotic confidence interval for $\psi$

---

The ML estimator of  $\psi$  is obtained as usual by taking the partial derivative of the (2.1) and equating to zero. That is, solving

$$(2.15) \quad \psi - \frac{n}{\log A} \left(\frac{\text{BesselK}[0, 2\sqrt{n\psi \log A}]}{\text{BesselK}[1, 2\sqrt{n\psi \log A}]}\right)^2 = 0$$

for  $\psi$ , the ML estimator  $\tilde{\psi}_{\text{syn}}$  of  $\psi$  can be obtained. It is well known that under certain regularity conditions  $\tilde{\psi}_{\text{syn}}$  follows an asymptotic normal distribution ([18, Chapter 6.3]) with

mean  $\psi$  and variance  $\sigma^2(\tilde{\psi}_{\text{syn}}) = I(\psi)^{-1}$  where  $I(\psi) = -E\left[\left\{\frac{\partial^2 \log g_{\psi}(y)}{\partial \psi^2}\right\}\right]$  is the information at the true value of  $\psi$ . Since  $\sigma^2(\tilde{\psi}_{\text{syn}})$  depends on unknown  $\psi$ , an estimate of  $\sigma^2(\tilde{\psi}_{\text{syn}})$  is given by  $\hat{\sigma}^2(\tilde{\psi}_{\text{syn}}) = -\left\{\frac{\partial^2 \log g_{\psi}(y)}{\partial \psi^2}\right\}\Big|_{\psi=\tilde{\psi}_{\text{syn}}}$  ([23, Chapter 35]). Therefore, an asymptotic  $100(1 - \gamma)\%$  CI for  $\psi$  is given by  $(\tilde{\psi}_{\text{syn}} \pm z_{\gamma/2} \hat{\sigma}(\tilde{\psi}_{\text{syn}}))$ .

2.1.2. Inference on  $C$  when  $\psi$  is known

Under this scenario, a synthetic data  $y$  is generated from Pareto distribution with the scale parameter  $\hat{C} = X_{(1)}$  and the shape parameter as  $\psi$ . The goal is to derive inference on  $C$  based on  $y$ . Central to this goal is the joint pdf  $g_C(y)$  which can be used to obtain the likelihood function  $L(C|y)$ . Let us define  $\tilde{C} = Y_{(1)} = \min\{Y_1, \dots, Y_n\}$  and  $B = \prod_{i=1}^n y_i$ .

**Theorem 2.2.** *The joint pdf of  $Y$  is given by*

$$(2.16) \quad g_C(y) = \frac{n \psi^{n+1} C^{n\psi}}{B^{\psi+1}} \times \log\left(\frac{\tilde{C}}{C}\right),$$

where  $y_i > C > 0$  for  $i = 1, \dots, n$ ,  $\tilde{C} > C$  and  $\psi > 0$ .

**Proof:** Note that  $\tilde{C} > \hat{C} > C$ . Let  $g_3(y|\hat{C})$  and  $g_4(\hat{C}|C)$  be the conditional pdfs of  $y$  given  $\hat{C}$  and  $\hat{C}$  given  $C$ , respectively. Also,  $g_4(\hat{C}|C)$  is Pareto with parameters  $C$  and  $n\psi$ . For  $\tilde{C} > C$ , the joint pdf of  $Y$  is expressed as

$$(2.17) \quad \begin{aligned} g_C(y) &= \int_C^{\tilde{C}} g_3(y|\hat{C}) g_4(\hat{C}|C) d\hat{C} \\ &= \int_C^{\tilde{C}} \frac{\psi^n \hat{C}^{n\psi}}{(\prod y_i)^{\psi+1}} \times \frac{n \psi C^{n\psi}}{\hat{C}^{n\psi+1}} d\hat{C} \\ &= \frac{n \psi^{n+1} C^{n\psi}}{(\prod y_i)^{\psi+1}} \int_C^{\tilde{C}} \frac{d\hat{C}}{\hat{C}} \\ &= \frac{n \psi^{n+1} C^{n\psi}}{B^{\psi+1}} \times \log\left(\frac{\tilde{C}}{C}\right). \quad \square \end{aligned}$$

Uniformly minimum variance unbiased estimator and exact confidence interval for  $C$

Since  $\tilde{C} = Y_{(1)}$  is a complete sufficient statistic for  $C$  when  $\psi$  is known,  $C_u = \frac{(n\psi-1)^2}{(n\psi)^2} \tilde{C}$  is an unbiased estimator of  $C$  as shown below.

$$(2.18) \quad E\{\tilde{C}_u\} = E\{E\{\tilde{C}_u|\hat{C}\}\} = \frac{(n\psi-1)^2}{(n\psi)^2} E\{E\{\tilde{C}|\hat{C}\}\} = \frac{(n\psi-1)^2}{(n\psi)^2} E\left\{\frac{n\psi\hat{C}}{n\psi-1}\right\} = C.$$

By *Lehmann Scheffé theorem* ([4, Chapter 6]),  $\tilde{C}_u$  is the UMVUE of  $C$  when  $\psi$  is known. The variance of  $\tilde{C}_u$  is given by

$$\begin{aligned}
 V\{\tilde{C}_u\} &= V\{E\{\tilde{C}_u|\hat{C}\}\} + E\{V\{\tilde{C}_u|\hat{C}\}\} \\
 &= \frac{(n\psi - 1)^4}{(n\psi)^4} \times \left[ V\{E\{\tilde{C}|\hat{C}\}\} + E\{V\{\tilde{C}|\hat{C}\}\} \right] \\
 &= \frac{(n\psi - 1)^4}{(n\psi)^4} \times \left[ V\left\{ \frac{n\psi\hat{C}}{n\psi - 1} \right\} + E\left\{ \frac{n\psi\hat{C}^2}{(n\psi - 1)^2(n\psi - 2)} \right\} \right] \\
 (2.19) \quad &= \left\{ 2 - \frac{1}{(n\psi - 1)^2} \right\} \frac{C^2}{(n\psi)^2}.
 \end{aligned}$$

The development of an exact confidence interval for  $C$  involves construction of a pivot for  $C$  from its sufficient statistic  $\tilde{C} = Y_{(1)}$ . For  $\tilde{C} > C$ , the pdf of  $\tilde{C}$  is given by

$$\begin{aligned}
 h_C(\tilde{C}) &= \int_C^{\tilde{C}} g_4(\tilde{C}|\hat{C}) g_4(\hat{C}|C) d\hat{C} \\
 &= \int_C^{\tilde{C}} \frac{n\psi\hat{C}^{n\psi}}{\tilde{C}^{n\psi+1}} \times \frac{n\psi C^{n\psi}}{\hat{C}^{n\psi+1}} d\hat{C} \\
 (2.20) \quad &= \frac{n^2\psi^2 C^{n\psi}}{\tilde{C}^{n\psi+1}} \times \log\left(\frac{\tilde{C}}{C}\right).
 \end{aligned}$$

Let  $T = \log\left(\frac{\tilde{C}}{C}\right)$ , then the pdf of  $T$  as

$$\tilde{h}(t) = n^2\psi^2 t e^{-n\psi t}, \quad \text{for } t > 0$$

and  $\tilde{h}(t)$  is independent of  $C$ . For some  $\kappa_2 > \kappa_1 \geq 1$  and  $\gamma \in (0, 1)$ , we obtain

$$\int_{\kappa_1}^{\kappa_2} \tilde{h}(t) dt = 1 - \gamma.$$

Therefore, an exact  $100(1 - \gamma)\%$  CI for  $C$  is given by  $(\tilde{C}e^{-\kappa_2}, \tilde{C}e^{-\kappa_1})$ . Define  $\tilde{H}_C(c) = \int_0^c \tilde{h}(u) du$ . Following the steps as discussed in Section 2.1.1, the shortest length  $100(1 - \gamma)\%$  for  $C$  is obtained by solving

$$\begin{aligned}
 e^{\kappa_1} \tilde{h}(\kappa_1) - e^{\kappa_2} \tilde{h}(\kappa_2) &= 0, \\
 (2.21) \quad \tilde{H}_C(\kappa_2) - \tilde{H}_C(\kappa_1) - (1 - \gamma) &= 0,
 \end{aligned}$$

for  $\kappa_1$  and  $\kappa_2$ .

---

#### Maximum likelihood estimation of $C$

---

The usual method of derivative based ML estimation cannot be applied here to obtain the ML estimate of  $C$ . However, noting

$$\frac{\partial L(C|y)}{\partial C} = \frac{\partial g_C(y)}{\partial C} = -\frac{n\psi^{n+1} C^{-(n+1)}}{A^{\psi+1}} \left\{ 1 + \log\left(\frac{\tilde{C}}{C}\right) \right\} < 0,$$

i.e.,  $L(C|y)$  is decreasing in  $C$  with  $0 < C < \tilde{C}$ , the ML estimator of  $C$  is obtained as  $\tilde{C} = Y_{(1)}$ . The exact distribution of  $\tilde{C}$  is given by Equation (2.20). An estimate of the variance of  $\tilde{C}$  can be derived from (2.19) as

$$\widehat{V(\tilde{C})} = \left[ \frac{(n\psi)^2}{(n\psi - 1)^4} \left\{ 2 - \frac{1}{(n\psi - 1)^2} \right\} \right] \tilde{C}^2.$$

2.1.3. Inference on  $\theta = (C, \psi)^\top$  when both  $C$  and  $\psi$  are unknown

To develop inference on  $\theta$ , the joint pdf of  $y$  given  $\theta$  in Theorem 2.3, where  $y = (y_1, \dots, y_n)^\top$  is a synthetic data obtained from Pareto distribution with parameters  $\hat{C}$  and  $\hat{\psi}$ . Define  $\psi^* = \frac{n}{\sum_{i=1}^n \log(Y_i/Y_{(1)})}$ .  $\psi^*$  follows IG with parameters  $n - 1$  and  $n\hat{\psi}$ .

**Theorem 2.3.** *The joint pdf of  $Y$  is given by*

$$(2.22) \quad g_\theta(y) = \frac{(n\psi)^n C^{n\psi}}{\Gamma(n-1)} \int_0^\infty \left( \frac{\tilde{C}^{n(\hat{\psi}-\psi)} - C^{n(\hat{\psi}-\psi)}}{n(\hat{\psi}-\psi)} \right) \frac{\exp\left\{-\frac{n\psi}{\hat{\psi}}\right\}}{\left(\prod_{i=1}^n y_i\right)^{\hat{\psi}+1}} d\hat{\psi},$$

where  $\tilde{C} = \min\{y_1, \dots, y_n\} > C > 0$  and  $\psi > 0$ .

**Proof:** The conditional pdf of  $y$  given  $\hat{\theta} = (\hat{C}, \hat{\psi})^\top$  is expressed as

$$(2.23) \quad g_5(y|\hat{\theta}) = \frac{\hat{\psi}^n \hat{C}^{n\hat{\psi}}}{\left(\prod_{i=1}^n y_i\right)^{\hat{\psi}+1}},$$

where  $y_i > \hat{C} > C > 0$  for  $i = 1, \dots, n$  and  $\hat{\psi} > 0$ . Again, the conditional pdf of  $\hat{\theta}$  given  $\theta$  is

$$(2.24) \quad g_6(\hat{\theta}|\theta) = \frac{(n\psi)^n C^{n\psi} \exp\left\{-\frac{n\psi}{\hat{\psi}}\right\}}{\hat{C}^{n\hat{\psi}+1} \hat{\psi}^n \Gamma(n-1)},$$

for  $0 < C < \hat{C} < \tilde{C}$ ,  $\hat{\psi} > 0$  and  $\psi > 0$ . Equation (2.24) is obtained using the fact that  $\hat{C}$  and  $\hat{\psi}$  are stochastically independent where  $\hat{C}$  follows Pareto distribution with scale  $C$  and shape  $n\psi$ , and  $\hat{\psi}$  follows IG distribution with parameters  $n - 1$  and  $n\psi$ . Finally, the pdf of  $y$  is given by

$$(2.25) \quad \begin{aligned} g_\theta(y) &= \int_0^\infty \int_C^{\tilde{C}} g_5(y|\hat{\theta}) \times g_6(\hat{\theta}|\theta) d\hat{C} d\hat{\psi} \\ &= \frac{(n\psi)^n C^{n\psi}}{\Gamma(n-1)} \int_C^{\tilde{C}} \left( \hat{C}^{n(\hat{\psi}-\psi)-1} \right) \int_0^\infty \frac{\exp\left\{-\frac{n\psi}{\hat{\psi}}\right\}}{\left(\prod_{i=1}^n y_i\right)^{\hat{\psi}+1}} d\hat{\psi} d\hat{C}, \end{aligned}$$

which can be further simplified to

$$(2.26) \quad g_\theta(y) = \frac{(n\psi)^n C^{n\psi}}{\Gamma(n-1)} \int_0^\infty \left( \frac{\tilde{C}^{n(\hat{\psi}-\psi)} - C^{n(\hat{\psi}-\psi)}}{n(\hat{\psi}-\psi)} \right) \frac{\exp\left\{-\frac{n\psi}{\hat{\psi}}\right\}}{\left(\prod_{i=1}^n y_i\right)^{\hat{\psi}+1}} d\hat{\psi}. \quad \square$$

---

Construction of a pivot for  $\theta$

---

Let us define  $\tilde{\theta} = (\tilde{C}, \psi^*)^\top$ . The pdf of  $\tilde{\theta}$  is given by

$$\begin{aligned}
 h_{\tilde{\theta}}(\tilde{\theta}) &= \int_0^\infty \int_C g_6(\tilde{\theta}|\hat{\theta}) \times g_6(\hat{\theta}|\theta) d\hat{C} d\hat{\psi} \\
 &= \int_0^\infty \int_C \frac{(n\hat{\psi})^n \hat{C}^{n\hat{\psi}} \exp\left\{-\frac{n\hat{\psi}}{\psi^*}\right\}}{\tilde{C}^{n\hat{\psi}+1} \psi^{*n} \Gamma(n-1)} \times \frac{(n\psi)^n C^{n\psi} \exp\left\{-\frac{n\psi}{\psi}\right\}}{\hat{C}^{n\psi+1} \hat{\psi}^n \Gamma(n-1)} d\hat{C} d\hat{\psi} \\
 (2.27) \quad &= \frac{n^{2n} C^{n\psi} \psi^n}{\{\Gamma(n-1)\}^2 \psi^{*n}} \int_0^\infty \int_C \frac{\hat{C}^{n(\hat{\psi}-\psi)-1}}{\tilde{C}^{n\hat{\psi}+1}} \exp\left[-n\left\{\frac{\hat{\psi}}{\psi^*} + \frac{\psi}{\hat{\psi}}\right\}\right] d\hat{C} d\hat{\psi}.
 \end{aligned}$$

Substituting  $t = \frac{\hat{\psi}}{\psi}$ , we obtain

$$\begin{aligned}
 h_{\tilde{\theta}}(\tilde{\theta}) &= \frac{n^{2n} C^{n\psi} \psi^n}{\{\Gamma(n-1)\}^2 \psi^{*n}} \int_0^\infty \int_C \frac{\hat{C}^{n\psi(t-1)-1}}{\tilde{C}^{n\psi t+1}} \exp\left[-n\left\{\frac{\psi t}{\psi^*} + \frac{1}{t}\right\}\right] d\hat{C} \times \psi dt \\
 &= \frac{n^{2n} C^{n\psi} \psi^{n+1}}{\{\Gamma(n-1)\}^2 \psi^{*n}} \int_0^\infty \frac{1}{\tilde{C}^{n\psi t+1}} \exp\left[-n\left\{\frac{\psi t}{\psi^*} + \frac{1}{t}\right\}\right] \times \int_C \hat{C}^{n\psi(t-1)-1} d\hat{C} dt \\
 (2.28) \quad &= \frac{n^{2n} C^{n\psi} \psi^{n+1}}{\{\Gamma(n-1)\}^2 \psi^{*n}} \int_0^\infty \frac{1}{\tilde{C}^{n\psi t+1}} \exp\left[-n\left\{\frac{\psi t}{\psi^*} + \frac{1}{t}\right\}\right] \times \left[\frac{\tilde{C}^{n\psi(t-1)} - C^{n\psi(t-1)}}{n\psi(t-1)}\right] dt.
 \end{aligned}$$

Considering a bivariate transformation  $(\tilde{C}, \psi^*) \rightarrow (U, V)$  where

$$(2.29) \quad U = \left(\frac{\tilde{C}}{C}\right)^\psi \quad \text{and} \quad V = \frac{\psi^*}{\psi},$$

we obtain pdf of  $(U, V)$  which is independent of  $\theta$ . The Jacobian of the transformation is  $C u^{\frac{1}{\psi}-1}$ . From (2.28), the joint pdf of  $(U, V)$  is

$$(2.30) \quad h_{U,V}(u, v) = \frac{n^{2n-1}}{\{\Gamma(n-1)\}^2 v^n} \int_0^\infty \exp\left[-n\left\{\frac{t}{v} + \frac{1}{t}\right\}\right] \times \left[\frac{u^{n(t-1)} - 1}{u^{nt+1}(t-1)}\right] dt, \quad u > 1 \text{ and } v > 0,$$

which is independent of  $\theta$ . The marginal pdfs of  $U$  and  $V$  are obtained from (2.30) as follows:

$$(2.31) \quad h_U(u) = \frac{n^n}{\Gamma(n-1)} \int_0^\infty \frac{\{u^{n(t-1)} - 1\} \exp(-n/t)}{u^{nt+1} t^{n-1} (t-1)} dt, \quad u > 1,$$

and

$$(2.32) \quad h_V(v) = \frac{n^{2n-2}}{\{\Gamma(n-1)\}^2 v^n} \int_0^\infty t^{-1} \exp\left[-n\left\{\frac{t}{v} + \frac{1}{t}\right\}\right] dt, \quad v > 0.$$

The marginal cdfs  $U$  and  $V$  are, respectively,  $H_U(u) = \int_0^u h_U(a) da$  and  $H_V(v) = \int_0^v h_V(a) da$ . Now, we proceed as follows.

*Estimation of  $\psi$  when  $C$  is unknown*

The expected value of  $\psi^*$  is derived as

$$(2.33) \quad E\{\psi^*\} = E\{E\{\psi^*|\hat{\theta}\}\} = E\left\{\frac{n\hat{\psi}}{n-2}\right\} = \frac{n^2}{(n-2)^2} \psi.$$

Hence, an unbiased estimator  $\psi_u^*$  of  $\psi$  is  $\frac{(n-2)^2}{n^2} \psi^*$ . The variance of  $\psi_u^*$  is

$$(2.34) \quad V(\psi_u^*) = V(E\{\psi_u^*|\hat{\psi}\}) + E\{V(\psi_u^*|\hat{\psi})\} = \frac{(2n-5)}{(n-3)^2} \psi^2.$$

An estimate  $\widehat{V(\psi_u^*)}$  of  $V(\psi_u)$  is obtained by replacing  $\psi$  with  $\psi^*$  in (2.34). Mimicking steps in Section 2.1.1, a  $100(1-\gamma)\%$  CI for  $\psi$  has the following form:

$$(2.35) \quad \left(\frac{\psi^*}{\kappa_2}, \frac{\psi^*}{\kappa_1}\right),$$

where  $\kappa_1$  and  $\kappa_2$  are the roots of

$$(2.36) \quad \begin{aligned} \kappa_1^2 h_V(\kappa_1) - \kappa_2^2 h_V(\kappa_2) &= 0, \\ H_V(\kappa_2) - H_V(\kappa_1) - (1-\gamma) &= 0. \end{aligned}$$

*Estimation of  $C$  when  $\psi$  is unknown*

For  $C < \tilde{C}$ , we derive the marginal pdf of  $\tilde{C}$  from (2.28) as

$$(2.37) \quad \begin{aligned} q_{\tilde{C}}(\tilde{C}) &= \int_0^\infty h_{\tilde{C}}(\tilde{\theta}) d\psi^* \\ &= \frac{n^n C^{n\psi} \psi}{\Gamma(n-1)} \int_0^\infty \frac{\exp\{-n/t\}}{(t-1)t^{n-1}} \times \left[ \frac{\tilde{C}^{n\psi(t-1)} - C^{n\psi(t-1)}}{\tilde{C}^{n\psi t+1}} \right] dt. \end{aligned}$$

Note that  $U$  in (2.29) is not independent of  $\psi$ . Hence, in an effort to construct CI for  $C$ , we further take the transformation:

$$W^* = V \log U = \psi^* \log \frac{\tilde{C}}{C},$$

where the pdf of  $W^*$  is

$$(2.38) \quad h_{W^*}(w^*) = \frac{n^{(n-1)}(n-1)}{\Gamma(n-1)} \int_0^\infty \frac{\exp(-n/t)}{(t-1)} \left[ (t+w^*)^{-n} - (t(w^*+1))^{-n} \right] dt,$$

for  $w^* > 0$ . Therefore, a  $100(1-\gamma)\%$  CI for  $C$  is calculated using the following:

$$(2.39) \quad \left( \tilde{C} \exp\{-\kappa_2/\psi^*\}, \tilde{C} \exp\{-\kappa_1/\psi^*\} \right).$$

$\kappa_1$  and  $\kappa_2$  are calculated from  $\int_0^{\kappa_1} h_{W^*}(w^*) dw^* = \gamma/2$  and  $\int_{\kappa_2}^\infty h_{W^*}(w^*) dw^* = \gamma/2$ .

---

## 2.2. Posterior Predictive Sampling

---

This is the second method of sampling to draw synthetic data based on original data. Under a Bayesian setting, the synthetic data  $z = (z_1, \dots, z_n)^\top$  comes from the posterior predictive distribution of  $\theta$  given  $x$ . Here, we discuss the method of drawing inference on  $\psi$  when  $C$  is known.

---

### 2.2.1. Inference on $\psi$ when $C$ is known

---

We utilize the fact that the posterior distribution of  $\psi$  given  $U$  is Gamma with parameters  $(n + c_0, u + d)$  as given by [2]. Here,  $c_0 > 0$  and  $d > 0$  are the hyper parameters obtained using a Gamma prior with parameters  $c_0$  and  $d$ , and  $U = \sum_{i=1}^n \log(X_i/C)$ . Below we discuss the procedure for posterior predictive sampling:

**Step 1:** Draw  $\psi^*$  from the posterior distribution of  $\psi$  given  $u$ .

**Step 2:** Given value of  $\psi^*$  in Step 1, draw  $z = z_1, \dots, z_n$  as iid from the Pareto density

$$f_\theta(z_i) = \frac{\psi^* C^{\psi^*}}{z_i^{\psi^*+1}}.$$

For the purpose of analysis based on  $z$ , we develop the joint pdf of  $z$  in Theorem 2.4. In order to prove the theorem, the following three facts are used:

- $z_i | \psi^*$ ,  $i = 1, \dots, n$  are iid with each following Pareto distribution with parameters  $C$  and  $\psi^*$ ;
- $\psi^* | u$  follows Gamma distribution with parameters  $(n + c_0, u + d)$ ;
- $U | \psi$  is Gamma distribution with parameters  $(n, \psi)$ .

**Theorem 2.4.** *The joint pdf of  $z$  is given by*

$$(2.40) \quad f_\psi(z) = \frac{\psi^n}{\Gamma n \Gamma(n + c_0) (\prod_{i=1}^n z_i)} \times \int_0^\infty \left[ \int_0^\infty \psi^{*(2n+c_0-1)} \left\{ \frac{c^n}{(\prod_{i=1}^n z_i)} e^{-(u+d)} \right\}^{\psi^*} d\psi^* \right] u^{n-1} (u+d)^{n+c_0} e^{-u\psi} du,$$

where  $\psi > 0$  and  $z_i > C$ ,  $i = 1, \dots, n$ .

**Proof:** The above theorem can be proved by considering

$$f_\psi(z) = \int_0^\infty \int_0^\infty f(z|\psi^*) \times f(\psi^*|u) \times f(u|\psi) d\psi^* du,$$

where  $f$  denotes the corresponding pdfs as usual. □

Define  $\tilde{\psi} = \frac{n}{\sum \log(Z_i/C)}$  as an estimator of  $\psi$  and  $\tilde{\psi}|\psi^*$  follows IG distribution with parameters  $n$  and  $n\psi^*$ . The expected value of  $\tilde{\psi}$  is obtained as

$$\begin{aligned} E\{\tilde{\psi}\} &= E\{E\{\tilde{\psi}|\psi^*\}\} = E\left\{\frac{n}{(n-1)}\psi^*\right\} = \frac{n}{(n-1)} E\left\{\frac{n+c_0}{u+d}\right\} \\ &= \frac{n(n+c_0)\psi^n}{(n-1)\Gamma n} \int_0^\infty \frac{u^{n-1}e^{-u\psi}}{(u+d)} du = \frac{n(n+c_0)\psi^n}{(n-1)\Gamma n} M_1(\psi, n, d), \end{aligned}$$

where the term

$$M_1(\psi, n, d) = \int_0^\infty \frac{u^{n-1}e^{-u\psi}}{(u+d)} du.$$

Further, the variance of  $\tilde{\psi}$  is computed as follows:

$$V(\tilde{\psi}) = V(E\{\tilde{\psi}|\psi^*\}) + E\{V(\tilde{\psi}|\psi^*)\},$$

where

$$\begin{aligned} V(E\{\tilde{\psi}|\psi^*\}) &= \frac{n^2(n+c_0)(n+c_0+1)\psi^n}{(n-1)^2(n-2)\Gamma n} \int_0^\infty \frac{u^{n-1}\exp(-u\psi)}{(u+d)^2} du \\ &= \frac{n^2(n+c_0)(n+c_0+1)\psi^n}{(n-1)^2(n-2)\Gamma n} M_2(\psi, n, d), \end{aligned}$$

with

$$M_2(\psi, n, d) = \int_0^\infty \frac{u^{n-1}e^{-u\psi}}{(u+d)^2} du$$

and

$$E\{V(\tilde{\psi}|\psi^*)\} = \frac{n^2(n+c_0)\psi^n}{(n-1)^2\Gamma n} \left[ (n+c_0+1)M_2(\psi, n, d) - \frac{(n+c_0)}{\Gamma n} \psi^n M_1^2(\psi, n, d) \right].$$

Hence, we can express the variance of  $\tilde{\psi}$  as

$$(2.41) \quad V(\tilde{\psi}) = \frac{n^2(n+c_0)\psi^n}{(n-1)^2\Gamma n} \left[ \frac{(n-1)(n+c_0+1)}{(n-2)} M_2(\psi, n, d) - \frac{(n+c_0)}{\Gamma n} \psi^n M_1^2(\psi, n, d) \right].$$

#### Shortest confidence interval for $\psi$

Applying the same concept used in Theorem 2.4 and considering  $\tilde{\psi}|\psi^*$  follows IG distribution with parameters  $n$  and  $n\psi^*$ , the pdf of  $\tilde{\psi}$  is given by

$$(2.42) \quad f_{\tilde{\psi}}(\tilde{\psi}) = \frac{n^n \psi^n \Gamma(2n+c_0)}{(\Gamma n)^2 \Gamma(n+c_0) \tilde{\psi}^{n+1}} \int_0^\infty \frac{u^{n-1} e^{-u\psi} (u+d)^{n+c_0}}{(n/\tilde{\psi} + u + d)^{2n+c_0}} du.$$



For computational convenience, we consider  $d = 0$  which leads to the prior density of the parameter  $\psi$  to be a Jeffreys prior. However, the posterior density of  $\psi$  still follows a Gamma distribution. For a detailed discussion, refer to Section 2.1 in [2]. Henceforth, we assign  $d = 0$ . For  $\omega > 0$ , considering the transformations  $t = u \left[ \frac{n}{\psi} + u \right]^{-1}$  and  $\omega = \frac{\tilde{\psi}}{\psi}$  sequentially in (2.42), we get the pdf

$$(2.43) \quad f_W(\omega) = \frac{n^n \Gamma(2n + c_0)}{(\Gamma n)^2 \Gamma(n + c_0) \omega^{n+1}} \int_0^1 \frac{t^{2n+c_0-1} \exp\left\{-\frac{1}{\omega} \left[\frac{n-t}{1-t}\right]\right\}}{t-1} dt$$

independent of  $\psi$ . Hence,  $\omega$  is a pivotal quantity and the shortest distance  $(1 - \gamma)$  100% CI for  $\psi$  is

$$(2.44) \quad \left( \tilde{\psi} \omega_2^{-1}, \tilde{\psi} \omega_1^{-1} \right),$$

where  $\omega_1$  and  $\omega_2$  are obtained by solving

$$(2.45) \quad \begin{aligned} \omega_1^2 f_W(\omega_1) - \omega_2^2 f_W(\omega_2) &= 0, \\ F_W(\omega_2) - F_W(\omega_1) - (1 - \gamma) &= 0, \end{aligned}$$

and  $F_W(\omega) = \int_0^\omega f_W(u) du$ . The discussion on constructing the shortest CI can be found in Section 2.1.1.

**Remark:** In practice, it is unrealistic to assume that the shape parameter  $\psi$  is known and  $C$  is not known. Once we have data then the minimum value in the data is sufficient for  $C$ . As per [2] the posterior distribution of  $C$  given the original data is a power function distribution with two hyper parameters namely  $\delta \geq 0$  and  $\sigma_0 > 0$ . One of the parameters of the posterior distribution of  $C$  depends on  $\min\{\sigma_0, x_{(1)}\}$ . While computing the unconditional pdf of  $\tilde{C}$ , an explicit expression could not be obtained since the integrals involved in the derivation often have limits depending on the original data  $x$ . Hence we do not discuss this case here. On the other hand, the case of joint posterior distribution when both parameters are unknown, becomes extremely complex due to the same issue, and hence, it is not discussed either in this paper.

---

### 3. SIMULATION STUDY AND RESULTS

---

To study the performance of the proposed estimation methods, we carry out an extensive simulation study. For all scenarios, viz., only  $\psi$  unknown (Scenario 1), only  $C$  unknown (Scenario 2), both  $C$  and  $\psi$  unknown (Scenario 3) in case of plug-in sampling, and only  $\psi$  unknown in case of posterior predictive sampling (Scenario 4), few candidate true values of  $C$  and  $\psi$  are chosen. True values of  $C$  are taken as 1 and 100, while true values of  $\psi$  are selected to be 1.5 and 3. To study the effect of smaller and larger sample sizes on estimation,  $n = 50$  and  $n = 100$  are considered. Under these parameter settings, we examine the performance and robustness of our estimation methods with respect to singly imputed synthetic data based on one thousand Monte-Carlo simulation runs. `Mathematica 12.2` and `R-4.0.1` [24] software packages are employed for coding.

For these settings, parameter estimate (EST), empirical standard error (ESE), average model based standard error (ASE), average bias of the estimator (BIAS), average root mean squared error (RMSE), and coverage rate (CR) of 90% and 95% nominal level are provided in Tables 1–6. Based on simulation results, estimates are found to be accurate with low bias and low standard errors in all cases. As one would expect, increasing sample size results in more precise estimates with improved coverage probabilities, and with noticeable reduction in BIAS, ASE and RMSE. Estimates are less precise for estimating  $C$  when  $\psi$  is unknown, and for estimating  $\psi$  when  $C$  is unknown than their corresponding known counterparts.

This can be attributed to the fact that estimating associated parameter instead of using their known values introduces more variability to the data, resulting in less accuracy in estimation of the primary parameter. ASE and RMSE obtained for estimating  $C$  are high when the true value of  $C = 100$  than when the true value of  $C = 1$ . A similar trend is observed for estimating  $\psi$  as well; ASE and RMSE are high when true  $\psi = 3$  as compared to the case when true  $\psi = 1.5$ .

The coverage rates are mostly close to the nominal level throughout all scenarios, further suggesting the estimation method is robust and the estimates are accurate. More specifically, CRs corresponding to  $C$  behave quite well for both cases when  $\psi$  is known or unknown. However, though rare, there are some instances of slight under-coverage for  $\psi$  when employing our estimation method, specifically when  $C$  is unknown (see Table 4). A probable reason can be the mathematical dependence of the estimator of  $\psi$  on  $C$  (known or unknown). But, we would like emphasize that this under-coverage reduces as the sample size increases, validating that for large enough sample size confidence intervals provided by our estimation method are quite precise and reliable.

In Tables 5 and 6, we list the estimation results on  $\psi$  when  $C$  is known under posterior predictive sampling. Throughout, we assign  $d = 0$  that results in unbiased estimates of  $\psi$ . Simulation results corresponding to  $c_0 = 0$  and  $c_0 = 1$  are presented in Tables 5 and 6, respectively. The bias in the estimates are of the order of  $10^{-2}$  and coverage rates are close to the specified values of confidence level. Impact of increase in sample size can be seen in the reduction of BIAS and RMSE.

**Table 1:** EST, ESE, ASE, BIAS, RMSE and CR for  $C$  when  $\psi$  is known.

$C$	$\psi$	$n$	$1 - \gamma$	EST	ESE	ASE	BIAS	RMSE	CR
1	1.5	50	0.90	1.001	0.020	0.019	-0.001	0.020	0.893
			0.95	1.000	0.019	0.019	0.000	0.019	0.948
		100	0.90	0.999	0.008	0.009	0.001	0.009	0.911
			0.95	1.001	0.009	0.009	-0.001	0.010	0.948
	3.0	50	0.90	1.000	0.009	0.009	0.000	0.009	0.892
			0.95	1.000	0.009	0.009	0.000	0.009	0.946
		100	0.90	1.000	0.004	0.004	0.000	0.005	0.899
			0.95	1.000	0.004	0.004	0.000	0.004	0.948
100	1.5	50	0.90	99.950	1.818	1.962	0.050	2.673	0.913
			0.95	100.011	1.948	1.963	-0.011	2.764	0.948
		100	0.90	99.950	0.917	0.961	0.050	1.326	0.904
			0.95	100.011	0.982	0.962	-0.011	1.371	0.952
	3.0	50	0.90	100.044	0.968	0.962	-0.044	1.363	0.892
			0.95	99.992	0.941	0.961	0.008	1.345	0.949
		100	0.90	100.020	0.501	0.476	-0.020	0.685	0.892
			0.95	100.009	0.475	0.476	-0.009	0.670	0.951

**Table 2:** EST, ESE, ASE, BIAS, RMSE and CR for  $C$  when  $\psi$  is unknown.

$C$	$\psi$	$n$	$1 - \gamma$	EST	ESE	ASE	BIAS	RMSE	CR
1	1.5	50	0.90	1.001	0.018	0.019	0.001	0.026	0.909
			0.95	1.002	0.019	0.019	0.002	0.027	0.954
		100	0.90	1.000	0.010	0.009	0.000	0.014	0.897
			0.95	1.000	0.009	0.009	0.000	0.013	0.951
	3.0	50	0.90	1.000	0.009	0.009	0.000	0.013	0.902
			0.95	1.000	0.009	0.009	0.000	0.013	0.951
		100	0.90	1.000	0.005	0.005	0.000	0.007	0.904
			0.95	1.000	0.005	0.005	0.000	0.007	0.952
100	1.5	50	0.90	100.096	1.963	1.888	0.096	2.753	0.904
			0.95	100.026	1.886	1.883	0.026	2.692	0.951
		100	0.90	100.034	0.958	0.943	0.034	1.350	0.910
			0.95	100.008	0.960	0.938	0.008	1.349	0.950
	3.0	50	0.90	100.010	0.900	0.922	0.010	1.301	0.910
			0.95	100.078	0.956	0.919	0.079	1.341	0.949
		100	0.90	100.011	0.486	0.468	0.011	0.678	0.901
			0.95	100.015	0.480	0.467	0.015	0.673	0.947

**Table 3:** EST, ESE, ASE, BIAS, RMSE and CR for  $\psi$  when  $C$  is known.

$C$	$\psi$	$n$	$1 - \gamma$	EST	ESE	ASE	BIAS	RMSE	CR	
1	1.5	50	0.90	1.486	0.291	0.305	0.014	0.426	0.857	
			0.95	1.494	0.309	0.307	0.006	0.440	0.953	
	100	50	0.90	1.497	0.212	0.214	0.003	0.303	0.917	
			0.95	1.494	0.217	0.214	0.006	0.306	0.966	
	3.0	50	50	0.90	3.020	0.604	0.620	-0.020	0.874	0.860
				0.95	3.009	0.610	0.617	-0.009	0.877	0.952
100	3.0	100	0.90	2.999	0.427	0.429	0.001	0.609	0.910	
			0.95	3.018	0.439	0.432	-0.018	0.619	0.965	
100	1.5	50	0.90	1.488	0.285	0.305	0.012	0.422	0.868	
			0.95	1.482	0.298	0.304	0.018	0.430	0.953	
	100	50	0.90	1.498	0.215	0.215	0.002	0.306	0.910	
			0.95	1.497	0.227	0.214	0.003	0.314	0.963	
	3.0	50	50	0.90	2.988	0.573	0.613	0.012	0.847	0.864
				0.95	3.011	0.596	0.618	-0.011	0.867	0.958
	100	3.0	100	0.90	2.988	0.427	0.428	0.012	0.608	0.903
				0.95	3.019	0.448	0.432	-0.019	0.626	0.968

**Table 4:** EST, ESE, ASE, BIAS, RMSE and CR for  $\psi$  when  $C$  is unknown.

$C$	$\psi$	$n$	$1 - \gamma$	EST	ESE	ASE	BIAS	RMSE	CR	
1	1.5	50	0.90	1.494	0.307	0.310	0.006	0.440	0.852	
			0.95	1.506	0.313	0.312	-0.006	0.447	0.955	
	100	50	0.90	1.490	0.202	0.214	0.010	0.296	0.913	
			0.95	1.505	0.231	0.217	-0.005	0.318	0.952	
	3.0	50	50	0.90	2.997	0.587	0.621	0.003	0.863	0.859
				0.95	3.027	0.656	0.628	-0.027	0.918	0.949
100	3.0	100	0.90	2.994	0.430	0.431	0.006	0.612	0.907	
			0.95	3.013	0.450	0.434	-0.013	0.628	0.959	
100	1.5	50	0.90	1.500	0.289	0.311	0.000	0.429	0.860	
			0.95	1.498	0.306	0.311	0.002	0.441	0.953	
	100	50	0.90	1.508	0.216	0.217	-0.008	0.308	0.899	
			0.95	1.493	0.217	0.215	0.007	0.307	0.958	
	3.0	50	50	0.90	2.996	0.617	0.621	0.004	0.884	0.855
				0.95	3.004	0.615	0.623	-0.004	0.884	0.951
	100	3.0	100	0.90	2.983	0.421	0.429	0.017	0.605	0.900
				0.95	2.986	0.446	0.430	0.014	0.623	0.956

**Table 5:** Inference for  $\psi$  when  $C$  is known, under Bayesian predictive sampling with hyper parametric values  $d = 0$  and  $c_0 = 0$ .

$C$	$\psi$	$n$	$1 - \gamma$	UEST	ESE	ASE	BIAS	RMSE	CR
1	1.5	50	0.90	1.540	0.379	0.409	0.040	0.345	0.905
			0.95	1.544	0.394	0.410	0.044	0.359	0.947
		100	0.90	1.513	0.266	0.273	0.013	0.152	0.895
			0.95	1.521	0.265	0.274	0.021	0.153	0.952
	3.0	50	0.90	3.086	0.794	0.820	0.086	1.445	0.908
			0.95	3.048	0.755	0.810	0.048	1.356	0.959
		100	0.90	3.000	0.515	0.541	0.000	0.585	0.907
			0.95	3.035	0.522	0.547	0.035	0.601	0.963
100	1.5	50	0.90	1.554	0.385	0.413	0.054	0.355	0.926
			0.95	1.549	0.401	0.411	0.049	0.367	0.946
		100	0.90	1.513	0.257	0.273	0.013	0.147	0.908
			0.95	1.494	0.261	0.269	-0.006	0.148	0.951
	3.0	50	0.90	3.061	0.807	0.813	0.061	1.368	0.920
			0.95	3.116	0.809	0.827	0.116	1.441	0.955
		100	0.90	3.053	0.546	0.550	0.053	0.646	0.891
			0.95	2.997	0.537	0.540	-0.003	0.593	0.951

**Table 6:** Inference for  $\psi$  when  $C$  is known, under Bayesian predictive sampling with hyper parametric values  $d = 0$  and  $c_0 = 1$ .

$C$	$\psi$	$n$	$1 - \gamma$	UEST	ESE	ASE	BIAS	RMSE	CR
1	1.5	50	0.90	1.535	0.367	0.408	0.035	0.334	0.918
			0.95	1.543	0.387	0.410	0.043	0.353	0.953
		100	0.90	1.514	0.267	0.273	0.014	0.153	0.909
			0.95	1.524	0.264	0.275	0.024	0.153	0.955
	3.0	50	0.90	3.062	0.755	0.813	0.062	1.365	0.904
			0.95	3.060	0.766	0.813	0.060	1.382	0.950
		100	0.90	3.032	0.522	0.547	0.032	0.601	0.904
			0.95	3.021	0.531	0.545	0.021	0.607	0.949
100	1.5	50	0.90	1.521	0.369	0.404	0.021	0.331	0.921
			0.95	1.537	0.379	0.408	0.037	0.344	0.960
		100	0.90	1.521	0.268	0.274	0.021	0.154	0.904
			0.95	1.518	0.278	0.274	0.018	0.160	0.953
	3.0	50	0.90	3.065	0.775	0.814	0.065	1.368	0.909
			0.95	3.055	0.774	0.811	0.055	1.449	0.947
		100	0.90	3.046	0.539	0.549	0.046	0.647	0.898
			0.95	3.013	0.538	0.543	0.013	0.601	0.948

---

#### 4. CONCLUDING REMARKS

---

In this paper, we have derived likelihood based methods of inference for synthetic data when the original data comes from a two parameter Pareto model. To this end, synthetic data were generated by two different methods, viz., plug-in sampling and posterior predictive sampling. For the plug-in sampling method, we have developed unbiased estimators for the parameters, and obtained the expressions of the corresponding variances and shortest distance CIs under three possible scenarios (inference on  $\psi$  when  $C$  is known, inference on  $C$  when  $\psi$  is known and inference on  $\theta$  when both parameters are unknown). On the other hand, under posterior predictive sampling, inference has been drawn only for the shape parameter  $\psi$  when  $C$  is known. The methods have been discussed based on a single synthetic data set.

Results from the simulation study have shown that the plug-in sampling exhibits less bias, ASE and RMSE than posterior predictive sampling. A similar observation has been reported by [15] for a synthetic data from exponential distribution.

The developed estimators are unbiased in nature, and have been developed based on sufficient statistics. Exact shortest distance confidence intervals for parameters have been constructed for all methods of sampling, except for  $C$  when  $\psi$  is unknown in plug-in sampling. The primary strength of these methods is that they are based on a single synthetic data set, which is advantageous when release of multiple data sets is not allowed due to privacy concerns.

Despite observing actual microlevel data, the methodologies developed in this paper would allow researchers and policy makers to gain insights into the extent of financial burden tax payers face by filing income tax, or distribution of income or wealth across various strata of the society. The mathematical expressions provided in the paper would enable them to estimate the key parameters of the distribution relatively accurately, thereby, necessitating appropriate economic policy changes, or identifying gaps in a financial program or strategy. Computations of confidence intervals may require evaluating implicit integrals or solving non-linear simultaneous equations. However, these can be carried out easily by any established statistical software. It is recommended that users should carry out proper hypothesis test to verify whether a Pareto model fits data for a particular location or period. For researchers in government agencies, who have the access to actual data, could verify the precision of our estimates, and assess merits in our techniques. Future work may include developing estimation procedures in case of posterior predictive sampling with different informative priors.

---

#### ACKNOWLEDGMENTS

---

This work has not been supported by any grant or external funding source. We acknowledge the valuable suggestions provided by the referees.

---

**REFERENCES**

---

- [1] ARNOLD, B.C. (2015). *Pareto Distributions*, Chapman and Hall/CRC, Boca Raton, FL, USA.
- [2] ARNOLD, B.C. and PRESS, S.J. (1983). Bayesian inference for Pareto populations, *Journal of Econometrics*, **21**(3), 287–306.
- [3] BOWEN, C.M.; BRYANT, V.; BURMAN, L.; KHITATRAKUN, S.; MCCLELLAND, R.; STALLWORTH, P.; UHEYAMA, K. and WILLIAMS, A.R. (2020, September). *A synthetic supplemental public use file of low-income information return data: methodology, utility, and privacy implications*. In *International Conference on Privacy in Statistical Databases*, Springer, Cham, 257–270.
- [4] CASELLA, G. and BERGER, R.L. (2002). *Statistical Inference*, 2nd ed., Duxbury Pacific Grove, CA, USA.
- [5] DALENIUS, T. and REISS, S.P. (1982). Data-swapping: a technique for disclosure control, *Journal of Statistical Planning and Inference*, **6**(1), 73–85.
- [6] DRECHSLER, J. (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, Vol. 201, Springer Science & Business Media.
- [7] EVANS, T.; ZAYATZ, L. and SLANTA, J. (1996). Using noise for disclosure limitation of establishment tabular data, *Proceedings of the Annual Research Conference, US Bureau of the Census, Washington, DC*, **20233**(4), 65–86.
- [8] HAGSTROEM, K.G. (1960). Remarks on Pareto distributions, *Scandinavian Actuarial Journal*, **1**, 59–71.
- [9] KELLY, J.P.; GOLDEN, B.L. and ASSAD, A.A. (1992). Cell suppression: disclosure protection for sensitive tabular data, *Journal of Political Economy*, **22**(4), 397–417.
- [10] KIM, J.J. (1986). *A method for limiting disclosure in microdata based on random noise and transformation*. In “Proceedings of the American Statistical Association, Section on Survey Research Methods”, Alexandria, VA, American Statistical Association, 370–374.
- [11] KIM, J.J. and WINKLER, W. (1995). *Masking microdata files*. In “Proceedings of the American Statistical Association, Section on Survey Research Methods”, Alexandria, VA, American Statistical Association, 114–119.
- [12] KIM, J. and WINKLER, W. (2003). Multiplicative noise for masking continuous data, *Statistical Research Division, Research Report Series, U.S. Census Bureau*.
- [13] KINNEY, S.K.; REITER, J.P.; REZNEK, A.P.; MIRANDA, J.; JARMIN, R.S. and ABOWD, J.M. (2011). Towards unrestricted public use business microdata: the synthetic longitudinal business database, *International Statistical Review*, **79**(3), 362–384.
- [14] KLEIN, M.; MATHEW, T. and SINHA, B. (2014). Likelihood based inference under noise multiplication, *Thailand Statistician: Journal of the Thai Statistical Association*, **12**, 1–23.
- [15] KLEIN, M. and SINHA, B. (2015a). Likelihood-based finite sample inference for synthetic data based on exponential model, *Thailand Statistician*, **13**(1), 33–47.
- [16] KLEIN, M. and SINHA, B. (2015b). Inference for singly imputed synthetic data based on posterior predictive sampling under multivariate normal and multiple linear regression models, *Sankhya B*, **77**(2), 293–311.
- [17] KLEIN, M. and SINHA, B. (2015c). Likelihood-based inference for singly and multiply imputed synthetic data under a normal model, *Statistics & Probability Letters*, **105**, 168–175.
- [18] LEHMANN, L. and CASELLA, G. (2006). *Theory of Point Estimation*, Springer Science & Business Media.

- [19] LITTLE, R.J.A. (1993). Statistical analysis of masked data, *Journal of Official Statistics*, **9**(2), 407–426.
- [20] MALIK, H.J. (1970). Estimation of the parameters of the Pareto distribution, *Metrika*, **15**(1), 126–132.
- [21] MANDELBROT, B. (1963). New methods in statistical economics, *Journal of Political Economy*, **71**(5), 421–440.
- [22] NAYAK, T.K.; SINHA, B. and ZAYATZ, L. (2011). Statistical properties of multiplicative noise masking for confidentiality protection, *Journal of Official Statistics*, **27**(3), 527–544.
- [23] NEWEY, W.K. and MCFADDEN, D. (1994). Large sample estimation and hypothesis testing, *Handbook of Econometrics*, **4**, 2111–2245.
- [24] R CORE TEAM (2020). *R: a language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- [25] RAGHUNATHAN, T.E.; REITER, J.P. and RUBIN, D.B. (2003). Multiple imputation for statistical disclosure limitation, *Journal of Official Statistics*, **19**, 1–16.
- [26] REITER, J.P. (2003). Inference for partially synthetic, public use microdata sets, *Survey Methodology*, **29**(2), 181–188.
- [27] REITER, J.P. (2004). Releasing multiply imputed, synthetic public-use microdata: an illustration and empirical, *Journal of the Royal Statistical Society, Series A*, **168**, 185–205.
- [28] REITER, J.P. and KINNEY, S.K. (2012). Inferentially valid, partially synthetic data: generating from posterior predictive distributions not necessary, *Journal of Official Statistics*, **28**(4), 583–590.
- [29] REITER, J.P. and RAGHUNATHAN, T.E. (2007). The multiple adaptations of multiple imputation, *Journal of the American Statistical Association*, **102**, 1462–1471.
- [30] RUBIN, D.B. (1993). Statistical disclosure limitation, *Journal of Official Statistics*, **9**(2), 461–468.
- [31] RUBIN, D.B. (2004). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, Hoboken, NJ, USA.
- [32] SAKSENA, S.K. and JOHNSON, A.M. (1984). Best unbiased estimators for the parameters of a two-parameter Pareto distribution, *Metrika*, **31**(1), 77–83.
- [33] SINHA, B.; NAYAK, T.K. and ZAYATZ, L. (2011). Privacy protection and quantile estimation from noise multiplied data, *Sankhya B*, **73**(2), 297–315.
- [34] SOUMA, W. (1970). Universal structure of the personal income distribution, *Fractals*, **9**(4), 463–470.
- [35] WILLENBORG, L. and DE WAAL, T. (2012). *Elements of Statistical Disclosure Control*, Springer Science & Business Media, NY, USA.
- [36] WOLFRAM RESEARCH, INC. (2020). *Mathematica*, Version 12.2, Champaign, IL.



# 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](http://revstat.ine.pt) based in Open Journal System (OJS). The only working language allowed is English. Authors intending to submit any work must register, login and follow the guidelines.

There are no fees for publishing accepted manuscripts that will be made available in open access.

All articles consistent with REVSTAT aims and scope will undergo scientific evaluation by at least two reviewers, one from the Editorial Board and another external. Authors can suggest an editor or reviewer who is expert on the paper subject providing her/his complete information, namely: name, affiliation, email and, if possible, personal URL or ORCID number.

All published works are Open Access (CC BY 4.0) which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. Also, in the context of archiving policy, REVSTAT is a *blue* journal welcoming authors to deposit their works in other scientific repositories regarding the use of the published edition and providing its source.

Journal prints may be ordered at expenses of the author(s), and prior to publication.

### Abstract and Indexing services

REVSTAT-Statistical Journal is covered by *Journal Citation Reports - JCR (Clarivate)*; *DOAJ-Directory Of Open Access Journals*; *Current Index to Statistics*; *Google Scholar*; *Mathematical Reviews® (MathSciNet®)*; *Zentralblatt für Mathematic*; *Scimago Journal & Country Rank*; *Scopus*

### Author guidelines

The whole submission and review processes for REVSTAT are conducted exclusively online on the journal's webpage <https://revstat.ine.pt/> based in Open Journal System (OJS). Authors intending to submit any work must *register*, *login* and follow the indications choosing *Submissions*.

REVSTAT - Statistical Journal adopts the COPE guidelines on publication ethics.

### Work presentation

- the only working language is English;
- the first page should include the name, ORCID iD (optional), Institution, country, and mail-address of the author(s);
- a summary of fewer than one hundred words, followed by a maximum of six keywords and the MSC 2020 subject classification should be included also in the first page;
- manuscripts should be typed only in black, in double-spacing, with a left margin of at least 3 cm, with numbered lines, and a maximum of 25 pages;
- the title should be with no more than 120 characters (with spaces);
- figures must be a minimum of 300dpi and will be reproduced online as in the original work, however, authors should take into account that the printed version is always in black and grey tones;
- authors are encouraged to submit articles using LaTeX which macros are available at *REVSTAT style*;
- citations in text should be included in the text by name and year in parentheses, as in the following examples: § article title in lowercase (Author 1980); § This

theorem was proved later by AuthorB and AuthorC (1990); § This subject has been widely addressed (AuthorA 1990; AuthorB et al. 1995; AuthorA and AuthorB 1998).

- references should be listed in alphabetical order of the author's scientific surname at the end of the article;
- acknowledgments of people, grants or funds should be placed in a short section before the References title page. Note that religious beliefs, ethnic background, citizenship and political orientations of the author(s) are not allowed in the text;
- authors are welcome to suggest one of the Editors or Associate Editors or yet other reviewer expert on the subject providing a complete information, namely: name, affiliation, email and personal URL or ORCID number in the Comments for the Editor (submission form).

### Accepted papers

After final revision and acceptance of an article for publication, authors are requested to provide the corresponding LaTeX file, as in REVSTAT style.

Supplementary files may be included and submitted separately in .tiff, .gif, .jpg, .png .eps, .ps or .pdf format. These supplementary files may be published online along with an article, containing data, programming code, extra figures, or extra proofs, etc; however, REVSTAT is not responsible for any supporting information supplied by the author(s).

### Copyright Notice

Upon acceptance of an article, the author(s) will be asked to transfer copyright of the article to the publisher, Statistics Portugal, in order to ensure the widest possible dissemination of information.

According to REVSTAT's *archiving policy*, after assigning the copyright form, authors may cite and use limited excerpts (figures, tables, etc.) of their works accepted/published in REVSTAT in other publications and may deposit only the published edition in scientific repositories providing its source as REVSTAT while the original place of publication. The Executive Editor of the Journal must be notified in writing in advance.

## EDITORIAL BOARD 2019-2023

### Editor-in-Chief

Isabel FRAGA ALVES, University of Lisbon, Portugal

### Co-Editor

Giovani L. SILVA, University of Lisbon, Portugal

### Associate Editors

Marília ANTUNES, University of Lisbon, Portugal

Barry ARNOLD, University of California, USA

Narayanaswamy BALAKRISHNAN, McMaster University, Canada

Jan BEIRLANT, Katholieke Universiteit Leuven, Belgium

Graciela BOENTE, University of Buenos Aires, Argentina

Paula BRITO, University of Porto, Portugal

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

Alan GELFAND, Duke University, USA

Stéphane GIRARD, Inria Grenoble Rhône-Alpes, France

Marie KRATZ, ESSEC Business School, France

Victor LEIVA, Pontificia Universidad Católica de Valparaíso, Chile

Artur LEMONTE, Federal University of Rio Grande do Norte, Brazil

Shuangzhe LIU, University of Canberra, Australia

Maria Nazaré MENDES-LOPES, University of Coimbra, Portugal

Fernando MOURA, Federal University of Rio de Janeiro, Brazil

John NOLAN, American University, USA

Paulo Eduardo OLIVEIRA, University of Coimbra, Portugal

Pedro OLIVEIRA, University of Porto, Portugal

Carlos Daniel PAULINO, University of Lisbon, Portugal

Arthur PEWSEY, University of Extremadura, Spain

Gilbert SAPORTA, Conservatoire National des Arts et Métiers, France

Alexandra M. SCHMIDT, McGill University, Canada

Manuel SCOTTO, University of Lisbon, Portugal

Lisete SOUSA, University of Lisbon, Portugal

Milan STEHLÍK, University of Valparaíso, Chile and LIT-JK University Linz, Austria

María Dolores UGARTE, Public University of Navarre, Spain

### Executive Editor

José A. PINTO MARTINS, Statistics Portugal

### Assistant Editor

Olga BESSA MENDES, Statistics Portugal