CHILEAN JOURNAL OF **STATISTICS** Volume 11 Number 2 - December 2020

www.soche.cl/chjs

Volume 11 Number 2 - December 2020

CHILEAN JOURNAL OF **STATISTICS**

CHILEAN JOURNAL OF STATISTICS

Edited by Víctor Leiva and Carolina Marchant

A free open access journal indexed by

Volume 11 Number 2
December 2020
ISSN: 0718-7912 (print)
ISSN: 0718-7920 (online)

Published by the
Chilean Statistical Society
SOCHE
SOCIEDAD CHILENA DE ESTADÍSTICA

Aims

The Chilean Journal of Statistics (ChJS) is an official publication of the Chilean Statistical Society (`www.soche.cl`). The ChJS takes the place of *Revista de la Sociedad Chilena de Estadística*, which was published from 1984 to 2000.

The ChJS covers a broad range of topics in statistics, as well as in artificial intelligence, big data, data science, and machine learning, focused mainly on research articles. However, review, survey, and teaching papers, as well as material for statistical discussion, could be also published exceptionally. Each paper published in the ChJS must consider, in addition to its theoretical and/or methodological novelty, simulations for validating its novel theoretical and/or methodological proposal, as well as an illustration/application with real data.

The ChJS editorial board plans to publish one volume per year, with two issues in each volume. On some occasions, certain events or topics may be published in one or more special issues prepared by a guest editor.

Chilean Journal of Statistics

Volume 11, Number 2
December 2020

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

Contents

Chilean Journal of Statistics
Vol. 11, No. 2, December 2020, 69–72

Eleventh Volume – Second Number
Editorial Paper

Confirming our international presence with publications and submissions from all continents in COVID-19 pandemic

Víctor Leiva[1] and Carolina Marchant[2]

[1]School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
[2]Faculty of Basic Sciences, Universidad Católica del Maule, Talca, Chile

Welcome to the second issue of the eleventh volume of the Chilean Journal of Statistics (ChJS), published on 30 December 2020. The ChJS celebrates eleven years of life during very testing times due to the coronavirus pandemic. The new coronavirus has affected the whole world in different ways, which was first identified in November 2019 in Wuhan, China. This is an ongoing pandemic of coronavirus disease caused by severe acute respiratory syndrome named coronavirus 2, also known as SARS-CoV-2 or COVID-19. Such a pandemic is also a key period for statistics, because its use has allowed governments around the planet to establish regulations aimed at stopping the spread of the virus. The World Health Organization declared the outbreak an international public health emergency in January 2020 and a pandemic in March 2020. Today, more than 82.1 million of cases have been confirmed and more than 1.79 million of deaths have been attributed to COVID-19 around the world. Medicine, science, statistics, and generation of new knowledge have played a fundamental role during this year, with scientific journals playing a relevant role as they publish research of quality. We believe the world will overcome this situation, but we are sure new customs acquired during this period, such as interconnectivity, telework, teleconferencing, virtuality, among others, will remain forever. Although their development was initiated much before this pandemic, areas of big data, data science, machine learning, and statistics have been emphasized in 2020. All of these areas play an important role not only in artificial intelligence, science, and engineering, but in practically all areas of knowledge.

The scientific and editorial production of this volume of the ChJS would not have been achieved without the valuable contributions of many people. Renowned international researchers from the all five continents have honored us by publishing their interesting works in our journal so that we thank their contributions. In addition, we also thank all the anonymous reviewers who have contributed to keeping the high quality standards of the ChJS. Furthermore, our editorial board is grown from this issue. Welcome on board Dr. Yolanda Gómez from Chile, Dr. Diego Gallardo from Chile, and Dr. Danilo Alvares from Brazil/Chile. We are sure that with their enthusiasm, dynamism and talent, our journal will benefit greatly. We are honored by their acceptance to be part of the ChJS. Moreover, we are obliged and pleased to thank our prestigious editorial board, composed of colleagues from all five continents and listed in http://chjs.mat.utfsm.cl/board.html, who have collaborated from different perspective to increase our visibility and quality of the papers published by the ChJS. Of course, we must also thank the President and the Board of Directors of the Chilean Statistics Society (listed in https://soche.cl/quienes-somos) and the entire Chilean statistical community for placing in us, the Editors-in-Chief of the ChJS, their trust in our work.

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

The second issue of the eleventh volume of the ChJS comprises six papers authored by researchers from Algeria, Brazil, Chile, Colombia, India, Netherlands, Saudi Arabia, and United Kingdom (UK):

(i) Our first paper is authored by Ibrahim M. Almanjahie, Mohammed Kadi Attouch, Omar Fetitah, and Hayat Louhab, who derived a family of robust nonparametric estimators for a regression function with unknown scale parameter based on the kernel density method and functional data. The authors accompanied their theoretical results with a simulation study to evaluate the good performance of their proposal, and applied it to real functional ergodic data concerning air pollution.

(ii) The second paper is authored by Ricardo Puziol de Oliveira, Marcos Vinicius de Oliveira Peres, Jorge Alberto Achcar, and Nasser Davarzani, who developed a trivariate Marshall-Olkin-Weibull distribution for the modeling of right-censored data. In addition, the authors performed simulations to assess the performance of the corresponding model estimators, and applied their results to real data from industrial engineering.

(iii) In the third paper, Henrique José de Paula Alves and Daniel Furtado Ferreira introduced a new alternative test to the Hotelling T2 and likelihood ratio methods for multivariate normal and non-normal population mean vectors. In this paper, the authors accompanied their methodological findings evaluating the performance of the proposed new tests by Monte Carlo simulations, calculating the type I error probabilities and power of the tests. Addition, an illustration with real data of contents of sand and clay from Amazon, Brazil, was provided to show potential applications.

(iv) The fourth paper is authored by Josmar Mazucheli, André F.B. Menezes, Sanku Dey, and Saralees Nadarajah, who developed and characterized a reparametrize Chaudhry-Ahmad distribution. In this paper, the authors accompanied their theoretical results with a Monte Carlo simulation study to compare the proposed estimators and their bias-corrected versions, obtained from the Cox-Snell formula and parametric bootstrapping. This paper closes with an application by using wind speed data from Brazil.

(v) The fifth paper is authored by André Leite, Abel Borges, Geiza Silva, and Raydonal Ospina, who presented an integer programming formulation to handle a real instance of a courses-to-lecturers timetabling problem based on a case study with real data from Brazil.

(vi) This issue of the ChJS closes with a paper authored by Jorge Figueroa-Zúñiga, Rodrigo Sanhueza-Parkes, Bernardo Lagos-Álvarez, and Germán Ibacache-Pulgar, who presented and characterized an ingenious distribution to model bounded data with equal and unequal proportions of cases at the tails. The authors named this the trapezoidal Kumaraswamy distribution and applied it to data sets from education and engineering.

An aspect of which we are proud at the Chilean Statistical Society and ChJS is that we are and will continue to be an open access journal, publishing works free of any article processing charges (APC). In addition, we are working on implementing an online platform for submissions, which should facilitate the work of Editors and Reviewers, as well as the submission process for authors. Moreover, we are pleased and happy to count with a new indexation for the ChJS corresponding to the Scopus system of Elsevier, which joins the emerging sources citation index published by the Clarivate Analytics from the Institute for Scientific Information (ISI) belonging to Web of Science Group. We feel very motivated because this year we received 51 submissions, coming from 30 countries (of all five continents) and 132 authors. We detail the information generated from these data on countries of authors and submissions in Figures 1 and 2.

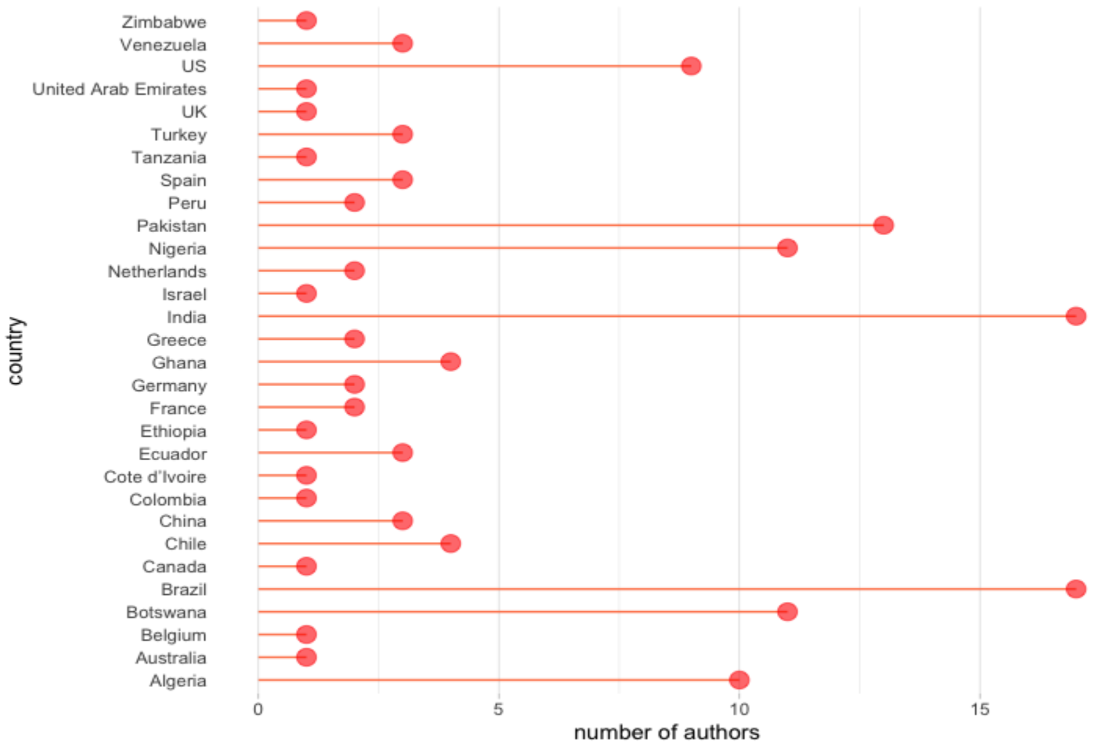


Figure 1. Distribution of authors of the articles submitted to the ChJS during 2020 with respect to their country.

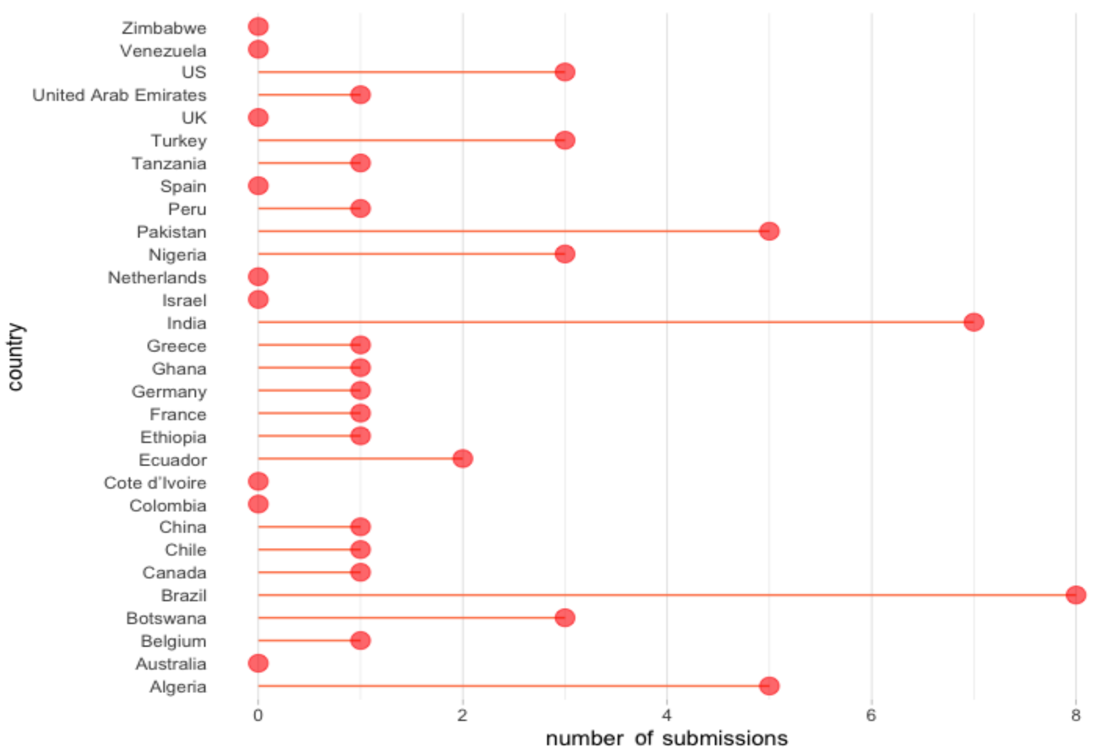


Figure 2. Distribution of papers submitted to the ChJS during 2020 according to the country of the corresponding author.

Finally, we would like the statistics and data science communities around the world, our Editorial Board, and authors who have published with us, to champion the ChJS as a high quality open access journal, free of APC, that cares for gender equality, and is independent from global systems. We are proud to be an eleven-year-old international journal with high and fair standards of review, so we encourage past and new authors to submit their works to the ChJS. As mentioned above, currently we are indexed by several international systems, including the ISI Web of Science and the Scopus database of Elsevier for abstracting and citations. We face important challenges for the future, such as implementing our online platform for submissions, reaching the Science Citation Index, and looking for partnerships with scientific publishers, societies and associations. Just as with any other scientific endeavor, the success of the ChJS depends on a team effort. We are all important to meet these challenges, so we need all of you.

Víctor Leiva and Carolina Marchant
Editors-in-Chief
Chilean Journal of Statistics
http://soche.cl/chjs

Chilean Journal of Statistics
Vol. 11, No. 2, December 2020, 73–93

Functional and Nonparametric Statistics
Research Paper

Robust kernel regression estimator of the scale parameter for functional ergodic data with applications

Ibrahim M. Almanjahie[1], Mohammed Kadi Attouch[2], Omar Fetitah[2], and Hayat Louhab[2]

[1]College of Science, Department of Mathematics, King Khalid University, Saudi Arabia
[2]Laboratoire de Probabilités Statistique Processus Stochastiques, Sidi Bel Abbes University, Algeria

(Received: 19 October 2020 · Accepted in final form: 22 December 2020)

Abstract

In this paper, we propose a family of robust nonparametric estimators for a regression function with unknown scale parameter based on the kernel method. We establish the asymptotic normality of the estimators for functional explanatory variables when the observations exhibit some kind of dependence (stationary ergodic process). This approach can be used for predicting and for building confidence regions. A simulation study is conducted to support our theoretical results and to exhibit the good behavior of the proposed estimator for finite samples with different rates of dependency, and particularly in the presence of several outliers in the data set. In addition, a real data study is provided to illustrate the good behavior of our estimator.

Keywords: Confidence bands · Functional data · Lindeberg condition · Nonparametric kernel estimate · Robust equivariant regression.

Mathematics Subject Classification: Primary 62G35 · Secondary 62G20.

1. Introduction

Nonparametric kernel regression estimation is a familiar tool to explore the underlying relation between the response variable and covariates. In the functional data studies, these estimators are largely studied in Ramsay and Silverman (2002), and Ferraty and Vieu (2006). As in parametric regression estimation, the kernel estimator may be affected by outliers and then it is needed to consider robustness estimation.

Recall that robust regression modeling is an old subject in statistics. It was started by Huber (1964) who studied estimation of a location parameter. We cite Collomb and Hardle (1986) and Laïb and Ould Saïd (2000) for some results on multivariate time series (mixing and ergodicity conditions). Robust regression is widely studied in nonparametric functional statistics. Indeed, it was firstly introduced by Azzedine et al. (2008) who proved the almost complete convergence of this model in the independent and identically distributed case. Since their work, several results on nonparametric robust functional regression were considered. Key references on this topic are Crambes et al. (2008), Chen and Zhang (2009), Attouch et

*Corresponding author. Email: attou_kadi@yahoo.fr

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

al. (2009), Attouch et al. (2010), Gheriballah et al. (2013), Boente and Fraiman (1990) and the references therein.

Notice that all these results are obtained when the scale parameter is known. Boente and Vahnovan (2015, 2017) proposed robust equivariant M-estimators for regression and partial linear models. In this paper, we consider the more general case, that is, when the scale is unknown and the data are dependent. Specifically we model ergodic functional time series.

It is well known that ergodicity is a fundamental hypothesis in statistical physics, thermodynamics and signal processing. In all these areas, ergodicity is studied on a continuous path. Thus, it is necessary to develop statistical tools allowing one to treat the continuous ergodic process in its own dimension by exploring its functional character. This is the general framework of the present work.

Note that the ergodicity assumption is less restrictive than the mixing condition usually assumed in functional time series studies. In particular, this is implied by most mixing conditions. The literature on ergodic functional time series data is still limited. The few existing results are in Laïb and Louani (2011, 2010), Gheriballah et al. (2013), Benziadi et al. (2016a,b). Among the extensive literature on functional data analysis, we only refer to the overviews for parametric models given by Bosq (2000), Ramsay and Silverman (2002) and to the monograph of Ferraty and Vieu (2006) for nonparametric models.

The main objective of this paper is to generalize the results of Boente and Vahnovan (2015) from the independent case to the ergodic case. Specifically we prove the asymptotic normality of an estimator constructed by combining the concepts of robustness with those of unknown scale parameter. This result is obtained under standard conditions allowing us to explore the different structural axes of the subject, such as the robustness of the regression function and the correlation between the observations. We point out that, unlike the case of fixed scale, here the scale parameter must be estimated, which makes the establishment of its asymptotic properties more difficult.

The reminder of this paper is organized as follows. Section 2 is dedicated to the presentation of the robust estimator with unknown scale parameter. The needed assumptions and notations are given in Section 3. We state and proof our main results in Section 4. Some simulation results are reported in Section 5 to compare the M-estimator (for known and unknown scale parameter) with the kernel regression estimator. Section 6 deals with a real data application. The proofs of the main results are relegated to the Appendix. In Section 7, the main conclusions of this study and ideas for future research are provided.

2. The robust equivariant estimators and their related functional

Let $(X_i, Y_i)_{i=1,\dots,n}$ be a sequence of strictly stationary dependent random variables and identically distributed as (X, Y), which is a random pair valued in $\mathcal{F} \times \mathbb{R}$, where $(\mathcal{F}, d)$ is a semi-metric space. We study the nonparametric estimation of the robust regression $\theta(x)$, when the scale parameter is unknown and strong dependencies are present (ergodicity). In fact, for any $x \in \mathcal{F}$, $\theta(x)$ is defined as a zero with respect to the parameter a by means of

$$\Psi(x, a, \sigma) = \mathrm{E}\left[\psi_x\left(\frac{Y-a}{\sigma}\right)|X = x\right] = 0,$$

where ψ_x is a real valued function which satisfies some regularity conditions, to be stated below, and σ is a robust measure of the conditional scale. In what follows, we assume, for all $x \in \mathcal{F}$, that the robust regression $\theta(x)$ exists and is unique; see, for example, Boente and Fraiman (1989).

Consider a functional stationary ergodic process $Z_i = (X_i, Y_i)_{i=1,\dots,n}$; see Laïb and Louani (2011) for some definitions and examples. When the scale parameter is unknown, a robust estimator may be constructed following two steps. Firstly, we estimate the scale parameter σ by the local median of the absolute deviation from the conditional median (MED), $\widehat{m}_{\mathrm{MED}}(x)$, of the conditional distribution of Y given $X = x$, denoted $F(y|X = x) = \mathrm{E}(\mathbb{1}_{(-\infty,y]}(Y)|X = x)$, for any $y \in \mathrm{R}$, where $\mathbb{1}_A$ denotes the indicator function on the set A. Then, for $x \in \mathcal{F}$, the kernel estimator $\widehat{s}(x)$ of $\sigma(x)$ is the zero of the equation given by $\widehat{F}\,(s|X = x) = 1/2$, with

$$\widehat{F}\,(y|X = x) = \frac{\sum_{i=1}^n K\left(\frac{d(x,X_i)}{h}\right)\mathbb{1}_{(-\infty,y]}\,(Y_i)}{\sum_{i=1}^n K\left(\frac{d(x,X_i)}{h}\right)},$$

where K is a kernel function, $d(x, X_i)$ denotes the distance between the fixed point x and the realization of the functional random variable X_i, and the bandwidth parameter $h = h_n$ is a sequence of positive numbers which goes to zero as n goes to infinity. Next, the kernel estimator $\widehat{\theta}(x)$, of the robust regression $\theta(x)$, is the zero, with respect to a, of $\widehat{\Psi}(x, a, \widehat{s}) = 0$, where

$$\widehat{\Psi}\,(x, a, \widehat{s}) = \frac{\sum_{i=1}^n K\left(\frac{d(x,X_i)}{h}\right)\psi_x\left(\frac{Y_i-a}{\widehat{s}}\right)}{\sum_{i=1}^n K\left(\frac{d(x,X_i)}{h}\right)}.$$

3. Notations, hypotheses and comments

Throughout the paper, when no confusion is possible, C and $C^{'}$ are some strictly positive generic constants, x is a fixed point in $\mathcal{F}$ and $\mathcal{N}_x$ is a fixed neighborhood of x. For $r > 0$, let $B(x, r) := \{x^{'} \in \mathcal{F}/d(x^{'}, x) < r\}$. Moreover, for $i = 1, \dots, n$, $\mathcal{F}_k$ is the $\sigma-$field generated by $((X_1, Y_1), \dots, (X_k, Y_k))$ and we pose $\mathcal{B}_k$ is the $\sigma-$field generated by $((X_1, Y_1), \dots, (X_k, Y_k), X_{k+1})$.

Our basic assumptions are:

- (A1) The function ψ_x is continuous and monotone in the second component.
- (A2) The processes $(X_i, Y_i)_{i\in N}$ satisfies: (i) $\phi\,(x, r) = \mathrm{P}\,(X \in B\,(x, r)) > 0$, and $\phi_i\,(x, r) = \mathrm{P}\,(X_i \in B\,(x, r)\,|\mathcal{F}_{i-1}) > 0$, $\forall r > 0$; and (ii) for all $r > 0, 1/(n\phi\,(x, r)) \sum_{i=1}^n \phi_i\,(x, r) \xrightarrow{\mathrm{P}} 1$, and $n\phi\,(x, h) \to \infty$ as $h \to 0$, with $\xrightarrow{\mathrm{P}}$ meaning convergence in probability.
- (A3) The function Ψ is such that: (i) the function $\Psi\,(x, ., \sigma)$ is of class $\mathcal{C}^1$ in $\mathcal{N}_x$, a fixed neighborhood of $\theta(x)$; (ii) for each fixed t in $\mathcal{N}_x$, the functions $\Psi\,(., t, \sigma)$, and $\lambda_2\,(\cdot, t, \sigma) = \mathrm{E}\left[\psi_x^2\,((Y - t)/\sigma)\,|X = \cdot\right]$ are continuous at x; and (iii) the derivative of $\Phi\,(x, z, \sigma) = \mathrm{E}\,[\Psi\,(X_1, z, \sigma) - \Psi\,(x, z, \sigma)\,|d\,(x, X_1) = s]$ exists at $s = 0$, and is continuous in the second component in $\mathcal{N}_x$.
- (A4) For each fixed t in the neighborhood of $\theta(x)$, and $\forall j \geq 2$,

$$\mathrm{E}\left[\psi_x^j\,((Y - t)/\sigma)\,|\mathcal{B}_{i-1}\right] = \mathrm{E}\left[\psi_x^j\,((Y - t)/\sigma)\,|X_i\right] < c < \infty, \text{a.s.},$$

  with "a.s." meaning almost sure convergence.
- (A5) The kernel K is a positive function with support in $(0, 1)$, its derivative $K^{'}$ exists in $(0, 1)$, and satisfies $K^{'}\,(t) < 0$ for $0 < t < 1$.
- (A6) There exists a function τ_x, such that $\forall t \in [0, 1]$, $\lim_{h\to 0} \phi\,(x, th)/\phi\,(x, h) = \tau_x\,(t)$, $K^2\,(1) - \int_0^1 \left(K^2\,(u)\right)^{'} \tau_x\,(u)\,\mathrm{d}u > 0$ and $K\,(1) - \int_0^1 K^{'}\,(u)\,\tau_x\,(u)\,\mathrm{d}u \neq 0$.

(A7) The functions $f_x(x)$ and $p(x)$ are bounded on $\mathcal{S}$ such that $A_p = \inf_{x\in\mathcal{S}} p(x) > 0$, and $A_f = \inf_{x\in\mathcal{S}} f_x(x) > 0$. Moreover, $p(x)$ is a continuous function in a neighborhood of $\mathcal{S}$.

(A8) First, we have that: (i) $F(y|X=x)$ is a continuous function of x in a neighborhood of $\mathcal{S}$ and besides it satisfies the equicontinuity condition $\forall\varepsilon > 0$, $\exists\delta > 0 : |u-v| < \delta \Longrightarrow \sup_{x\in\mathcal{S}} (|F(U|X=x) - F(v|X=x)|) < \varepsilon$; and second (ii) $F(y|X=x)$ is symmetric around $\theta(x)$ and a continuous function of y for each fixed x.

(A9) The sequence $h = h_n$ is such that $h_n \to 0$, $n\phi(h) \to \infty$ and $(n\phi(h))/n \to \infty$.

(A10) The sequence $k = k_n$ is such that $k_n/n \to 0$, $k_n \to \infty$ and $k_n/\log(n) \to \infty$.

Remark It is well known that a fundamental property of robust M-estimators is the convexity and the boundedness of the score function. Convexity is important for the existence and uniqueness of the estimate, whereas the boundedness is essential for reducing the influence of atypical values. In this work, convexity is controlled by means of the monotonicity condition (A1). However, we opt for a presentation without the boundedness condition to cover, for example, the classical regression, which is studied under the ergodic process framework by Laïb and Louani (2011). Assumptions (A2) and (A3) are the same conditions used in Gheriballah et al. (2013), while conditions (A4), (A5) and (A6) are very similar to those used by Ferraty et al. (2010). In addition, (A7) and (A8) are the regularity conditions on the marginal density of X and on the conditional distribution function which imply that, for any set $\mathcal{S} \in \mathcal{F}$, $0 < \inf_{x\in\mathcal{S}} s(x) \leq \sup_{x\in\mathcal{S}} s(x) < \infty$ and that $\theta(x)$ is a continuous function of x. Assumptions (A9) and (A10) are standard conditions imposed for brevity of proofs.

4. Asymptotic results

The result in Proposition 4.1 ensures the uniform consistency on a set $\mathcal{S} \in \mathcal{F}$, for both kernel or nearest neighbor with kernel estimates. Theorem 4.2 deals with the asymptotic normality of the proposed estimator.

Proposition 4.1 Assume that assumptions (A5), (A7) and (A8) holds. Moreover, assume that (A9) hold for kernel weights, and that (A10) holds for nearest neighbor with kernel weights. Then, for any set $\mathcal{S}$, we have that

(a) Under (A1) and (A8-ii), we have that $\sup_{x\in\mathcal{S}} |\widehat{\theta}(x) - \theta(x)| \xrightarrow{\text{a.s.}} 0$.

(b) If $F(y|X=x)$ has a unique median at $\theta(x)$, then we reach $\sup_{x\in\mathcal{S}} |\widehat{m}_{\mathrm{MED}}(x) - \theta(x)| \xrightarrow{\text{a.s.}} 0$.

Theorem 4.2 Assume that (A1)-(A6), and (A8-ii) hold. Then, as $\widehat{\theta}(x) \xrightarrow{\text{p}} \theta(x)$ and $\widehat{s}(x) \xrightarrow{\text{p}} \sigma(x)$, we have that

$$\left(\frac{n\phi(x,h)}{\sigma^2(x,\theta(x))}\right)^{1/2} \left(\widehat{\theta}(x) - \theta(x) - B_n(x)\right) \xrightarrow{\text{d}} \mathrm{N}(0,1) \quad \text{as} \quad n \to \infty,$$

where $\xrightarrow{\text{d}}$ meaning convergence in distribution, $B_n(x) = h\Phi'(0,\theta(x))\beta_0/\beta_1 + o(h)$ and $\sigma^2(x,\theta(x)) = \beta_2\lambda_2(x,\theta(x),\sigma)/(\beta_1^2(\Gamma_1(x,\theta(x),\sigma))^2)$, with $\beta_0 = -\int_0^1 (sK(s))'\beta_x(s)\mathrm{d}s$, $\beta_j = -\int_0^1 (K^j)'(s)\beta_x(s)\mathrm{d}s$, for $j = 1,2$, $\Gamma_1(x,\theta(x),\sigma) = \partial\Psi(x,\theta(x),\sigma)/\partial t$, and $\mathcal{A} = \{z \in \mathcal{F}, \lambda_2(z,\theta(z),\sigma)\Gamma_1(z,\theta(z),\sigma) \neq 0\}$.

In order to remove the bias term B_n, we need an additional condition on the bandwidth parameter h.

Corollary 4.3 Under the assumptions of Theorem 4.2, and if the bandwidth parameter h satisfies $nh^2\phi(x,h) \to 0$ as $n \to \infty$, then

$$\left(\frac{n\phi(x,h)}{\sigma^2(x,\theta(x))}\right)^{1/2}\left(\widehat{\theta}(x)-\theta(x)\right) \xrightarrow{\mathrm{d}} \mathrm{N}(0,1) \quad \text{as} \quad n \to \infty.$$

5. Simulation study

Next, we show the efficiency of the proposed estimator in terms of consistency.

The first direct use of Theorem 4.2 is to predict a functional time series process. Let $(Z_t)_{t\in[0,b[}$ be a continuous-time real-valued random process. From the process Z_t, we may construct N functional random variables $(X_i)_{i=1,\dots,N}$ defined by $X_i(t) = Z_{N^{-1}((i-1)b+c)}$, $\forall t \in [0,b]$. The predictor estimator of Y is defined by $\widehat{Y} = \widehat{\theta}(X_N)$. Then, by applying the above results, we obtain the following corollary.

Corollary 5.1 Under the assumptions of Corollary 4.3, we have

$$\left(\frac{N\phi(x,h_N)}{\sigma^2(X_N,\theta(X_N))}\right)^{1/2}\left(\widehat{\theta}(X_N)-\theta(X_N)\right) \xrightarrow{\mathrm{d}} \mathrm{N}(0,1) \quad \text{as} \quad N \to \infty.$$

The second direct result obtained in Theorem 4.2 is to build the conditional confidence curve. Note that an important application of the asymptotic normality result is the construction of confidence intervals for the true value of $\theta(x)$ given that $X = x$. However, the latter requires an estimation of the bias $B_n(x)$ term and of the standard deviation $\sigma(x,\theta(x))$. For the sake of shortness, we neglect the bias term and we estimate $\sigma(x,\theta(x))$ by plug-in method as follows. Effectively, if ψ_x is of class $\mathcal{C}^1$, with respect to the second component, the quantities $\lambda_2(x,\theta(x),s)$ and $\Gamma_1(x,\theta(x),s)$ can be estimated by

$$\widehat{\lambda}_2\left(x,\widehat{\theta}(x),\widehat{s}\right) = \frac{\sum_{i=1}^n K\left(\frac{d(x,X_i)}{h}\right)\psi_x^2\left(\frac{Y_i-\widehat{\theta}(x)}{\widehat{s}}\right)}{\sum_{i=1}^n K\left(\frac{d(x,X_i)}{h}\right)},$$

$$\widehat{\Gamma}_1\left(x,\widehat{\theta}(x),\widehat{s}\right) = \frac{\sum_{i=1}^n K\left(\frac{d(x,X_i)}{h}\right)\frac{\partial}{\partial t}\psi_x\left(\frac{Y_i-\widehat{\theta}(x)}{\widehat{s}}\right)}{\sum_{i=1}^n K\left(\frac{d(x,X_i)}{h}\right)}.$$

We estimate β_1 and β_2 by

$$\widehat{\beta}_1 = \frac{1}{n\phi(x,h)}\sum_{i=1}^n K\left(\frac{d(x,X_i)}{h}\right), \quad \widehat{\beta}_2 = \frac{1}{n\phi(x,h)}\sum_{i=1}^n K^2\left(\frac{d(x,X_i)}{h}\right).$$

It follows that $\widehat{\sigma}(x,\widehat{\theta}(x)) = (\widehat{\beta}_2\widehat{\lambda}_2(x,\widehat{\theta}(x),\widehat{s})/(\widehat{\beta}_1)^2\widehat{\Gamma}_1^2(x,\widehat{\theta}(x),\widehat{s}))^{1/2}$.

Then, from the asymptotic normality result in Section 4, we have

$$\Lambda_n = \left(\frac{n\phi(x,h)}{\sigma^2(x,\theta(x))}\right)^{1/2}\left(\widehat{\theta}(x)-\theta(x)\right) \xrightarrow{\mathrm{d}} \mathrm{N}(0,1) \quad \text{as} \quad n \to \infty.$$

Therefore, we get an approximate $(1-\zeta)100\%$ confidence interval for $\theta(x)$ stated as

$$\widehat{\theta}(x) \pm t_{1-\zeta/2} \times \left(\frac{\widehat{\sigma}_n^2 \left(x, \widehat{\theta}(x)\right)}{n\phi\left(x,h\right)} \right)^{1/2},$$

where $t_{1-\zeta/2}$ denotes the $(1-\zeta/2)100$th standard normal quantile.

To verify the theoretical results, it is possible to visualize the data histogram and then compare its shape to the normal density. The histogram of Λ_n is almost symmetric around zero and to well-shaped like the standard normal density. To do that, we consider the functional nonparametric model given by

$$Y_i = r\left(X_i\right) + \epsilon_i, \quad i = 1, \ldots, n,$$

where the ϵ_is are generated independently according to a normal distribution with mean 0.

Now, we describe how our functional ergodic data are generated. Firstly, we use an `R` routine named `simul.far` of the `far` package to generate the functional explanatory variables $(X_i)_{i=1,\ldots,n}$. This routine simulates a functional autoregressive process white Wiener noise.

For this simulation experiments, we consider sinusoidal basis, with five functional axis, of the continuous functions from $[0,1]$ to R. Recall that, as it is shown in Laïb and Louani (2011), this kind of process satisfies the ergodicity condition. The curves X_is are discretized in the same grid composed by 100 points and are plotted in Figure 1.

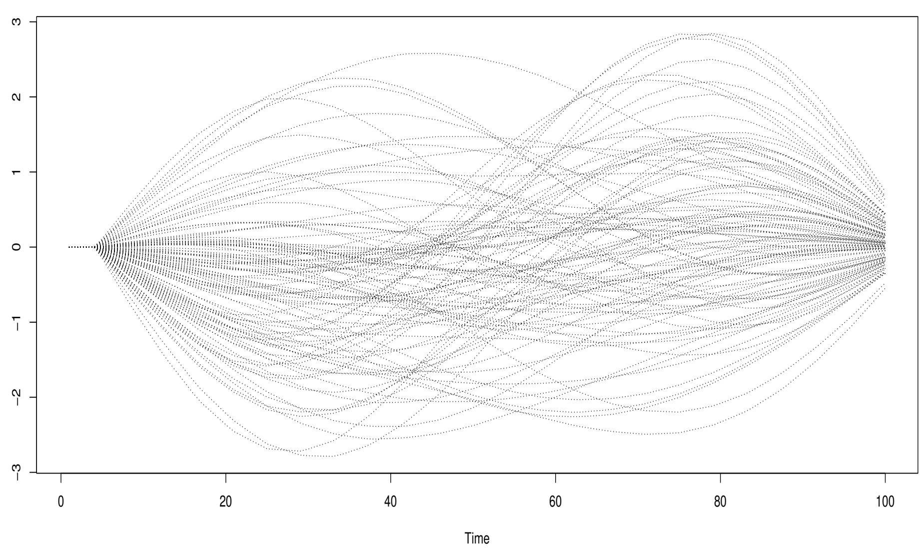


Figure 1. A sample of 100 curves, for $d_\rho = (0.45, 0.90, 0.34, 0.45)$

Secondly, the scalar response Y_i is computed by considering the operator defined as

$$r(x) = 5\int_0^1 \exp\left\{x\left(t\right)\right\} \mathrm{d}t.$$

We compare our estimator (robust equivariant regression –RER–) $\widehat{\theta}(x)$ with the kernel robust regression (KRR) $\tilde{\theta}(x)$ introduced by (Azzedine et al., 2008) and the functional kernel regression (FKR) (Ferraty and Vieu, 2006), where $\widehat{\theta}(x)$, $\tilde{\theta}(x)$ and $\widehat{m}(x)$ are defined as $\widehat{\theta}(x)$ is the zero with respect to a of

$$\frac{\sum_{i=1}^{n} K\left(\frac{d(x,X_i)}{h}\right)\psi_x\left(\frac{Y_i-a}{\widehat{s}(x)}\right)}{\sum_{i=1}^{n} K\left(\frac{d(x,X_i)}{h}\right)} = 0,$$

and $\widetilde{\theta}(x)$ is the zero with respect to a of

$$\frac{\sum_{i=1}^{n} K\left(\frac{d(x,X_i)}{h}\right)\psi_x\left(Y_i-a\right)}{\sum_{i=1}^{n} K\left(\frac{d(x,X_i)}{h}\right)} = 0, \quad \widehat{m}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{d(x,X_i)}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{d(x,X_i)}{h}\right)}.$$

The efficiency of the predictors is evaluated by the empirical mean square errord (MSEs) expressed as

$$\begin{aligned}
\mathrm{MSE}_{\widehat{\theta}} &= n^{-1}\sum_{i=1}^{n}\left(\theta(X_i)-\widehat{\theta}(X_i)\right)^2,\\
\mathrm{MSE}_{\tilde{\theta}} &= n^{-1}\sum_{i=1}^{n}\left(\theta(X_i)-\tilde{\theta}(X_i)\right)^2,\\
\mathrm{MSE}_{\widehat{m}} &= n^{-1}\sum_{i=1}^{n}\left(\theta(X_i)-\widehat{m}(X_i)\right)^2.
\end{aligned}$$

Through this simulation study, we chose the quadratic kernel K defined as $K(u) = (3/4)(1-u^2)\mathbb{1}_{[0,1]}(u)$. The choice of bandwidth parameter h is a crucial question in nonparametric estimation, we propose to choose the optimal bandwidth by using cross-validation (CV) procedure. We adopt the selection rule proposed by Ferraty and Vieu (2006) and given by $h = \arg\min_h \mathrm{CV}(h)$, where $\mathrm{CV}(h) = \sum_{i=1}^{n}(Y_i - \widehat{\theta}^{-i}(X_i))^2$, with $\widehat{\theta}^{-i}(\cdot)$ being the leave-one-out CV –values of the estimator $\widehat{\theta}(\cdot)$ calculate at X_i–; see Ferraty and Vieu (2006) for more details.

We use the semi-metric given by he first derivative of sample curves stated as

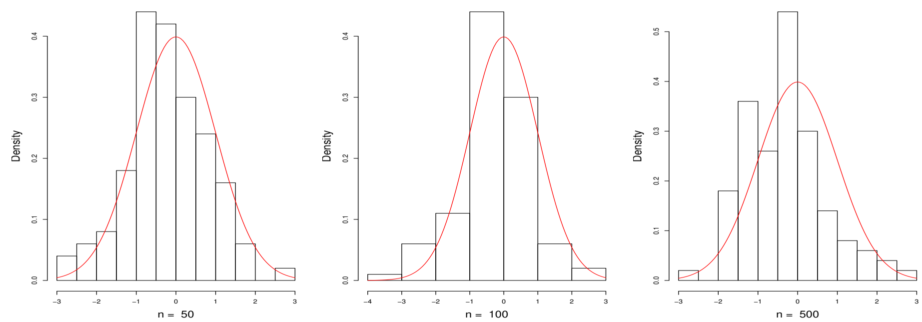
$$d\left(X_i, X_j\right) = \sqrt{\int\left(X_i'(t) - X_j'(t)\right)^2 \mathrm{d}t}.$$

For this comparison study, we treat three estimators in the same conditions.

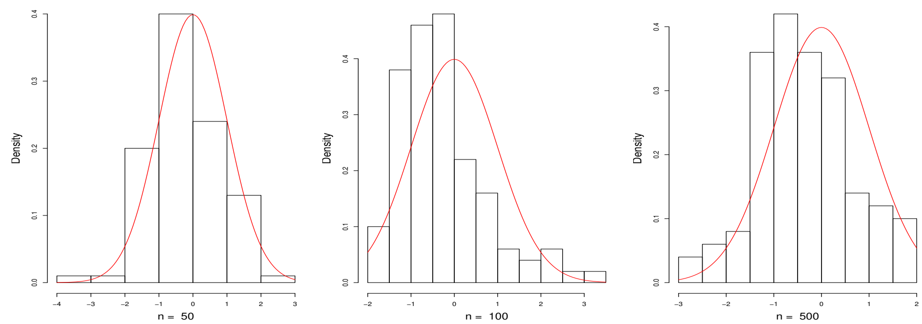
The first illustration concerns the asymptotic normality of $\widehat{\theta}\,(x)$. In order to conduct a Monte Carlo study of the asymptotic normality, we fix one curve, X_0 say, from the previous data. Then, we draw 100 independent replication with samples of size $n = 50, 100, 500$ of the same data and we compute, for each sample a quantity established as

$$\widehat{\Lambda}_n = \left(\frac{\left(\widehat{\beta}_1\right)^2 \widehat{\Gamma}_1^2\left(X_0, \widehat{\theta}\left(X_0\right), \widehat{s}\right)}{\widehat{\beta}_2\widehat{\lambda}_2\left(X_0, \widehat{\theta}\left(X_0\right), \widehat{s}\right)}\right)^{1/2}\left(\widehat{\theta}\left(X_0\right) - \theta\left(X_0\right)\right).$$

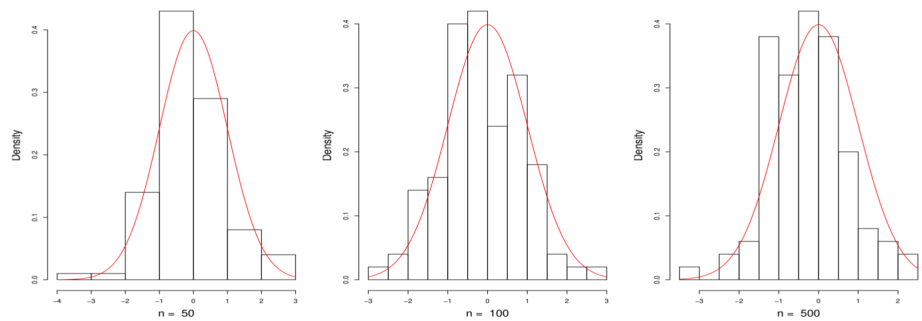
We point out that the functions $\phi(x,h)$ did not intervene in the computation of the normalized deviation by simplification. Thus, the simulation results indicate that $\widehat{\Lambda}_n$ obeys the standard normal law when n is large; see Figure 2 (a)-(c).



(a) $\mathrm{d}_\rho = \mathrm{diag}(0.225, 0.45, 0.17, 0.225)$



(b) $\mathrm{d}_\rho = \mathrm{diag}(0.45, 0.90, 0.34, 0.45)$



(c) $\mathrm{d}_\rho = \mathrm{diag}(0.90, 1.80, 0.68, 0.90)$

Figure 2. Histograms and density curves.

Now, in order to explore the two structural axes of our study, such as the correlation of data and the robustness of the estimate, we compare the performance of our estimator with various values of n and various parameters of the functional autoregressive X_i. Typically, we consider three values of $n = 50, 100, 500$, and three matrix $\mathrm{d}_\rho = \mathrm{diag}(0.225, 0.45, 0.17, 0.225)$, $\mathrm{d}_\rho = \mathrm{diag}(0.45, 0.90, 0.34, 0.45)$ and $\mathrm{d}_\rho = \mathrm{diag}(0.90, 1.80, 0.68, 0.90)$. We emphasize that the results of our simulation study are evaluated over 100 independent replications. The most significant results are gathered in Figure 2 (a)-(c). Note the performance of the estimator is closely related to the degree of correlation expressed by $\|\rho\|$. In sense that the histogram density converge significantly with respect to the value of $\|\rho\|$.

The second result concern the confidence intervals presented in Figures 3 and 4, where three curves corresponding to the predicted interval (green and blue curves) the estimated value (red curve) are drawn. Note that Figure 4 shows the good behavior of our functional forecasting procedure for the robust method in presence of outliers.

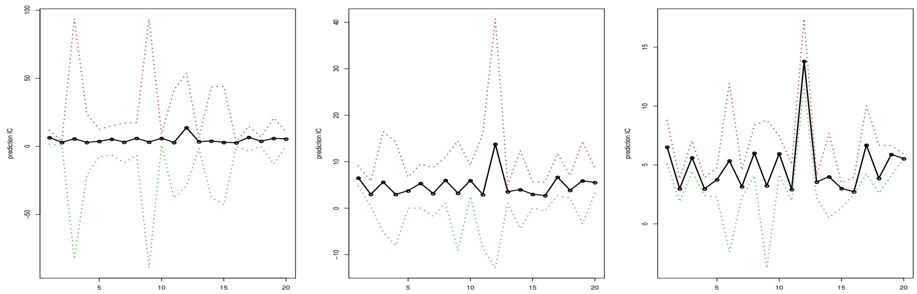


Figure 3. Extremities of the predicted values versus the true values and the confidence bands for the FKR, KRR and RER models respectively (simulation data without outliers).

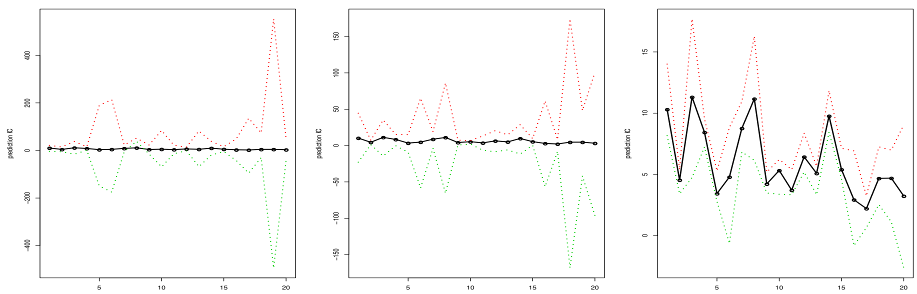


Figure 4. Extremities of the predicted values versus the true values and the confidence bands for the FKR, KRR and RER models respectively (simulation data with 7% of outliers).

Table 1. Comparison between the both methods in the presence of outliers.

Number of the perturbed observations by M	$\text{MSE}_{\widehat{\theta}}$	$\text{MSE}_{\tilde{\theta}}$	$\text{MSE}_{\widehat{m}}$
2	2.3118	2.8112	36.786
14	5.5197	21.8335	2513.116
100	50.716	220.506	331002.4

6. A real data application

Air pollution is one of the most influential factors in human health. Many different chemical substances contribute to the air quality. These substances come from a variety of sources. On the one hand, there are natural sources such as forest fires, volcanic eruptions, wind erosion, pollen dispersal, evaporation of organic compounds, and natural radioactivity. Furthermore, on the other hand, human industrial activity represents the artificial air pollution sources. Ozone (O_3), nitric oxide (NO) and nitrogen dioxide (NO_2) are among the most important contaminants in urban areas, as they have been associated with adverse effects on human health and the natural environment.

We apply the theoretical results obtained in the previous sections to real data. More specifically, in functional prediction context, we examine the performance of the proposed estimator by the robust equivariant approach $\widehat{\theta}(x)$.

In this real data example, we are interested in the prediction of the future O_3, NO and NO_2 concentrations given the curve of it is previous days. For this purpose application, we consider hourly concentrations of the 3 air pollution gases for the year 2018 $(Z_t)_{t\in[0,8760[}$. We consider the data collected from the Leicester University monitoring site in the UK. These observations are available on the following website: https://uk-air.defra.gov.uk. Table 2 gives descriptive statistics of these.

Table 2. Descriptive statistics of the air pollution data.

	O_3	NO	NO_2
Minimum	0.00	0.000	0.00
1st quartile	25.35	1.288	12.26
Median	42.11	3.346	19.56
Mean	42.52	7.027	23.07
3rd quartile	57.73	7.347	30.75
Maximum	149.58	190.765	115.43

We assume that the observations are linked by the model defined as

$$Y_i = r\left(X_i\right) + \epsilon_i, \quad i = 1, \ldots, n-1,$$

where $n = 365$, the functional random variables $\left(X_i\right)_{i=1,\ldots,n}$ defined by $X_i(t) = Z_{(24(i-1)+t)}$, $\forall t \in [0, 24[$, and the scalar response variable Y is defined by $Y_i = \left(Z_{24i+s}\right)_{i=1,\ldots,n-1}$ for a fixed $s \in [0, 24[$. Indeed, Z_t designs the O_3, NO and NO_2 concentrations for 8760 hours between January 01^{st}, 2018 and 31 December 2018. We cut this functional time series in $n - 1 = 364$ pieces X_i of 24 hours (one day). These functionals variables X_i are presented in Figure 5.

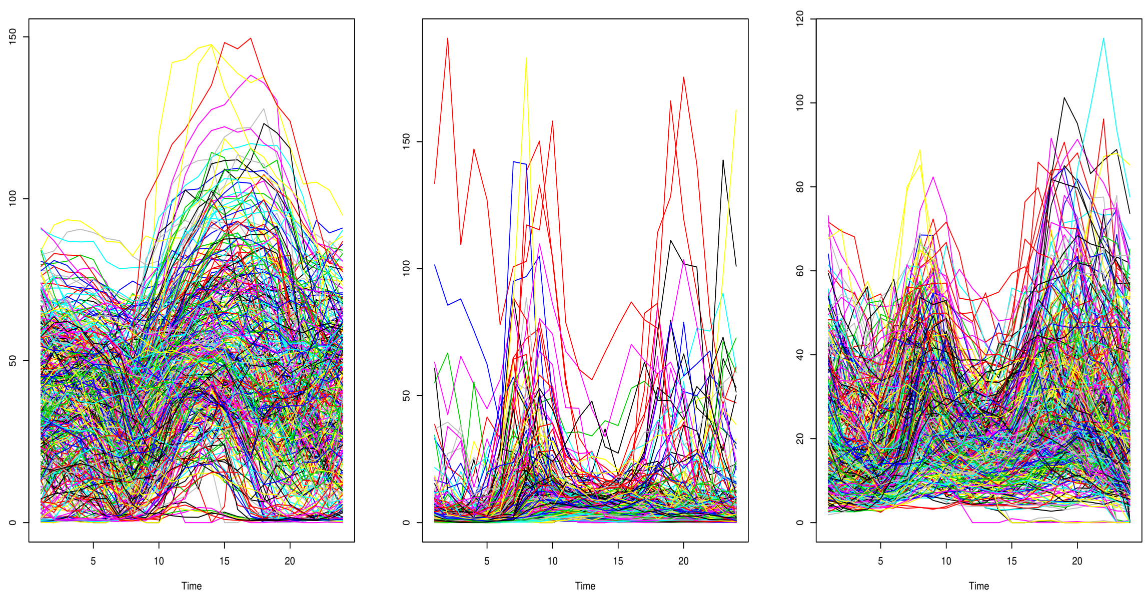


Figure 5. Hourly O_3 (left), NO (center) and NO_2 (right) concentrations of the year 2018.

We want to compare our proposed estimator $\widehat{\theta}(x)$ (RER) with the robust one $\widetilde{\theta}(x)$ (KRR), and the (FKR) $\widehat{m}\left(x\right)$. The kernel K is chosen to be quadratic defined as

$$K(u) = \frac{3}{4}\left(1 - u^2\right) \mathbb{1}_{[0,1]}(u).$$

The choice of bandwidth parameter h is a crucial question in nonparametric estimation. We propose to choose the optimal bandwidth by using the CV procedure. As mentioned, we adopt the selection rule proposed by Ferraty and Vieu (2006). Regarding the shape of the curves X_i, we suggest to use standard functional principal components analysis semi-metrics

(Ferraty and Vieu, 2006), and we adapt it to the data set under analysis obtaining

$$d_q\left(X_i, X_j\right) = \sqrt{\sum_{k=1}^{q}\left(\int\left[X_i(t) - X_j(t)\right] v_k(t)\mathrm{d}t\right)^2}.$$

Here, we take $q = 4$, and v_k is selected among the eigenfunctions of the empirical covariance operator defined as

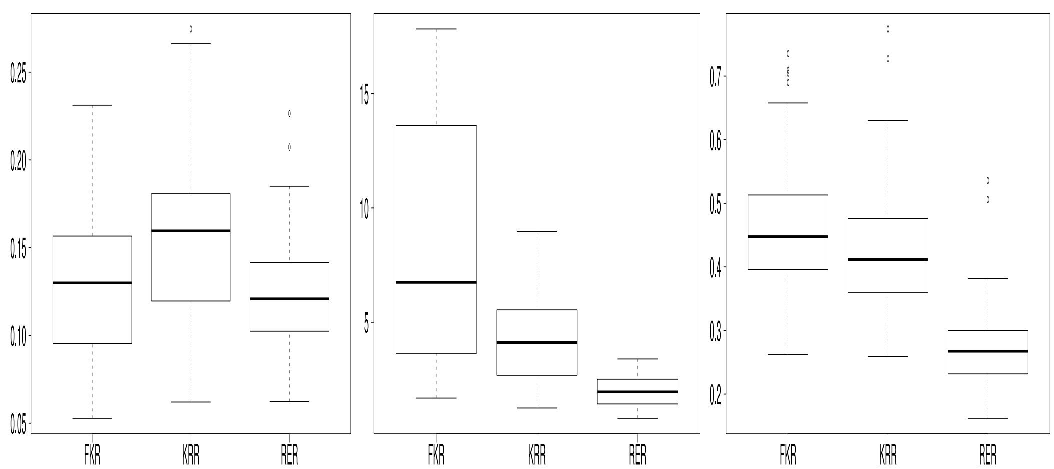
$$\Gamma_X^n(s,t) = \frac{1}{n}\sum_{i=1}^{n} X_i(s)X_i(t).$$

We randomly split our data set $(X_i, Y_i)_{i=1\ldots 364}$ into two subsets, that is, in (i) a learning sample $(T_i, X_i, Y_i)_{i\in I}$ (75% of the observations); and (ii) in a test sample $(X_i, Y_i)_{i\in I'}$, corresponding to a 25% of the observations. We use the relative mean square error RMSE as accuracy measure defined as

$$\mathrm{RMSE} = \frac{1}{\#(I')}\sum_{i\in I'}\left(\frac{Y_i - \widetilde{Y}_i}{Y_i}\right)^2,$$

where $\widetilde{Y}_i$ is the estimator for the three FKR, KRR and RER methods, and $\#(I')$ is the size of I'.

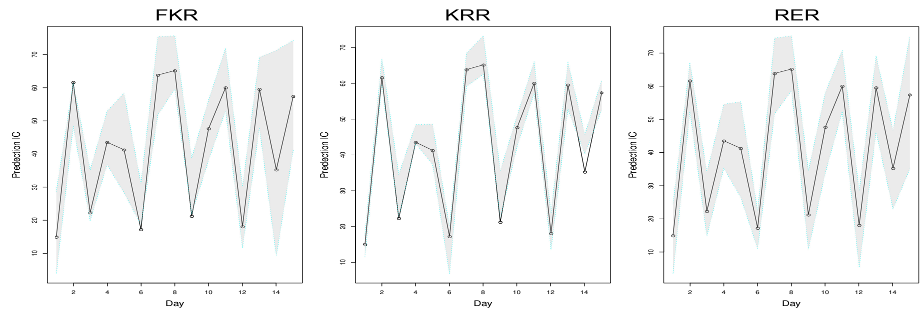
To further explore the performances of our models, we carry out $M = 100$ independent replications which allows us to compute 100 values for RMSE and to display their distribution by means of a scatter-plots. Figures 6 (a)-(c) shows the scatter-plots of the RMSE of the prediction values for the O_3, NO and NO_2, respectively.



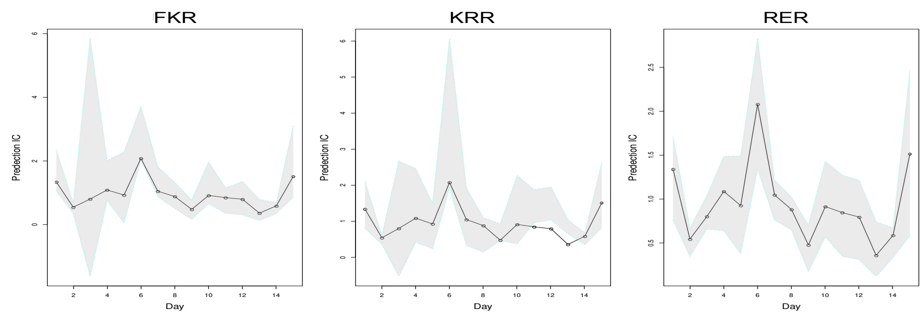
(a) O_3 (b) NO (c) NO_2

Figure 6. Comparison of the RMSE among the FKR, KRR and RER methods for the variable indicated.

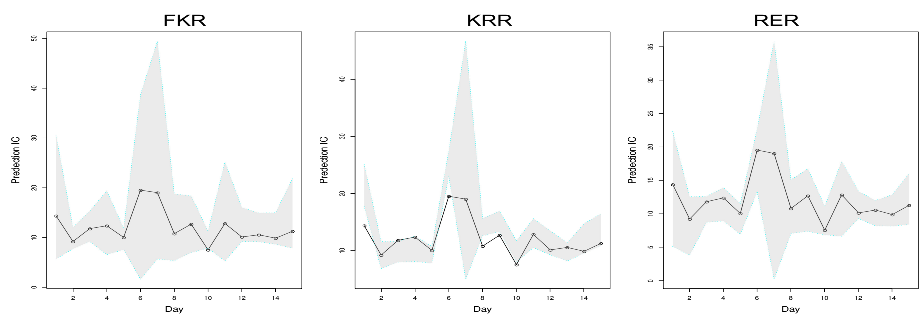
The obtained results of the scatter-plots of the RMSE proves that the Robust equivariant regression gives better results than the Classical and the robust methods. In addition, we give in Figure 7 (a)-(c) the 90% predictive intervals of the concentrations for the three gases of the last 15 values in the sample test by using the three modeles FKR, KRR and RER. The solid black curve the true values. the gray area represents the confidence zone between the dashed Blue curves which represents the lower and upper predicted values.



(a) O_3



(b) NO



(c) NO_2

Figure 7. Comparison of the 90% predictive intervals among the listed methods for the variable indicated.

7. Conclusion and future research

We have provided in this work a generalization of the results given in Boente and Vahnovan (2015) to the functional ergodic data. More precisely, we have proven the asymptotic normality of the robust regression function in the case of unknown scale parameter. These results were obtained under sufficient standard conditions that allowed us to explore different structural axes of the subject, such as the functional naturalness of the model and the data as well as the robustness of the regression function and the correlation of the observation.

Based on the results of this paper on robust regression with unknown scale parameter, we guess that most of the techniques using nonparametric functional kernel smothers could be efficiently extended. For instance, challenging open questions in this sense could concern as extensions to other forms of nonparametric predictors (like functional local linear ones, functional kNN ones, and many other ones). Extensions to other kinds of prediction models in which a preliminary kernel stage plays a crucial role. This would include many semiparametric regression models like functional single index models, and partial linear models, and many other ones. In addition, we see the possibility of extending our asymptotic result to other kinds of dependency data, more particularly the data associated positively (Azevedo and Oliveira, 2011).

Acknowledgements

The authors would like to thank the Editors-in-Chief and the anonymous reviewers for their valuable comments and suggestions which improved substantially the quality of this paper.

References

Attouch, M. Laksaci, A., and Ould Saïd E., 2009. Asymptotic distribution of robust estimator for functional nonparametric models. Communications in Statistics: Theory and Methods, 38, 1317–1335.

Attouch, M. Laksaci, A., and Ould Saïd E., 2010. Asymptotic normality of a robust estimator of the regression function for functional time series data. Journal of the Korean Statistical Society, 39, 489–500.

Azevedo, C. and Oliveira, P. E., 2011. On the kernel estimation of a multivariate distribution function under positive dependence. Chilean Journal of Statistics, 2, 99–113.

Azzedine, N., Laksaci, A., and Ould-Saïd, E., 2008. On the robust nonparametric regression estimation for functional regressor. Statistics and Probability Letters, 78, 3216–3221.

Benziadi, F., Gheriballah, A., and Laksaci, A., 2016a. Asymptotic normality of kernel estimator of ψ-regression functional ergodic data. New Trends in Mathematical Sciences, 1, 268–282.

Benziadi, F., Laksaci, A., and Tebboune, F., 2016b. Recursive kernel estimate of the conditional quantile for functional ergodic data. Communications in Statistics: Theory and Methods, 45, 3097–3113.

Boente, G. and Fraiman, R., 1989. Nonparametric regression estimation. J ournal of Multivariate Analysis, 29, 180–198.

Boente, G. and Fraiman, R., 1990. Asymptotic distribution of robust estimators for nonparametric models from mixing processes. Annals of Statistics, 18, 891–906.

Boente, G. and Vahnovan, A., 2015. Strong convergence of robust equivariant nonparametric functional regression estimators. Statistics and Probability Letters, 100, 1–11.

Boente, G. and Vahnovan, A., 2017. Robust estimators in semi-functional partial linear regression models. Journal of Multivariate Analysis, 154, 59–84.

Bosq, D., 2000. Linear Processes in Function Spaces. Springer, New York.

Chen, J. and Zhang, L., 2009. Asymptotic properties of nonparametric M-estimation for mixing functional data. Journal of Statistical Planning and Inference, 139, 533–546.

Collomb, G. and Härdle, W., 1986. Strong uniform convergence rates in robust nonparametric time series analysis and prediction: Kernel regression estimation from dependent observations. Stochastic Processes and their Applications, 23, 77–89.

Crambes, C., Delsol, L., and Laksaci, A., 2008. Robust nonparametric estimation for functional data. Journal of Nonparametric Statistics, 20, 573–598.

Ferraty. F. and Vieu. P., 2006. Nonparametric Functional Data Analysis: Theory and Practice. Springer, New York.

Ferraty, F., Laksaci, A., Tadj, A., and Vieu, P. 2010. Rate of uniform consistency for nonparametric estimates with functional variables. Journal of Statistical Planning and Inference, 140, 335–352.

P. Gaenssler, P., Strobel, J., and Stute, W., 1978. On central limit theorems for martingale triangular arrays. Acta Mathematica Academiae Scientiarum Hungarica, 31, 205–216.

Gheriballah, A. Laksaci, A., and Sekkal, S., 2013. Nonparametric M-regression for functional ergodic data. Statistics and Probability Letters, 83, 902–908.

Huber, P.J., 1964. Robust estimation of a location parameter. Annals of Mathematical Statistics, 35, 73–101.

Laïb, N. and Louani, D., 2010. Nonparametric kernel regression estimation for functional stationary ergodic data: Asymptotic properties. Journal of Multivariate Analysis, 101, 2266–2281.

Laïb, N. and Louani D., 2011. Rates of strong consistencies of the regression function estimator for functional stationary ergodic data. Journal of Statistical Planning and Inference, 141, 359–372.

Laïb, N. and Ould Saïd, E., 2000. A robust nonparametric estimation of the autoregression function under ergodic hypothesis. Canadian Journal of Statistics, 28, 817–828.

Ramsay, J.O. and Silverman, B.W., 2002. Applied Functional Data Analysis: Methods and Case Studies. Springer, New York.

Appendix

Proof of Proposition 4.1 In order to prove Proposition 4.1, we begin by fixing some notation. We prove that, for any measurable $A \subset R$, $\widehat{\phi}_A(x) = \widehat{r}_A(x)/\widehat{p}(x)$, where

$$\widehat{r}_A(x) = \sum_{i=1}^{n} W_{i,n}(x) I_A(y_i), \widehat{p}(x) = \sum_{i=1}^{n} W_{i,n}, W_{i,n}(x) = \frac{K\left(\frac{d(x_i,x)}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{d(x_j,x)}{h_n}\right)}, \tag{1}$$

denote the kernel weights. Next, we prove (a) and (b). Note that:

(a) Arguing as in Theorem 3.3 in Boente and Fraiman (1990), we only need to prove that

$$\sup_{x\in\mathcal{S}}\sup_{y\in R}\left|\widehat{F}(y|X=x) - F(y|X=x)\right| \overset{\text{a.s.}}{\to} 0.$$

Theorems 3.1 or 3.2 from Boente and Fraiman (1990) entail that

$$\sup_{x\in\mathcal{S}}\sup_{y\in R}\left|\widehat{r}(y,x) - r(y,x)\right| \overset{\text{a.s.}}{\to} 0, \quad \sup_{x\in\mathcal{S}}\left|\widehat{p}(x) - p(x)\right| \overset{\text{a.s.}}{\to} 0, \tag{2}$$

where $r(y,x) = \phi_{(-\infty,y]}(x) = p(x)F(y|X=x)$ and $\widehat{r}(y,x) = \widehat{\phi}_{(-\infty,y]}(x)$, with $\widehat{\phi}_{(-\infty,y]}(x)$ and $\widehat{p}(x)$ being defined in Equation (1).

Note that Equation (2) can be derived for kernel weights using Proposition 2 in Collomb and Hardle (1986). Now, Equation (2) follows from using (A7) and the inequality

$$\sup_{x\in\mathcal{S}}\sup_{y\in R}\left|\widehat{F}(y|X=x) - F(y|X=x)\right| \le \frac{\sup_{x\in\mathcal{S}}\sup_{y\in R}\left|\widehat{r}(y,x) - r(y,x)\right| + \sup_{x\in\mathcal{S}}\left|\widehat{p}(x) - p(x)\right|}{A_p\widehat{A}_p},$$

where $A_p = \inf_{x\in\mathcal{S}} p(x)$ and $\widehat{A}_p = \inf_{x\in\mathcal{S}} \widehat{p}(x)$.

(b) The equicontinuity condition given in (A8), and the uniqueness of the conditional median, imply that $\theta(x)$ is a continuous function of x. Thus, for any fixed $a \in \mathrm{R}$, the function $h_a(x) = F(a+\theta(x)|X=x)$ also is continuous with respect to x.

Given $\epsilon > 0$, let $0 < \delta < \epsilon$, such that

$$|u-v| < \delta \Longrightarrow \sup_{x\in\mathcal{S}}\left(|F(U|X=x) - F(v|X=x)|\right) < \frac{\epsilon}{2}. \tag{3}$$

Then, from the uniqueness of the conditional median and Equation (3), we get that

$$\frac{1}{2} < F(\theta(x)+\delta|X=x) < \frac{1}{2} + \frac{\epsilon}{2}, \tag{4}$$

$$\frac{1}{2} - \frac{\epsilon}{2} < F(\theta(x)-\delta|X=x) < \frac{1}{2}. \tag{5}$$

Consider $\iota(\delta) = \inf_{x\in\mathcal{S}} F(\theta(x)+\delta|X=x)$ and $\nu(\delta) = \sup_{x\in\mathcal{S}} F(\theta(x)-\delta|X=x)$. The continuity of $h_\delta(x)$ and $h_{-\delta}(x)$ together with Equations (4) and (5), entail that $\nu(\delta) < 1/2 < \iota(\delta)$, and so $\eta = \min(\iota(\delta) - 1/2, 1/2 - \nu(\delta)) > 0$. If Equation (2) holds, $\mathrm{P}(\mathcal{N}) = 0$, and $\sup_{x\in\mathcal{S}}\sup_{y\in R}|\widehat{F}(y|X=x) - F(y|X=x)| \to 0$, then, for n large enough, we have that $\sup_{x\in\mathcal{S}}\sup_{y\in R}|\widehat{F}(y|X=x) - F(y|X=x)| < \min(\eta/2, \epsilon/2) = \epsilon_1$. Then, for $x \in \mathcal{S}$, we get

$$F(\theta(x)+\delta|X=x) - \epsilon_1 < \widehat{F}(\theta(x)+\delta|X=x) < F(\theta(x)+\delta|X=x) + \epsilon_1,$$
$$F(\theta(x)-\delta|X=x) - \epsilon_1 < \widehat{F}(\theta(x)-\delta|X=x) < F(\theta(x)-\delta|X=x) + \epsilon_1,$$

which entails that

$$\frac{1}{2} < \widehat{F}(\theta(x)+\delta|X=x) < \frac{1}{2} + \epsilon, \quad \frac{1}{2} - \epsilon < \widehat{F}(\theta(x)-\delta|X=x) < \frac{1}{2},$$

and hence, $\sup_{x\in\mathcal{S}}|\widehat{m}_{\mathrm{MED}}(x) - \theta(x)| \le \delta < \epsilon$, which concludes the proof. □

Proof of Theorem 4.2 and Corollary 4.3 We give the proof for the case of increasing ψ_x, with the decreasing case being obtained by considering $-\psi_x$. Thus, we define, for all $u \in \mathrm{R}$, $z = \theta(x) - B_n(x) + u\left[n\phi(x,h)\right]^{-1/2}\sigma(x,\theta(x))$. Notice that

$$\mathrm{P}\left(\left(\frac{n\phi(x,h)}{\sigma^2(x,\theta(x))}\right)^{1/2}\left(\widehat{\theta}(x) - \theta(x) + B_n(x)\right) < u\right) = \mathrm{P}\left(\widehat{\theta}(x) < \theta(x) - B_n(x) + u\left[n\phi(x,h)\right]^{-1/2}\sigma(x,\theta(x))\right)$$
$$= \mathrm{P}\left(0 < \widehat{\Psi}(x,z,\widehat{s})\right).$$

In addition, we have that

$$\widehat{\Psi}(x,t,\widehat{s}) = B_n(x,t,\widehat{s}) + \frac{R_n(x,t,\widehat{s})}{\widehat{\Psi}_D(x)} + \frac{Q_n(x,t,\widehat{s})}{\widehat{\Psi}_D(x)},$$

where

$$Q_n(x,t,\widehat{s}) = \left(\widehat{\Psi}_N(x,t,\widehat{s}) - \bar{\Psi}_N(x,t,\widehat{s})\right) - \Psi(x,t,\widehat{s})\left(\widehat{\Psi}_D(x) - \bar{\Psi}_D(x)\right),$$
$$R_n(x,t,\widehat{s}) = -\left(\frac{\bar{\Psi}_N(x,t,\widehat{s})}{\bar{\Psi}_D(x)} - \Psi(x,t,\widehat{s})\right)\left(\widehat{\Psi}_N(x,t,\widehat{s}) - \bar{\Psi}_N(x,t,\widehat{s})\right),$$
$$B_n(x,t,\widehat{s}) = \frac{\bar{\Psi}_N(x,t,\widehat{s})}{\bar{\Psi}_D(x)},$$

with

$$\widehat{\Psi}_N(x,a,\widehat{s}) = \frac{1}{n\mathrm{E}\left[K\left(h^{-1}d(x,X_1)\right)\right]}\sum_{i=1}^{n} K\left(h^{-1}d(x,X_i)\right)\psi_x\left(\frac{Y_i - a}{\widehat{s}}\right),$$
$$\bar{\Psi}_N(x,a,\widehat{s}) = \frac{1}{n\mathrm{E}\left[K\left(h^{-1}d(x,X_1)\right)\right]}\sum_{i=1}^{n} \mathrm{E}\left[K\left(h^{-1}d(x,X_i)\right)\psi_x\left(\frac{Y_i - a}{\widehat{s}}\right)/\mathcal{F}_{i-1}\right],$$
$$\widehat{\Psi}_D(x) = \frac{1}{n\mathrm{E}\left[K\left(h^{-1}d(x,X_1)\right)\right]}\sum_{i=1}^{n} K\left(h^{-1}d(x,X_i)\right),$$
$$\bar{\Psi}_D(x) = \frac{1}{n\mathrm{E}\left[K\left(h^{-1}d(x,X_1)\right)\right]}\sum_{i=1}^{n} \mathrm{E}\left[K\left(h^{-1}d(x,X_i)\right)|\mathcal{F}_{i-1}\right].$$

Then, it follows that

$$\mathrm{P}\left(\left(\frac{n\phi(x,h)}{\sigma^2(x,\theta_x)}\right)^{1/2}\left(\widehat{\theta}(x) - \theta(x) + B_n(x)\right) < u\right) =$$
$$\mathrm{P}\left(-\widehat{\Psi}_D(x)B_n(x,z,\widehat{s}) - R_n(x,z,\widehat{s}) < Q_n(x,z,\widehat{s})\right).$$

Therefore, our main result is a consequence of the following intermediate results. □

Lemma 7.1 Under the assumptions of Theorem 4.2, we have, for any $x \in \mathcal{A}$,

$$\left(\frac{n\phi(x,h)\beta_1^2}{\beta_2\lambda_2(x,\theta(x),\widehat{s})}\right)^{1/2} Q_n(x,z,\widehat{s}) \xrightarrow{\mathrm{d}} \mathrm{N}(0,1), \quad \text{as} \quad n \to \infty.$$

Proof of Lemma 7.1 For all $i = 1,\ldots,n$, we denote by $K_i(x) = K\left(h^{-1}d(x,X_i)\right)$,

$$\eta_{ni} = \left(\frac{\phi(x,h)\beta_1^2}{\beta_2\lambda_2(x,\theta(x),\widehat{s})}\right)^{1/2}\left(\psi_x\left(\frac{Y_i - z}{\widehat{s}}\right) - \Psi(x,z,\widehat{s})\right)\frac{K_i(x)}{\mathrm{E}\left[K_1(x)\right]},$$

and we define $\zeta_{ni} = \eta_{ni} - \mathrm{E}\left[\eta_{ni}|\mathcal{F}_{i-1}\right]$. Then, we obtain

$$\left(\frac{n\phi\left(x,h\right)\beta_1^2}{\beta_2\lambda_2\left(x,\theta(x),\widehat{s}\right)}\right)^{1/2} Q_n\left(x,z,\widehat{s}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\zeta_{ni}.$$

Since ζ_{ni} is a triangular array of martingale differences according the $\sigma-$field $\left(\mathcal{F}_{i-1}\right)$, we can apply the Central Limit Theorem based on the unconditional Lindeberg condition (Gaenssler et al., 1978). More precisely, we must verify conditions:

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\zeta_{ni}^2|\mathcal{F}_{i-1}\right] \xrightarrow{\mathrm{P}} 1, \tag{6}$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\zeta_{ni}^2 I_{\zeta_{ni}^2>\epsilon n}\right] \xrightarrow{\mathrm{P}} 0, \forall\epsilon>0, \tag{7}$$

We begin by proving Equation (6). In order to do that, we write

$$\mathrm{E}\left[\zeta_{ni}^2|\mathcal{F}_{i-1}\right] = \mathrm{E}\left[\eta_{ni}^2|\mathcal{F}_{i-1}\right] - \mathrm{E}^2\left[\eta_{ni}|\mathcal{F}_{i-1}\right].$$

Therefore, it suffices to prove that

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}^2\left[\eta_{ni}|\mathcal{F}_{i-1}\right] \xrightarrow{\mathrm{P}} 0,$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\eta_{ni}^2|\mathcal{F}_{i-1}\right] \xrightarrow{\mathrm{P}} 1. \tag{8}$$

For the first convergence, we have

$$\begin{aligned}
\left|\mathrm{E}\left[\eta_{ni}|\mathcal{F}_{i-1}\right]\right| &= \frac{1}{\mathrm{E}K_1\left(x\right)}\left(\frac{\phi\left(x,h\right)\beta_1^2}{\beta_2\lambda_2\left(x,\theta(x),\widehat{s}\right)}\right)^{1/2}\left|\mathrm{E}\left[\left(\Psi\left(X_i,t,\widehat{s}\right)-\Psi\left(x,t,\widehat{s}\right)K_i\left(x\right)\right)|\mathcal{F}_{i-1}\right]\right| \\
&\leq \frac{1}{\mathrm{E}\left[K_1\left(x\right)\right]}\left(\frac{\phi\left(x,h\right)\beta_1^2}{\beta_2\lambda_2\left(x,\theta(x),\widehat{s}\right)}\right)^{1/2}\sup_{u\in B(x,h)}\left|\Psi\left(u,t,\widehat{s}\right)-\Psi\left(x,t,\widehat{s}\right)\right|\mathrm{E}\left[K_i\left(x\right)|\mathcal{F}_{i-1}\right].
\end{aligned}$$

Obviously, under (A2) and (A5), we have $C\phi_i\left(x,h\right) \leq \mathrm{E}\left[K_i|\mathcal{F}_{i-1}\right] \leq C^{'}\phi_i\left(x,h\right)$ and $C\phi\left(x,h\right) \leq \mathrm{E}\left[\Delta_i\left(x\right)\right] \leq C^{'}\phi\left(x,h\right)$. In addition, condition (A3-ii) implies that

$$\sup_{u\in B(x,h)}\left|\Psi\left(u,t,\widehat{s}\right)-\Psi\left(x,t,\widehat{s}\right)\right| = o\left(1\right).$$

Combining the lasts three results, we obtain

$$\begin{aligned}
\left(\left|\mathrm{E}\left[\eta_{ni}|\mathcal{F}_{i-1}\right]\right|\right)^2 &\leq \sup_{u\in B(x,h)}\left|\Psi\left(u,t,\widehat{s}\right)-\Psi\left(x,t,\widehat{s}\right)\left(\frac{\beta_1^2}{\beta_2\lambda_2\left(x,\theta(x),\widehat{s}\right)}\right)\right|\frac{1}{\phi\left(x,h\right)}\phi_i^2\left(x,h\right) \\
&\leq \sup_{u\in B(x,h)}\left|\Psi\left(u,t,\widehat{s}\right)-\Psi\left(x,t,\widehat{s}\right)\left(\frac{\beta_1^2}{\beta_2\lambda_2\left(x,\theta(x),\widehat{s}\right)}\right)\right|\frac{1}{\phi\left(x,h\right)}\phi_i\left(x,h\right).
\end{aligned}$$

Thus, by using the fact that

$$\frac{1}{n\phi\left(x,h\right)}\sum_{i=1}^{n}\phi_i\left(x,h\right) \xrightarrow{\mathrm{P}} 1,$$

we obtain

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{E}\left[\eta_{ni}|\mathcal{F}_{i-1}\right]\right)^2 &= \sup_{u\in B(x,h)}\left|\Psi\left(u,t,\widehat{s}\right)-\Psi\left(x,t,\widehat{s}\right)\right|\left(\frac{\beta_1^2}{\beta_2\lambda_2\left(x,\theta_x,\widehat{s}\right)}\right)\left(\frac{1}{n\phi\left(x,h\right)}\sum_{i=1}^{n}\phi_i\left(x,h\right)\right) \\
&= o_p\left(1\right).
\end{aligned}$$

Now, we analyze to the convergence in Equation (8). Consider

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\eta_{ni}^{2}|\mathcal{F}_{i-1}\right] &= \frac{1}{n\left(\mathrm{E}K_1(x)\right)^2}\left(\frac{\phi\left(x,h\right)\beta_1^2}{\beta_2\lambda_2\left(x,\theta(x),\widehat{s}\right)}\right)\\
&\quad\times\sum_{i=1}^{n}\mathrm{E}\left[\left(\psi_x\left(\frac{Y_i-z}{\widehat{s}}\right)-\Psi\left(x,z,\widehat{s}\right)\right)^2 K_i^2\left(x\right)|\mathcal{F}_{i-1}\right]\\
&= \frac{1}{n\left(\mathrm{E}K_1\left(x\right)\right)^2}\left(\frac{\phi\left(x,h\right)\beta_1^2}{\beta_2\lambda_2\left(x,\theta_x,\widehat{s}\right)}\right)\left(\sum_{i=1}^{n}\mathrm{E}\left[\psi_x^2\left(\frac{Y_i-z}{\widehat{s}}\right)\Delta_i^2\left(x\right)|\mathcal{F}_{i-1}\right]\right)\\
&\quad-\frac{2\Psi\left(x,z,\widehat{s}\right)}{n\left(\mathrm{E}K_1\left(x\right)\right)^2}\left(\frac{\phi\left(x,h\right)\beta_1^2}{\beta_2\lambda_2\left(x,\theta_x,\widehat{s}\right)}\right)\sum_{i=1}^{n}\mathrm{E}\left[\psi_x\left(\frac{Y_i-z}{\widehat{s}}\right)\Delta_i^2\left(x\right)|\mathcal{F}_{i-1}\right]\\
&\quad+\frac{1}{n\left(\mathrm{E}K_1\left(x\right)\right)^2}\left(\frac{\phi\left(x,h\right)\beta_1^2}{\beta_2\lambda_2\left(x,\theta_x,\widehat{s}\right)}\right)\Psi^2\left(x,z,\widehat{s}\right)\sum_{i=1}^{n}\mathrm{E}\left[\Delta_i^2\left(x\right)|\mathcal{F}_{i-1}\right].
\end{aligned}
$$

Let $D_1=\sum_{i=1}^{n}\mathrm{E}\left[\psi_x^2\left((Y_i-z)/\widehat{s}\right)\Delta_i^2\left(x\right)|\mathcal{F}_{i-1}\right]$, $D_2=\sum_{i=1}^{n}\mathrm{E}\left[\psi_x\left((Y_i-z)/\widehat{s}\right)\Delta_i^2\left(x\right)|\mathcal{F}_{i-1}\right]$, and $D_3=\sum_{i=1}^{n}\mathrm{E}\left[\Delta_i^2\left(x\right)|\mathcal{F}_{i-1}\right]$. Observe that

$$
\begin{aligned}
D_1 &= \lambda_2\left(x,z,\widehat{s}\right)\sum_{i=1}^{n}\mathrm{E}\left[K_i^2\left(x\right)|\mathcal{F}_{i-1}\right]+\sum_{i=1}^{n}\left[\mathrm{E}\left[K_i^2\left(x\right)\mathrm{E}\left[\psi_x^2\left(\frac{Y_i-z}{\widehat{s}}\right)|\mathcal{B}_{i-1}\right]|\mathcal{F}_{i-1}\right]\right]\\
&\quad-\sum_{i=1}^{n}\lambda_2\left(x,z,\widehat{s}\right)\mathrm{E}\left[K_i^2\left(x\right)|\mathcal{F}_{i-1}\right]\\
&= \lambda_2\left(x,z,\widehat{s}\right)\sum_{i=1}^{n}\mathrm{E}\left[K_i^2\left(x\right)|\mathcal{F}_{i-1}\right]+\sum_{i=1}^{n}\mathrm{E}\left[K_i^2\left(x\right)\mathrm{E}\left[\psi_x^2\left(\frac{Y_i-z}{\widehat{s}}\right)|X_i\right]|\mathcal{F}_{i-1}\right]\\
&\quad-\sum_{i=1}^{n}\left[\lambda_2\left(x,z,\widehat{s}\right)\mathrm{E}\left[K_i^2\left(x\right)|\mathcal{F}_{i-1}\right]\right].
\end{aligned}
$$

To evaluate the second term, we have

$$
\begin{aligned}
&\frac{1}{n\mathrm{E}\left[K_1\left(x\right)\right]}\sum_{i=1}^{n}\left[\mathrm{E}\left[K_i^2\left(x\right)\mathrm{E}\left[\psi_x^2\left(\frac{Y_i-z}{\widehat{s}}\right)|X_i\right]|\mathcal{F}_{i-1}\right]-\lambda_2\left(x,z,\widehat{s}\right)\mathrm{E}\left[K_i^2\left(x\right)|\mathcal{F}_{i-1}\right]\right]\\
&\qquad\leq\sup_{u\in B(x,h)}\left|\lambda_2\left(x,u,\widehat{s}\right)-\lambda_2\left(x,z,\widehat{s}\right)\right|\left(\frac{1}{n\phi\left(x,h\right)}\sum_{i=1}^{n}\mathrm{P}\left(X_i\in B\left(x,h\right)|\mathcal{F}_{i-1}\right)\right).
\end{aligned}
$$

Moreover, we use the continuity of $\lambda_2\left(x,.,\widehat{s}\right)$ to write

$$
\lambda_2\left(x,z,\widehat{s}\right)=\lambda_2\left(x,\theta(x),\widehat{s}\right)+o\left(1\right).
$$

Thus, we get

$$
\frac{1}{n\mathrm{E}\left[K_1\left(x\right)\right]}D_1=\lambda_2\left(x,\theta(x),\widehat{s}\right)\frac{1}{n\mathrm{E}\left[K_1\left(x\right)\right]}\sum_{i=1}^{n}\mathrm{E}\left[K_i^2\left(x\right)|\mathcal{F}_{i-1}\right]+o\left(1\right),
$$

and similarly, we can obtain

$$
\frac{1}{n\mathrm{E}\left[K_1\left(x\right)\right]}D_2=\Psi\left(x,\theta(x),\widehat{s}\right)\frac{1}{n\mathrm{E}\left[K_1\left(x\right)\right]}\sum_{i=1}^{n}\mathrm{E}\left[K_i^2\left(x\right)|\mathcal{F}_{i-1}\right]+o\left(1\right)=o\left(1\right).
$$

Hence, we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\eta_{ni}^{2}|\mathcal{F}_{i-1}\right]=\frac{1}{n\left(\mathrm{E}\left[K_{1}\left(x\right)\right]\right)^{2}}\left(\frac{\phi\left(x,h\right)\beta_{1}^{2}}{\beta_{2}}\right)\sum_{i=1}^{n}\mathrm{E}\left[K_{i}^{2}\left(x\right)|\mathcal{F}_{i-1}\right]+o\left(1\right).$$

In what follows, we employ the same ideas used in Ferraty et al. (2010) to reach

$$\mathrm{E}\left[K_{i}^{2}\left(x\right)|\mathcal{F}_{i-1}\right]=K^{2}\left(1\right)\phi_{i}\left(x,h\right)-\int_{0}^{1}\left(K^{2}(u)\right)^{\prime}\phi_{i}\left(x,uh\right)\mathrm{d}u,$$

and $\mathrm{E}\left[K_1\left(x\right)\right]=K\left(1\right)\phi\left(x,h\right)-\int_0^1\left(K(u)\right)^{\prime}\phi\left(x,uh\right)\mathrm{d}u$. Then, it follows that

$$\begin{aligned}\frac{1}{n\phi\left(x,h\right)}\sum_{i=1}^{n}\mathrm{E}\left[K_{i}^{2}\left(x\right)|\mathcal{F}_{i-1}\right]&=\frac{K^{2}\left(1\right)}{n\phi\left(x,h\right)}\sum_{i=1}^{n}\phi_{i}\left(x,h\right)\\&\quad-\int_{0}^{1}\left(K^{2}(u)\right)^{\prime}\frac{\phi\left(x,uh\right)}{n\phi\left(x,h\right)\phi\left(x,uh\right)}\sum_{i=1}^{n}\phi_{i}\left(x,uh\right)\mathrm{d}u\\&=K^{2}\left(1\right)-\int_{0}^{1}\left(K^{2}(u)\right)^{\prime}\tau_{x}\left(u\right)\mathrm{d}u+o_{p}\left(1\right)=\beta_{2}+o_{p}\left(1\right),\end{aligned}$$

and

$$\frac{1}{n\phi\left(x,h\right)}\mathrm{E}\left[K_{1}\left(x\right)\right]=\beta_{1}+o\left(1\right).$$

We deduce that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\eta_{ni}^{2}|\mathcal{F}_{i-1}\right]=1,$$

which completes the proof of Equation (6).

Concerning Equation (7), we write

$$\zeta_{ni}^{2}I_{\zeta_{ni}^{2}>\varepsilon n}\leq\frac{\left|\zeta_{ni}\right|^{2+\delta}}{\sqrt{\left(\varepsilon n\right)^{\delta}}},\quad\forall\delta>0.$$

Observe that

$$\begin{aligned}\mathrm{E}\left[\zeta_{ni}^{2+\delta}\right]&=\mathrm{E}\left[\left|\eta_{ni}\left(x\right)-\mathrm{E}\left[\eta_{ni}\left(x\right)|\mathcal{F}_{i-1}\right]\right|^{2+\delta}\right]\\&\leq2^{1+\delta}\mathrm{E}\left[\left|\eta_{ni}\left(x\right)\right|^{2+\delta}\right]+2^{1+\delta}\left|\mathrm{E}\left[\mathrm{E}\left[\eta_{ni}|\mathcal{F}_{i-1}\right]^{2+\delta}\right]\right|.\end{aligned}$$

Using the Jensen inequality, we obtain $\mathrm{E}[\zeta_{ni}^{2+\delta}]\leq C\mathrm{E}[|\eta_{ni}(x)|^{2+\delta}]$. Thus, it remains to evaluate $\mathrm{E}[|\eta_{ni}(x)|^{2+\delta}]$. To that end, we once again use the C_r-inequality obtaining

$$\begin{aligned}\mathrm{E}\left[\left|\eta_{ni}\left(x\right)\right|^{2+k}\right]&\leq C\left(\frac{\phi\left(x,h\right)\beta_{1}^{2}}{\beta_{2}\lambda_{2}\left(x,\theta(x),\widehat{s}\right)\mathrm{E}^{2}\left[K_{1}\left(x\right)\right]}\right)^{1+\delta/2}\mathrm{E}\left[K_{i}^{2+\delta}\left(x\right)\psi_{x}^{2+\delta}\left(\frac{Y_{i}-t}{\widehat{s}}\right)\right]\\&\quad+\quad\Psi^{2+\delta}\left(x,z,\widehat{s}\right)\mathrm{E}\left[K_{i}^{2+\delta}\left(x\right)\right].\end{aligned}$$

We condition on X_i and use the fact that

$$\mathrm{E}\left[\psi_{x}^{2+\delta}\left(\frac{Y_{i}-t}{\widehat{s}}\right)|X_{i}\right]<\infty,$$

to obtain

$$\mathrm{E}\left[\left|\eta_{ni}\left(x\right)\right|^{2+\delta}\right]\leq C\left(\frac{1}{\phi\left(x,h\right)}\right)^{1+\delta/2}\mathrm{E}\left(\left[K_{i}\left(x\right)\right]^{2+\delta}\right)\leq C\left(\frac{1}{\phi\left(x,h\right)}\right)^{\delta/2}.$$

Consequently, we get

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\zeta_{ni}^{2}I_{\zeta_{ni}^{2}>\varepsilon n}\right]\leq C\left(\frac{1}{n\phi\left(x,h\right)}\right)^{\delta/2}\rightarrow 0,$$

and the proof is complete. □

Lemma 7.2 (Laïb and Louani, 2010) Under assumptions (A1), (A2), (A5), and (A6), we have $\widehat{\Psi}_{D}\left(x\right)-1=o_{p}\left(1\right)$.

Lemma 7.3 Under assumptions (A1)-(A3), (A5), and (A6), we have

$$\left(\frac{n\phi\left(x,h\right)\beta_{1}^{2}}{\beta_{2}\lambda_{2}\left(x,\theta(x),\widehat{s}\right)}\right)^{1/2}B_{n}\left(x,z,\widehat{s}\right)=u+o\left(1\right),\quad \text{as}\quad n\rightarrow\infty.$$

Proof of Lemma 7.3 From a simple manipulation, we obtain

$$\begin{aligned}\frac{\bar{\Psi}_{N}\left(x,z,\widehat{s}\right)}{\bar{\Psi}_{D}\left(x\right)}&=\frac{1}{\sum_{i=1}^{n}\mathrm{E}\left[K_{i}\left(x\right)|\mathcal{F}_{i-1}\right]}\sum_{i=1}^{n}\mathrm{E}\left[\ K_{i}\left[\mathrm{E}\left[\psi_{x}\left(\frac{Y-z}{\widehat{s}}\right)|X_{1}\right]\right.\right.\\&\quad-\left.\left.\mathrm{E}\left[\psi_{x}\left(\frac{Y-z}{\widehat{s}}\right)|X=x\right]\right]|\mathcal{F}_{i-1}\ \right]+\mathrm{E}\left[\psi_{x}\left(\frac{Y-z}{\widehat{s}}\right)|X=x\right]\\&\quad-\mathrm{E}\left[\psi_{x}\left(\frac{Y-\theta(x)}{\widehat{s}}\right)|X=x\right]=D_{1}\left(x\right)+D_{2}\left(x\right).\end{aligned}$$

For $D_{1}\left(x\right)$, the main idea of the proof follows from Ferraty et al. (2010). Under (A3-iii), obtaining

$$\begin{aligned}A_{i}&=\mathrm{E}\left[K_{i}\left[\mathrm{E}\left[\psi_{x}\left(\frac{Y-z}{\widehat{s}}\right)|X_{i}\right]-\mathrm{E}\left[\psi_{x}\left(\frac{Y-z}{\widehat{s}}\right)|X=x\right]\right]|\mathcal{F}_{i-1}\right]\\&=\mathrm{E}\left[K_{i}\left[\mathrm{E}\left[\Psi\left(X_{i},z,\widehat{s}\right)-\Psi\left(x,z,\widehat{s}\right)|d\left(x,X_{i}\right)|\mathcal{F}_{i-1}\right]\right]\right]\\&=\mathrm{E}\left[K_{i}\Phi\left(d\left(x,X_{i}\right),z\right)|\mathcal{F}_{i-1}\right]\\&=\int\Phi\left(th,z\right)K\left(t\right)\mathrm{dP}^{\mathcal{F}_{i-1}}\left(th\right)=h\Phi^{'}\left(0,z\right)\int tK\left(t\right)\mathrm{dP}^{\mathcal{F}_{i-1}}\left(th\right).\end{aligned}$$

We use the continuity of $\Phi^{'}\left(0,\cdot\right)$, and the fact that

$$\int tK\left(t\right)d\mathrm{P}^{\mathcal{F}_{i-1}}\left(th\right)=K\left(1\right)\phi_{i}\left(x,h\right)-\int_{0}^{1}\left(sK\left(s\right)\right)^{'}\phi_{i}\left(x,sh\right)\mathrm{d}s,$$

to obtain

$$\frac{1}{n}\sum_{i=1}^{n}A_{i}=h\Phi^{'}\left(0,\theta(x)\right)\left(K\left(1\right)-\int_{0}^{1}\left(sK\left(s\right)\right)^{'}\tau_{x}\left(s\right)\mathrm{d}s\right)+o_{p}\left(h\right).$$

In similar way, we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[K_{i}\left(x\right)|\mathcal{F}_{i-1}\right]=\left(K\left(1\right)-\int_{0}^{1}K^{'}\left(s\right)\tau_{x}\left(s\right)\mathrm{d}s\right)+o_{p}\left(1\right).$$

Thus, we have $D_{1}=B_{n}\left(x\right)+o\left(h\right).$ Concerning D_{2}, we use a Taylor expansion to get, under (A3),

$$D_{2}=-B_{n}\left(x\right)+u\left[n\phi\left(x,h\right)\right]^{-1/2}\sigma\left(x,\theta(x)\right)\frac{\partial}{\partial t}\Psi\left(x,\theta(x),\widehat{s}\right)+o\left(\left[n\phi\left(x,h\right)\right]^{-1/2}\right).$$

This completes the proof. □

Lemma 7.4 Under assumptions (A1)-(A3), (A5), and (A6), we have

$$\left(\frac{n\phi\left(x,h\right)\beta_{1}^{2}}{\beta_{2}\lambda_{2}\left(x,\theta(x),\widehat{s}\right)}\right)^{1/2}R_{n}\left(x,z,\widehat{s}\right)=o\left(1\right).$$

Proof of Lemma 7.4 Here, it suffices to prove that

$$\frac{\bar{\Psi}_N(x,t,\widehat{s})-\Psi(x,t,\widehat{s})\,\bar{\Psi}_D(x)}{\bar{\Psi}_D(x)}=o_p(1)$$

and

$$\left|\widehat{\Psi}_N(x,t,\widehat{s})-\bar{\Psi}_N(x,t,\widehat{s})\right|=o_p(1).$$

In addition, we have that

$$\begin{aligned}\frac{\bar{\Psi}_N(x,t,\widehat{s})-\Psi(x,t,\widehat{s})\,\bar{\Psi}_D(x)}{\bar{\Psi}_D(x)} &= \frac{1}{n\mathrm{E}\left[K_1(x)\right]\bar{\Psi}_D(x)}\sum_{i=1}^{n}\mathrm{E}\left[K_i(x)\,\mathrm{E}\left[\psi_x\left(\frac{Y_i-t}{\widehat{s}}\right)|\mathcal{B}_{i-1}\right]|\mathcal{F}_{i-1}\right]\\ &\quad-\Psi(x,t,\widehat{s})\,\mathrm{E}\left[K_i(x)\,|\mathcal{F}_{i-1}\right]\\ &= \frac{1}{n\mathrm{E}\left[K_1(x)\right]\bar{\Psi}_D(x)}\sum_{i=1}^{n}\mathrm{E}\left[K_i(x)\,\mathrm{E}\left[\psi_x\left(\frac{Y_i-t}{\widehat{s}}\right)|X_i\right]|\mathcal{F}_{i-1}\right]\\ &\quad-\Psi(x,t,\widehat{s})\,\mathrm{E}\left[K_i(x)\,|\mathcal{F}_{i-1}\right]\\ &\le \frac{1}{n\mathrm{E}\left[K_1(x)\right]\bar{\Psi}_D(x)}\sum_{i=1}^{n}\mathrm{E}\left[K_i(x)\,|\Psi(X_i,t,\widehat{s})-\Psi(x,t,\widehat{s})|\,|\mathcal{F}_{i-1}\right].\end{aligned}$$

Using (A2-ii), we deduce that

$$\left|\frac{\bar{\Psi}_N(x,t,\widehat{s})-\Psi(x,t,\widehat{s})\,\bar{\Psi}_D(x)}{\bar{\Psi}_D(x)}\right|\le\sup_{x'\in B(x,h)}\left|\Psi\left(x',t,\widehat{s}\right)-\Psi(x,t,\widehat{s})\right|\to 0.$$

Furthermore, we get $\widehat{\Psi}_N(x,z,\widehat{s})-\bar{\Psi}_N(x,z,\widehat{s})=o_p(1)$. Now, we must prove that $\mathrm{E}[\widehat{\Psi}_N(x,z,\widehat{s})-\bar{\Psi}_N(x,z,\widehat{s})]\to 0$ and $\mathrm{Var}[\widehat{\Psi}_N(x,z,\widehat{s})-\bar{\Psi}_N(x,z,\widehat{s})]\to 0$. The first one is a consequence of the definitions of $\widehat{\Psi}_N(x,z,\widehat{s})$, and $\bar{\Psi}_N(x,z,\widehat{s})$. For the second one, we obtain $\widehat{\Psi}_N(x,z,\widehat{s})-\bar{\Psi}_N(x,z,\widehat{s})=\sum_{i=1}^{n}\delta_i(x,z,\widehat{s})$, where

$$\delta_i(x,z,\widehat{s})=\frac{1}{n\mathrm{E}\left[K_1\right]}K_i\psi_x\left(\frac{Y_i-z}{\widehat{s}}\right)-\mathrm{E}\left[K_i\psi_x\left(\frac{Y_i-z}{\widehat{s}}\right)|\mathcal{F}_{i-1}\right].$$

By the Burkholder inequality, we have

$$\mathrm{E}\left[\sum_{i=1}^{n}\delta_i(x,z,\widehat{s})\right]^2\le\sum_{i=1}^{n}\mathrm{E}\left[\delta_i(x,z,\widehat{s})\right]^2.$$

In addition, by the Jensen inequality, we arrive at

$$\mathrm{E}^2\left[\delta_i(x,z,\widehat{s})\right]\le\frac{1}{n^2\mathrm{E}^2\left[K_1\right]}\mathrm{E}\left[K_i^2\psi_x^2\left(\frac{Y_i-z}{s}\right)\right]\le\frac{1}{n^2\mathrm{E}^2\left[K_1\right]}\mathrm{E}\left[K_i^2\right]\le\frac{1}{n\phi^2(x,h)}\phi_i(x,h).$$

Now, (A2) yields $\mathrm{Var}\left[\widehat{\Psi}_N(x,z,\widehat{s})-\bar{\Psi}_N(x,z,\widehat{s})\right]\to 0.$ □

Lemma 7.5 Under assumptions (A1), (A2), (A5), and (A6), $\widehat{\theta}(x)$ exists a.s. for all sufficiently large n.

Proof of Lemma 7.5
From the monotonicity of $\psi_x(Y-./\widehat{s})$, for all $\varepsilon>0$, we have

$$\Psi(x,\theta(x)-\varepsilon,\widehat{s})\le\Psi(x,\theta(x),\widehat{s})\le\Psi(x,\theta(x)+\varepsilon,\widehat{s}).$$

By using a similar argument as those used in the previous Lemmas, we prove that

$$\widehat{\Psi}(x,t,\widehat{s})\xrightarrow{\mathrm{P}}\Psi(x,t,\widehat{s})\,m\quad\forall t\in N_x.$$

Thus, for sufficiently large n and for all ε small enough, we reach $\widehat{\Psi}\,(x, \theta(x) - \varepsilon, \widehat{s}) \leq 0 \leq \widehat{\Psi}\,(x, \theta(x) + \varepsilon, \widehat{s})$, which holds with probability tending to one.

Since ψ_x is a continuous function, it follows that $\widehat{\Psi}\,(x, t, \widehat{s})$ is a continuous function of t and, there exists $\widehat{\theta}\,(x) \in [\theta(x) - \varepsilon, \theta(x) + \varepsilon]$ such that $\widehat{\Psi}(x, \widehat{\theta}(x), \widehat{s}) = 0$. Hence, the uniqueness of $\widehat{\theta}\,(x)$ is a direct consequence of the strict monotonicity of ψ_x in the second component and the fact that

$$\mathrm{P}\left(\sum_{i=1}^{n} K_i = 0\right) = \mathrm{P}\left(\widehat{\Psi}_D\,(x) = 0\right) \to 0 \quad \text{as} \quad n \to \infty,$$

which implies $\sum_{i=1}^{n} K_i \neq 0$ with probability tending to 1. Moreover, since $\widehat{\theta}(x) \in [\theta(x) - \varepsilon, \theta(x) + \varepsilon]$ in probability, it follows that $\widehat{\theta}(x) \xrightarrow{\mathrm{P}} \theta(x)$, as $n \to \infty$. □

Chilean Journal of Statistics
Vol. 11, No. 2, December 2020, 95–116

Multivariate statistics
Research Paper

Inference for the trivariate Marshall-Olkin-Weibull distribution in presence of right-censored data

Ricardo Puziol de Oliveira[1,*], Marcos Vinicius de Oliveira Peres[1], Jorge Alberto Achcar[1], and Nasser Davarzani[2]

[1]Medical School, University of São Paulo, Ribeirão Preto, SP, Brazil,
[2]Department of Data Science and Knowledge Engineering, Maastricht University, the Netherlands

(Received: 20 June 2020 · Accepted in final form: 02 December 2020)

Abstract

Multivariate lifetime data are common in many applications, especially in medical and engineering studies. In this paper, we consider a trivariate Marshall-Olkin-Weibull distribution to model trivariate data in presence of right censored data.Maximum likelihood and Bayesian methods are used to get the parameter estimators of interest. An extensive simulation study was performed to verify the effectiveness of the maximum likelihood estimators. Reliability data sets related to fiber failure strengths were considered to illustrate the performance of the proposed model under the classical and Bayesian approaches. As a result, note that the trivariate Marshall-Olkin-Weibull model could be considered as a good alternative to model trivariate lifetime data, especially under a Bayesian approach which could be of interest for the reliability analysis, as observed with the real data application in industrial engineering presented in the study or any other area of interest.

Keywords: Bayesian approach · Censored data · Maximum likelihood method · Monte Carlo simulation · Multivariate distributions.

Mathematics Subject Classification: Primary 62-XX · Secondary 62Hxx.

1. Introduction

Lifetime distributions have been studied extensively in the literature due to its medical and engineering applications. Usually it is possible to have two or more lifetimes associated with each subject as for example in medical recurrent events. In these situations, it is needed statistical models which capture the dependence among the lifetimes related to each unit. These lifetime data may be censored at a fixed time point due to the limitation of the follow-up period or withdrawal of the subject from the study. Assuming two lifetime observations, Arnold and Strauss (1988); Sarkar (1987); Hawkes (1972); Downton (1970); Gumbel (1960) introduced some bivariate distributions with exponential conditionals. Block and Basu (1974); Marshall and Olkin (1967a); Freund (1961) proposed extensions of the bivariate exponential distribution. In other direction, Basu and Dhar (1995) and Arnold (1975) introduced some bivariate geometric distributions. Pellerey (2008) modeled

*Corresponding author. Email: rpuziol.oliveira@gmail.com

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

dependent lifetimes using Archimedean survival copulas. Moreover, assuming three or more lifetimes, Gultekin and Bairamov (2013); De Oliveira et al. (2021) introduced trivariate geometric distributions. Hougaard (1986) proposed a class of multivariate failure time distributions. Marshall and Olkin (1967b) introduced a multivariate exponential distribution. Arellano-Valle and Genton (2010) introduced multivariate unified skew-elliptical distributions and Richter and Venz (2014) proposed geometric representations of multivariate skewed elliptically contoured distributions.

Considering the univariate situation, a distribution which is widely considered in the lifetime data analysis is the Weibull distribution (Weibull, 1951) given the flexibility of fit for the data. The mathematical properties and its applicability and generalizations have been studied by many authors (see for example, Cohen, 1965; Philip, 1974; Lai et al., 2003; Thoman et al., 1969; Stevens and Smulders, 1979; Rinne, 2008; Mudholkar et al., 1996; Brown and Wohletz, 1995; Pinder III et al., 1978; Cao, 2004; Pham and Lai, 2007; Saraiva and Suzuki, 2017; among many others). In this study, we explore a multivariate exponential distribution introduced by Marshall and Olkin (1967b) given as an extension of the fatal shock model to a multi-component system to build a new trivariate lifetime distribution denoted as the trivariate Marshall-Olkin-Weibull (TMOW) distribution.

We assume three lifetime random variables denoted following this new distribution in presence of right censored data. Maximum likelihood (ML) inference methods using numerical iterative techniques and Bayesian methods using Markov chain Monte Carlo (MC) methods are used to get the inferences of interest. Under the classical approach, the inferences of interest are obtained using standard asymptotically normality of the likelihood function considering the observed Fisher information matrix in place of the usual expected Fisher information matrix given the complexity of the likelihood function. An extensive simulation study is also performed to verify the effectiveness of the considered inference method assuming different fixed values for the parameters of the model and different sample sizes. An application for real data is also presented in order to verify the usefulness of the proposed model.

The paper is organized as follows: in Section 2, it is introduced the TMOW along with some mathematical properties. The estimation procedures assuming complete and censored data are introduced in Section 3 and 4. In Section 5, the results of the MC simulation study are presented to evaluate the biases, the root of the mean squared error and the asymptotic normality of the ML estimators for the TMOW distribution. Section 6 presents an application to reliability data related to fiber failure strengths. Section 7 provides some concluding remarks.

2. The TMOW distribution

The TMOW distribution is constructed considering k-independent Poisson processes governing the occurrence of shocks to components $1, \ldots, k$, respectively; governing the occurrence of shocks to components pairs 1 and 2, 1 and 3, …, $k-1$ and k, respectively; and so on. This construction of the TMOW distribution plays a central role in life testing and reliability analysis since it has exponential marginal distributions, a useful property in many applications.

It is worth mentioning that an important property of the TMOW distribution is that it is not absolutely continuous since it has singular parts (Marshall and Olkin, 1967b)). In addition, the TMOW distribution could be also represented in terms of independent exponentials since there exist independent exponential random variables Z_s such that $X_i = \min_{s_i=1} \quad Z_s$, for $i = 1, \ldots, k$ obtained from the fatal shock model.

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_k)$ be a random vector and consider the occurrence of simultaneous shocks to all k-components assuming the fatal shock model. Then, the survival function

(SF) of this special case of the TMOW distribution with $k+1$ parameters is given by

$$\begin{aligned} S(y_1,\ldots,y_k) &= \mathrm{P}(Y_1 > y_1,\ldots,Y_k > y_k) \\ &= \exp\{-\lambda_1 y_1 - \cdots - \lambda_k y_k - \lambda_{k+1}\max(y_1,\ldots,y_k)\}, \end{aligned} \tag{1}$$

where $\lambda_j > 0$ and $y_j > 0$, for $j = 1,\ldots,k+1$. Notice that the TMOW distribution is, mathematically, a fairly simple distribution, however, its marginal distributions could be inappropriate to model the behavior of units which have no constant failure rates. In this way, an alternative is the use of a Weibull distribution which is the most commonly used distribution to model reliability data since it is easy to interpret, has great flexibility of fit and is an extension of the exponential distribution.

The probability density function (PDF) of a continuous random variable X with a Weibull distribution is given by $f_W(x;\alpha,\beta) = \alpha\beta^\alpha x^{\alpha-1}\exp\{-\beta x^\alpha\}$, where $x \geq 0$, $\beta > 0$ is the scale parameter and $\alpha > 0$ is the shape parameter. Their corresponding cumulative distribution function (CDF) and SF are given respectively by $F_W(x;\alpha,\beta) = 1 - \exp\{-\beta x^\alpha\}$ and $S_W(x;\alpha,\beta) = \exp\{-\beta x^\alpha\}$. Assuming the fatal shock model previously described and considering Equation (1), it is possible to define the multivariate Marshall-Olkin Weibull (MMOW) distribution as an extension of the TMOW distribution. A comprehensive discussion about the MMOW$_k$ model is presented by Kundu and Dey (2009) and a discussion assuming dependent right censorship is presented by Davarzani et al. (2015).

Definition 2.1. (Model formulation) Consider the transformation $Y_j = X_j^\sigma$, that is, $X_j = Y_j^{1/\sigma}$, for $j = 1,\ldots,k$; $\sigma > 0$. Let $\boldsymbol{X} = (X_1,\ldots,X_k)$ be a random vector following a MMOW distribution denoted by MMOW$_k(\lambda_1,\ldots,\lambda_{k+1},\sigma)$ with multivariate SF given by

$$\begin{aligned} S(x_1,x_2,\ldots,x_k) &= \mathrm{P}(X_1 > x_1,\ldots,X_k > x_k) \\ &= \exp\{-\lambda_1 x_1^\sigma - \ldots - \lambda_k x_k^\sigma - \lambda_{k+1}\max(x_1^\sigma,\ldots,x_k^\sigma)\}. \end{aligned} \tag{2}$$

Note that if $\sigma = 1$ in Equation (2), we obtain the multivariate Marshall-Olkin exponential distribution. In this paper, we assume the special case of $k = 3$ lifetimes, that is, the TMOW distribution, assuming a 3-component system. The SF for the lifetimes X_1, X_2 and X_3 is given by

$$\begin{aligned} S(x_1,x_2,x_3) &= \mathrm{P}(X_1 > x_1, X_2 > x_2, X_3 > x_3) \\ &= \exp\{-\lambda_1 x_1^\sigma - \lambda_2 x_2^\sigma - \lambda_3 x_3^\sigma - \lambda_4\max(x_1^\sigma,x_2^\sigma,x_3^\sigma)\}, \end{aligned} \tag{3}$$

that is,

$$S(\boldsymbol{x}) = \begin{cases} S_1(\boldsymbol{x}) = \exp\{-\lambda_{14}x_1^\sigma - \lambda_2 x_2^\sigma - \lambda_3 x_3^\sigma\}, & \text{if} \quad x_2 < x_3 < x_1 \quad \text{or} \quad x_3 < x_2 < x_1, \\ S_2(\boldsymbol{x}) = \exp\{-\lambda_1 x_1^\sigma - \lambda_{24}x_2^\sigma - \lambda_3 x_3^\sigma\}, & \text{if} \quad x_1 < x_3 < x_2 \quad \text{or} \quad x_3 < x_1 < x_2, \\ S_3(\boldsymbol{x}) = \exp\{-\lambda_1 x_1^\sigma - \lambda_2 x_2^\sigma - \lambda_{34}x_3^\sigma\}, & \text{if} \quad x_2 < x_1 < x_3 \quad \text{or} \quad x_1 < x_2 < x_3, \\ S_4(\boldsymbol{x}) = \exp\{-\lambda_1 x_1^\sigma - (\lambda-\lambda_1)x^\sigma\}, & \text{if} \quad x_1 < x_2 = x_3 = x, \\ S_5(\boldsymbol{x}) = \exp\{-\lambda_2 x_2^\sigma - (\lambda-\lambda_2)x^\sigma\}, & \text{if} \quad x_2 < x_1 = x_3 = x, \\ S_6(\boldsymbol{x}) = \exp\{-\lambda_3 x_3^\sigma - (\lambda-\lambda_3)x^\sigma\}, & \text{if} \quad x_3 < x_1 = x_2 = x, \\ S_7(\boldsymbol{x}) = \exp\{-\lambda x^\sigma\}, & \text{if} \quad x_1 = x_2 = x_3 = x, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where $\lambda = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4$, $\lambda_{14} = \lambda_1 + \lambda_4$, $\lambda_{24} = \lambda_2 + \lambda_4$ and $\lambda_{34} = \lambda_3 + \lambda_4$. In addition, the PDF for the random vector $\boldsymbol{X} = (X_1, X_2, X_3)$ is obtained from $f(\boldsymbol{x}) = f(x_1,x_2,x_3) =$

$-\partial^3 S(x_1,x_2,x_3)/\partial x_1\partial x_2\partial x_3$, where $S(x_1,x_2,x_3)$ is given in Equation (4), that is,

$$f(\boldsymbol{x}) = \begin{cases} f_1(\boldsymbol{x}) = \lambda_{14}\lambda_2\lambda_3\sigma^3(x_1x_2x_3)^{\sigma-1}\exp\{-\lambda_{14}x_1^\sigma - \lambda_2x_2^\sigma - \lambda_3x_3^\sigma\}, & \text{if} \quad x_2<x_3<x_1 \quad \text{or} \quad x_3<x_2<x_1, \\ f_2(\boldsymbol{x}) = \lambda_1\lambda_{24}\lambda_3\sigma^3(x_1x_2x_3)^{\sigma-1}\exp\{-\lambda_1x_1^\sigma - \lambda_{24}x_2^\sigma - \lambda_3x_3^\sigma\}, & \text{if} \quad x_1<x_3<x_2 \quad \text{or} \quad x_3<x_1<x_2, \\ f_3(\boldsymbol{x}) = \lambda_1\lambda_2\lambda_{34}\sigma^3(x_1x_2x_3)^{\sigma-1}\exp\{-\lambda_1x_1^\sigma - \lambda_2x_2^\sigma - \lambda_{34}x_3^\sigma\}, & \text{if} \quad x_2<x_1<x_3 \quad \text{or} \quad x_1<x_2<x_3, \\ f_4(\boldsymbol{x}) = \lambda_1\lambda_4\sigma^2(x_1x)^{\sigma-1}\exp\{-\lambda_1x_1^\sigma - (\lambda-\lambda_1)x^\sigma\}, & \text{if} \quad x_1<x_2=x_3=x, \\ f_5(\boldsymbol{x}) = \lambda_2\lambda_4\sigma^2(x_2x)^{\sigma-1}\exp\{-\lambda_2x_2^\sigma - (\lambda-\lambda_2)x^\sigma\}, & \text{if} \quad x_2<x_1=x_3=x, \\ f_6(\boldsymbol{x}) = \lambda_3\lambda_4\sigma^2(x_3x)^{\sigma-1}\exp\{-\lambda_3x_3^\sigma - (\lambda-\lambda_3)x^\sigma\}, & \text{if} \quad x_3<x_1=x_2=x, \\ f_7(\boldsymbol{x}) = \lambda_4\sigma x^{\sigma-1}\exp\{-\lambda x^\sigma\}, & \text{if} \quad x_1=x_2=x_3=x, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

3. Classical inference for the TMOW with complete data

Let $(X_{11},X_{21},X_{31}),\ldots,(X_{1n},X_{2n},X_{3n})$ be a random sample of size n from a TMOW distribution with PDF given in (5). Consider the indicator variables stated as

$$v_1 = \begin{cases} 1, & \text{if} \quad x_2<x_3<x_1 \quad \text{or} \quad x_3<x_2<x_1, \\ 0, & \text{otherwise;} \end{cases}$$

$$v_2 = \begin{cases} 1, & \text{if} \quad x_1<x_3<x_2 \quad \text{or} \quad x_3<x_1<x_2, \\ 0, & \text{otherwise;} \end{cases}$$

$$v_3 = \begin{cases} 1, & \text{if} \quad x_2<x_1<x_3 \quad \text{or} \quad x_1<x_2<x_3, \\ 0, & \text{otherwise;} \end{cases}$$

$$v_4 = \begin{cases} 1, & \text{if} \quad x_1<x_2=x_3=x, \\ 0, & \text{otherwise;} \end{cases}$$

$$v_5 = \begin{cases} 1, & \text{if} \quad x_2<x_1=x_3=x, \\ 0, & \text{otherwise;} \end{cases}$$

$$v_6 = \begin{cases} 1, & \text{if} \quad x_3<x_1=x_2=x, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Then, we have seven possible situations considering the indicator variables defined as

- $r_1 = v_1(1-v_2)(1-v_3)(1-v_4)(1-v_5)(1-v_6)$, if $x_2<x_3<x_1$ or $x_3<x_2<x_1$;
- $r_2 = v_2(1-v_1)(1-v_3)(1-v_4)(1-v_5)(1-v_6)$, if $x_1<x_3<x_2$ or $x_3<x_1<x_2$;
- $r_3 = v_3(1-v_2)(1-v_1)(1-v_4)(1-v_5)(1-v_6)$, if $x_2<x_1<x_3$ or $x_1<x_2<x_3$;
- $r_4 = v_4(1-v_2)(1-v_3)(1-v_1)(1-v_5)(1-v_6)$, if $x_1<x_2=x_3=x$;
- $r_5 = v_5(1-v_2)(1-v_3)(1-v_4)(1-v_1)(1-v_6)$, if $x_2<x_1=x_3=x$;
- $r_6 = v_6(1-v_2)(1-v_3)(1-v_4)(1-v_5)(1-v_1)$, if $x_3<x_1=x_2=x$;
- $r_7 = (1-v_1)(1-v_2)(1-v_3)(1-v_4)(1-v_5)(1-v_6)$, if $x_1=x_2=x_3=x$.

From Equation (12), the log-likelihood function assuming a TMOW distribution and a random sample of size n of lifetimes X_1, X_2 and X_3 is given by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^n r_{1i}\log\left(\lambda_{14}\lambda_2\lambda_3\sigma^3\right) + \sum_{i=1}^n r_{1i}(\sigma-1)\log(x_{1i}x_{2i}x_{3i}) + \sum_{i=1}^n r_{7i}(\sigma-1)\log(x_i)$$

$$
\begin{aligned}
&+\sum_{i=1}^{n} r_{1i}\left[-\lambda_{14}x_{1i}^{\sigma}-\lambda_{2}x_{2i}^{\sigma}-\lambda_{3}x_{3i}^{\sigma}\right]+\sum_{i=1}^{n} r_{2i}\log\left(\lambda_{1}\lambda_{24}\lambda_{3}\sigma^{3}\right)+\sum_{i=1}^{n} r_{7i}\log\left(\lambda_{4}\sigma\right)\\
&+\sum_{i=1}^{n} r_{2i}(\sigma-1)\log(x_{1i}x_{2i}x_{3i})+\sum_{i=1}^{n} r_{2i}\left[-\lambda_{1}x_{1i}^{\sigma}-\lambda_{24}x_{2i}^{\sigma}-\lambda_{3}x_{3i}^{\sigma}\right]-\lambda\sum_{i=1}^{n} r_{7i}x_{i}^{\sigma}\\
&+\sum_{i=1}^{n} r_{3i}\log\left(\lambda_{1}\lambda_{2}\lambda_{34}\sigma^{3}\right)+\sum_{i=1}^{n} r_{3i}(\sigma-1)\log(x_{1i}x_{2i}x_{3i})+\sum_{i=1}^{n} r_{4i}\log\left(\lambda_{1}\lambda_{4}\sigma^{2}\right)\\
&+\sum_{i=1}^{n} r_{4i}(\sigma-1)\log(x_{1i}x_{i})+\sum_{i=1}^{n} r_{4i}\left[-\lambda_{1}x_{1i}^{\sigma}-(\lambda-\lambda_{1})x_{i}^{\sigma}\right]+\sum_{i=1}^{n} r_{5i}\log\left(\lambda_{2}\lambda_{4}\sigma^{2}\right)\\
&+\sum_{i=1}^{n} r_{5i}(\sigma-1)\log(x_{2i}x_{i})+\sum_{i=1}^{n} r_{5i}\left[-\lambda_{2}x_{2i}^{\sigma}-(\lambda-\lambda_{2})x_{i}^{\sigma}\right]+\sum_{i=1}^{n} r_{6i}\log\left(\lambda_{3}\lambda_{4}\sigma^{2}\right)\\
&+\sum_{i=1}^{n} r_{6i}(\sigma-1)\log(x_{3i}x_{i})+\sum_{i=1}^{n} r_{6i}\left[-\lambda_{3}x_{3i}^{\sigma}-(\lambda-\lambda_{3})x_{i}^{\sigma}\right]\\
&+\sum_{i=1}^{n} r_{3i}\left[-\lambda_{1}x_{1i}^{\sigma}-\lambda_{2}x_{2i}^{\sigma}-\lambda_{34}x_{3i}^{\sigma}\right]. \qquad (7)
\end{aligned}
$$

The equations for the ML estimators are presented in Appendix 1. Since the ML estimators do not have closed form, it is needed to use numerical methods as the Newton-Raphson, the Nelder-Mead or the quasi-Newton methods to get the ML estimators for each parameter of the model.

4. Classical inference for the TMOW with censored data

A particularity in the analysis of lifetime data is the presence of censored data, that could be right, left or interval censoring. In this section, we assume the presence of right censored data, that is, associated with each lifetime X_j, for $j = 1, 2, 3$, we have a fixed censoring time C_j and the data are given by $T_1 = \min(X_1, C_1)$, $T_2 = \min(X_2, C_2)$ and $T_3 = (X_3, C_3)$. In this way, the likelihood function for the parameters of the TMOW distribution has the data set classified in eight regions stated as

- B_1: T_1, T_2 and T_3 are complete observations;
- B_2: T_1 is complete, T_2 and T_3, are censored observations;
- B_3: T_1 is censored, T_2 is complete and T_3 is a censored observation;
- B_4: T_1 and T_2 are censored and T_3 is a complete observation;
- B_5: T_1 and T_2 are complete and T_3 is a censored observation;
- B_6: T_1 is complete, T_2 is censored and T_3 is a complete observation;
- B_7: T_1 is censored, T_2 and T_3 are complete observations;
- B_8: T_1, T_2 and T_3 are censored observations.

Thus, the likelihood function for $\boldsymbol{\theta} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \sigma)$ based on n observations $\boldsymbol{t}_i = (t_{1i}, t_{2i}, t_{3i})$, for $i = 1, \ldots, n$, is given by

$$
\begin{aligned}
L(\boldsymbol{\theta}) = &\prod_{i\in B_1} f(\boldsymbol{t}_i)\prod_{i\in B_2}\left(-\frac{\partial S(\boldsymbol{t}_i)}{\partial \boldsymbol{t}_{1i}}\right)\prod_{i\in B_3}\left(-\frac{\partial S(\boldsymbol{t}_i)}{\partial \boldsymbol{t}_{2i}}\right)\prod_{i\in B_4}\left(-\frac{\partial S(\boldsymbol{t}_{3i})}{\partial \boldsymbol{t}_{3i}}\right)\prod_{i\in B_5}\left(\frac{\partial^2 S(\boldsymbol{t}_i)}{\partial \boldsymbol{t}_{1i}\partial \boldsymbol{t}_{2i}}\right)\\
&\times\prod_{i\in B_6}\left(\frac{\partial^2 S(\boldsymbol{t}_i)}{\partial \boldsymbol{t}_{1i}\partial \boldsymbol{t}_{3i}}\right)\prod_{i\in B_7}\left(\frac{\partial^2 S(\boldsymbol{t}_i)}{\partial \boldsymbol{t}_{2i}\partial \boldsymbol{t}_{3i}}\right)\prod_{i\in B_8} S(\boldsymbol{t}_i), \qquad (8)
\end{aligned}
$$

where $S(\boldsymbol{t})$ is defined by Equations (3) and (4). Define the indicator variables for the censored data as

$$\delta_{ji} = \begin{cases} 1, & \text{if} \quad T_{ji} \leq C_{ji}, \\ 0, & \text{if} \quad T_{ji} > C_{ji}, \end{cases} \tag{9}$$

with $j = 1, 2, 3$ and $i = 1, \ldots, n$. In this way, the logarithm of the likelihood function stated in Equation (8) using the results $(a), (b), \ldots, (h)$ presented in Appendix 1 for the TMOW distribution in presence of right censored data is given by

$$\begin{aligned}
\ell(\boldsymbol{\theta}) = & \sum_{i=1}^{n} \delta_{1i}\delta_{2i}\delta_{3i}r_{1i} \log f_1(\boldsymbol{t}_i) + \sum_{i=1}^{n} \delta_{1i}\delta_{2i}\delta_{3i}r_{2i} \log f_2(\boldsymbol{t}_i) + \sum_{i=1}^{n} \delta_{1i}\delta_{2i}\delta_{3i}r_{3i} \log f_3(\boldsymbol{t}_i) \\
& + \sum_{i=1}^{n} \delta_{1i}\delta_{2i}\delta_{3i}r_{4i} f_4(\boldsymbol{t}_i) + \sum_{i=1}^{n} \delta_{1i}\delta_{2i}\delta_{3i}r_{5i} \log f_5(\boldsymbol{t}_i) + \sum_{i=1}^{n} \delta_{1i}\delta_{2i}\delta_{3i}r_{6i} \log f_6(\boldsymbol{t}_i) \\
& + \sum_{i=1}^{n} \delta_{1i}\delta_{2i}\delta_{3i}r_{7i} \log f_7(\boldsymbol{t}_i) + \sum_{i=1}^{n} \delta_{1i}(1-\delta_{2i})(1-\delta_{3i})r_{1i} \log g_{11}(\boldsymbol{t}_i) \\
& + \sum_{i=1}^{n} \delta_{1i}(1-\delta_{2i})(1-\delta_{3i})r_{2i} \log g_{12}(\boldsymbol{t}_i) + \sum_{i=1}^{n} \delta_{1i}(1-\delta_{2i})(1-\delta_{3i})r_{3i} \log g_{13}(\boldsymbol{t}_i) \\
& + \sum_{i=1}^{n} \delta_{1i}(1-\delta_{2i})(1-\delta_{3i})r_{4i} \log g_{14}(\boldsymbol{t}_i) + \sum_{i=1}^{n} \delta_{1i}(1-\delta_{2i})(1-\delta_{3i})r_{5i} \log g_{15}(\boldsymbol{t}_i) \\
& + \sum_{i=1}^{n} \delta_{1i}(1-\delta_{2i})(1-\delta_{3i})r_{7i} \log g_{17}(\boldsymbol{t}_i) + \sum_{i=1}^{n} (1-\delta_{1i})\delta_{2i}(1-\delta_{3i})r_{1i} \log g_{21}(\boldsymbol{t}_i) \\
& + \sum_{i=1}^{n} (1-\delta_{1i})\delta_{2i}(1-\delta_{3i})r_{3i} \log g_{23}(\boldsymbol{t}_i) + \sum_{i=1}^{n} (1-\delta_{1i})\delta_{2i}(1-\delta_{3i})r_{4i} \log g_{24}(\boldsymbol{t}_i) \\
& + \sum_{i=1}^{n} (1-\delta_{1i})(1-\delta_{2i})\delta_{3i}r_{2i} \log g_{32}(\boldsymbol{t}_i) + \sum_{i=1}^{n} (1-\delta_{1i})(1-\delta_{2i})\delta_{3i}r_{3i} \log g_{33}(\boldsymbol{t}_i) \\
& + \sum_{i=1}^{n} (1-\delta_{1i})(1-\delta_{2i})(1-\delta_{3i})r_{6i} \log S_5(\boldsymbol{t}_i) + \sum_{i=1}^{n} \delta_{1i}\delta_{2i}(1-\delta_{3i})r_{3i} \log g_{43}(\boldsymbol{t}_i) \\
& + \sum_{i=1}^{n} (1-\delta_{1i})(1-\delta_{2i})(1-\delta_{3i})r_{3i} \log S_3(\boldsymbol{t}_i) + \sum_{i=1}^{n} \delta_{1i}\delta_{2i}(1-\delta_{3i})r_{2i} \log g_{42}(\boldsymbol{t}_i) \\
& + \sum_{i=1}^{n} \delta_{1i}\delta_{2i}(1-\delta_{3i})r_{4i} \log g_{44}(\boldsymbol{t}_i) + \sum_{i=1}^{n} \delta_{1i}\delta_{2i}(1-\delta_{3i})r_{5i} \log g_{45}(\boldsymbol{t}_i) \\
& + \sum_{i=1}^{n} \delta_{1i}\delta_{2i}(1-\delta_{3i})r_{7i} \log g_{47}(\boldsymbol{t}_i) + \sum_{i=1}^{n} \delta_{1i}(1-\delta_{2i})\delta_{3i}r_{3i} \log g_{53}(\boldsymbol{t}_i) \\
& + \sum_{i=1}^{n} (1-\delta_{1i})(1-\delta_{2i})(1-\delta_{3i})r_{7i} \log S_7(\boldsymbol{t}_i) + \sum_{i=1}^{n} \delta_{1i}(1-\delta_{2i})\delta_{3i}r_{2i} \log g_{52}(\boldsymbol{t}_i)
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{i=1}^{n}\delta_{1i}(1-\delta_{2i})\delta_{3i}r_{6i}\log g_{56}(\boldsymbol{t}_i)+\sum_{i=1}^{n}(1-\delta_{1i})\delta_{2i}\delta_{3i}r_{1i}\log g_{61}(\boldsymbol{t}_i)\\
&+\sum_{i=1}^{n}\delta_{1i}(1-\delta_{2i})\delta_{3i}r_{1i}\log g_{51}(\boldsymbol{t}_i)+\sum_{i=1}^{n}(1-\delta_{1i})(1-\delta_{2i})(1-\delta_{3i})r_{2i}\log S_{2}(\boldsymbol{t}_i)\\
&+\sum_{i=1}^{n}(1-\delta_{1i})\delta_{2i}\delta_{3i}r_{3i}\log g_{63}(\boldsymbol{t}_i)+\sum_{i=1}^{n}(1-\delta_{1i})(1-\delta_{2i})(1-\delta_{3i})r_{1i}\log S_{1}(\boldsymbol{t}_i)\\
&+\sum_{i=1}^{n}\delta_{1i}\delta_{2i}(1-\delta_{3i})r_{1i}\log g_{41}(\boldsymbol{t}_i)+\sum_{i=1}^{n}(1-\delta_{1i})(1-\delta_{2i})(1-\delta_{3i})r_{4i}\log S_{4}(\boldsymbol{t}_i)\\
&+\sum_{i=1}^{n}(1-\delta_{1i})(1-\delta_{2i})\delta_{3i}r_{6i}\log g_{36}(\boldsymbol{t}_i)+\sum_{i=1}^{n}(1-\delta_{1i})(1-\delta_{2i})(1-\delta_{3i})r_{6i}\log S_{6}(\boldsymbol{t}_i)\\
&+\sum_{i=1}^{n}(1-\delta_{1i})\delta_{2i}\delta_{3i}r_{2i}\log g_{62}(\boldsymbol{t}_i)+\sum_{i=1}^{n}(1-\delta_{1i})\delta_{2i}(1-\delta_{3i})r_{2i}\log g_{22}(\boldsymbol{t}_i)\\
&+\sum_{i=1}^{n}\delta_{1i}(1-\delta_{2i})(1-\delta_{3i})r_{6i}\log g_{16}(\boldsymbol{t}_i)+\sum_{i=1}^{n}(1-\delta_{1i})\delta_{2i}(1-\delta_{3i})r_{5i}\log g_{25}(\boldsymbol{t}_i).
\end{aligned}
$$

5. Simulation study

This section reports the results of a MC simulation study carried out to assess the performance of the ML estimators of the TMOW distribution assuming complete data.

The computations for classical approach were performed using `maxLik` package (Henningsen and Toomet, 2011) from the `R` software (R Core Team, 2015) with the option `optim.method = ''BFGS''` for `maxLik` function. To apply the proposed Bayesian approach, we have considered the Gibbs Sampling algorithm available in the package `R2jags` (Su and Yajima, 2012) from the `R` and `JAGS` software. A chain with $N = 100,000$ values was generated for each parameter, considering a burn-in of 5% of the size of the chain. In addition, a value generated for every 100 was considered, resulting in chains of size $1,000$ for each parameter. Furthermore, using trace plots and Geweke's diagnostic, the convergence of the chains was monitored, and their stationarity was revealed. Computer codes are available under request.

To estimate the parameters of the TMOW distribution, based on the squared error loss function, $L(\eta, a) = (\eta - a)^2$, we consider that the joint posterior PDF of the parameter $\boldsymbol{\Phi} = (\lambda_1,\ \lambda_2,\ \lambda_3,\ \lambda_4, \sigma)$ is obtained directly from the Bayes formula assuming independent non-informative gamma prior distributions with hyperparameters equals to $\alpha = 0.0001$ and $\beta = 0.0001$ for each parameter and is written as $\pi(\boldsymbol{\theta}; \text{data}) = L(\boldsymbol{\Phi})\pi(\sigma)\prod\pi(\lambda_i)/\int L(\boldsymbol{\theta})\pi(\sigma)\prod\pi(\lambda_i)\,\mathrm{d}\sigma\mathrm{d}\theta_i$, for $= 1, 2, 3, 4$.

The generation of the random values X_1, X_2 and X_3 from the TMOW distribution follows the steps: (1) generate $U_1 \sim \text{Weibull}(\lambda_1, \sigma)$, $U_2 \sim \text{Weibull}(\lambda_2, \sigma)$, $U_3 \sim \text{Weibull}(\lambda_3, \sigma)$, $U_4 \sim \text{Weibull}(\lambda_4, \sigma)$; (2) define $X_1 = \min(U_1, U_4), X_2 = \min(U_2, U_4)$ and $X_3 = \min(U_3, U_4)$; and (3) return the observed values (x_1, x_2, x_3) of (X_1, X_2, X_3).

The simulation study was performed under five scenarios and reported in Table 1, assuming the sample sizes equal to $n = 10, 20, 30, \ldots, 100$. In addition, it was considered 1000 MC replications for each scenario from which were computed the biases and the root of mean squared error (RMSE) as given in Equation (10). Specifically, the bias and RMSE were calculated using the expressions given by

$$\text{Bias}(\widehat{\boldsymbol{\Psi}}) = \frac{1}{B}\sum_{i=1}^{B}(\widehat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i), \quad \text{RMSE}(\widehat{\boldsymbol{\Psi}}) = \left(\frac{1}{B}\sum_{i=1}^{B}(\widehat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i)^2\right)^{1/2}, \tag{10}$$

where $B = 1000$ is the number of simulations and $\boldsymbol{\Psi}$ denotes each parameter $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ or σ. The obtained results are presented in Tables 2 and 3 from where note that:

- The biases and RMSE for parameters λ_1, λ_2 and λ_3 are high and decrease slowly to zero when $n \to \infty$ when compared to the others parameters, but, in general, the average biases and RMSE decrease when $n \to \infty$ that show the consistency property of the ML estimators. That is, we have $\mathrm{E}(\lambda_i) \approx \lambda_i, i = 1, 2, 3, 4$ and $\mathrm{E}(\sigma) \approx \sigma$ when $n \to \infty$.
- In the scenarios 3, 4 and 5, the biases for λ_4 are negatives and close to zero. The same happens to σ in scenario 4. However, for the others parameters and scenarios, the biases are positives for $\lambda_i, i = 1, 2, 3$.
- The results presented in scenario 3 has the higher values for the the biases and RMSE for $\lambda_i, i = 1, 2, 3$ and σ. For λ_4, this occur in the scenario 2. In contrast, the smaller values for the the biases and RMSE are presented in scenarios 1 and 4.
- It is important to point out that the simulation also could be made using a Bayesian approach with different prior distributions for the parameters of the TMOW distribution. The coverage probability and the coverage length could be also computed;
- The simulation results could be improved considering other random variable generation methods and using a better approach for the correlation structures of X_1, X_2 and X_3. Moreover, we conclude that the TMOW distribution could be used as a good alternative model to describe trivariate lifetimes with good accuracy in applications.

As a numeric experiment, let us consider a complete simulated data set and a censored data set (cut point equal to 2.5 for censored lifetimes) that consists of $n = 50$ trivariate lifetimes generated from the TMOW distribution assuming the parameter values presented in scenario 4 (see Table 1) for illustrative purposes of the model performance. The data sets are presented in Table 4. The inference results of interest were obtained using the `maxLik` package of the `R` software with `optim.method = 'SANN'` and are presented in Table 5 as well the asymptotic 95% confidence intervals (CIs) which were obtained using the asymptotic normal distribution given by $\mathrm{N}_5(\boldsymbol{\theta}, \boldsymbol{\Sigma}^{-1})$.

From the results presented in Table 5, we conclude that the TMOW model has a good accuracy for both simulated data sets due to the small values (< 0.5) for the standard error (SE) and the small length of the CI for each parameter which is expected since the data set was generated from the TMOW distribution.

Table 1. True parameters values for each scenario.

Scenario	λ_1	λ_2	λ_3	λ_4	σ
1	1.20	1.30	1.30	0.18	0.10
2	1.20	1.30	1.50	1.45	1.20
3	0.40	0.50	0.60	1.50	2.00
4	0.40	0.50	0.60	0.70	0.80
5	0.80	0.90	0.70	0.35	0.20

Table 2. Bias for each parameter for the considered scenarios.

	Sample Size	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
Bias(λ_1)	10	0.9406	1.8494	2.0303	0.8178	1.2065
	20	0.7687	1.5009	1.8726	0.7527	1.0163
	30	0.7132	1.4897	1.7802	0.7260	0.9889
	40	0.6948	1.4287	1.7758	0.7225	0.9519
	50	0.6727	1.3632	1.7367	0.7131	0.9457
	60	0.6712	1.3575	1.7278	0.7090	0.9405
	70	0.6592	1.3499	1.7270	0.7059	0.9365
	80	0.6561	1.3471	1.7236	0.7053	0.9248
	90	0.6380	1.3221	1.7235	0.7041	0.9110
	100	0.6323	1.3108	1.7017	0.7006	0.9042
Bias(λ_2)	10	0.8385	2.0096	2.1121	0.8607	1.1388
	20	0.6767	1.6119	1.8778	0.7709	0.9247
	30	0.6528	1.5602	1.7946	0.7562	0.9112
	40	0.5746	1.4537	1.7654	0.7217	0.8730
	50	0.5722	1.4269	1.7236	0.7126	0.8663
	60	0.5653	1.4038	1.7084	0.7037	0.8543
	70	0.5563	1.3780	1.6938	0.7005	0.8493
	80	0.5389	1.3703	1.6889	0.6929	0.8414
	90	0.5226	1.3242	1.6825	0.6887	0.8145
	100	0.5139	1.3121	1.6395	0.6701	0.8067
Bias(λ_3)	10	0.8184	2.1178	2.4012	0.8813	1.2464
	20	0.6996	1.7690	1.9862	0.7785	1.0996
	30	0.6114	1.6260	1.8556	0.7233	1.0481
	40	0.5824	1.5543	1.8552	0.7165	1.0096
	50	0.5809	1.5376	1.7908	0.7095	1.0092
	60	0.5749	1.5191	1.7693	0.7092	1.0052
	70	0.5709	1.5141	1.7593	0.7087	1.0040
	80	0.5691	1.4873	1.7591	0.7072	1.0006
	90	0.5524	1.4720	1.7536	0.6985	0.9865
	100	0.5509	1.4612	1.7181	0.6846	0.9768
Bias(λ_4)	10	0.0248	0.0454	-0.0180	-0.0179	-0.0462
	20	0.0126	0.0378	-0.0170	-0.0169	-0.0458
	30	0.0106	0.0360	-0.0157	-0.0157	-0.0454
	40	0.0089	0.0344	-0.0145	-0.0147	-0.0451
	50	0.0085	0.0339	-0.0132	-0.0135	-0.0447
	60	0.0071	0.0338	-0.0130	-0.0123	-0.0446
	70	0.0069	0.0336	-0.0112	-0.0119	-0.0445
	80	0.0069	0.0332	-0.0104	-0.0104	-0.0432
	90	0.0064	0.0329	-0.0092	-0.0093	-0.0403
	100	0.0057	0.0326	-0.0080	-0.0073	-0.0335
Bias(σ)	10	0.0055	0.0855	0.1077	-0.0571	0.0097
	20	0.0028	0.0591	0.0666	-0.0553	0.0041
	30	0.0026	0.0519	0.0340	-0.0526	0.0031
	40	0.0021	0.0483	0.0247	-0.0523	0.0025
	50	0.0010	0.0300	0.0205	-0.0522	0.0023
	60	0.0008	0.0264	0.0097	-0.0500	0.0019
	70	0.0004	0.0244	0.0070	-0.0408	0.0016
	80	0.0002	0.0228	0.0040	-0.0372	0.0010
	90	0.0002	0.0215	0.0037	-0.0330	0.0005
	100	0.0001	0.0178	0.0014	-0.0228	0.0003

Table 3. RMSE for each parameter for the considered scenarios.

	Sample Size	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
RMSE(λ_1)	10	1.4104	2.5790	2.3998	1.0009	1.5558
	20	0.9728	1.7555	2.0090	0.8267	1.1534
	30	0.8109	1.6609	1.8660	0.7745	1.0626
	40	0.7820	1.5619	1.8437	0.7601	1.0087
	50	0.7421	1.4531	1.7772	0.7316	0.9897
	60	0.7308	1.4406	1.7664	0.7314	0.9711
	70	0.7120	1.4122	1.7587	0.7269	0.9612
	80	0.6875	1.3879	1.7580	0.7258	0.9581
	90	0.6843	1.3820	1.7502	0.7208	0.9409
	100	0.6828	1.3725	1.7370	0.7165	0.9382
RMSE(λ_2)	10	1.1970	2.5686	2.4845	1.0251	1.4077
	20	0.8903	1.8512	2.0086	0.8477	1.0475
	30	0.7808	1.7373	1.8814	0.8081	0.9955
	40	0.6412	1.5749	1.8152	0.7424	0.9172
	50	0.6402	1.5241	1.7751	0.7403	0.9114
	60	0.6385	1.4822	1.7545	0.7391	0.9037
	70	0.6313	1.4668	1.7356	0.7286	0.8941
	80	0.5879	1.4361	1.7329	0.7119	0.8775
	90	0.5623	1.3677	1.7116	0.7020	0.8407
	100	0.5580	1.3556	1.6671	0.6855	0.8307
RMSE(λ_3)	10	1.1615	2.8146	3.0221	1.1047	1.5354
	20	0.9039	2.0753	2.1512	0.8608	1.2066
	30	0.7067	1.7595	1.9633	0.7638	1.0943
	40	0.7057	1.7371	1.9256	0.7566	1.0678
	50	0.6599	1.6315	1.8414	0.7444	1.0415
	60	0.6471	1.6293	1.8226	0.7403	1.0409
	70	0.6247	1.5752	1.7931	0.7287	1.0327
	80	0.6124	1.5382	1.7873	0.7271	1.0206
	90	0.6013	1.5333	1.7858	0.7196	1.0087
	100	0.5989	1.5181	1.7425	0.7021	0.9973
RMSE(λ_4)	10	0.0335	0.0476	0.0181	0.0180	0.0463
	20	0.0170	0.0386	0.0170	0.0170	0.0360
	30	0.0140	0.0363	0.0168	0.0168	0.0236
	40	0.0118	0.0346	0.0157	0.0158	0.0189
	50	0.0108	0.0341	0.0144	0.0146	0.0153
	60	0.0094	0.0340	0.0132	0.0135	0.0128
	70	0.0092	0.0337	0.0125	0.0121	0.0107
	80	0.0089	0.0333	0.0119	0.0117	0.0079
	90	0.0083	0.0329	0.0100	0.0108	0.0072
	100	0.0073	0.0327	0.0098	0.0100	0.0057
RMSE(σ)	10	0.0178	0.2115	0.3536	0.1174	0.0328
	20	0.0122	0.1541	0.2546	0.0908	0.0230
	30	0.0099	0.1318	0.1998	0.0800	0.0187
	40	0.0080	0.1044	0.1657	0.0799	0.0162
	50	0.0075	0.1022	0.1485	0.0723	0.0150
	60	0.0071	0.0871	0.1362	0.0692	0.0135
	70	0.0063	0.0779	0.1241	0.0665	0.0120
	80	0.0056	0.0707	0.1216	0.0661	0.0112
	90	0.0055	0.0690	0.1066	0.0655	0.0108
	100	0.0046	0.0617	0.1044	0.0639	0.0104

Table 4. Simulated data sets assuming the true parameters presented in scenario 4 for TMOW distribution.

Complete data		2.9050	1.6158	0.0360	1.4561	0.3848	1.6423	1.7376	2.4365	0.2093	0.8411	1.4982	0.3554	5.4551
	X_1:	0.9944	0.8322	0.2012	0.0784	0.7711	0.0184	1.7222	0.6561	2.3231	2.2312	2.8298	0.1304	1.8389
		0.0566	0.7413	0.5567	0.2910	3.7102	1.1128	2.5874	0.2352	0.2228	1.1922	0.8506	1.1641	0.1689
		1.9354	1.3771	0.8613	0.7622	2.5632	0.0054	1.0581	1.3065	0.3640	0.0045	1.8680		
		2.9050	1.3486	0.8815	0.3376	0.3848	0.7980	1.7376	0.1366	0.2093	0.8411	2.1215	0.4391	2.3127
	X_2:	1.0389	0.6509	0.2012	0.0784	0.7711	2.5113	1.7222	0.1462	0.0569	1.9140	1.2382	1.5333	0.0373
		0.3128	2.0642	0.1847	0.2930	0.5026	1.1128	1.4667	0.0727	0.2228	0.0352	0.8506	2.5642	0.1689
		2.3867	1.3771	1.0442	0.7622	2.7941	0.0054	0.3199	0.5004	1.5470	1.4413	2.9805		
		1.5124	1.4561	2.5641	0.5822	0.3848	0.3794	1.7376	0.5608	0.2093	0.1380	0.4593	0.9403	0.0971
	X_3:	1.4204	0.4511	0.2012	0.0784	0.7711	1.7915	1.7222	0.6561	0.7871	0.4913	2.8522	0.2301	3.5971
		0.3128	0.7610	0.5567	0.2930	1.4540	0.0052	2.0376	0.7102	0.0016	2.8227	0.8506	0.3387	0.1689
		3.4390	0.0107	0.6489	0.7622	1.9607	0.0054	1.2430	1.3065	1.5470	1.2935	0.9869		
Censored data		2.9050	1.6158	0.0360	1.4561	0.3848	1.6423	1.7376	2.4365	0.2093	0.8411	1.4982	0.3554	5.4551
	X_1:	0.9944	0.8322	0.2012	0.0784	0.7711	0.0184	1.7222	0.6561	2.3231	2.2312	2.8298	0.1304	1.8389
		0.0566	0.7413	0.5567	0.2910	3.7102	1.1128	2.5874	0.2352	0.2228	1.1922	0.8506	1.1641	0.1689
		1.9354	1.3771	0.8613	0.7622	2.5632	0.0054	1.0581	1.3065	0.3640	0.0045	1.8680		
		2.9050	1.3486	0.8815	0.3376	0.3848	0.7980	1.7376	0.1366	0.2093	0.8411	2.1215	0.4391	2.3127
	X_2:	1.0389	0.6509	0.2012	0.0784	0.7711	2.5113	1.7222	0.1462	0.0569	1.9140	1.2382	1.5333	0.0373
		0.3128	2.0642	0.1847	0.2930	0.5026	1.1128	1.4667	0.0727	0.2228	0.0352	0.8506	2.5642	0.1689
		2.3867	1.3771	1.0442	0.7622	2.7941	0.0054	0.3199	0.5004	1.5470	1.4413	2.9805		
		1.5124	1.4561	2.5641	0.5822	0.3848	0.3794	1.7376	0.5608	0.2093	0.1380	0.4593	0.9403	0.0971
	X_3:	1.4204	0.4511	0.2012	0.0784	0.7711	1.7915	1.7222	0.6561	0.7871	0.4913	2.8522	0.2301	3.5971
		0.3128	0.7610	0.5567	0.2930	1.4540	0.0052	2.0376	0.7102	0.0016	2.8227	0.8506	0.3387	0.1689
		3.4390	0.0107	0.6489	0.7622	1.9607	0.0054	1.2430	1.3065	1.5470	1.2935	0.9869		

Table 5. ML estimates and the corresponding SE for the model parameters (both simulated data sets).

Parameter	Complete data			Censored data		
	ML	SE	95% CI	ML	SE	95% CI
λ_1	0.4276	0.0915	(0.2482, 0.6069)	0.3831	0.0847	(0.2172, 0.5491)
λ_2	0.5623	0.1164	(0.3342, 0.7904)	0.5705	0.1030	(0.3687, 0.7723)
λ_3	0.6489	0.1282	(0.3976, 0.9001)	0.4703	0.0970	(0.2801, 0.6604)
λ_4	0.7489	0.1441	(0.4664, 1.0314)	0.7629	0.1366	(0.4951, 1.0306)
σ	0.8109	0.0660	(0.6815, 0.9402)	0.7792	0.0633	(0.6553, 0.9032)

6. Application to real reliability data

To illustrate the proposed model, let us assume a reliability data set introduced by Crowder et al. (1994). This data set consists of fiber failure strengths. The four values in each row give the breaking strengths of fiber sections of lengths 5, 12, 30 and 75mm. The values are right-censored at 4.0 and a zero indicates accidental breakage prior to testing; the zeros have been treated as missing data. The data sets are available in Table 7.2 from Crowder et al. (1994). In view of the apparent heterogeneity between fibers a model allowing individual random levels would be appropriate and the proposed TMOW model could be useful in the data analysis. In this way, we assume as response lifetimes the length equals 12mm as X_1, the length equals 30mm as X_2 and the length equals 75mm as X_3. Firstly, to apply the proposed methodology under a right-censored scheme, we have considered the Classical approach. The inference results of interest were obtained using the `maxLik` package of the `R` software with the option `optim.method = ‘‘BFGS’’` for `maxLik` function and are presented in Table 6 as well the asymptotic 95% CIs which were obtained using the asymptotic normal distribution given by $\mathrm{N}_5(\boldsymbol{\theta}, \boldsymbol{\Sigma}^{-1})$.

Table 6. ML estimates for fiber failure strengths data sets.

Parameter	Data set 1			Data set 2			Data set 3		
	ML	SE	95% CI	ML	SE	95% CI	ML	SE	95% CI
λ_1	0.00008	0.00008	(-0.00008, 0.00024)	0.00019	5.93164	(-11.6256, 11.6260)	0.00013	0.00013	(-0.00012, 0.00038)
λ_2	0.00095	0.00071	(-0.00044, 0.00234)	0.00845	0.00131	(0.00588, 0.01102)	0.00192	0.00119	(-0.00041, 0.00425)
λ_3	0.00353	0.00209	(-0.00057, 0.00763)	0.02057	0.00331	(0.01408, 0.02706)	0.00420	0.00245	(-0.00061, 0.00910)
λ_4	0.00002	0.00003	(-0.00004, 0.00008)	0.00108	5.93164	(-11.6247, 11.6269)	0.00001	0.00017	(-0.00033, 0.00033)
σ	7.08472	0.69575	(5.72108, 8.44836)	6.58032	0.47135	(5.65649, 7.50415)	6.86184	0.63028	(5.62652, 8.09716)

From the results displayed in Table 6, one can notice that there is an instability using the classical approach (negative bounds for 95% CI, high values for standard errors), especially for Data set 2. This fact may be related to the complexity of the likelihood in presence of right-censored data. Thus, to avoid this problem, a Bayesian method was considered (see Appendix 1). The inference results of interest for each data set are presented in Table 7 and the plots of the marginal posterior densities for the parameters of the model considering each data set are presented in Figure 1.

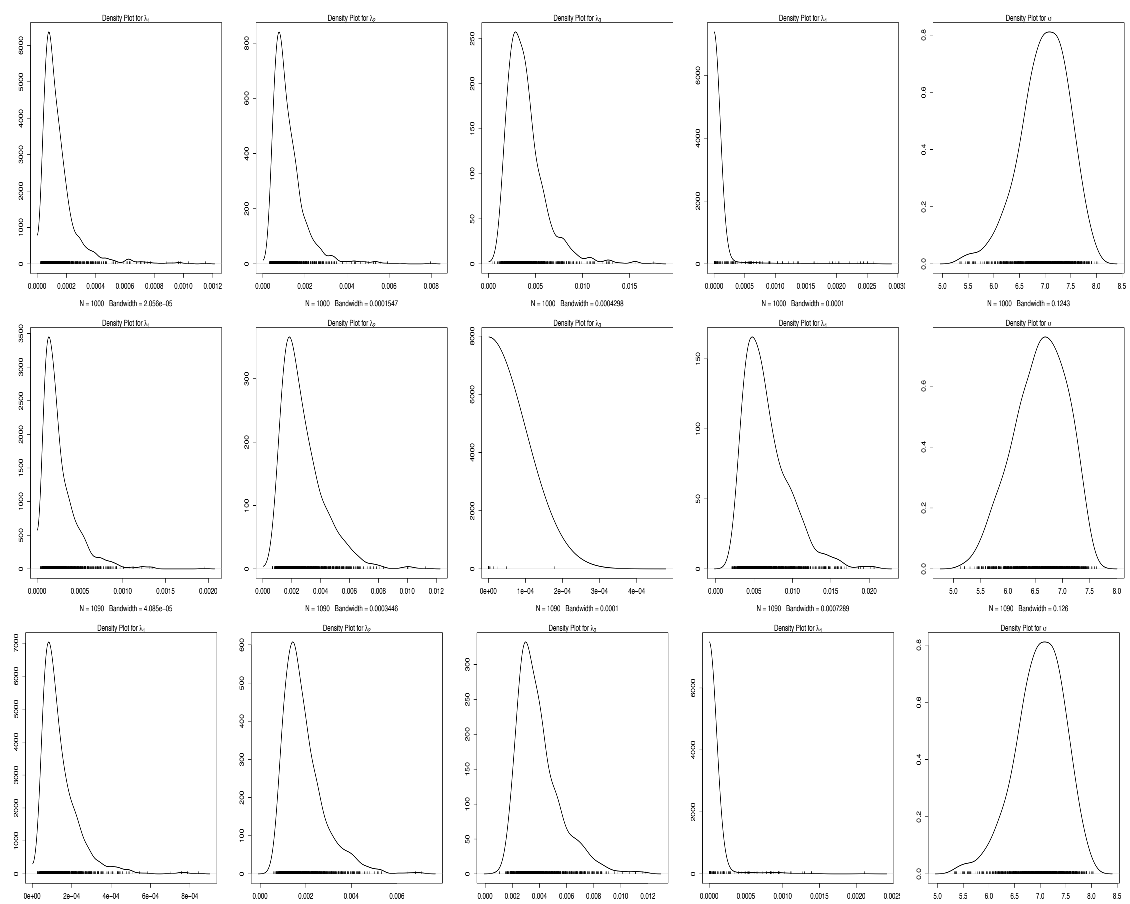


Figure 1. Posterior PDF plots for the parameters of the model assuming the three failure strength data sets (top: Data set 1; middle: Data set 2; bottom: Data set 3).

Table 7. Bayesian estimate, credible interval (CrI) and corresponding standard deviation (SD) for fiber failure strengths data sets.

Parameter	Data set 1			Data set 2			Data set 3		
	Mean	SD	95% CrI	Mean	SD	95% CrI	Mean	SD	95% CrI
λ_1	0.00015	0.00013	(0.00003, 0.00052)	0.00026	0.00017	(0.00006, 0.00064)	0.00014	0.00010	(0.00004, 0.00041)
λ_2	0.00123	0.00081	(0.00040, 0.00333)	0.00273	0.00135	(0.00104, 0.00581)	0.00191	0.00091	(0.00082, 0.00426)
λ_3	0.00399	0.00211	(0.00147, 0.00936)	0.00002	0.00001	(0.00001, 0.00003)	0.00395	0.00161	(0.00190, 0.00800)
λ_4	0.00013	0.00035	(0.00001, 0.00120)	0.00661	0.00287	(0.00263, 0.01291)	0.00001	0.00007	(921E-08, 0.00015)
σ	6.98866	0.47595	(5.92113, 7.79890)	6.57077	0.48138	(5.68461, 7.39571)	6.93945	0.39701	(6.08930, 7.53128)

From the results obtained, note that, for Data sets 1 and 3, the estimate of the parameter λ_4 is very close to zero that means its contribution for the likelihood function is very small. The same happens to the parameter λ_3 in Data set 2. In general, we conclude that the posterior SD values approach to zero and the 95% CrI have reasonable lengths.

7. Concluding remarks

In this paper, we introduced a new trivariate distribution obtained as a special case of the multivariate Marshall-Olkin Weibull distribution. For this new model, we presented some inference properties and an extensive simulation study was performed to verify the performance of the maximum likelihood estimators assuming different fixed values for the parameters of the model and different sample sizes.

The obtained results from Monte Carlo studies showed that the bias and root of mean squared error of the estimators of the trivariate Marshall-Olkin-Weibull distribution are asymptotically non-biased and approaches to zero when the sample size increases even assuming negative values for the biases in some scenarios. From these results, it is possible to conclude that using the proposed model, the obtained inference results are reasonable accurate considering complete data sets and with good performance of the computational algorithm used to get the inferences of interest. However, in the application of fiber strengths, there was a problem with maximum likelihood estimators leading to negative bound for the 95% confidence intervals which could be related of the likelihood function under a right-censoring scheme. To avoid this problem, we considered a Bayesian estimator that provide a better accuracy and good convergence of the simulation algorithm used to get the inference results of interest even using approximately non-informative prior distributions.

In conclusion, the trivariate Marshall-Olkin-Weibull distribution could be used as an alternative to model trivariate data which could be interesting for the reliability analysis (as the fiber strength application) used in engineering applications, or others areas of interest, especially considering a Bayesian approach to estimate the parameters. It is important to point out that other approaches also could be used to get inferences of the proposed model using the expectation-maximization algorithm (Kundu and Dey, 2009) but this topic will be the goal of other study.

Acknowledgments

The authors are grateful to the Editors and Reviewers for several suggestions that led to an improvement of the manuscript.

References

Arellano-Valle, R.B., and Genton, M.G., 2010. Multivariate unified skew-elliptical distributions. Chilean Journal of Statistics, 1, 17-33.

Arnold, B.C., 1975. A characterization of the exponential distribution by multivariate geometric compounding. Sankhyā A, 37, 164–173.

Arnold, B.C. and Strauss, D., 1988. Bivariate distributions with exponential conditionals. Journal of the American Statistical Association, 83, 522–527.

Basu, A.P. and Dhar, S., 1995. Bivariate geometric distribution. Journal of Applied Statistical Science, 2, 33–44.

Block, H.W. and Basu, A., 1974. A continuous, bivariate exponential extension. Journal of the American Statistical Association, 69, 1031–1037.

Brown, W.K. and Wohletz, K.H., 1995. Derivation of the Weibull distribution based on physical principles and its connection to the Rosin–Rammler and lognormal distributions. Journal of Applied Physics, 78, 2758–2763.

Cao, Q.V., 2004. Predicting parameters of a Weibull function for modeling diameter distribution. Forest Science, 50, 682–685.

Cohen, A.C., 1965. Maximum likelihood estimation in the Weibull distribution based on complete and on censored samples. Technometrics, 7, 579–588.

Crowder, M.J., Kimber, A., Sweeting, T., and Smith, R., 1994. Statistical Analysis of Reliability Data. CRC Press, Boca Raton, FL, US.

Davarzani, N., Achcar, J.A., Smirnov, E.N., and Peeters, R., 2015. Bivariate lifetime geometric distribution in presence of cure fractions. Journal of Data Science, 13, 755–770.

De Oliveira, R.P., Mazucheli, J., and Achcar, J.A., 2021. A generalization of Basu-Dhars bivariate geometric distribution to the trivariate case. Communications in Statistics-Simulation and Computation, in press available http://dx.doi.org/10.1080/03610918.2019.1643881.

Downton, F., 1970. Bivariate exponential distributions in reliability theory. Journal of the Royal Statistical Society B, 32, 408–417.

Freund, J.E., 1961. A bivariate extension of the exponential distribution. Journal of the American Statistical Association, 56, 971–977.

Gultekin, O.E. and Bairamov, I., 2013. A trivariate geometric distribution: characterization and asymptotic distribution. EGE University Journal of the Faculty of Science, 37, 1–18.

Gumbel, E.J., 1960. Bivariate exponential distributions. Journal of the American Statistical Association, 55, 698–707.

Hawkes, A.G., 1972. A bivariate exponential distribution with applications to reliability. Journal of the Royal Statistical Society B, 34, 129–131.

Henningsen, A. and Toomet, O., 2011. maxlik: A package for maximum likelihood estimation in R. Computational Statistics, 26, 443–458.

Hougaard, P., 1986. A class of multivariate failure time distributions. Biometrika, 73, 671–678.

Kundu, D. and Dey, A.K., 2009. Estimating the parameters of the Marshall–Olkin bivariate Weibull distribution by EM algorithm. Computational Statistics and Data Analysis, 53, 956–965.

Lai, C., Xie, M., and Murthy, D., 2003. A modified Weibull distribution. IEEE Transactions on Reliability, 52, 33–37.

Marshall, A.W. and Olkin, I., 1967a. A generalized bivariate exponential distribution. Journal of Applied Probability, 4, 291–302.

Marshall, A.W. and Olkin, I., 1967b. A multivariate exponential distribution. Journal of the American Statistical Association, 62, 30–44.

Mudholkar, G.S., Srivastava, D.K., and Kollia, G.D., 1996. A generalization of the Weibull distribution with application to the analysis of survival data. Journal of the American Statistical Association, 91, 1575–1583.

Pellerey, F., 2008. On univariate and bivariate aging for dependent lifetimes with Archimedean survival copulas. Kybernetika, 44, 795–806.

Pham, H. and Lai, C.-D., 2007. On recent generalizations of the Weibull distribution. IEEE Transactions on Reliability, 56, 454–458.

Philip, G.C., 1974. A generalized EOQ model for items with Weibull distribution deterioration. AIIE Transactions, 6, 159–162.

Pinder III, J.E., Wiener, J.G., and Smith, M.H., 1978. The Weibull distribution: A new method of summarizing survivorship data. Ecology, 59, 175–179.

R Core Team, 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Richter, W. D. and Venz, J., 2014. Geometric representations of multivariate skewed elliptically contoured distributions. Chilean Journal of Statistics, 5, 71–90.

Rinne, H., 2008. The Weibull distribution: A handbook. Chapman and Hall/CRC.

Saraiva, E. F. and Suzuki, A. K., 2017. Bayesian computational methods for estimation of two-parameters Weibull distribution in presence of right-censored data. Chilean Journal of Statistics, 8, 25-43.

Sarkar, S.K., 1987. A continuous bivariate exponential distribution. Journal of the American Statistical Association, 82, 667–675.

Stevens, M. and Smulders, P., 1979. The estimation of the parameters of the Weibull wind speed distribution for wind energy utilization purposes. Wind Engineering, 3, 132–145.

Su, Y.S. and Yajima, M., 2012. R2jags: A package for running JAGS from R. R package version 0.03-08.

Thoman, D.R., Bain, L.J., and Antle, C.E., 1969. Inferences on the parameters of the Weibull distribution. Technometrics, 11, 445–460.

Weibull, W., 1951. A statistical distribution function of wide applicability. Journal of Applied Mechanics, 18, 293–297.

Appendix

ML estimators with complete data

From Equations (5) and (6), the likelihood function for $\boldsymbol{\theta} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \sigma)$ assuming a TMOW distribution and a random sample of size n of the lifetimes X_1, X_2 and X_3 is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [f_1(\boldsymbol{x}_i)]^{r_{1i}} \prod_{i=1}^{n} [f_2(\boldsymbol{x}_i)]^{r_{2i}} \prod_{i=1}^{n} [f_3(\boldsymbol{x}_i)]^{r_{3i}} \prod_{i=1}^{n} [f_4(\boldsymbol{x}_i)]^{r_{4i}} \prod_{i=1}^{n} [f_5(\boldsymbol{x}_i)]^{r_{5i}} \prod_{i=1}^{n} [f_6(\boldsymbol{x}_i)]^{r_{6i}} \times \prod_{i=1}^{n} [f_7(\boldsymbol{x}_i)]^{r_{7i}}. \tag{11}$$

where $\boldsymbol{x}_i = (x_{1i}, x_{2i}, x_{3i})$, for $i = 1, \ldots, n$; $f_1(\boldsymbol{x}_i)$, $f_2(\boldsymbol{x}_i)$ and $f_3(\boldsymbol{x}_i)$ are given in Equation (5). In this way, the likelihood function stated in Equation (11) can be rewritten as

$$\begin{aligned} L(\boldsymbol{\theta}) = {} & \left(\lambda_{14}\lambda_2\lambda_3\sigma^3\right)^{\sum_{i=1}^{n} r_{1i}} \prod_{i=1}^{n} (x_{1i}x_{2i}x_{3i})^{r_{1i}(\sigma-1)} \exp\left\{ - \lambda_{14} \sum_{i=1}^{n} r_{1i}x_{1i}^{\sigma} - \lambda_2 \sum_{i=1}^{n} r_{1i}x_{2i}^{\sigma} \right. \\ & \left. -\lambda_3 \sum_{i=1}^{n} r_{1i}x_{3i}^{\sigma} \right\} \left(\lambda_1\lambda_{24}\lambda_3\sigma^3\right)^{\sum_{i=1}^{n} r_{2i}} \prod_{i=1}^{n} (x_{1i}x_{2i}x_{3i})^{r_{2i}(\sigma-1)} \exp\left\{ - \lambda_1 \sum_{i=1}^{n} r_{2i}x_{1i}^{\sigma} \right. \\ & \left. -\lambda_{24} \sum_{i=1}^{n} r_{2i}x_{2i}^{\sigma} - \lambda_3 \sum_{i=1}^{n} r_{2i}x_{3i}^{\sigma} \right\} \left(\lambda_1\lambda_2\lambda_{34}\sigma^3\right)^{\sum_{i=1}^{n} r_{3i}} \prod_{i=1}^{n} (x_{1i}x_{2i}x_{3i})^{r_{3i}(\sigma-1)} \\ & \times \exp\left\{ -\lambda_1 \sum_{i=1}^{n} r_{3i}x_{1i}^{\sigma} - \lambda_2 \sum_{i=1}^{n} r_{3i}x_{2i}^{\sigma} - \lambda_{34} \sum_{i=1}^{n} r_{3i}x_{3i}^{\sigma} \right\} \left(\lambda_1\lambda_4\sigma^2\right)^{\sum_{i=1}^{n} r_{4i}} \\ & \times \prod_{i=1}^{n} (x_{1i}x_i)^{r_{4i}(\sigma-1)} \exp\left\{ -\lambda_1 \sum_{i=1}^{n} r_{4i}x_{1i}^{\sigma} - (\lambda - \lambda_1) \sum_{i=1}^{n} r_{4i}x_i^{\sigma} \right\} \left(\lambda_2\lambda_4\sigma^2\right)^{\sum_{i=1}^{n} r_{5i}} \\ & \times \prod_{i=1}^{n} (x_{2i}x_i)^{r_{5i}(\sigma-1)} \exp\left\{ -\lambda_2 \sum_{i=1}^{n} r_{5i}x_{2i}^{\sigma} - (\lambda - \lambda_2) \sum_{i=1}^{n} r_{5i}x_i^{\sigma} \right\} \left(\lambda_3\lambda_4\sigma^2\right)^{\sum_{i=1}^{n} r_{6i}} \\ & \times \prod_{i=1}^{n} (x_{3i}x_i)^{r_{6i}(\sigma-1)} \exp\left\{ -\lambda_3 \sum_{i=1}^{n} r_{6i}x_{3i}^{\sigma} - (\lambda - \lambda_3) \sum_{i=1}^{n} r_{6i}x_i^{\sigma} \right\} (\lambda_4\sigma)^{\sum_{i=1}^{n} r_{7i}} \\ & \times \prod_{i=1}^{n} x_i^{r_{7i}(\sigma-1)} \exp\left\{ -\lambda \sum_{i=1}^{n} r_{7i}x_i^{\sigma} \right\}. \end{aligned} \tag{12}$$

The ML estimators for the parameters λ_1, λ_2, λ_3, λ_4 and σ are obtained solving the equations $\partial\ell/\partial\lambda_1 = 0$, $\partial\ell/\partial\lambda_2 = 0$, $\partial\ell/\partial\lambda_3 = 0$, $\partial\ell/\partial\lambda_4 = 0$ and $\partial\ell/\partial\sigma = 0$. From the log-likelihood given in Equation (7), the first derivatives of $\ell(\boldsymbol{\theta})$ with respect to λ_1, λ_2, λ_3,

λ_4 and σ are given respectively by

$$
\begin{aligned}
\frac{\partial \ell}{\partial \sigma} &= \frac{3}{\sigma}\sum_{i=1}^{n}[r_{1i}+r_{2i}+r_{3i}] + \frac{2}{\sigma}\sum_{i=1}^{n}[r_{4i}+r_{5i}+r_{6i}] + \frac{1}{\sigma}\sum_{i=1}^{n} r_{7i} \\
&\quad + \sum_{i=1}^{n}[r_{1i}+r_{2i}+r_{3i}]\log(x_{1i},x_{2i},x_{3i}) + \sum_{i=1}^{n} r_{4i}\log(x_{1i},x_i) \\
&\quad + \sum_{i=1}^{n} r_{5i}\log(x_{2i},x_i) + \sum_{i=1}^{n} r_{6i}\log(x_{3i},x_i) \\
&\quad + \sum_{i=1}^{n} r_{7i}\log(x_i) - \sum_{i=1}^{n}[\lambda_{14}r_{1i}+\lambda_1(r_{2i}+r_{3i}+r_{4i})]x_{1i}^{\sigma}\log(x_{1i}) \\
&\quad - \sum_{i=1}^{n}[\lambda_{24}r_{2i}+\lambda_2(r_{1i}+r_{3i}+r_{5i})]x_{2i}^{\sigma}\log(x_{2i}) \\
&\quad - \sum_{i=1}^{n}[\lambda_{34}r_{3i}+\lambda_3(r_{1i}+r_{2i}+r_{6i})]x_{1i}^{\sigma}\log(x_{1i}) \\
&\quad - \sum_{i=1}^{n}[(\lambda-\lambda_1)r_{4i}+(\lambda-\lambda_2)r_{5i}+(\lambda-\lambda_3)r_{6i}+\lambda r_{7i}]x_i^{\sigma}\log(x_i), \\
\frac{\partial \ell}{\partial \lambda_1} &= \frac{1}{\lambda_{14}}\sum_{i=1}^{n} r_{1i} + \frac{1}{\lambda_1}\sum_{i=1}^{n}[r_{2i}+r_{3i}+r_{4i}] - \sum_{i=1}^{n}\left[(r_{1i}+r_{2i}+r_{3i}+r_{4i})x_{1i}^{\sigma}\right] \\
&\quad - \sum_{i=1}^{n}\left[(r_{5i}+r_{6i}+r_{7i})x_i^{\sigma}\right], \\
\frac{\partial \ell}{\partial \lambda_2} &= \frac{1}{\lambda_{24}}\sum_{i=1}^{n} r_{2i} + \frac{1}{\lambda_2}\sum_{i=1}^{n}[r_{1i}+r_{3i}+r_{5i}] - \sum_{i=1}^{n}\left[(r_{1i}+r_{2i}+r_{3i}+r_{5i})x_{2i}^{\sigma}\right] \\
&\quad - \sum_{i=1}^{n}\left[(r_{4i}+r_{6i}+r_{7i})x_i^{\sigma}\right], \\
\frac{\partial \ell}{\partial \lambda_3} &= \frac{1}{\lambda_{34}}\sum_{i=1}^{n} r_{3i} + \frac{1}{\lambda_3}\sum_{i=1}^{n}[r_{1i}+r_{2i}+r_{6i}] - \sum_{i=1}^{n}\left[(r_{1i}+r_{2i}+r_{3i}+r_{6i})x_{3i}^{\sigma}\right] \\
&\quad - \sum_{i=1}^{n}\left[(r_{4i}+r_{5i}+r_{7i})x_i^{\sigma}\right], \\
\frac{\partial \ell}{\partial \lambda_4} &= \frac{1}{\lambda_{14}}\sum_{i=1}^{n} r_{1i} + \frac{1}{\lambda_{24}}\sum_{i=1}^{n} r_{2i} + \frac{1}{\lambda_{34}}\sum_{i=1}^{n} r_{3i} + \frac{1}{\lambda_4}\sum_{i=1}^{n}[r_{4i}+r_{5i}+r_{6i}+r_{7i}] \\
&\quad - \sum_{i=1}^{n}\left[r_{1i}x_{1i}^{\sigma}+r_{2i}x_{2i}^{\sigma}+r_{3i}x_{3i}^{\sigma}\right] - \sum_{i=1}^{n}\left[(r_{4i}+r_{5i}+r_{6i}+r_{7i})x_i^{\sigma}\right].
\end{aligned}
$$

Under standard asymptotic ML theory, confidence intervals and hypothesis tests for $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and σ could be obtained from the asymptotic normality of the ML estimators $\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4$ and $\hat{\sigma}$, that is, $\widehat{\boldsymbol{\theta}} = (\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4, \hat{\sigma}) \sim \mathrm{N}_5(\boldsymbol{\theta}, \boldsymbol{\Sigma}^{-1})$, where N_5 denotes a multivariate normal distribution of dimension 5 assuming large sample sizes and $\boldsymbol{\Sigma}$ is the observed Fisher information matrix given by

$$\boldsymbol{\Sigma} = \begin{pmatrix} -\frac{\partial^2 \ell}{\partial \lambda_1^2} & -\frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_2} & -\frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_3} & -\frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_4} & -\frac{\partial^2 \ell}{\partial \lambda_1 \partial \sigma} \\ -\frac{\partial^2 \ell}{\partial \lambda_2 \partial \lambda_1} & -\frac{\partial^2 \ell}{\partial \lambda_2^2} & -\frac{\partial^2 \ell}{\partial \lambda_2 \partial \lambda_3} & -\frac{\partial^2 \ell}{\partial \lambda_2 \partial \lambda_4} & -\frac{\partial^2 \ell}{\partial \lambda_2 \partial \sigma} \\ -\frac{\partial^2 \ell}{\partial \lambda_3 \partial \lambda_1} & -\frac{\partial^2 \ell}{\partial \lambda_3 \partial \lambda_2} & -\frac{\partial^2 \ell}{\partial \lambda_3^2} & -\frac{\partial^2 \ell}{\partial \lambda_3 \partial \lambda_4} & -\frac{\partial^2 \ell}{\partial \lambda_3 \partial \sigma} \\ -\frac{\partial^2 \ell}{\partial \lambda_4 \partial \lambda_1} & -\frac{\partial^2 \ell}{\partial \lambda_4 \partial \lambda_2} & -\frac{\partial^2 \ell}{\partial \lambda_4 \partial \lambda_3} & -\frac{\partial^2 \ell}{\partial \lambda_4^4} & -\frac{\partial^2 \ell}{\partial \lambda_4 \partial \sigma} \\ -\frac{\partial^2 \ell}{\partial \sigma \partial \lambda_1} & -\frac{\partial^2 \ell}{\partial \sigma \partial \lambda_2} & -\frac{\partial^2 \ell}{\partial \sigma \partial \lambda_3} & -\frac{\partial^2 \ell}{\partial \sigma \partial \lambda_4} & -\frac{\partial^2 \ell}{\partial \sigma^2} \end{pmatrix}, \tag{13}$$

where all components of Equation (13) are calculated at the obtained ML estimators for the parameters of the model. The second derivatives of the log-likelihood function $\ell(\boldsymbol{\theta})$ required in the observed Fisher information matrix are given by

$$\frac{\partial^2 \ell}{\partial \lambda_1^2} = -\frac{1}{\lambda_{14}^2} \sum_{i=1}^{n} r_{1i} - \frac{1}{\lambda_1^2} \sum_{i=1}^{n} [r_{2i} + r_{3i} + r_{4i}],$$

$$\frac{\partial^2 \ell}{\partial \lambda_2^2} = -\frac{1}{\lambda_{24}^2} \sum_{i=1}^{n} r_{2i} - \frac{1}{\lambda_2^2} \sum_{i=1}^{n} [r_{1i} + r_{3i} + r_{5i}],$$

$$\frac{\partial^2 \ell}{\partial \lambda_3^2} = -\frac{1}{\lambda_{34}^2} \sum_{i=1}^{n} r_{3i} - \frac{1}{\lambda_3^2} \sum_{i=1}^{n} [r_{1i} + r_{2i} + r_{6i}],$$

$$\frac{\partial^2 \ell}{\partial \lambda_4^2} = -\frac{1}{\lambda_{14}^2} \sum_{i=1}^{n} r_{1i} - \frac{1}{\lambda_{24}^2} \sum_{i=1}^{n} r_{2i} - \frac{1}{\lambda_{34}^2} \sum_{i=1}^{n} r_{3i} - \frac{1}{\lambda_4^2} \sum_{i=1}^{n} [r_{4i} + r_{5i} + r_{6i} + r_{7i}],$$

$$\frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_4} = \frac{\partial^2 \ell}{\partial \lambda_4 \partial \lambda_1} = -\frac{1}{\lambda_{14}^2} \sum_{i=1}^{n} r_{1i},$$

$$\frac{\partial^2 \ell}{\partial \lambda_2 \partial \lambda_4} = \frac{\partial^2 \ell}{\partial \lambda_4 \partial \lambda_2} = -\frac{1}{\lambda_{24}^2} \sum_{i=1}^{n} r_{2i},$$

$$\frac{\partial^2 \ell}{\partial \lambda_3 \partial \lambda_4} = \frac{\partial^2 \ell}{\partial \lambda_4 \partial \lambda_3} = -\frac{1}{\lambda_{34}^2} \sum_{i=1}^{n} r_{3i},$$

$$\frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_2} = \frac{\partial^2 \ell}{\partial \lambda_2 \partial \lambda_1} = \frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_3} = \frac{\partial^2 \ell}{\partial \lambda_3 \partial \lambda_1} = \frac{\partial^2 \ell}{\partial \lambda_2 \partial \lambda_3} = \frac{\partial^2 \ell}{\partial \lambda_3 \partial \lambda_2} = 0,$$

$$\frac{\partial^2 \ell}{\partial \lambda_1 \partial \sigma} = \frac{\partial^2 \ell}{\partial \sigma \partial \lambda_1} = -\sum_{i=1}^{n} [(r_{1i} + r_{2i} + r_{3i} + r_{4i}) x_{1i}^{\sigma} \log(x_{1i})] - \sum_{i=1}^{n} [(r_{5i} + r_{6i} + r_{7i}) x_i^{\sigma} \log(x_i)],$$

$$\frac{\partial^2\ell}{\partial\lambda_2\partial\sigma}=\frac{\partial^2\ell}{\partial\sigma\partial\lambda_2}=-\sum_{i=1}^{n}\left[(r_{1i}+r_{2i}+r_{3i}+r_{5i})x_{2i}^{\sigma}\log(x_{2i})\right]-\sum_{i=1}^{n}\left[(r_{4i}+r_{6i}+r_{7i})x_{i}^{\sigma}\log(x_{i})\right],$$

$$\frac{\partial^2\ell}{\partial\lambda_3\partial\sigma}=\frac{\partial^2\ell}{\partial\sigma\partial\lambda_3}=-\sum_{i=1}^{n}\left[(r_{1i}+r_{2i}+r_{3i}+r_{6i})x_{2i}^{\sigma}\log(x_{3i})\right]-\sum_{i=1}^{n}\left[(r_{4i}+r_{5i}+r_{7i})x_{i}^{\sigma}\log(x_{i})\right],$$

$$\begin{aligned}\frac{\partial^2\ell}{\partial\lambda_4\partial\sigma}=\frac{\partial^2\ell}{\partial\sigma\partial\lambda_4}=&-\sum_{i=1}^{n}\left[r_{1i}x_{1i}^{\sigma}\log(x_{1i})+r_{2i}x_{2i}^{\sigma}\log(x_{2i})+r_{3i}x_{3i}^{\sigma}\log(x_{3i})\right]\\&-\sum_{i=1}^{n}\left[(r_{4i}+r_{5i}+r_{6i}+r_{7i})x_{i}^{\sigma}\log(x_{i})\right],\end{aligned}$$

$$\begin{aligned}\frac{\partial^2\ell}{\partial\sigma^2}=&-\frac{3}{\sigma^2}\sum_{i=1}^{n}[r_{1i}+r_{2i}+r_{3i}]-\frac{2}{\sigma^2}\sum_{i=1}^{n}[r_{4i}+r_{5i}+r_{6i}]-\frac{1}{\sigma^2}\sum_{i=1}^{n}r_{7i}\\&-\sum_{i=1}^{n}[\lambda_{14}r_{1i}+\lambda_1(r_{2i}+r_{3i}+r_{4i})]2x_{1i}^{\sigma}\log(x_{1i})\\&-\sum_{i=1}^{n}[\lambda_{24}r_{2i}+\lambda_2(r_{1i}+r_{3i}+r_{5i})]2x_{2i}^{\sigma}\log(x_{2i})\\&-\sum_{i=1}^{n}[\lambda_{34}r_{3i}+\lambda_3(r_{1i}+r_{2i}+r_{6i})]2x_{1i}^{\sigma}\log(x_{1i})\\&-\sum_{i=1}^{n}[(\lambda-\lambda_1)r_{4i}+(\lambda-\lambda_2)r_{5i}+(\lambda-\lambda_3)r_{6i}+\lambda r_{7i}]2x_{i}^{\sigma}\log(x_{i}).\end{aligned}$$

Terms of the likelihood function with censored data

From Equations (9) and (6), we obtain expressions for the terms of the likelihood function defined in Equation (8) as

(a)

$$\begin{aligned}\prod_{i\in B_1}f(\boldsymbol{t}_i)=\Big[&\prod_{i=1}^{n}[f_1(\boldsymbol{t}_i)]^{r_{1i}}\prod_{i=1}^{n}[f_2(\boldsymbol{t}_i)]^{r_{2i}}\prod_{i=1}^{n}[f_3(\boldsymbol{t}_i)]^{r_{3i}}\prod_{i=1}^{n}[f_4(\boldsymbol{t}_i)]^{r_{4i}}\\&\times\prod_{i=1}^{n}[f_5(\boldsymbol{t}_i)]^{r_{5i}}\prod_{i=1}^{n}[f_6(\boldsymbol{t}_i)]^{r_{6i}}\prod_{i=1}^{n}[f_7(\boldsymbol{t}_i)]^{r_{7i}}\Big]^{\delta_{1i}\delta_{2i}\delta_{3i}},\end{aligned}$$

where $f_1(\boldsymbol{t}_i)$, $f_2(\boldsymbol{t}_i)$, $f_3(\boldsymbol{t}_i)$, $f_4(\boldsymbol{t}_i)$, $f_5(\boldsymbol{t}_i)$, $f_6(\boldsymbol{t}_i)$ and $f_7(\boldsymbol{t}_i)$ are defined by Equation (5).

(b)

$$\begin{aligned}\prod_{i\in B_2}\left(-\frac{\partial S(\boldsymbol{t}_i)}{\partial \boldsymbol{t}_{1i}}\right)=\Big[&\prod_{i=1}^{n}[g_{11}(\boldsymbol{t}_i)]^{r_{1i}}\prod_{i=1}^{n}[g_{12}(\boldsymbol{t}_i)]^{r_{2i}}\prod_{i=1}^{n}[g_{13}(\boldsymbol{t}_i)]^{r_{3i}}\prod_{i=1}^{n}[g_{14}(\boldsymbol{t}_i)]^{r_{4i}}\\&\times\prod_{i=1}^{n}[g_{15}(\boldsymbol{t}_i)]^{r_{5i}}\prod_{i=1}^{n}[g_{16}(\boldsymbol{t}_i)]^{r_{6i}}\prod_{i=1}^{n}[g_{17}(\boldsymbol{t}_i)]^{r_{7i}}\Big]^{\delta_{1i}(1-\delta_{2i})(1-\delta_{3i})},\end{aligned}$$

where

$$\begin{aligned}
g_{11}(\boldsymbol{t}_i) &= \lambda_{14}\sigma t_{1i}^{\sigma-1}\exp\{-\lambda_{14}t_{1i}^{\sigma} - \lambda_2 t_{2i}^{\sigma} - \lambda_3 t_{3i}^{\sigma}\},\\
g_{12}(\boldsymbol{t}_i) &= \lambda_1\sigma t_{1i}^{\sigma-1}\exp\{-\lambda_1 t_{1i}^{\sigma} - \lambda_{24}t_{2i}^{\sigma} - \lambda_3 t_{3i}^{\sigma}\},\\
g_{13}(\boldsymbol{t}_i) &= \lambda_1\sigma t_{1i}^{\sigma-1}\exp\{-\lambda_1 t_{1i}^{\sigma} - \lambda_2 t_{2i}^{\sigma} - \lambda_{34}t_{3i}^{\sigma}\},\\
g_{14}(\boldsymbol{t}_i) &= \lambda_1\sigma t_{1i}^{\sigma-1}\exp\{-\lambda_1 t_{1i}^{\sigma} - (\lambda-\lambda_1)t_{2i}^{\sigma}\},\\
g_{15}(\boldsymbol{t}_i) &= (\lambda-\lambda_2)\sigma t_{1i}^{\sigma-1}\exp\{-\lambda_2 t_{2i}^{\sigma} - (\lambda-\lambda_2)t_{1i}^{\sigma}\},\\
g_{16}(\boldsymbol{t}_i) &= (\lambda-\lambda_3)\sigma t_{1i}^{\sigma-1}\exp\{-\lambda_3 t_{3i}^{\sigma} - (\lambda-\lambda_2)t_{1i}^{\sigma}\},\\
g_{17}(\boldsymbol{t}_i) &= \lambda\sigma t_{1i}^{\sigma-1}\exp\{-\lambda t_{1i}^{\sigma}\}.
\end{aligned}$$

(c)

$$\begin{aligned}
\prod\nolimits_{i\in B_3}\left(-\frac{\partial S(\boldsymbol{t}_i)}{\partial \boldsymbol{t}_{2i}}\right) = \Bigg[&\prod\nolimits_{i=1}^{n}[g_{21}(\boldsymbol{t}_i)]^{r_{1i}}\prod\nolimits_{i=1}^{n}[g_{22}(\boldsymbol{t}_i)]^{r_{2i}}\prod\nolimits_{i=1}^{n}[g_{23}(\boldsymbol{t}_i)]^{r_{3i}}\prod\nolimits_{i=1}^{n}[g_{24}(\boldsymbol{t}_i)]^{r_{4i}}\\
&\times\prod\nolimits_{i=1}^{n}[g_{25}(\boldsymbol{t}_i)]^{r_{5i}}\prod\nolimits_{i=1}^{n}[g_{26}(\boldsymbol{t}_i)]^{r_{6i}}\prod\nolimits_{i=1}^{n}[g_{27}(\boldsymbol{t}_i)]^{r_{7i}}\Bigg]^{(1-\delta_{1i})\delta_{2i}(1-\delta_{3i})},
\end{aligned}$$

where

$$\begin{aligned}
g_{21}(\boldsymbol{t}_i) &= \lambda_2\sigma t_{2i}^{\sigma-1}\exp\{-\lambda_{14}t_{1i}^{\sigma} - \lambda_2 t_{2i}^{\sigma} - \lambda_3 t_{3i}^{\sigma}\},\\
g_{22}(\boldsymbol{t}_i) &= \lambda_{24}\sigma t_{2i}^{\sigma-1}\exp\{-\lambda_1 t_{1i}^{\sigma} - \lambda_{24}t_{2i}^{\sigma} - \lambda_3 t_{3i}^{\sigma}\},\\
g_{23}(\boldsymbol{t}_i) &= \lambda_2\sigma t_{2i}^{\sigma-1}\exp\{-\lambda_1 t_{1i}^{\sigma} - \lambda_2 t_{2i}^{\sigma} - \lambda_{34}t_{3i}^{\sigma}\},\\
g_{24}(\boldsymbol{t}_i) &= (\lambda-\lambda_1)\sigma t_{2i}^{\sigma-1}\exp\{-\lambda_1 t_{1i}^{\sigma} - (\lambda-\lambda_1)t_{i}^{\sigma}\},\\
g_{25}(\boldsymbol{t}_i) &= \lambda_2\sigma t_{2i}^{\sigma-1}\exp\{-\lambda_2 t_{2i}^{\sigma} - (\lambda-\lambda_2)t_{1i}^{\sigma}\}, g_{26}(\boldsymbol{t}_i) = 0, g_{27}(\boldsymbol{t}_i) = 0.
\end{aligned}$$

(d)

$$\begin{aligned}
\prod\nolimits_{i\in B_3}\left(-\frac{\partial S(\boldsymbol{t}_i)}{\partial \boldsymbol{t}_{3i}}\right) = \Bigg[&\prod\nolimits_{i=1}^{n}[g_{31}(\boldsymbol{t}_i)]^{r_{1i}}\prod\nolimits_{i=1}^{n}[g_{32}(\boldsymbol{t}_i)]^{r_{2i}}\prod\nolimits_{i=1}^{n}[g_{33}(\boldsymbol{t}_i)]^{r_{3i}}\prod\nolimits_{i=1}^{n}[g_{34}(\boldsymbol{t}_i)]^{r_{4i}}\\
&\times\prod\nolimits_{i=1}^{n}[g_{35}(\boldsymbol{t}_i)]^{r_{5i}}\prod\nolimits_{i=1}^{n}[g_{36}(\boldsymbol{t}_i)]^{r_{6i}}\prod\nolimits_{i=1}^{n}[g_{37}(\boldsymbol{t}_i)]^{r_{7i}}\Bigg]^{(1-\delta_{1i})(1-\delta_{2i})\delta_{3i}},
\end{aligned}$$

where

$$\begin{aligned}
g_{31}(\boldsymbol{t}_i) &= \lambda_3\sigma t_{3i}^{\sigma-1}\exp\{-\lambda_{14}t_{1i}^{\sigma} - \lambda_2 t_{2i}^{\sigma} - \lambda_3 t_{3i}^{\sigma}\},\\
g_{32}(\boldsymbol{t}_i) &= \lambda_3\sigma t_{3i}^{\sigma-1}\exp\{-\lambda_1 t_{1i}^{\sigma} - \lambda_{24}t_{2i}^{\sigma} - \lambda_3 t_{3i}^{\sigma}\},\\
g_{33}(\boldsymbol{t}_i) &= \lambda_{34}\sigma t_{3i}^{\sigma-1}\exp\{-\lambda_1 t_{1i}^{\sigma} - \lambda_2 t_{2i}^{\sigma} - \lambda_{34}t_{3i}^{\sigma}\},\\
g_{34}(\boldsymbol{t}_i) &= 0, g_{35}(\boldsymbol{t}_i) = 0, g_{37}(\boldsymbol{t}_i) = 0, g_{36}(\boldsymbol{t}_i) = \lambda_3\sigma t_{3i}^{\sigma-1}\exp\{-\lambda_3 t_{3i}^{\sigma} - (\lambda-\lambda_3)t_{i}^{\sigma}\},
\end{aligned}$$

(e)

$$\prod_{i\in B_5}\left(\frac{\partial^2 S(\boldsymbol{t}_i)}{\partial \boldsymbol{t}_{1i}\partial \boldsymbol{t}_{2i}}\right)=\left[\prod_{i=1}^{n}[g_{41}(\boldsymbol{t}_i)]^{r_{1i}}\prod_{i=1}^{n}[g_{42}(\boldsymbol{t}_i)]^{r_{2i}}\prod_{i=1}^{n}[g_{43}(\boldsymbol{t}_i)]^{r_{3i}}\prod_{i=1}^{n}[g_{44}(\boldsymbol{t}_i)]^{r_{4i}}\right.$$
$$\left.\times\prod_{i=1}^{n}[g_{45}(\boldsymbol{t}_i)]^{r_{5i}}\prod_{i=1}^{n}[g_{46}(\boldsymbol{t}_i)]^{r_{6i}}\prod_{i=1}^{n}[g_{47}(\boldsymbol{t}_i)]^{r_{7i}}\right]^{\delta_{1i}\delta_{2i}(1-\delta_{3i})},$$

where

$$\begin{aligned}
g_{41}(\boldsymbol{t}_i)&=\lambda_{14}\lambda_2\sigma^2(t_{1i}t_{2i})^{\sigma-1}\exp\{-\lambda_{14}t_{1i}^{\sigma}-\lambda_2 t_{2i}^{\sigma}-\lambda_3 t_{3i}^{\sigma}\},\\
g_{42}(\boldsymbol{t}_i)&=\lambda_{1}\lambda_{24}\sigma^2(t_{1i}t_{2i})^{\sigma-1}\exp\{-\lambda_{1}t_{1i}^{\sigma}-\lambda_{24} t_{2i}^{\sigma}-\lambda_3 t_{3i}^{\sigma}\},\\
g_{43}(\boldsymbol{t}_i)&=\lambda_{1}\lambda_2\sigma^2(t_{1i}t_{2i})^{\sigma-1}\exp\{-\lambda_{1}t_{1i}^{\sigma}-\lambda_2 t_{2i}^{\sigma}-\lambda_{34} t_{3i}^{\sigma}\},\\
g_{44}(\boldsymbol{t}_i)&=\lambda_{1}\sigma^2(t_{1i}t_{2i})^{\sigma-1}(\lambda-\lambda_1)\exp\{-\lambda_{1}t_{1i}^{\sigma}-(\lambda-\lambda_1) t_{2i}^{\sigma}\},\\
g_{45}(\boldsymbol{t}_i)&=(\lambda-\lambda_2)\lambda_2\sigma^2(t_{1i}t_{2i})^{\sigma-1}\exp\{-\lambda_{2}t_{2i}^{\sigma}-(\lambda-\lambda_2) t_{1i}^{\sigma}\}, g_{46}(\boldsymbol{t}_i)=0, g_{47}(\boldsymbol{t}_i)=0.
\end{aligned}$$

(f)

$$\prod_{i\in B_6}\left(\frac{\partial^2 S(\boldsymbol{t}_i)}{\partial \boldsymbol{t}_{1i}\partial \boldsymbol{t}_{3i}}\right)=\left[\prod_{i=1}^{n}[g_{51}(\boldsymbol{t}_i)]^{r_{1i}}\prod_{i=1}^{n}[g_{52}(\boldsymbol{t}_i)]^{r_{2i}}\prod_{i=1}^{n}[g_{53}(\boldsymbol{t}_i)]^{r_{3i}}\prod_{i=1}^{n}[g_{54}(\boldsymbol{t}_i)]^{r_{4i}}\right.$$
$$\left.\times\prod_{i=1}^{n}[g_{55}(\boldsymbol{t}_i)]^{r_{5i}}\prod_{i=1}^{n}[g_{56}(\boldsymbol{t}_i)]^{r_{6i}}\prod_{i=1}^{n}[g_{57}(\boldsymbol{t}_i)]^{r_{7i}}\right]^{\delta_{1i}(1-\delta_{2i})\delta_{3i}},$$

where

$$\begin{aligned}
g_{51}(\boldsymbol{t}_i)&=\lambda_{14}\lambda_3\sigma^2(t_{1i}t_{3i})^{\sigma-1}\exp\{-\lambda_{14}t_{1i}^{\sigma}-\lambda_2 t_{2i}^{\sigma}-\lambda_3 t_{3i}^{\sigma}\},\\
g_{52}(\boldsymbol{t}_i)&=\lambda_{1}\lambda_3\sigma^2(t_{1i}t_{3i})^{\sigma-1}\exp\{-\lambda_{1}t_{1i}^{\sigma}-\lambda_{24} t_{2i}^{\sigma}-\lambda_3 t_{3i}^{\sigma}\},\\
g_{53}(\boldsymbol{t}_i)&=\lambda_{1}\lambda_{34}\sigma^2(t_{1i}t_{3i})^{\sigma-1}\exp\{-\lambda_{1}t_{1i}^{\sigma}-\lambda_2 t_{2i}^{\sigma}-\lambda_{34} t_{3i}^{\sigma}\},\\
g_{54}(\boldsymbol{t}_i)&=, g_{55}(\boldsymbol{t}_i)=0, g_{57}(\boldsymbol{t}_i)=0,\\
g_{56}(\boldsymbol{t}_i)&=(\lambda-\lambda_3)\lambda_3\sigma^2(t_{1i}t_{3i})^{\sigma-1}\exp\{-\lambda_{3}t_{3i}^{\sigma}-(\lambda-\lambda_2) t_{1i}^{\sigma}\}.
\end{aligned}$$

(g)

$$\prod_{i\in B_7}\left(\frac{\partial^2 S(\boldsymbol{t}_i)}{\partial \boldsymbol{t}_{2i}\partial \boldsymbol{t}_{3i}}\right)=\left[\prod_{i=1}^{n}[g_{61}(\boldsymbol{t}_i)]^{r_{1i}}\prod_{i=1}^{n}[g_{62}(\boldsymbol{t}_i)]^{r_{2i}}\prod_{i=1}^{n}[g_{63}(\boldsymbol{t}_i)]^{r_{3i}}\prod_{i=1}^{n}[g_{64}(\boldsymbol{t}_i)]^{r_{4i}}\right.$$
$$\left.\times\prod_{i=1}^{n}[g_{65}(\boldsymbol{t}_i)]^{r_{5i}}\prod_{i=1}^{n}[g_{66}(\boldsymbol{t}_i)]^{r_{6i}}\prod_{i=1}^{n}[g_{67}(\boldsymbol{t}_i)]^{r_{7i}}\right]^{(1-\delta_{1i})\delta_{2i}\delta_{3i}},$$

where

$$
\begin{aligned}
g_{61}(\boldsymbol{t}_i) &= \lambda_2\lambda_3\sigma^2(t_{2i}t_{3i})^{\sigma-1}\exp\{-\lambda_{14}t_{1i}^{\sigma} - \lambda_2 t_{2i}^{\sigma} - \lambda_3 t_{3i}^{\sigma}\},\\
g_{62}(\boldsymbol{t}_i) &= \lambda_{24}\lambda_3\sigma^2(t_{2i}t_{3i})^{\sigma-1}\exp\{-\lambda_1 t_{1i}^{\sigma} - \lambda_{24} t_{2i}^{\sigma} - \lambda_3 t_{3i}^{\sigma}\},\\
g_{63}(\boldsymbol{t}_i) &= \lambda_2\lambda_{34}\sigma^2(t_{2i}t_{3i})^{\sigma-1}\exp\{-\lambda_1 t_{1i}^{\sigma} - \lambda_2 t_{2i}^{\sigma} - \lambda_{34} t_{3i}^{\sigma}\},\\
g_{64}(\boldsymbol{t}_i) &= 0, g_{65}(\boldsymbol{t}_i) = 0, g_{66}(\boldsymbol{t}_i) = 0, g_{67}(\boldsymbol{t}_i) = 0.
\end{aligned}
$$

(h)

$$
\begin{aligned}
\prod_{i\in B_8} S(\boldsymbol{t}_i) = &\left[\prod_{i=1}^{n}[S_1(\boldsymbol{t}_i)]^{r_{1i}}\prod_{i=1}^{n}[S_2(\boldsymbol{t}_i)]^{r_{2i}}\prod_{i=1}^{n}[S_3(\boldsymbol{t}_i)]^{r_{3i}}\prod_{i=1}^{n}[S_4(\boldsymbol{t}_i)]^{r_{4i}}\right.\\
&\left.\times\prod_{i=1}^{n}[S_5(\boldsymbol{t}_i)]^{r_{5i}}\prod_{i=1}^{n}[S_6(\boldsymbol{t}_i)]^{r_{6i}}\prod_{i=1}^{n}[S_7(\boldsymbol{t}_i)]^{r_{7i}}\right]^{(1-\delta_{1i})(1-\delta_{2i})(1-\delta_{3i})},
\end{aligned}
$$

where $S_1(\boldsymbol{t}_i)$, $S_2(\boldsymbol{t}_i)$, $S_3(\boldsymbol{t}_i)$, $S_4(\boldsymbol{t}_i)$, $S_5(\boldsymbol{t}_i)$, $S_6(\boldsymbol{t}_i)$ and $S_7(\boldsymbol{t}_i)$ are defined by Equation (4).

Chilean Journal of Statistics
Vol. 11, No. 2, December 2020, 117–136

Multivariate Statistical Inference
Research Paper

On new robust tests for the multivariate normal mean vector with high-dimensional data and applications

Henrique José de Paula Alves[1,*] and Daniel Furtado Ferreira[1]

[1]Department of Statistics, Federal University of Lavras, Lavras, Brazil

(Received: 04 August 2020 · Accepted in final form: 10 December 2020)

Abstract

New alternative tests to the Hotelling T2 and the likelihood ratio tests for the multivariate normal and non-normal population mean vector are proposed here. These new tests are based on the ordinary and robust comedian covariance matrix estimator. The new adapted likelihood ratio test overcomes the high dimensional issue that occurs with both T2 and likelihood ratio tests. The asymptotic and parametric bootstrap distributions for test statistics are used and the performance of these new tests based on normal and non-normal distributions is evaluated through Monte Carlo simulations. Contaminated normal multivariate populations are also considered to evaluate the effects of outliers on test performances. Type I error probabilities and power in all simulations are computed using the R software. The non-robust parametric bootstrap version of the likelihood ratio test performs better and is recommended since it is easy to implement and computationally fast. An application of the proposed new and T2 tests to a real data set is provided. We use an R package of our authorship to perform the tests described here.

Keywords: Bootstrapping · Hotelling and likelihood ratio tests · Types I-II errors.

Mathematics Subject Classification: Primary 62G10 · Secondary 62F05.

1. Introduction

A big challenge in statistics is based on verifying if a p-variate normal mean vector $\boldsymbol{\mu}$ is equal to a known vector $\boldsymbol{\mu}_0$ when the dimensionality p is greater than the sample size n. The Hotelling T^2 test is widely used to test the null hypothesis. However, under non-normal distributions or in the presence of outliers, the use of the Hotelling T^2 test is not recommended. First, this statistic is built under multivariate normality. Second, even under normality, this statistic considers the average sample vector $\overline{\boldsymbol{X}}$ and the covariance matrix $\boldsymbol{S}$ that are strongly influenced by outliers, as in the univariate case (Willems et al., 2002). Third, T^2 cannot be calculated when the number of variables p is greater than or equal to the number of observations n, since the sample covariance matrix, that is present in this statistic, is a singular matrix (Bai and Saranadasa, 1996). In addition, Bai and Saranadasa (1996) noted that the test power based on the T^2 statistic has low power when under these same conditions, as also shown in Pan and Zhou (2011).

*Corresponding author. Email: jpahenrique@gmail.com

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

Another widely used statistical method is the likelihood ratio (LR) test (Ferreira, 2018; Wagala, A., 2020). In addition, testing the null hypothesis on a vector of population means becomes a challenge under non-normal asymmetric distributions or in the presence of outliers. The Hotelling or likelihood ratio tests consider the sample estimators of the vector of means and the covariance matrix in their expressions, which are highly influenced by outliers. Some robust testing proposals can be found by Tiku (1982); Mudholkar and Srivastava (2000); Willems et al. (2002). In contrast, Srivastava and Du (2008); Srivastava (2009); Chen et al. (2010); Lee et al. (2012); Srivastava et al. (2013); Wang et al. (2013); Marozzi (2015) proposed alternative nonparametric tests for the LR test in non-normal populations, the number of variables p is greater than or equal to the number of observations n.

In this article, new statistical tests are proposed for the null hypothesis involving tests on the vector of multivariate population means. The idea is to obtain robust adaptations of the T^2 and LR statistics using robust comedian estimators (Sajesh and Srinivasan, 2012) for the vector of averages and population covariance matrix. The fundamental concept is to replace ordinary estimators with the mean vector and covariance matrix with their respective robust comedian estimators and provide accurate tests for the mean vector considering the parametric bootstrap distribution under the null hypothesis. In addition, for the LR test statistic with the original and robust comedian estimator of the covariance matrix, the determinants (generalized variances) are replaced by the trace operator, which represents the total sample variance. These new tests are potentially more advantageous than the adapted tests mentioned above, since they can perform better under non-normality and in the presence of outliers. In addition, they are computationally easy to implement and apply.

The performance of these proposed new tests is evaluated by Monte Carlo simulations calculating the type I error probabilities and power of the tests. In Section 2, the new proposed tests are introduced. The results regarding the type I error probability and power of the tests are shown in Section 3. The exact binomial test proposed by Cardoso de Oliveira and Ferreira (2010) is evaluated by Monte Carlo simulations. Section 4 applies the results obtained in this work to real data. In Section 5, the conclusions are presented.

2. Methods

2.1 General context

Consider the problem of testing the hypotheses given by

$$\mathrm{H}_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{versus} \quad \mathrm{H}_1\colon \boldsymbol{\mu} \neq \boldsymbol{\mu}_0. \tag{1}$$

In order to do this, let $\boldsymbol{X}_j = [X_{\mathrm{j}1}, \ldots, X_{jp}]^\top$, with $j = 1, \ldots, n$, be a random sample of size n from a p-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Here, n refers to the number of observations, and p refers to the number of variables (dimensionality) in each random vector. In general, the p components in the random vectors are correlated variables, where its $p \times p$ covariance matrix $\boldsymbol{\Sigma}$ is positive definite.

Under the null hypothesis H_0 as in Equation (1) the test statistic Hotelling T^2 is given by

$$T_{\mathrm{c}}^2 = n(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \boldsymbol{S}^{-1}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0). \tag{2}$$

where $\overline{\boldsymbol{X}} = \sum_{\mathrm{j}=1}^{n} \boldsymbol{X}_{\mathrm{j}}/n$ is the sample mean vector, $\boldsymbol{S} = (1/(n-1)) \sum_{\mathrm{j}=1}^{n}(\boldsymbol{X}_{\mathrm{j}} - \overline{\boldsymbol{X}})(\boldsymbol{X}_{\mathrm{j}} - \overline{\boldsymbol{X}})^\top$ is the sample covariance matrix, and n is the sample size. Under H_0 and with the assumption of normality and homoscedastic covariance matrix, the T_c^2 given in Equation (2) follows a Hotelling T^2 distribution given by $(n-1)pF_{p,n-p}/(n-p)$, where $F_{p,n-p}$ is the F distribution with p and $n-p$ degrees of freedom.

Considering the hypotheses given in Equation (1), another statistical method used is the LR test. Let $\boldsymbol{X} \sim \mathrm{N_p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is unknown. Therefore, the LR statistic is given by the expression stated as

$$-2\log(\Lambda) = n[\log(|\boldsymbol{S}+\boldsymbol{H}|) - \log(|\boldsymbol{S}|)], \tag{3}$$

where $\boldsymbol{H} = (\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0)(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top$ and log is the natural logarithm. Consider $\Omega \in \mathrm{R}^s$ the unrestricted parametric space and $\Omega_0 \subseteq \mathrm{R}^r$ the restricted parametric space, with $\Omega_0 \subset \Omega$. In general, under certain conditions of regularity, Equation (3) follows an asymptotic chi-square distribution with $r-s$ degrees of freedom (Ferreira, 2018) under the null hypothesis H_0. Thus, the rejection region of H_0 is given by $\mathrm{R} = \{\boldsymbol{x}| -2\log(\Lambda(\boldsymbol{x})) > \chi^2_{1-\alpha}(r-s)\}$, where α is the nominal significance level and $\chi^2_{1-\alpha}(r-s)$ is the $100(1-\alpha)\%$ percentile of a chi-square distribution with $r-s$ degrees of freedom. In this case, for testing hypothesis about normal mean vector, the degrees of freedom $r-s$ are equal to p.

Here, the proposed new tests based on the modifications of the T^2 statistic defined in Equation (2) and LR statistic stated in Equation (3) are shown. Also, the adopted Monte Carlo simulation procedure to assess their performance is described. For this, consider a sample of size n from the normal p-variate distribution with a mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ to test the null hypothesis H_0 given in Equation (1). In all cases, the original and transformed LR expressed in Equation (3) are applied, consider the trace operator replacing the determinant operator and the robust estimator replacing the traditional estimator of the covariance matrix. Only in the cases where $p < n$, the T^2 test and its modifications that use the comedian estimators. The theoretical justification can be seen in Section 1. To evaluate test performance, first, type I error probabilities are calculated by generating sample sizes from populations under H_0, with a mean vector of $\boldsymbol{\mu}_0$. Second, random samples are generated under H_1, with $\boldsymbol{\mu} \neq \boldsymbol{\mu_0}$. In both cases, samples from normal and non-normal populations are generated. The p-variate Student-t distributions with 5 degrees of freedom for the non-normal distribution case. We also consider contaminated normal (CN) populations for generating outliers. Some factorial combinations of the number of variables p and sample size n are considered.

Without loss of generality, the population covariance matrix $\boldsymbol{\Sigma}$ with the compound symmetry structure given by

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix} = \sigma^2[(1-\rho)\boldsymbol{I} + \rho\boldsymbol{J}] \tag{4}$$

is considered, where $\boldsymbol{J}$ is a $p \times p$ matrix with all entries equal to 1 and $\boldsymbol{I}$ is an identity array of the same order as $\boldsymbol{J}$. Also, without loss of generality, $\sigma^2 = 1$ and $\rho = 0.9$ since the test statistics are invariant under the true covariance structure. The p-variate Student-t distribution is used to pick a distribution that has heavier tails than the multivariate normal distribution and violate the assumptions of the T^2 and LR statistics. The p-variate CN distribution is $\omega\mathrm{N}_1(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1) + (1-\omega)\mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$. Again, without loss of generality, $\omega = 0.9$, $\boldsymbol{\Sigma}_1$ is defined by Equation (4) and $\boldsymbol{\Sigma}_2$ is constructed using the constraint: $|\boldsymbol{\Sigma}_2|/|\boldsymbol{\Sigma}_1| = \Delta$, and thus, $\boldsymbol{\Sigma}_2 = \Delta^{1/p}\boldsymbol{\Sigma}_1$, where $\Delta = 2$. Sample sizes are $n = 10, 50, 70, 100$ and 200 and the nominal significance level α is $\alpha = 5\%$. The number of variables is $p = 2, 5$ and 200 and 2 000 Monte Carlo simulations to evaluate the empirical estimates of the type I error probabilities and power of each test are considered. The parametric bootstrap null distribution are generated with 2 000 resamples from a $\mathrm{N}(\boldsymbol{0}, \boldsymbol{S}^\bullet)$ distribution, where the null hypothesis H_0 is imposed by considering $\boldsymbol{\mu} = \boldsymbol{0}$ to generate the null distribution of the statistic. Also, the covariance matrix used to generate the bootstrap null distribution, given $\boldsymbol{S}^{\bullet}$, is the sample covariance matrix computed by using the traditional or robust comedian estimator in the original sample for traditional and robust bootstrap tests, respectively.

Without loss of generality, to evaluate the type I error probabilities the $\boldsymbol{\mu}_0$ vector under the null hypothesis given in Equation (1) is the p-dimensional $\mathbf{0}$ null vector and the true vector of population mean is also $\boldsymbol{\mu} = \mathbf{0}$. Under the alternative hypothesis H_1 stated in Equation (1), for the power study, the true population mean vector $\boldsymbol{\mu}$ is chosen considering a fixed generalized Mahalanobis distance $\delta(\boldsymbol{\mu}, \boldsymbol{\mu}_0)$ between $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_0$ given by

$$\delta(\boldsymbol{\mu}, \boldsymbol{\mu}_0) = n(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0). \tag{5}$$

In this case, $\boldsymbol{\mu}_0 = \mathbf{0}$, and the true population mean vector is calculated by trial and error in Equation (5) by taken a fixed δ value and the final value is used as a parameter in each of the population distributions considered under H_1. The chosen values from δ are $0, 0.5, 1, 1.5, 3, 5, 10$. Therefore, since the values of the mean vector change as n changes, keeping fixed the value of the distance of Mahalanobis $\delta(\boldsymbol{\mu}, \boldsymbol{\mu}_0) = n(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$, the power does not change as n increases.

Thus, five new tests based on the traditional T^2 statistic defined in Equation (2) and in the LR statistic defined in Equation (3) are proposed, including the traditional and robust versions that use the comedian mean vector and covariance matrix estimators (Falk, 1997; Maronna and Zamar, 2002; Sajesh and Srinivasan, 2012). Some tests are based on parametric bootstrap versions as well as the asymptotic chi-square distribution. However, some asymptotic chi-square tests have not been shown since they did not control the type I error probabilities. The performance of these new tests is evaluated by Monte Carlo simulations. Below, each of the proposed new tests for testing the null hypothesis H_0: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ is described. The new LR test has a chi-square asymptotic distribution with p degrees of freedom, as the original LR test.

One special case is considered regarding the distributions and for some values of n, p and δ. A shifted zero mean exponential distribution with parameter $\boldsymbol{\lambda} = \mathbf{1}_{\mathrm{p}}$, where $\mathbf{1}_{\mathrm{p}}$ is a p-dimensional vector with 1 in all entries. The latter is a case of a skewed distribution. For the exponential distribution, a p-dimension random vector $\boldsymbol{Z}$ is generated from a $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. A p-dimensional random vector $\boldsymbol{Y}$ from this distribution is obtained considering for the ith entry the random variable stated as $Y_i = F^{-1}(\Phi(\mathrm{Z}_i); \lambda_i) - 1/\lambda_i$, for $i = 1, \ldots, p$, where $F^{-1}(\mathrm{x}; \lambda_i)$ is the quantile function of the exponential distribution of parameter λ_i evaluated at x and $\Phi(x)$ is the cumulative distribution function of the standard normal distribution evaluated at x.

2.2 The parametric bootstrap T^2 test

We construct the parametric bootstrap T^2 test, called T^2_{PB} (T2PB), where PB stands for parametric bootstrap, adopting the following steps:

(1) From the original sample, the parameters $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ are estimated, respectively, by $\boldsymbol{S}^*$ and $\overline{\boldsymbol{X}}^*$, where $\boldsymbol{S}^*$ and $\overline{\boldsymbol{X}}^*$ are the traditional sample covariance matrix and vector mean, respectively. The test statistic is computed by

$$T^{*2} = n\left(\overline{\boldsymbol{X}}^* - \boldsymbol{\mu}_0\right)^\top \boldsymbol{S}^{*-1}\left(\overline{\boldsymbol{X}}^* - \boldsymbol{\mu}_0\right). \tag{6}$$

(2) By using the original covariance estimates $\boldsymbol{S}^*$, a random sample of size n is generated from a p-variate normal distribution imposing H_0, that is, by setting $\boldsymbol{\mu} = \boldsymbol{\mu}_0$. Also $\boldsymbol{\Sigma} = \boldsymbol{S}^*$. Therefore, a sample of size n is generated from a $\mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{S}^*)$ distribution.

(3) In each parametric bootstrap sample, the sample mean $\overline{\boldsymbol{X}}_{\mathrm{PB}}$ and the sample covariance matrix $\boldsymbol{S}_{\mathrm{PB}}$ are estimated.

(4) In each parametric bootstrap sample, compute the test statistic by means of

$$T^2_{\mathrm{PB}} = n(\overline{\boldsymbol{X}}_{\mathrm{PB}} - \boldsymbol{\mu}_0)^\top \boldsymbol{S}^{-1}_{\mathrm{PB}}(\overline{\boldsymbol{X}}_{\mathrm{PB}} - \boldsymbol{\mu}_0). \tag{7}$$

(5) Steps (2) to (4) are repeated B times and a set of size $B+1$ is constructed with the test statistic values computed in Equation (7) and the original value calculated in Equation (6). The null distribution of the parametric bootstrap test is constituted by this set. Therefore, if the ith member of this set is represented by T_i^2, for $i = 1, \ldots, B+1$, then the p-value is computed by

$$p\text{-value} = \frac{\sum_{i=1}^{B+1} I(T_i^2 \geq T^{*2})}{B+1}, \tag{8}$$

where $I(T_i^2 \geq T^{*2})$ is the indicator function.

(6) The null hypothesis given in Equation (1) is rejected at the significance level α if the p-value defined in Equation (8) is less than α.

Note that the traditional T^2 test is also considered, by computing the p-value direct from the Hotelling T^2 distribution of the test statistic value obtained in Equation (6). This is named by T2 and is considered the benchmark test.

2.3 The robust parametric bootstrap T^2 test

The robust parametric bootstrap T^2 test, called T^2_{RPB} (T2RPB), in which RPB stands for robust parametric bootstrap, are performed by adopting the same steps described for the previous test. However, some of them are modified as in the following sequence.

In Step 1, the estimators $\boldsymbol{S}^*$ and $\overline{\boldsymbol{X}}^*$ are replaced by comedian estimators $\boldsymbol{S}_{\mathrm{R}}$ and $\overline{\boldsymbol{X}}_{\mathrm{R}}$ the test statistic in the original sample is computed by $T^{*2} = n(\overline{\boldsymbol{X}}_{\mathrm{R}} - \boldsymbol{\mu}_0)^\top \boldsymbol{S}_{\mathrm{R}}^{-1}(\overline{\boldsymbol{X}}_{\mathrm{R}} - \boldsymbol{\mu}_0)$.

In Step 2, the sample of size n is generated from a $\mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{S}_{\mathrm{R}})$ distribution, where again, the null hypothesis is imposed by considering the multivariate normal mean equal to $\boldsymbol{\mu}_0$, the null value of the population mean.

In Step 3, the mean and the sample covariance in each parametric bootstrap sample are denoted, respectively, by $\overline{\boldsymbol{X}}_{\mathrm{RPB}}$ and $\boldsymbol{S}_{\mathrm{RPB}}$.

In Step 4, the test statistic is computed by $T^2_{\mathrm{RPB}} = n(\overline{\boldsymbol{X}}_{\mathrm{RPB}} - \boldsymbol{\mu}_0)^\top \boldsymbol{S}^{-1}_{\mathrm{RPB}}(\overline{\boldsymbol{X}}_{\mathrm{RPB}} - \boldsymbol{\mu}_0)$.

Steps 5 and 6 are identically as described in the previous test, with T_i^2 replaced now by the ith value from the bootstrap null distribution of T^2_{RPB}. In this case, the asymptotic chi-square distribution with p degrees of freedom is not considered as an alternative test, since the corresponding robust T^2 test did not control the type I error probability (results omitted here). More details on the performance of the above tests can be seen in Alves and Ferreira (2019).

We have shown in Section 1, in addition to the problems presented for their use in data following non-normal distributions, that the traditional T^2 test is not valid for high dimensional data ($p > n$) due to the singularity of the sample covariance matrix $\boldsymbol{S}$. The LR test has the same limitations of the T^2 to be implemented in high dimensional data sets and non-normal circumstances. Considering to Ledoit and Wolf (2002) as reference, we propose an alternative test to the LR test that is based on replacing the determinants of the matrices $\boldsymbol{S}$ and $\boldsymbol{S} + \boldsymbol{H}$ for their respective traces. Here, $\boldsymbol{H} = n(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0)(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top$. In this way, we have obtained a new test that applies to high-dimensional ($p > n$) data sets and that maintains the same distributional properties of LR test. The validity of the asymptotic null distribution of the new test with traditional sample estimators of the mean vector and covariance matrix is evaluated by Monte Carlo simulations. Also, the parametric bootstrap and robust parametric bootstrap versions for this latest proposal are built, as we have done for the T^2 test. In the following subsections, we present the new procedures.

2.4 The asymptotic LR trace test

The asymptotic trace version of the LR test, named asymptotic trace LR (ATLR) test, is obtained by directly replacing the determinant given in the expression of the LR by the trace operator tr. Let the null hypothesis be H_0: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$, then the ATLR test statistic is $T^2_{\mathrm{ATLR}} = n\{\log[\mathrm{tr}(\boldsymbol{S}^* + \boldsymbol{H})] - \mathrm{tr}(\boldsymbol{S}^*)\}$, where $\boldsymbol{H} = (\overline{\boldsymbol{X}}^* - \boldsymbol{\mu}_0)(\overline{\boldsymbol{X}}^* - \boldsymbol{\mu}_0)^\top$, that under the null hypothesis H_0 and normality has a chi-square distribution with p degrees of freedom, since the plim (probability limit) of $\boldsymbol{H}$ is $\mathbf{0}$ as $n \to \infty$, still under the null H_0. The null hypothesis should be rejected if the $T^2_{\mathrm{ATLR}} \geq \chi^2_{1-\alpha}(p)$.

2.5 The TLR parametric bootstrap test

There is no guarantee the T^2_{ATLR} has an asymptotic chi-square distribution with p degrees of freedom under H_0 and multivariate normality. To overcome this issue we proposed the TLR parametric bootstrap test, named trace likelihood ratio parametric bootstrap (TLRPB) test. The steps to apply for this test are the same as described previously for T^2_{PB}, except for some details explained as follows.

In Step 1, the estimators $\boldsymbol{S}^*$ and $\overline{\boldsymbol{X}}^*$ are computed in the original sample and the test statistic is $T^{*2} = n\{\log[\mathrm{tr}(\boldsymbol{S}^* + \boldsymbol{H}^*)] - \mathrm{tr}(\boldsymbol{S}^*)\}$, where $\boldsymbol{H}^* = (\overline{\boldsymbol{X}}^* - \boldsymbol{\mu}_0)(\overline{\boldsymbol{X}}^* - \boldsymbol{\mu}_0)^\top$.

In Step 2, the sample of size n is generated from a $\mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{S}^*)$ distribution, where the null hypothesis is imposed by considering the multivariate normal mean equal to $\boldsymbol{\mu}_0$, the null value of the population mean.

In Step 3, the mean and the sample covariance in each parametric bootstrap sample are denoted respectively by $\overline{\boldsymbol{X}}_{\mathrm{TLRPB}}$ and $\boldsymbol{S}_{\mathrm{TLRPB}}$. In Step 4, the test statistic is $T^2_{\mathrm{TLRPB}} = n\{\log[\mathrm{tr}(\boldsymbol{S}_{\mathrm{TLRPB}} + \boldsymbol{H}_{\mathrm{TLRPB}})] - \mathrm{tr}(\boldsymbol{S}_{\mathrm{TLRPB}})\}$, with $\boldsymbol{H}_{\mathrm{TLRPB}} = (\overline{\boldsymbol{X}}_{\mathrm{TLRPB}} - \boldsymbol{\mu}_0)(\overline{\boldsymbol{X}}_{\mathrm{TLRPB}} - \boldsymbol{\mu}_0)^\top$.

Steps 5 and 6 are identically as described in the previous test, with T_i^2 replaced now by the ith value from the bootstrap null distribution of T^2_{TLRPB}.

2.6 The robust TLR parametric bootstrap test

For overcoming problems with outliers the robust parametric bootstrap version of the previous TLRPB, called robust trace likelihood ratio parametric bootstrap (RTLRPB) is constructed. The steps necessary for this test to be applied are the same as the previous steps described for the T^2_{PB}, except for some details explained below.

In Step 1, the comedian estimators $\boldsymbol{S}_{\mathrm{R}}$ and $\overline{\boldsymbol{X}}_{\mathrm{R}}$ are computed in the original sample and the test statistic is $T^{*2} = n\{\log[\mathrm{tr}(\boldsymbol{S}_{\mathrm{R}} + \boldsymbol{H}_{\mathrm{R}})] - \mathrm{tr}(\boldsymbol{S}_{\mathrm{R}})\}$, where $\boldsymbol{H}_{\mathrm{R}} = (\overline{\boldsymbol{X}}_{\mathrm{R}} - \boldsymbol{\mu}_0)(\overline{\boldsymbol{X}}_{\mathrm{R}} - \boldsymbol{\mu}_0)^\top$.

In Step 2, the sample of size n is generated from the $\mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{S}_{\mathrm{R}})$ null distribution.

In Step 3, the comedian sample mean and sample covariance in each parametric bootstrap sample are denoted respectively by $\overline{\boldsymbol{X}}_{\mathrm{RTLRPB}}$ and $\boldsymbol{S}_{\mathrm{RTLRPB}}$.

In Step 4, the test statistic is $T^2_{\mathrm{RTLRPB}} = n\{\log[\mathrm{tr}(\boldsymbol{S}_{\mathrm{RTLRPB}} + \boldsymbol{H}_{\mathrm{RTLRPB}})] - \mathrm{tr}(\boldsymbol{S}_{\mathrm{RTLRPB}})\}$, where $\boldsymbol{H}_{\mathrm{RTLRPB}} = (\overline{\boldsymbol{X}}_{\mathrm{RTLRPB}} - \boldsymbol{\mu}_0)(\overline{\boldsymbol{X}}_{\mathrm{RTLRPB}} - \boldsymbol{\mu}_0)^\top$.

Steps 5 and 6 are identically as described in the previous test, with T_i^2 replaced now by the ith value from the bootstrap null distribution of T^2_{RTLRPB}.

2.7 The exact binomial test

The test type I error probabilities are evaluated by Monte Carlo simulations, and according to Cardoso de Oliveira and Ferreira (2010), these estimates are not error-free. Therefore, an exact binomial test is used to decide whether each of the modified or the original statistical test is considered accurate, liberal or conservative. In this sense, considering a nominal level of significance of 1%, the hypotheses to be tested are defined as

$$\mathrm{H}_0\colon \alpha = 5\% \quad \text{versus} \quad \mathrm{H}_1\colon \alpha \neq 5\%. \tag{9}$$

The statistic of the exact binomial test is given by

$$F_{\mathrm{c}} = \left(\frac{z+1}{N-z}\right)\left(\frac{1-\alpha}{\alpha}\right), \tag{10}$$

where z is the number of rejection of the null hypothesis accounted by one of the tests considering the nominal significance level of α and N is the number of Monte Carlo simulations performed. Under the null hypothesis defined in Equation (9), the F_{c} statistic defined in (10) follows a F distribution with $\nu_1 = 2(N-z)$ and $\nu_2 = 2(z+1)$ degrees of freedom. If the null hypothesis is rejected and the type I error probability is considered significantly less than the nominal level adopted of the 5%, the test can be considered conservative; if the null hypothesis is rejected and the type I error is considered significantly higher than the nominal level adopted of the 5%, the test can be considered liberal; and if the null hypothesis is not rejected the test can be considered accurate.

A computer with a Core-I7 processor with 4 cores and 8 GB of RAM is used. The simulations are performed in `R` with functions developed by the authors, except for the `mvrnomr`, `var`, `pf`, `pchisq`, `ginv` and `covComed` functions of the `MASS`, `statistics` and `robustbase` packages. The execution time of each simulation is 12 hours on average considering small sample sizes (10, 50, and 70) and, in the case of larger sample sizes (100, and 200), the average duration of the simulations is 2 days, regardless of the dimension considered.

3. Monte Carlo simulations

3.1 General context

The performance results for the proposed new tests are presented in two stages. First, the results regarding type I error probability control and power for cases where $p = 2$ and $p = 5$ where shown. Second, the results for the special case of high dimension ($p = 200$) are shown. The performance of these new tests is evaluated considering the multivariate normal, Student-t with 5 degrees of freedom and CN distributions.

3.2 Type I error probabilities

The type I error probabilities for the five new tests proposed via Monte Carlo simulations are shown in Table 1, considering the dimension $p = 2$ at the significance level of $\alpha = 0.05$. The exact binomial test is used to classify these tests as exact, liberal, or conservative (see Section 2). The traditional and ordinary Hotelling T^2 test is also applied in each circumstance and it is invoked as a benchmark test. We note that for all n sample sizes considered, as well as for all evaluated multivariate distributions, the proposed tests T2, T2PB, and T2RPB are exact since they showed test size equal to the nominal significance level α. The traditional T^2 test constituted an exception when considering the bivariate Normal distribution with a sample size of 50. In this case, this test is conservative but still acceptable. The same is not the case for LR adaptations. The ATLR is conservative on all distributions and sample sizes considered. For this test, a substantial loss of power is expected to occur, which is an important fact to be taken into account. The TLRPB test and RTLRPB test (see Section 2), in all the evaluated scenarios, are all accurate. A single exception occurs for the TLRPB test, considering the multivariate normal distribution with $n = 50$, where it shows a liberal performance in the control of the type I error probability. In practice, a test is considered reliable if it has an exact size. Otherwise, if it is conservative, it can be considered acceptable. However, if it is a liberal test, then it must be discarded. It does not appear the case for the TLRPB test, once it showed a unique exception. Therefore, since in circumstances where the normality assumption is violated, the performance regarding the type I error probability control of the proposed new tests is acceptable.

Table 1. Type I error probabilities of the six tests with $\alpha = 5\%$ and p = 2, considering the multivariate normal (N), Student-t with 5 degrees of freedom (t_5) and CN distributions.

Test	Model	n 10	50	70	100	200
T2	N	0.0525^-	0.0380^-	0.0505	0.0485	0.0460
	t_5	0.0450	0.0495	0.0420	0.0485	0.0510
	CN	0.0520	0.0450	0.0420	0.0490	0.0570
T2PB	N	0.0510	0.0475	0.0455	0.0485	0.0475
	t_5	0.0510	0.0430	0.0405	0.0485	0.0560
	CN	0.0535	0.0435	0.0440	0.0430	0.0515
T2RPB	N	0.0430	0.0480	0.0525	0.0450	0.0535
	t_5	0.0430	0.0390	0.0520	0.0500	0.0485
	CN	0.0530	0.0455	0.0525	0.0490	0.0615
ATLR	N	0.0195^-	0.0085^-	0.0120^-	0.0110^-	0.0130^-
	t_5	0.0275^-	0.0085^-	0.0105^-	0.0140^-	0.0145^-
	CN	0.0200^-	0.0125^-	0.0120^-	0.0140^-	0.0120^-
TLRPB test	N	0.0535	0.0940^+	0.0555	0.0445	0.0535
	t_5	0.0500	0.0680	0.0460	0.0520	0.0580
	CN	0.0515	0.0520	0.0485	0.0445	0.0470
RTLPBT	N	0.0560	0.0490	0.0445	0.0430	0.0530
	t_5	0.0415	0.0415	0.0555	0.0510	0.0555
	CN	0.0550	0.0505	0.0480	0.0435	0.0570

$^-$: significantly ($p < 0.01$) less than the nominal significance level of 5%.

To evaluate if the patterns presented in Table 1 are maintained we decided to increase the dimensionality p and keep the sample sizes n fixed. Table 2 presents the results for the empirical type I error probabilities considering the dimension $p = 5$. Similar behavioral pattern to the control of the type I error probabilities showed for the case with $p = 2$ holds. The atypical situation (conservative) for the T^2 test no longer occurs. We note that the T2RPB for the multivariate Student-t distribution with 5 degrees of freedom and $n = 10$ and 200 is conservative in this case. The TLRPB test and RTLRPB test are considered exact tests. The ATLR remains conservative and therefore acceptable. Likewise, it expects that this test substantially loses power for this dimension. Next, we present the results of the power of the tests for these same cases.

3.3 Power

The powerful performance of the proposed new tests is evaluated in the same cases used to evaluate type I error probabilities: distributions, sample sizes, and dimensions. In all circumstances, the power curves are plotted against Mahalanobis distances (δ), given in 5, between the true population vector $\boldsymbol{\mu}$ and the hypothetical mean vector $\boldsymbol{\mu_0}$. These distances have been fixed (see Section 2) to establish the true value of the population mean vector.

Table 2. Type I error probabilities of the six tests with $\alpha = 5\%$ and p = 5, considering the multivariate Normal (N), Student-t with 5 degrees of freedom (t_5) and CN distributions.

Test	Model	n = 10	50	70	100	200
T2	N	0.0525	0.0545	0.0430	0.0500	0.0565
	t_5	0.0420	0.0545	0.0435	0.0440	0.0445
	CN	0.0475	0.0535	0.0465	0.0475	0.0520
T2PB	N	0.0520	0.0545	0.0430	0.0500	0.0560
	t_5	0.0545	0.0505	0.0440	0.0440	0.0465
	CN	0.0475	0.0540	0.0555	0.0470	0.0515
T2RPB	N	0.0505	0.0595	0.0465	0.0460	0.0605
	t_5	0.0315^-	0.0385	0.0430	0.0405	0.0365^-
	CN	0.0430	0.0525	0.0460	0.0475	0.0500
ATLR	N	0.0050^-	0.0000^-	0.0005^-	0.0005^-	0.0005^-
	t_5	0.0000^-	0.0005^-	0.0000^-	0.0005^-	0.0010^-
	CN	0.0005^-	0.0005^-	0.0005^-	0.0000^-	0.0105^-
TLRPB test	N	0.0555	0.0550	0.0490	0.0585	0.0570
	t_5	0.0445	0.0455	0.0610	0.0495	0.0530
	CN	0.0510	0.0575	0.0465	0.0545	0.0525
RTLRPB test	N	0.0555	0.0550	0.0490	0.0585	0.0570
	t_5	0.0490	0.0460	0.0535	0.0455	0.0405
	CN	0.0495	0.0550	0.0425	0.0450	0.0520

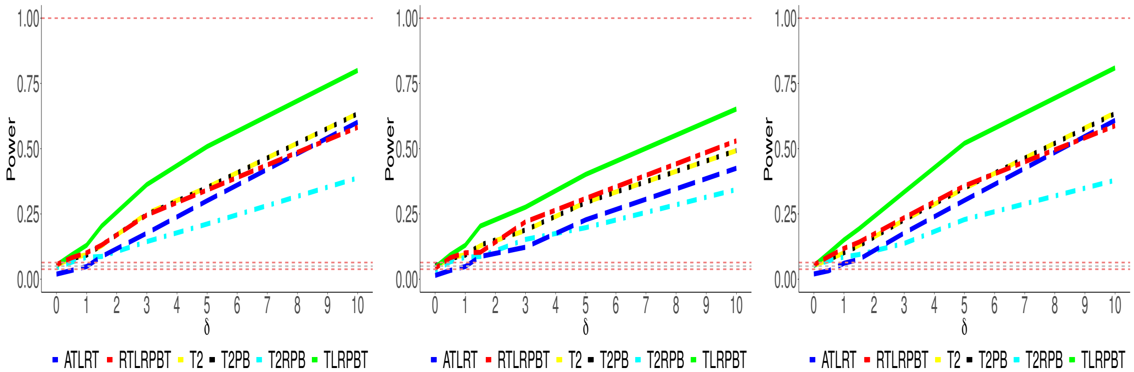
$^-$: significantly ($p < 0.01$) less than the nominal significance level of 5%.

Note in Figure 1 that the TLRPB test showed the best performance among all evaluated tests with $p = 2$. Under non-normality or in the presence of outliers, the performance of this test showed a loss of power only when the multivariate Student-t distribution with 5 degrees of freedom is considered (Figure 1 (b)). We also noticed that the asymptotic version ATLR has lower power as expected; see Tables 1 and 2.

In order to verify if the patterns observed in Figure 1 remain the dimensionality is fixed in $p = 2$ and the multivariate distributions are the same, but the number of observations varied in $n = 50$, 70, 100 and 200 for generating the power curves by Monte Carlo simulations. Under these circumstances, the behavioral power patterns shown in Figure 1 remain the same (Figures 2, 3, 4 and 5). The only exceptions are for the multivariate Student-t distribution with 5 degrees of freedom (Figures 2(b), 3(b), 4(b) and 5(b)). In this case, the performance of the TLRPB test and RTLRPB test is equivalent and higher than the performance of the other tests.

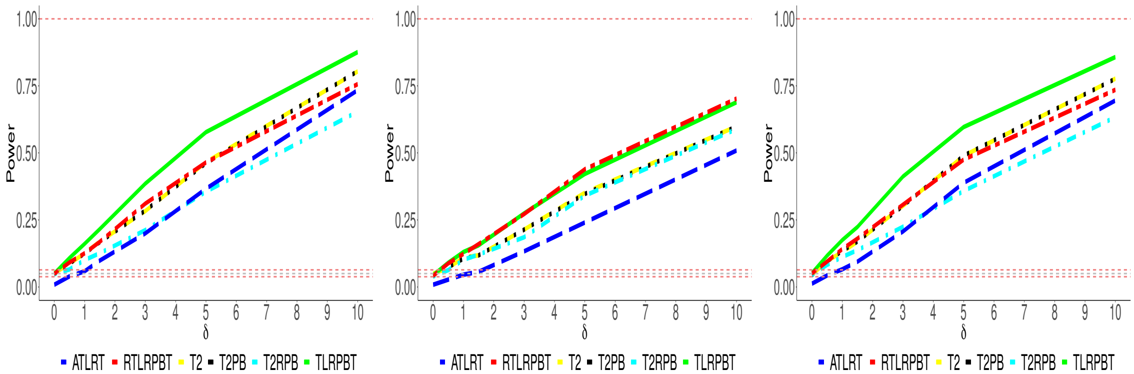
In general, for this dimension ($p = 2$), the parametric bootstrap version (TLRPB test) performed better when compared to the other tests. Also, all tests show a very robust behavioral pattern, since they control the type I error probabilities and show higher power when compared with the multivariate normal case.

We decide to increase the dimension to $p = 5$ and to maintain the same sample sizes n and distributions. Similar performance of all tests that presented at $p = 2$ remains. Considering the dimension $p = 5$, Figures 6, 7, 8, 9 and 10 show the results obtained for the power of the tests. Figures 6(b), 7(b), 8(b), 9(b) and 10(b) show that power of the parametric bootstrap version (TLRPB test) and the robust parametric bootstrap version (RTLRPB test) are equivalent and higher than the other tests in the multivariate Student-t distribution with 5 degrees of freedom. In general, the TLRPB test performs better. The ATLR continues to show substantial low power, once it has shown to be a conservative test, this pattern of behavior is expected; see Tables 1 and 2.



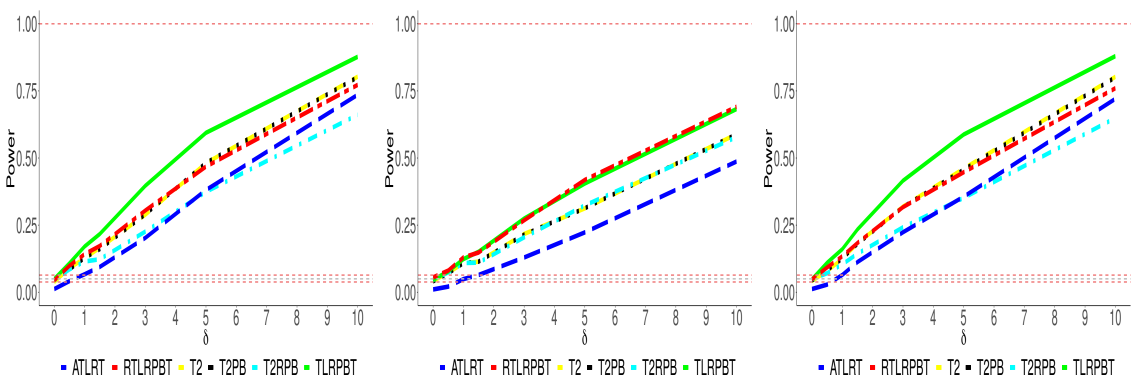
(a) Normal distribution (b) t_5 distribution (c) CN distribution

Figure 1. Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 10$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



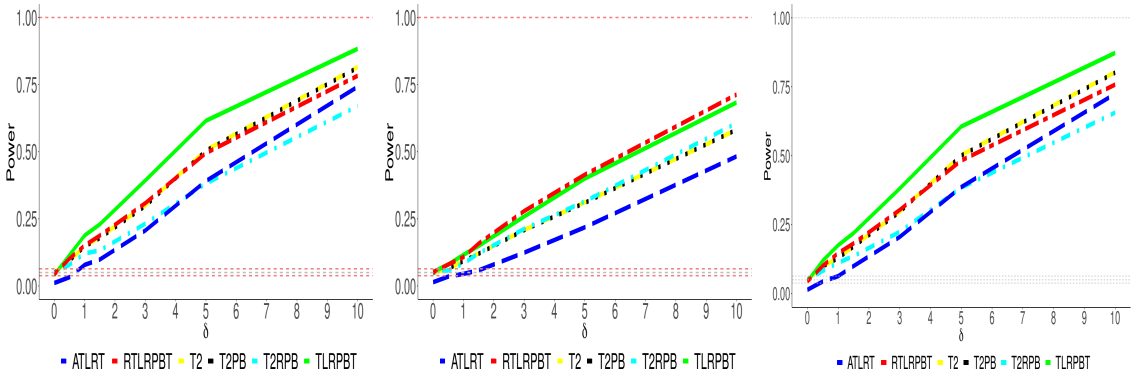
(a) Normal distribution (b) t_5 distribution (c) CN distribution

Figure 2. Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 50$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



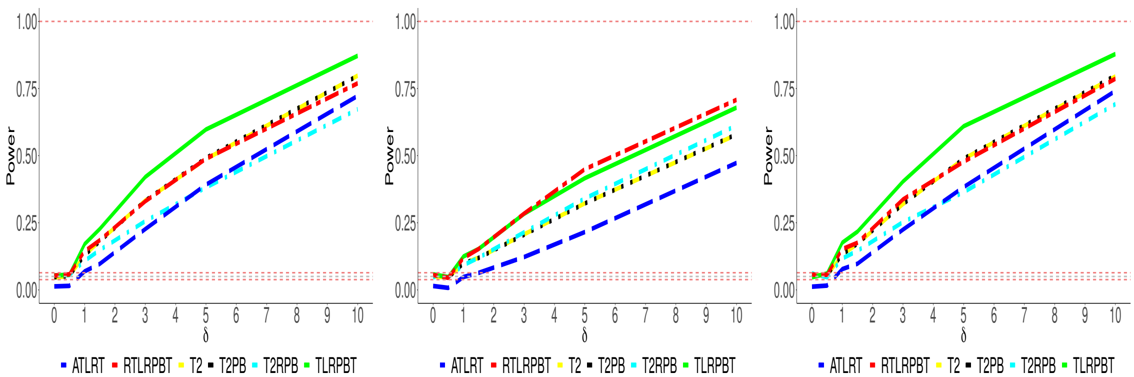
(a) Normal distribution (b) t_5 distribution (c) CN distribution

Figure 3. Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 70$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



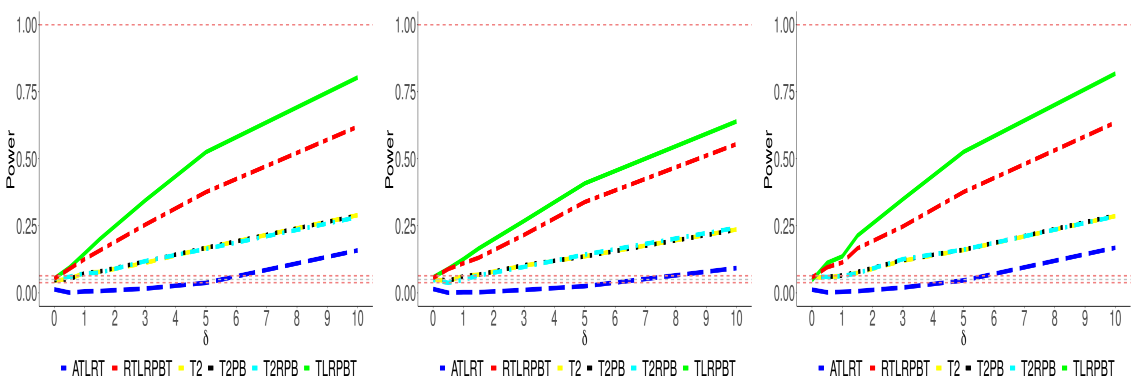
(a) Normal distribution (b) t_5 distribution (c) CN distribution

Figure 4. Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 100$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



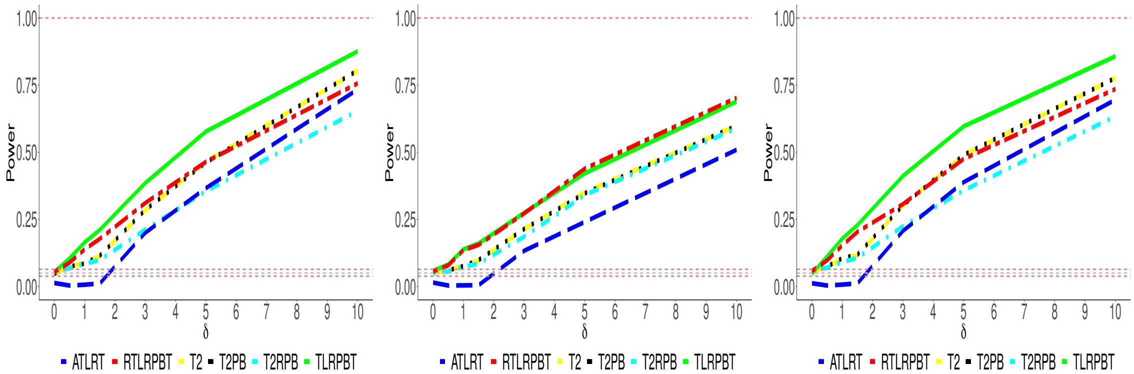
(a) Normal distribution (b) t_5 distribution (c) CN distribution

Figure 5. Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 200$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



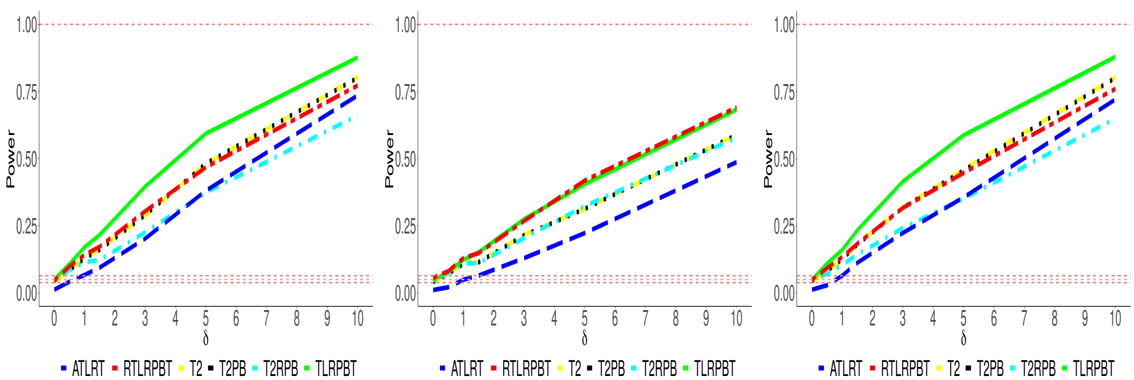
(a) Normal distribution (b) t_5 distribution (c) CN distribution

Figure 6. Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 10$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



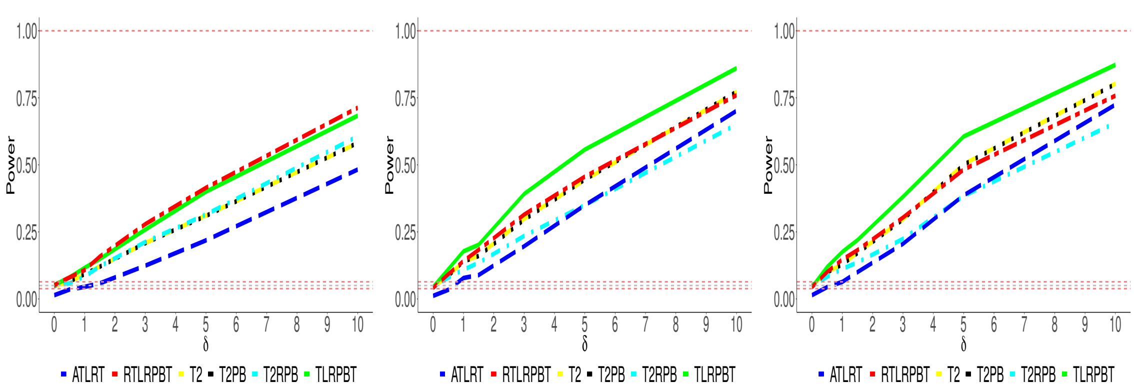
(a) Normal distribution (b) t_5 distribution (c) CN distribution

Figure 7. Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 50$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



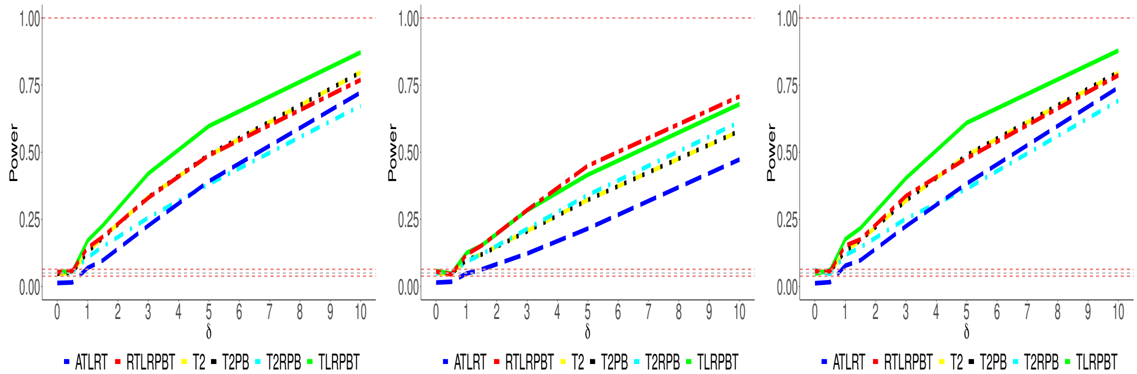
(a) Normal distribution (b) t_5 distribution (c) CN distribution

Figure 8. Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 70$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



(a) Normal distribution (b) t_5 distribution (c) CN distribution

Figure 9. Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 100$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



(a) Normal distribution (b) t_5 distribution (c) CN distribution

Figure 10. Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 200$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.

3.4 Special case of $p = 200$: type I error probabilities and power

Table 3 shows the results considering this case of high dimensionality ($p = 200$). It is considering sample sizes of $n = 10, 50, 70, 100$, and 200 and therefore deal with the cases where $p \geq n$. For this circumstance, the traditional T^2 test and the adapted T2PB, T2RPB can not be applied due to the high dimension issue (see Section 1). It is noticed that the ATLR is conservative in all scenarios considered. It is expected that this test shows substantial low power. It is also noticed that the other adapted tests, TLRPB test and RTLRPB test, are exact, according to the binomial test (see Section 2).

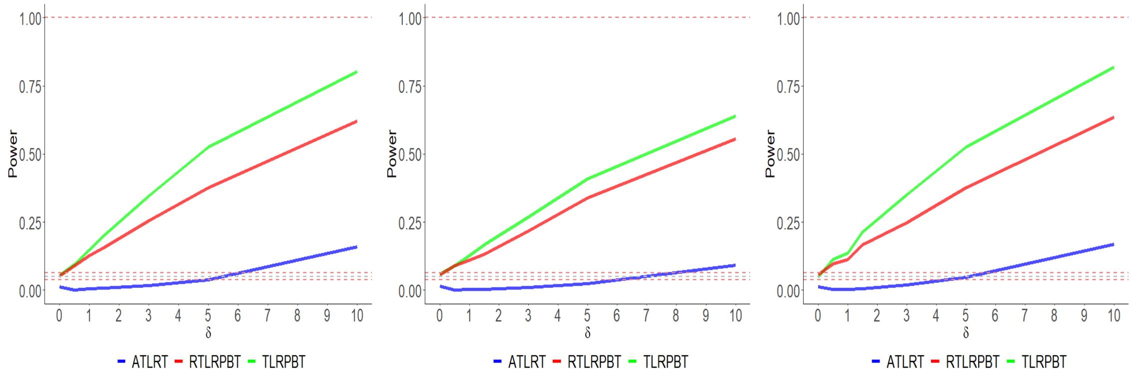
Table 3. Type I error probabilities of the three tests with $\alpha = 5\%$ and $p = 200$, considering the multivariate Normal (N), Student-t with 5 degrees of freedom (t_5) and CN distributions.

Test	Model	n 10	50	70	100	200
ATLR	N	0.0130^-	0.0130^-	0.0120^-	0.0110^-	0.0130^-
	t_5	0.0145^-	0.0145^-	0.0105^-	0.0140^-	0.0145^-
	CN	0.0120^-	0.0120^-	0.0120^-	0.0140^-	0.0120^-
TLRPB test	N	0.0535	0.0535	0.0495	0.0450	0.0535
	t_5	0.0580	0.0580	0.0460	0.0520	0.0580
	CN	0.0470	0.0470	0.0485	0.0445	0.0470
RTLRPB test	N	0.0530	0.0530	0.0445	0.0430	0.0530
	t_5	0.0555	0.0555	0.0555	0.0510	0.0555
	CN	0.0570	0.0570	0.0480	0.0435	0.0570

$^-$: significantly ($p < 0.01$) less than the nominal significance level of 5%.

The power results for $p = 200$ are shown in Figures 11 to 15. Considering $n = 10$, Figure 11 shows that the TLRPB test outperformed the other proposed tests. It is also noticed that, unlike the other dimensions considered ($p = 2$ and $p = 5$), the TLRBP and RTLRPB test perform similarly with the multivariate Student-t distribution with 5 degrees of freedom (Figure 11 (b)). It is also noticed for this scenario that the asymptotic version (ATLR) shows substantial low power. This fact is already expected, as it did control the type I error probabilities in a conservative way (Table 3).

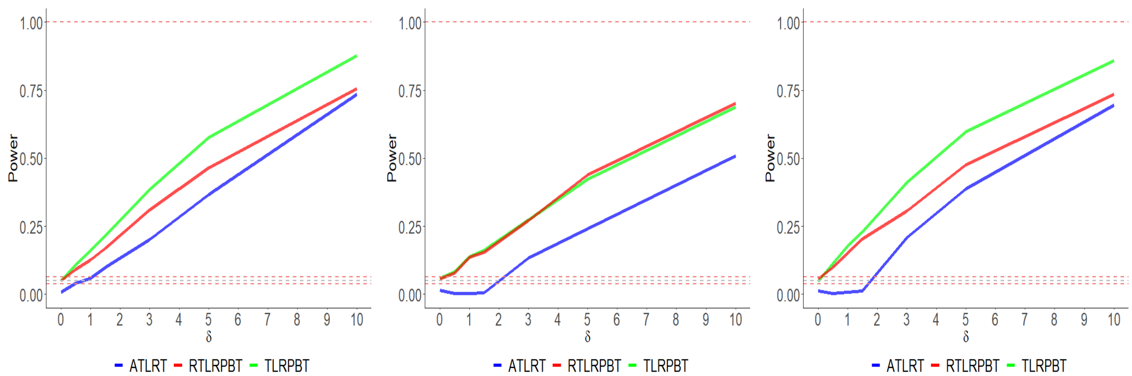
When the sample size n is increased, it is noticed that the TLRPB test continues to perform better (Figures 12, 13, 14 and 15). For the multivariate Student-t distribution with 5 degrees of freedom (Figures 12(b), 13(b), 14(b) and 15(b)), the TLRPB test and RTLRPB test have similar performance. This pattern is identical to those for the dimensions of $p = 2$ and $p = 5$. In contrast, the ATLR has substantial power gain for this dimension (p = 200).



(a) Normal distribution (b) t_5 distribution (c) CN distribution

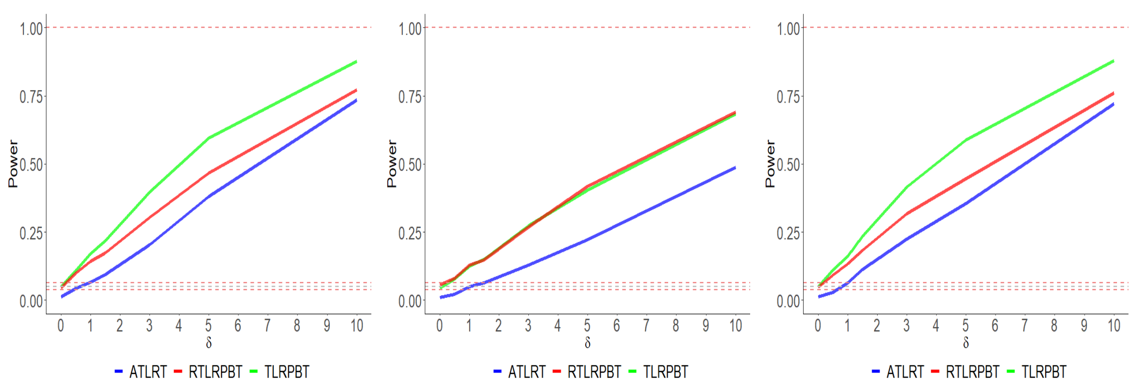
Figure 11. Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 10$ and $p = 200$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.

This fact can be noticed in Figures 13, 14 and 15. In general, the TLRPB test (parametric bootstrap version test) outperformed in power when confronted with other tests.



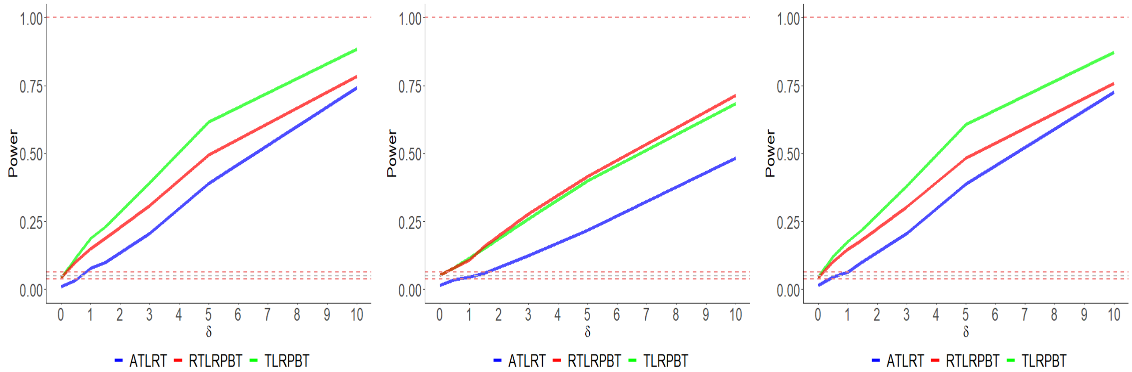
(a) Normal distribution (b) t_5 distribution (c) CN distribution

Figure 12. Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 50$ and $p = 200$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



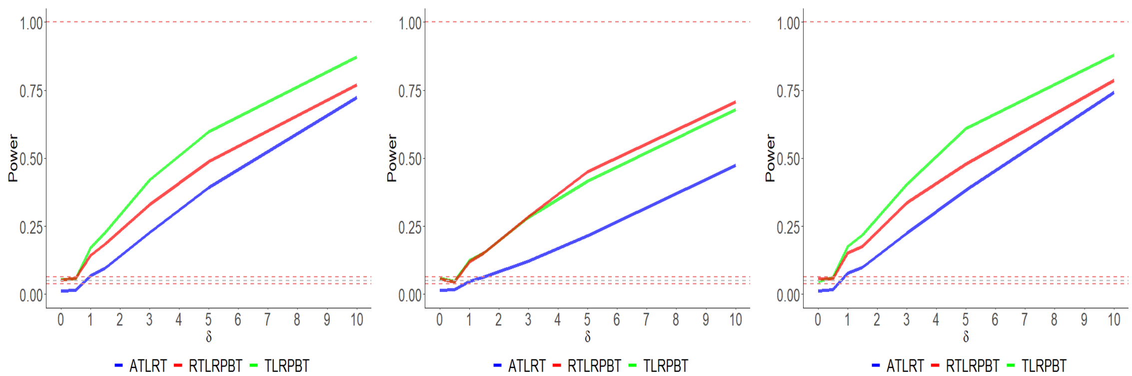
(a) Normal distribution (b) t_5 distribution (c) CN distribution

Figure 13. Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 70$ and $p = 200$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



(a) Normal distribution (b) t_5 distribution (c) CN distribution

Figure 14. Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 100$ and $p = 200$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



(a) Normal distribution (b) t_5 distribution (c) CN distribution

Figure 15. Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 200$ and $p = 200$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.

In a more general analysis, considering all scenarios evaluated ($n = 10, 50, 70, 100$, and 200 and $p = 2, 5, 200$), the Monte Carlo simulations for type I error probability and power showed that the TLRPB test performed better. We recommend using this test as it is not hard to implement and computationally fast.

Willems et al. (2002) concluded that his proposed new test T^2_{R} showed power losses when compared to the traditional Hotelling T^2 test for several configurations of n and p ($n > p$). Under p-variate CN populations, with 10% of contamination, the T^2_{R} has also less power than the traditional Hotelling T^2 test. Dong et al. (2016) indicated that his proposed new test for high dimensional data, based on a shrinkage process of the traditional Hotelling T^2 statistic test, showed high power under p-variate normal and Student-t with 4 degrees of freedom distributions, considering different values of $\boldsymbol{\mu}$.

In both tests (Willems et al., 2002; Dong et al., 2016), the effect of the sample size n influenced the power, which is an expected fact. In the results of the present work it does not occur, as presented before, because the values of the population means change when the sample sizes change, keeping fixed the value of the distance of Mahalanobis $\delta(\boldsymbol{\mu}, \boldsymbol{\mu}_0) = n(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$. To clarify, note that Willems et al. (2002) and Dong et al. (2016) fixed the population mean vector $\boldsymbol{\mu}$, and therefore, the value of Mahalanobis distance δ increases with increasing sample size n. Thus the power grew with an increase of n. On the contrary, in the present work $\boldsymbol{\mu}$ changed in each simulation when n varied, keeping the value of δ fixed, which keeps the power practically constant, as can be seen from Figures 1 to 15. Differences in the power values found under identical configuration, but at different values of n, are attributed to the Monte Carlo error.

4. Special cases and real data analysis

4.1 General context

Marozzi (2015) proposed an alternative multivariate test class for case-control studies for high dimensional data, considering heavy tails or skewed distributions. The proposed tests are based on the combination of tests on inter point distances. The Euclidean distance is utilized. These tests are exact, unbiased and consistent. The results showed that the proposed tests are very powerful under normality, heavy tails, and skewed distributions. Marozzi (2015) applied these same tests to magnetic resonance data which are usually with few observations and many variables, that is, high-dimensional data.

We decided to verify the behavior of our proposed tests regarding the type I error control and power considering heavy tails and skewed distributions. For this, we consider the multivariate exponential distributions with parameter $\boldsymbol{\lambda} = \mathbf{1}$. In the latter case, data are shifted to zero mean by subtracting the population exponential mean $\boldsymbol{\mu} = \mathbf{1}$. We consider in our simulations only the dimensions $p = 2$ and 5 and the same sample sizes n of Section 2.

4.2 Multivariate exponential distribution

For the multivariate exponential distribution, in general, the tests are very liberal. The exception occurred for ATLR and TLRBPT. The ATLR is exact for $n = 10$ and conservative for $n \geq 50$ with $p = 2$ or $p = 5$ (see Tables 4 and 5). The TLRPB test did not control the type I error for $n = 10$ and $n = 100$ with $p = 2$ and $p = 5$, showing a liberal behavior (see Tables 4 and 5, though with no expressive difference from the nominal significance level of 5%. In the other cases, it is exact. For large enough n, say $n = 200$, the T2 and T2BP tests showed either type I error probabilities control or inexpressive liberal behavior, although significant (see Tables 4 and 5).

Table 4. Type I error probabilities of the six tests with $\alpha = 5\%$ and $p = 2$, considering the multivariate exponential distribution.

Test	Model	n = 10	50	70	100	200
T2	exponential	0.1296^+	0.0778^+	0.0648^+	0.0658^+	0.0558
T2BP		0.5414^+	0.9651^+	0.0648^+	0.0658^+	0.0558
T2RBP		0.3470^+	0.9412^+	0.9821^+	1.0000^+	1.0000^+
ATLR		0.0459	0.0269^-	0.0160^-	0.0199^-	0.0070^-
TLRPB test		0.0927^+	0.0608	0.0439	0.0658^+	0.0489
RTLRPB test		0.2393^+	0.6112^+	0.7149^+	0.8335^+	0.9811^+

$^-$: significantly ($p < 0.01$) less than the nominal significance level of 5%.
$^+$: significantly ($p < 0.01$) greater than the nominal significance level of 5%.

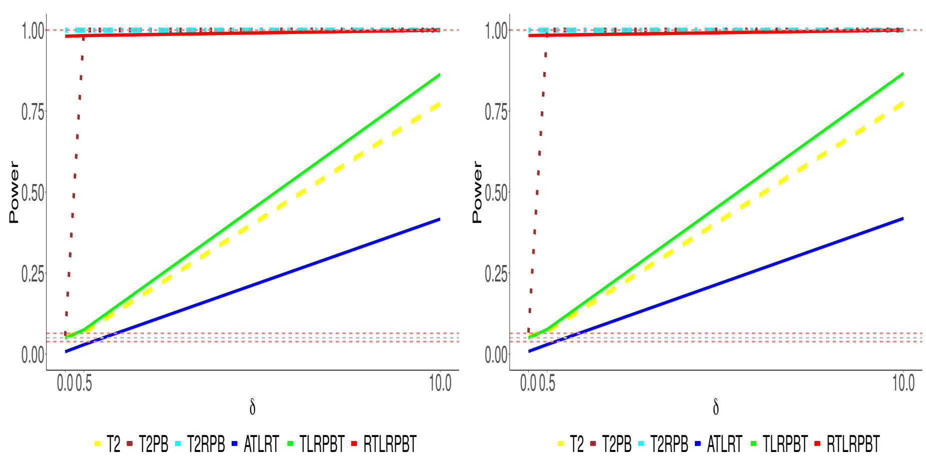
The power of the tests for exponential distribution with $n = 200$ and $p = 2$ and $p = 5$ can be seen in Figures 16(a) and (b). Only the T2, ATLR, and TLRPB test tests should be considered in the comparison, as they are those who controlled the Type I error probabilities. Again, the TLRPB test test is the most powerful, especially with $\delta = 10$, followed by the T2 and ATLR tests, in this order.

Table 6 shows the results of type I error probabilities for multivariate exponential distribution. We consider the dimension $p = 200$ and the sample size $n = 200$, with the nominal

Table 5. Type I error probabilities of the six tests with $\alpha = 5\%$ and $p = 5$, considering the multivariate exponential distribution.

Test	Model	n 10	50	70	100	200
T2	exponential	0.1296^+	0.0778^+	0.0648^+	0.0658^+	0.0558
T2BP		0.5517^+	0.9753^+	0.0676^+	0.0759^+	0.0658^+
T2RBP		0.3490^+	0.9423^+	0.9827^+	1.0000^+	1.0000^+
ATLR		0.0449	0.0267^-	0.0180^-	0.0299^-	0.0080^-
TLRPB test		0.0929^+	0.0618	0.0459	0.0755^+	0.0495
RTLRPB test		0.2397^+	0.6125^+	0.7193^+	0.8435^+	0.9831^+

$^-$: significantly ($p < 0.01$) less than the nominal significance level of 5%.
$^+$: significantly ($p < 0.01$) greater than the nominal significance level of 5%.



(a) $n = 200$, $p = 2$ (b) $n = 200$, $p = 5$

Figure 16. Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with multivariate exponential distribution, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.

significance level of 5%, that is, an extreme case. We realized that only the ATLR and TLRPB test tests controlled the type I error, being the first conservative and the second, exact. The margin of error of the Monte Carlo simulations is also presented (see Table 6). The RTLRPB test test showed all type I error probabilities close to 1 and is extremely liberal in this high-dimensional case.

Table 6. Type I error probabilities ($\delta = 0$) and power ($\delta = 0.5$ and 10) of the three tests with $\alpha = 5\%$, $p = 200$ and $n = 200$, considering the multivariate exponential distribution, where the values in parenthesis express the margin of error of Monte Carlo simulations.

Test	Model	δ 0	0.5	10
ATLR	exponential	0.0069^-(0.0051)	0.0000(0.0036)	0.0000(0.0036)
TLRPB test		0.0488 (0.0133)	0.0473(0.0131)	0.9909(0.0058)
RTLRPB test		0.9810^+(0.0084)	0.9673(0.0110)	1.0000(0.0036)

$^-$: significantly ($p < 0.01$) less than the nominal significance level of 5%.
$^+$: significantly ($p < 0.01$) greater than the nominal significance level of 5%.

We develop a package that is available in the `R` software (R Core Team , 2020) to assist the user in executing the mentioned methodology called multivariate tests for the vector of means (Alves and Ferreira, 2020). Then, we introduce the use of this package to one real data set.

4.3 Application to real data

In this section, the proposed methodology is applied to one real data set, that deals with the contents of sand and clay from capoeira nova, in the Amazon, Brazil, available in Ferreira (2018). The data set has two variables (sand and clay) and 30 observations ($p = 2$, $n = 30$). We want to verify that the new capoeira soil has an average sand and clay content equal to that of a forest population, at a level of 5% of significance. An exploratory analysis is previously carried out and we verified that the variables sand and clay are correlated and the data do not show normal p-variable according to the Royston test. There is also the presence of outliers in the data. Table 7 presents the data set of sand and clay contents in a new capoeira soil in the Amazon to be analyzed.

The vector of sample averages for the sand and clay contents takes on the values of 22 and 36.1, respectively, that is, $\overline{\boldsymbol{X}} = [22, 36.1]^\top$. According to Ferreira (2018), it is known that in a forest soil the average levels of sand and clay content have values equal to 14 and 42, respectively, that is, $\boldsymbol{\mu}_0 = [14, 42]^\top$. So, in possession of the samples collected of sand and clay contents in a new capoeira soil, in the Amazon, the hypotheses to be tested are $\mathrm{H}_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{versus} \quad \mathrm{H}_1\colon \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. The T2, T2PB, T2R, T2RPB, ATLR, TLRPB and RTLRPB tests have been applied; see Section 2.

Table 8 shows that all tests took the same decision to reject the null hypothesis H_0. However, since the assumption of p-variate normality is not met, we suggest choosing the result of the TLRPB test because this is the most powerful among all tests evaluated in Alves and Ferreira (2019). All tests provided the same decision to reject the null hypothesis.

Table 7. Sand and clay contents in a new capoeira soil in the Amazon.

sand	clay	sand	clay	sand	clay
11	38	20	32	13	47
24	25	18	34	28	32
16	49	17	39	11	45
18	34	30	32	27	36
5	64	45	24	7	59
11	40	11	50	42	23
17	38	41	21	21	35
9	40	22	36	48	21
13	40	14	32	12	36
53	21	25	28	31	32

Table 8. Tests for the vector of population means for the levels of sand and clay in a new capoeira soil, in the Amazon.

Test	Statistics	p-value	Decision
T2	11.93406	0.00802	Reject H_0
T2PB	11.93406	0.00899	Reject H_0
T2RPB	45.19158	0.00049	Reject H_0
ATLR	9.21556	0.00997	Reject H_0
TLRPB test	9.21556	0.00299	Reject H_0
RTLRPB test	7.11871	0.02848	Reject H_0

5. Conclusions

The trace likelihood ratio parametric bootstrap test is recommended for testing hypothesis about a multivariate population mean vector of normal and non-normal populations, including the presence of outliers. For the case of the contaminated multivariate normal distribution, the robust average and comedian covariance matrix estimators performed below tests that do not use these estimators. This fact occurred in all scenarios evaluated considering this distribution. It is possible to conclude that the use of robust comedian mean and covariance estimators is not helpful for testing hypotheses on a population mean vector.

These tests have some limitations, as in the multivariate lognormal distribution, where they did not perform well in controlling type I error probabilities, being considered liberal. As a future work, we will intend to adapt these tests to data from two or more populations.

References

Alves, H.J. and Ferreira, D.F., 2019. Proposition of new alternative tests adapted to the traditional T^2 test. Communications in Statistics: Simulation and Computation, 1-14.

Alves, H.J.P. and Ferreira, D.F., 2020. TVMM: Multivariate Tests for the Vector of Means. R package version 3.1. https://CRAN.R-project.org/package=TVMM.

Bai, Z. and Saranadasa, H., 1996. Effect of high dimension: by an example of a two sample problem. Statistica Sinica, 311–329.

Chen, S.X. and Qin, Y.L., 2010. A two-sample test for high-dimensional data with applications to gene-set testing. The Annals of Statistics, 38, 808–835.

Dong, K., Pang, H., Tong, T., and Genton, M.G., 2016. Shrinkage-based diagonal Hotelling's tests for high-dimensional small sample size data. Journal of Multivariate Analysis, 143, 127–142.

Falk, M., 1997. On MAD and comedians. Annals of the Institute of Statistical Mathematics, 49s, 615–644.

Ferreira, D.F., 2018. Estatística Multivariada. Ufla, Brazil.

Ledoit, O. and Wolf, M., 2002. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. The Annals of Statistics, 30, 1081–1102.

Lee, S.H., Lim, J., Li, E., Vannucci, M., and Petkova, E., 2012. Order test for high-dimensional two-sample means. Journal of Statistical Planning and Inference, 142, 2719–2725.

Maronna, R.A. and Zamar, R.H., 2002. Robust estimates of location and dispersion for high-dimensional datasets. Technometrics, 44, 307–317.

Marozzi, M., 2015. Multivariate multidistance tests for high-dimensional low sample size case-control studies. Statistics in Medicine, 34, 1511–1526.

Mudholkar, G.S. and Srivastava, D.K., 2000. A class of robust stepwise alternatives to Hotelling's T^2 tests. Journal of Applied Statistics, 27, 599-619.

Cardoso de Oliveira, I.R. and Ferreira, D.F., 2010. Multivariate extension of chi-squared univariate normality test. Journal of Statistical Computation and Simulation, 80, 513–526.

Pan, G.M. and Zhou, W., 2011. Central limit theorem for Hotelling's T^2 statistic under large dimension. The Annals of Applied Probability, 1860–1910.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Sajesh, T.A. and Srinivasan, M.R., 2012. Outlier detection for high dimensional data using the Comedian approach. Journal of Statistical Computation and Simulation, 82, 745–757.

Srivastava, M.S., 2009. A test for the mean vector with fewer observations than the dimension under non-normality. Journal of Multivariate Analysis, 100, 518–532.

Srivastava, M.S. and Du, M., 2008. A test for the mean vector with fewer observations than the dimension. Journal of Multivariate Analysis, 99, 386–402.

Srivastava, M.S., Katayama, S., and Kano, Y., 2013. A two sample test in high dimensional data. Journal of Multivariate Analysis, 114, 349–358.

Tiku, M.L., 1982. Robust statistics for testing equality of means or variances. Communications in Statistics: Theory and Methods, 11, 2543–2558.

Wang, R., Peng, L., and Qi, Y., 2013. Jackknife empirical likelihood test for equality of two high dimensional means. Statistica Sinica, 23, 667–690.

Wagala, A., (2020). A likelihood ratio test for correlated paired multivariate samples. Chilean Journal of Statistics, 11, 41–51.

Willems, G., Pison, G., Rousseeuw, P.J., and Van Aelst, S., 2002. A robust Hotelling test. Metrika, 55, 125–138.

Chilean Journal of Statistics
Vol. 11, No. 2, December 2020, 137–150

Statistical Inference
Research Paper

Improved parameter estimation of the Chaudhry and Ahmad distribution with climate applications

Josmar Mazucheli[1,*], André F.B. Menezes[2], Sanku Dey[3], and Saralees Nadarajah[4]

[1]Departmento of Estatística, Universidade Estadual de Maringá, Maringá-PR, Brazil,
[2]Departamento de Estatística, Universidade Estadual de Campinas, Campinas-SP, Brazil,
[3]Department of Statistics, St. Anthony's College, Shillong, Meghalaya, India
[4]School of Mathematics, University of Manchester, Manchester M13 9PL, UK

(Received: 04 August 2020 · Accepted in final form: 11 December 2020)

Abstract

The Chaudhry-Ahmad distribution is a two-parameter continuous probability distribution obtained as a solution to a generalized Pearson system of differential equation. Although its probability density curve resembles the inverse-Gaussian, gamma, log-normal, Weibull and other distributions, it has been neglected in the analysis of right-skewed data. The purpose of this paper is three folded. Firstly, to reparametrize the Chaudhry and Ahmad distribution and present some of its basic properties. Secondly to derive the analytical bias-corrected maximum likelihood estimators applying the Cox-Snell methodology and thirdly to study, by MC simulations, the small-sample properties of the maximum likelihood estimators and their bias-corrected versions, obtained from the Cox-Snell formula and by parametric bootstrap technique. The numerical results show, for both parameters, that the maximum likelihood estimators are highly biased, especially in small samples. On the other hand, both, the analytical and bootstrap methodologies, significantly reduce the biases and the mean-squared errors. It is apparent from the results that the analytical bias-correction is more efficient than bootstrap resamples. Finally, wind speed data from six weather stations distributed in the state of Tocantins in Brazil is used to illustrate the applicability of the proposed methods.

Keywords: Bootstrap bias correction · Cox-Snell bias-correction · Maximum likelihood estimation · Monte Carlo simulation · Wind speed data.

Mathematics Subject Classification: Primary 60E05 · Secondary 62F10.

1. Introduction

Chaudhry and Ahmad (1993) introduced a nonnegative two-parameter probability distribution, called the Chaudhry-Ahmad (CA) distribution as a solution of the generalized Pearson system of differential equation. It is noteworthy that from the generalized Pearson system of probability distributions, many continuous probability density functions (PDFs) can be generated (Sankaran et al., 2003; Stavroyiannis, 2014). Indeed, as discussed in Shakil et al. (2010, 2016), the well known families of distributions such as the normal and

*Corresponding author. Email: jmazucheli@gmail.com

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

Student-t (known as Pearson type VII), beta distribution (known as Pearson type I) and gamma distribution (known as Pearson type III), introduced by Karl Pearson during the late 19th century (Pearson, 1893, 1895, 1901, 1916), can be generated as a solution to Equation (1) by proper choice of its parameters.

Although the CA PDF curve resembles the inverse-Gaussian (IG), gamma, log-normal, Weibull and other distributions, it has not been widely explored in the statistical literature. Recently, Shakil et al. (2010) derived a family of distribution, which includes the CA distribution as a special case. To the best of our knowledge, there are only two real data analysis considering the CA distribution. Nanos and Montero (2001) showed that CA distribution fitted better than the Weibull distribution in a problem involving prediction of the diameter distribution of a stand. In Nanos et al. (2000) the Weibull and CA distributions were used to model resin production distributions for maritime pine stands.

It is important to point out that the CA distribution is capable of modeling increasing hazard rate funtions (HRFs). There are many situations where only increasing HRFs are used or observed: Woosley and Cossman (2007) observed that drugs during clinical development have increasing HRFs; Tsarouhas and Arvanitoyannis (2010) showed that machines of the bread production display increasing HRFs; Koutras (2011) observed that software degradation times have increasing HRFs; Lai (2013) investigated the optimum number of minimal repairs for systems have increasing hazard rates and so on.

Although the maximum likelihood (ML) estimators have many appealing properties (Edwards, 1992; Lehmann and Casella, 1998), it is also well known that ML estimators could be biased, especially when the study is being done in small samples. Owing to this reason, researchers strive to develop nearly unbiased estimators for the parameters of several probability distributions. Notable among them are Saha and Paul (2005), Lemonte et al. (2007), Giles and Feng (2009), Lagos-Àlvarez et al. (2011), Giles (2012a), Giles (2012b), Schwartz et al. (2013), Giles et al. (2013), Teimouri and Nadarajah (2013), Ling and Giles (2014), Zhang and Liu (2015), Teimouri and Nadarajah (2016), Reath (2016), Schwartz and Giles (2016), Wang and Wang (2017), Mazucheli and Dey (2018), Mazucheli et al. (2018), Mazucheli et al. (2020) and references cited therein.

The objective of this paper is to perform improved parameter estimation of the CA distribution. We consider the analytical methodology introduced by Cox-Snell (1968) and the parametric bootstrap resampling method (Efron, 1982). We describe two corrective approaches to bias-correction, both methods reduce the biases of the ML estimators to the second order magnitude.

After this introduction, the paper is organized as follows. In Section 2, we introduce the CA distribution and deduce expressions used to obtain the ML estimators of its parameters, calculating the expected Fisher information matrix. In Section 3, by using the Cox-Snell formula, we derive analytical expressions for the second order biases of the maximum likelihood estimators, and also discuss the bootstrap bias correction. A Monte Carlo (MC) simulation study is carried out in Section 4 to compare the ML estimators and their bias-corrected versions, obtained from the Cox-Snell formula and parametric bootstrap technique. An application by using wind speed data from Brazil is provided also in this section. As a result of this application, we are able to provide, for example, better estimates of most frequent wind speeds observed at various stations. Some concluding remarks are presented in Section 5.

2. Preliminaries, model description and estimation

In this section, we provide background on the CA distribution and the ML estimators of its parameters, as well as the corresponding expected Fisher information matrix.

2.1 Background on the Chaudhry-Ahmad distribution

Chaudhry and Ahmad (1993) developed a two-parameter probability distribution as a solution of the generalized Pearson system of differential equation

$$\frac{\mathrm{d}}{\mathrm{d}x} f(x) = \frac{c_0 + c_1\, x + c_2\, x^2 + \cdots + c_m\, x^m}{c'_0 + c'_1\, x + c'_2\, x^2 + \cdots + c'_n\, x^n}\, f(x), \tag{1}$$

where $m, n \geq 1$ are integers, and the coefficients c and c' are real numbers. These authors considered a special case of Equation (1) taking $m = 4$, $n = 3$, $c'_0 = c'_1 = c'_2 = 0$, $c_4/2\,c'_3 = -2\,\alpha$, $c_0/2\,c'_3 = 2\,\beta$ and $c'_3 \neq 0$. This distribution, which now bear their names, can also be obtained as the root reciprocal of the inverse Gaussian distribution, that is, the distribution of the random variable $X = 1\,/\,\sqrt{Y}$, where $Y \sim \mathrm{IG}(\mu, \lambda)$ with $\mu = (\alpha\,/\,\beta)^{1/2}$ and $\lambda = 2\,\alpha$.

The cumulative distribution function (CDF) of the CA distribution is given by

$$F(x; \alpha, \beta) = \Phi\left[\sqrt{2}\,\left(\sqrt{\alpha}\,x - \sqrt{\beta}\,x^{-1}\right)\right] - \exp\left(4\,\sqrt{\alpha\,\beta}\right)\,\Phi\left[-\sqrt{2}\,\left(\sqrt{\alpha}\,x + \sqrt{\beta}\,x^{-1}\right)\right], \tag{2}$$

where x, α, $\beta > 0$ and Φ denotes the CDF of a standard normal distribution.

Solving the orthogonality differential equation of Cox and Reid (1987), we consider in Equation (2) $\beta = \alpha\lambda^4$, such that λ will be the mode of the PDF. The advantage of such parametrization is that λ has a direct interpretation and it is orthogonal to α. Thus, from Equation (2), the PDF of a CA distributed random variable with parameters α and λ can be written as

$$f(x; \alpha, \lambda) = 2\,\sqrt{\frac{\alpha}{\pi}}\,\exp\left[-\left(\sqrt{\alpha}\,x - \lambda^2\sqrt{\alpha}\,x^{-1}\right)^2\right]. \tag{3}$$

Figure 1 displays the PDF and the HRF curves considering different values of α and $\lambda = 1$ (λ is a location parameter). We observe that the PDF is skewed to the right and unimodal with turning point at $x_{\max} = \lambda = 1$. We also observe that the HRF of CA distribution is monotone increasing.

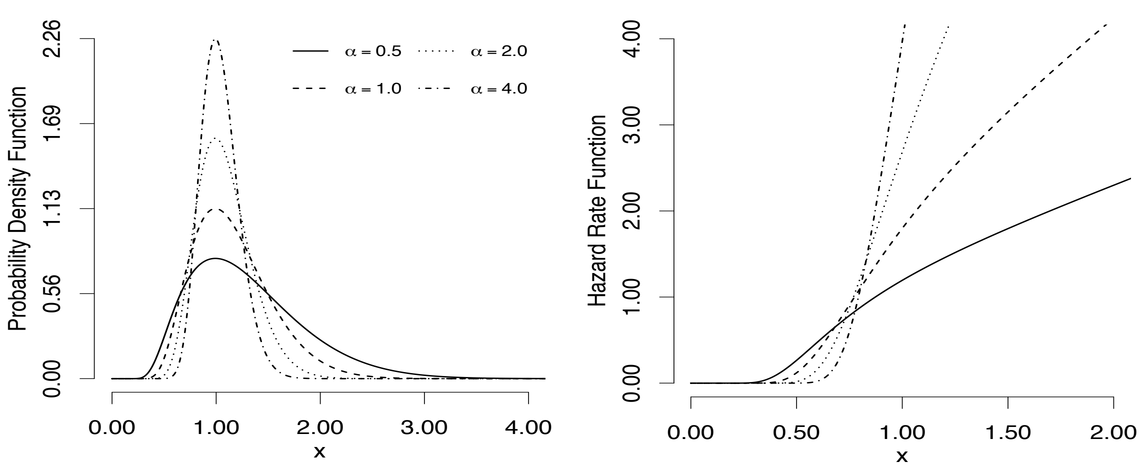


Figure 1. PDF and HRF of the CA distribution for $\alpha = (0.5, 1.0, 2.0$ and $4.0)$ and $\lambda = 1$.

The kth moment about the origin of CA distribution is given by

$$\mu'_k = 2\sqrt{\frac{\alpha}{\pi}}\, \exp\left(2\,\alpha\,\lambda^2\right)\, \lambda^{k+1}\, K_{\frac{k}{2}+\frac{1}{2}}(2\,\alpha\,\lambda^2), \tag{4}$$

where K_ν denotes the modified Bessel function of the second kind (Abramowitz and Stegun, 1974). In particular, from Equation (4), the first four moments about the origin are stated as

$$\begin{aligned}
\mu'_1 &= 2\sqrt{\frac{\alpha}{\pi}}\, \exp\left(2\,\alpha\,\lambda^2\right)\, \lambda^{2}\, K_{\frac{1}{2}+\frac{1}{2}}(2\,\alpha\,\lambda^2),\\
\mu'_2 &= \frac{2\,\alpha\,\lambda^2+1}{2\,\alpha},\\
\mu'_3 &= 2\sqrt{\frac{\alpha}{\pi}}\, \exp\left(2\,\alpha\,\lambda^2\right)\, \lambda^{4}\, K_{\frac{3}{2}+\frac{1}{2}}(2\,\alpha\,\lambda^2),\\
\mu'_4 &= \frac{4\,\alpha^2\,\lambda^4+6\,\alpha\,\lambda^2+3}{4\,\alpha^4}.
\end{aligned}$$

2.2 Maximum likelihood estimation

Suppose that $\boldsymbol{X} = (X_1,\ldots,X_n)^\top$ is a random sample of size n from CA distribution with PDF given by Equation (3) and $\boldsymbol{x} = (x_1,\ldots,x_n)^\top$ its observations. The log-likelihood function for $\boldsymbol{\theta} = (\alpha,\lambda)$ is given by

$$\ell(\boldsymbol{\theta};\boldsymbol{x}) \propto \frac{n}{2}\log(\alpha) + 2\,n\,\alpha\,\lambda^2 - \alpha\sum_{i=1}^{n} x_i^2 - \lambda^4\,\alpha\sum_{i=1}^{n} x_i^{-2}. \tag{5}$$

Differentiating in Equation (5) with respect to α and λ, we have the score vector $\boldsymbol{U_\theta} = (U_\alpha, U_\lambda)^\top$ with components given by

$$U_\alpha = \frac{n}{2\,\alpha} + 2\,n\,\lambda^2 - \sum_{i=1}^{n} x_i^2 - \lambda^4\sum_{i=1}^{n} x_i^{-2}, \tag{6}$$

$$U_\lambda = 4\,n\,\alpha\,\lambda - 4\,\lambda^3\alpha\sum_{i=1}^{n} x_i^{-2}. \tag{7}$$

After simple algebraic manipulation of Equations (6) and (7), note that the ML estimates of α and λ can be written as $\widehat{\lambda} = (m'_{-2})^{-1/2}$ and $\widehat{\alpha} = [2(m'_{-2}\,m'_2 - 1)\,m'_{-2}]^{-1}(m'_{-2})^2$, where $m'_2 = (1/n)\sum_{i=1}^{n} x_i^2$ and $m'_{-2} = (1/n)\sum_{i=1}^{n} x_i^{-2}$.

The expected Fisher information matrix of $\boldsymbol{\theta}$ is given by

$$\boldsymbol{I}(\boldsymbol{\theta}) = [I_{ij}] = -n\mathrm{E}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log(f(x_i;\boldsymbol{\theta}))\right) = \begin{bmatrix} -\dfrac{n}{2\,\alpha^2} & 0 \\ 0 & -8\,n\,\alpha \end{bmatrix}, \quad i,j = 1,2. \tag{8}$$

From Equation (8), we observe that the information matrix is diagonal, which means that the ML estimators are asymptotically independent. Hence, the asymptotic variance of $\widehat{\alpha}$ and $\widehat{\lambda}$ are given, respectively, by

$$\mathrm{Var}(\widehat{\alpha}) = \frac{2\,\alpha^2}{n}, \quad \mathrm{Var}(\widehat{\lambda}) = \frac{1}{8\,n\,\alpha}. \tag{9}$$

The asymptotic variance of $\widehat{\lambda}$ only depends on α. Thus, as α decreases, the variance of $\widehat{\lambda}$ increases. The asymptotic $100(1-\delta)$ confidence intervals for α and λ can be obtained respectively as

$$\widehat{\alpha} \pm z_{\delta/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\alpha})}, \quad \widehat{\lambda} \pm z_{\delta/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\lambda})}, \tag{10}$$

where $z_{\delta/2}$ indicated in Equation (10) denotes the $100(1-\delta/2)$ percentile of the standard normal distribution.

3. Bias-corrected maximum likelihood estimators

In this section, we derive analytical expressions for the second order biases of the maximum likelihood estimators by using the Cox-Snell formula, and also discuss the bootstrap bias correction.

3.1 Cox-Snell analytic bias correction

Let $\ell(\boldsymbol{\theta};\boldsymbol{x})$ denote the log-likelihood function of a p-dimensional parameter vector $\boldsymbol{\theta}$ based on a sample of observations $\boldsymbol{x}$. We assume the following regularity conditions on the behavior of the log-likelihood function (Cox and Hinkley, 1979):

(a) X_i, for $i = 1,\ldots,n$, are independent and identically distributed random variables.
(b) The parameter space of $\boldsymbol{\theta}$ is compact.
(c) The true but unknown parameter value $\boldsymbol{\theta}_0$ is identified, that is,

$$\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}} \mathrm{E}_{\boldsymbol{\theta}_0}\left[\log\left(f\left(x_i;\boldsymbol{\theta}\right)\right)\right].$$

(d) The likelihood function

$$\ell\left(\boldsymbol{\theta};\boldsymbol{x}\right) = \sum_{i=1}^{n}\log\left(f\left(x_i;\boldsymbol{\theta}\right)\right)$$

is continuous in $\boldsymbol{\theta}$.
(e) $\mathrm{E}_{\boldsymbol{\theta}_0}\left[\log\left(f\left(x_i;\boldsymbol{\theta}\right)\right)\right]$ exists.
(f) The log-likelihood function is such that $(1/n)\ell(\boldsymbol{\theta};\boldsymbol{x})$ converges almost surely (in probability) to $\mathrm{E}_{\boldsymbol{\theta}_0}[\log(f(x_i;\boldsymbol{\theta}))]$ uniformly in $\boldsymbol{\theta}$.

Conditions (a) to (d) are clearly satisfied for the CA distribution. Conditions (e) and (f) are also satisfied since, for all $\alpha > 0$ and $\beta > 0$,

$$\int_0^\infty x^2 \exp\left(-\alpha x^2 - \frac{\lambda^4\alpha}{x^2}\right)\mathrm{d}x < \infty,$$
$$\int_0^\infty x^{-2} \exp\left(-\alpha x^2 - \frac{\lambda^4\alpha}{x^2}\right)\mathrm{d}x < \infty.$$

The joint cumulants of the derivatives of ℓ are given by

$$\boldsymbol{I}_{ij} = \mathrm{E}\left[\frac{\partial^2 \ell}{\partial \boldsymbol{\theta}_i \, \partial \boldsymbol{\theta}_j}\right], \quad \boldsymbol{I}_{ijl} = \mathrm{E}\left[\frac{\partial^3 \ell}{\partial \boldsymbol{\theta}_i \, \partial \boldsymbol{\theta}_j \, \partial \boldsymbol{\theta}_l}\right], \quad \boldsymbol{I}_{ij,l} = \mathrm{E}\left[\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\theta}_i \, \partial \boldsymbol{\theta}_j}\right)\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}_l}\right)\right],$$

for $i, j, l = 1, \ldots, p$. All these expression are assumed to be of order $\mathcal{O}(n)$.

Cox-Snell (1968) showed that when the samples are independent, but not necessarily identically distributed, the bias of the rth element of the ML estimator of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}$, can be expressed as

$$\mathcal{B}(\widehat{\boldsymbol{\theta}}_r) = \sum_{i=1}^{p} \sum_{j=1}^{p} \sum_{l=1}^{p} \boldsymbol{I}^{ri} \, \boldsymbol{I}^{jl} \, \left[0.5 \boldsymbol{I}_{ijl} + \boldsymbol{I}_{ij,l}\right] + \mathcal{O}\left(n^{-2}\right), \tag{11}$$

where $r = 1, \ldots, p$ and $\boldsymbol{I}^{ij}$ denotes the (i, j)th element of the inverse of the expected Fisher information matrix.

In respect to the orthogonally parametrization of CA distribution, after extensive algebra, it can be shown that $\boldsymbol{I}_{111} = -24\, n\, \alpha/\lambda$, $\boldsymbol{I}_{122} = \boldsymbol{I}_{212} = \boldsymbol{I}_{221} = -8\, n$, $\boldsymbol{I}_{222} = n/\alpha^3$, $\boldsymbol{I}_{11,1} = 24\, n\, \alpha/\lambda$, $\boldsymbol{I}_{12,2} = \boldsymbol{I}_{21,2} = 8\, n$ and all other terms are equal to zero. Hence, the second-order bias of the ML estimators of α and λ are given respectively by

$$\mathcal{B}(\widehat{\alpha}) = \frac{3\, \alpha}{n} \tag{12}$$

and

$$\mathcal{B}(\widehat{\lambda}) = \frac{3}{16\, n\, \alpha\, \lambda}, \tag{13}$$

Using Equations (12) and (13), we define the bias-corrected (BC) estimator as

$$\widehat{\alpha}_{\mathrm{BC}} = \widehat{\alpha} - \widehat{\mathcal{B}}(\widehat{\alpha}), \quad \widehat{\lambda}_{\mathrm{BC}} = \widehat{\lambda} - \widehat{\mathcal{B}}(\widehat{\lambda}). \tag{14}$$

Note that $\widehat{\alpha}_{\mathrm{BC}}$ and $\widehat{\lambda}_{\mathrm{BC}}$ defined in Equation (14) have bias of order $\mathcal{O}(n^{-2})$ as indicated in (11). Thus, it is expected that they have superior sampling properties relative to $\widehat{\alpha}$ and $\widehat{\lambda}$. We also empathize that the bias-corrected ML estimator for β in the original parametrization Equation (2) is obtained from $\widehat{\beta} - (3\sqrt{\widehat{\alpha}}\, \widehat{\beta} + 1.5\sqrt{\widehat{\beta}})/\sqrt{\widehat{\alpha}}n$.

3.2 Parametric bootstrap bias correction

An alternative approach to analytically bias-corrected ML estimators is based on bootstrap resampling scheme (Efron and Tibshirani, 1993; Davison and Hinkley, 1997). In this method the bias correction is performed numerically without deriving analytical expression for the bias function. In fact, the parametric bootstrap bias correction (PB) estimates use the ML estimates of the data to generate pseudo-random samples from the distribution to estimate the bias and then subtract the bias from the ML estimates.

Let $\widehat{\boldsymbol{\theta}}_{(\cdot)}$ be the average value of the ML estimator from B bootstrap replications, based on a pseudo-sample of size n generated from Equation (3) using the parameters of the ML estimates $\widehat{\boldsymbol{\theta}}$. The estimated bias of $\widehat{\boldsymbol{\theta}}$ is defined as $\widehat{\mathcal{B}}(\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{\theta}}_{(\cdot)} - \widehat{\boldsymbol{\theta}}$. Then, the bootstrap bias-corrected estimator is $\widehat{\boldsymbol{\theta}}_{\mathrm{PB}} = 2\, \widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(\cdot)}$.

4. Numerical evaluations

In this section, we carry out a MC simulation study to compare the ML estimators and their bias-corrected versions. In addition, we illustrate the applicability of the CA distribution for bias corrections to the wind speed data.

4.1 Simulation study

Our MC simulation study is conducted to compare the finite-sample behavior of the ML estimators and their bias-corrections obtained by Cox-Snell methodology (BC) and parametric bootstrap scheme (PB) for the parameters that index the CA distribution. For this purpose, we generate samples of size $n = 10, 20, 30, 40$ and 50 from Equation (3) considering $\alpha = 0.5, 1.0, 1.5, 2.0$ and 4.0 and fixed $\lambda = 1$, since it is a location parameter and the estimators are scale invariant. The behavior of PDF and HRF for these parameters values were illustrated in Figure 1. It is important to note that the mean, variance, skewness and kurtosis of a CA distributed random variable decrease as α increases.

To simulate random variables from a CA distribution, we generated samples from a random variable Y with inverse Gaussian distribution and we used the transformation $X = 1/\sqrt{Y}$.

To assess the performance of the methods under consideration, we calculated the bias and root mean-squared error (RMSE). The number of MC simulations was fixed at $M = 10,000$ and $B = 1,000$ bootstrap replicates were used. All simulations were carried out in Ox Console which is a matrix programming language with object-oriented support developed by Jurgen Doornik (Doornik, 2007).

Table 1 depicts the estimated bias and root mean-squared error, in parentheses, for different values of α and $\lambda = 1$. We can observe that all the estimates show the property of consistency, that is, the RMSEs decrease as sample size increases. We also note that the ML estimates of α are highly biased, particularly when the sample size is small. For instance, the biases of the ML estimates of α for $(n, \alpha) = (10, 0.5)$ and $(n, \alpha) = (10, 4)$ are approximately 22% and 169%, respectively. Also the biases of the ML estimates of α for $(n, \alpha) = (20, 0.5)$ and $(n, \alpha) = (20, 4)$ are approximately 9% and 68%, respectively. The estimates $\widehat{\alpha}_{\rm BC}$ and $\widehat{\alpha}_{\rm PB}$ clearly outperform the ML estimates as far as the bias goes. For example, the biases of the BC estimates of α for $(n, \alpha) = (10, 0.5)$ and $(n, \alpha) = (10, 4)$ are approximately 0.3% and 1.5%, respectively. The biases of the PB estimates of α for $(n, \alpha) = (10, 0.5)$ and $(n, \alpha) = (10, 4)$ are approximately 8.9% and 74.6%, respectively. Thus, the proposed estimators achieve substantial bias reduction, especially for the small and moderate sample sizes and therefore, we consider them as better alternatives of the ML estimates of α. We also observe that the bias-corrected estimates are closer to the true parameter values than the unadjusted estimates as sample size increases. Additionally, the estimated root mean-squared errors for α of the bias corrected estimates are smaller than those of the uncorrected estimates. On the other hand, the RMSE of λ are very similar for all estimators.

Now, in order to evaluate the overall performance of each estimation method with respect to the bias and root mean squared error, for each value of n, we use two measures introduced by Cribari-Neto and Vasconcellos (2002). The authors called these quantities as integrated bias squared norm and average root mean squared error. They are calculated as follows

$$\mathrm{IBSQ}_{(k)} = \sqrt{\frac{1}{16}\sum_{h=1}^{16}(r_{h,k})^2}, \quad \mathrm{ARMSE}_{(k)} = \frac{1}{16}\sum_{h=1}^{16}\mathrm{RMSE}_{h,k},$$

Table 1. Estimated bias (root mean-squared error) for α and λ, ($\lambda = 1.0$).

α	n	Estimator of α ML	Estimator of α BC	Estimator of α PB	Estimator of λ ML	Estimator of λ BC	Estimator of λ PB
	10	0.2191 (0.5098)	0.0034 (0.3223)	-0.0891 (0.2778)	0.0351 (0.1646)	0.0018 (0.1667)	0.0029 (0.1662)
	20	0.0908 (0.2345)	0.0022 (0.1838)	-0.0135 (0.1787)	0.0173 (0.1142)	-0.0003 (0.1149)	-0.0000 (0.1149)
0.5	30	0.0571 (0.1686)	0.0014 (0.1428)	-0.0048 (0.1411)	0.0111 (0.0924)	-0.0009 (0.0929)	-0.0008 (0.0929)
	40	0.0410 (0.1369)	0.0004 (0.1208)	-0.0029 (0.1200)	0.0082 (0.0798)	-0.0009 (0.0801)	-0.0008 (0.0801)
	50	0.0324 (0.1174)	0.0005 (0.1061)	-0.0016 (0.1057)	0.0063 (0.0715)	-0.0010 (0.0717)	-0.0010 (0.0717)
	10	0.4330 (0.9830)	0.0031 (0.6178)	-0.1810 (0.5368)	0.0171 (0.1131)	0.0004 (0.1138)	0.0006 (0.1137)
	20	0.1775 (0.4620)	0.0009 (0.3625)	-0.0303 (0.3526)	0.0085 (0.0793)	-0.0004 (0.0795)	-0.0003 (0.0795)
1.0	30	0.1149 (0.3370)	0.0034 (0.2851)	-0.0090 (0.2818)	0.0059 (0.0648)	-0.0001 (0.0649)	-0.0001 (0.0650)
	40	0.0844 (0.2722)	0.0031 (0.2394)	-0.0035 (0.2380)	0.0046 (0.0559)	0.0000 (0.0560)	0.0000 (0.0560)
	50	0.0667 (0.2334)	0.0027 (0.2103)	-0.0015 (0.2094)	0.0034 (0.0501)	-0.0003 (0.0501)	-0.0003 (0.0501)
	10	0.6543 (1.5198)	0.0080 (0.9603)	-0.2688 (0.8293)	0.0121 (0.0927)	0.0010 (0.0931)	0.0011 (0.0930)
	20	0.2657 (0.6906)	0.0008 (0.5419)	-0.0459 (0.5272)	0.0064 (0.0653)	0.0005 (0.0654)	0.0005 (0.0654)
1.5	30	0.1653 (0.4970)	-0.0013 (0.4218)	-0.0197 (0.4173)	0.0040 (0.0530)	-0.0000 (0.0530)	-0.0000 (0.0530)
	40	0.1213 (0.4035)	-0.0003 (0.3559)	-0.0100 (0.3539)	0.0027 (0.0457)	-0.0004 (0.0458)	-0.0003 (0.0458)
	50	0.0963 (0.3489)	0.0005 (0.3152)	-0.0056 (0.3140)	0.0022 (0.0410)	-0.0002 (0.0410)	-0.0002 (0.0410)
	10	0.8784 (2.0228)	0.0149 (1.2756)	-0.3550 (1.1005)	0.0090 (0.0796)	0.0007 (0.0798)	0.0007 (0.0798)
	20	0.3588 (0.9360)	0.0050 (0.7348)	-0.0575 (0.7144)	0.0042 (0.0559)	-0.0003 (0.0560)	-0.0003 (0.0560)
2.0	30	0.2319 (0.6808)	0.0087 (0.5762)	-0.0161 (0.5694)	0.0026 (0.0456)	-0.0004 (0.0457)	-0.0004 (0.0457)
	40	0.1707 (0.5483)	0.0079 (0.4820)	-0.0054 (0.4788)	0.0020 (0.0397)	-0.0003 (0.0397)	-0.0003 (0.0397)
	50	0.1317 (0.4702)	0.0038 (0.4244)	-0.0044 (0.4227)	0.0015 (0.0354)	-0.0003 (0.0354)	-0.0003 (0.0354)
	10	1.6928 (3.8947)	-0.0150 (2.4553)	-0.7463 (2.1403)	0.0053 (0.0565)	0.0011 (0.0566)	0.0011 (0.0566)
	20	0.6965 (1.8473)	-0.0080 (1.4543)	-0.1329 (1.4157)	0.0027 (0.0398)	0.0004 (0.0398)	0.0004 (0.0398)
4.0	30	0.4358 (1.3260)	-0.0078 (1.1271)	-0.0573 (1.1144)	0.0018 (0.0325)	0.0003 (0.0325)	0.0003 (0.0325)
	40	0.3110 (1.0752)	-0.0123 (0.9521)	-0.0385 (0.9465)	0.0014 (0.0282)	0.0003 (0.0282)	0.0003 (0.0282)
	50	0.2406 (0.9169)	-0.0138 (0.8318)	-0.0301 (0.8289)	0.0011 (0.0253)	0.0002 (0.0253)	0.0002 (0.0253)

where $r_{h,k}$ and $\mathrm{RMSE}_{h,k}$ correspond to the 16 different values of the bias and the root mean squared errors of each estimator given in Table 1. The results are reported in Tables 2-3.

From Table 2, we see that integrated bias squared norm of the corrected estimates (BC and PB) are smaller than ML estimates for both parameter α and λ. From Table 2, we can see that the average root mean-squared error of the corrected estimates (BC and PB) are smaller than ML estimates for α, while for λ the ARMSE are quite similar. Therefore, these simulation results show that second-order bias reduction is quite successful in bringing the corrected estimates closer to their true values.

Table 2. Integrated bias squared norm.

n	Estimator of α ML	Estimator of α BC	Estimator of α PB	Estimator of λ ML	Estimator of λ BC	Estimator of λ PB
10	0.9274	0.0103	0.3990	0.0189	0.0011	0.0015
20	0.3806	0.0043	0.0695	0.0094	0.0004	0.0004
30	0.2398	0.0055	0.0284	0.0061	0.0005	0.0004
40	0.1728	0.0067	0.0181	0.0045	0.0005	0.0004
50	0.1342	0.0065	0.0139	0.0035	0.0005	0.0005

Table 3. Average root mean-squared error.

n	Estimator of α ML	Estimator of α BC	Estimator of α PB	Estimator of λ ML	Estimator of λ BC	Estimator of λ PB
10	2.1352	1.3464	1.1701	0.1077	0.1086	0.1084
20	1.0034	0.7892	0.7680	0.0752	0.0755	0.0755
30	0.7226	0.6135	0.6066	0.0611	0.0613	0.0613
40	0.5852	0.5172	0.5141	0.0528	0.0530	0.0530
50	0.5004	0.4532	0.4515	0.0473	0.0474	0.0474

4.2 Wind real data modeling

The data consist of annual maximum wind speed of six weather stations localized in state of Tocantins, Brazil. The data were obtained from the website http://www.inmet.gov.br/portal. Some descriptive statistics of the observed annual maximum wind speed for the stations are summarized in Table 4. Note that the values of skewness are positive for four stations, indicating that the data are right-skewed.

Table 4. Descriptive statistics of the wind speed data for all weather stations.

Station	Period	n	Min	Mean	Med	Max	SD	Skewn	Kurt
82659	1980-2016	34	1.4667	3.0441	3.0667	5.0000	1.1491	0.1541	1.5672
82863	1977-2016	40	2.5667	4.8862	4.6667	8.3333	1.6084	0.4240	2.0874
83033	1993-2016	21	3.6667	5.4817	5.4667	7.1000	1.0460	-0.1572	2.1114
83064	1961-2016	56	2.2667	3.9939	3.9750	6.0000	0.9736	0.1973	2.3523
83228	1975-2016	42	3.1000	4.3657	4.3333	5.6667	0.6096	-0.1140	2.6386
83235	1961-2016	56	2.3333	3.9994	3.6667	6.6667	0.9424	0.8900	3.4780

In Table 5 we report the ML estimates and the bias corrections estimates along with the asymptotic standard errors calculated from Equation (9). We can observe that the maximum likelihood estimates of α and λ are greater than the second order bias corrected estimates for all stations, this suggests that the ML estimates are overestimating the true value of the parameters. We also observe that the corrected ML estimates of α have smaller standard errors than the uncorrected estimates.

In order to test whether the data sets fits the CA distribution and whether the bias-corrected estimates yield better fits than the uncorrected estimates, we perform the goodness-of-fit tests based on Kolmogorov-Smirnov (KS), Cramér-von-Mises (CM) and Anderson-Darling (AD) statistics. The p-values of these statistics are shown in Table 8. We have used the function *mledist* from *fitdistrplus* library, (Delignette-Muller, 2015), available in *R* environment, (R Core Team, 2017), to find the ML estimates. The p-values associated with Kolmogorov-Smirnov, Anderson-Darling and Cramér-von Mises tests were calculated using 10,000 nonparametric bootstrap resamples applying the functions *ks.test*, *ad.test* and *cvm.test*, available in *goftest* (Faraway et al., 2014) R library. Here we are to note that CA distribution can be used to model the annual maximum wind speed data for the six stations. We can see from Table 8 that the p-values of KS, CM and AD computed from bias-corrected estimators are greater than the uncorrected estimator (except KS for one station). This means that bias corrected estimates provide better fits than the ML estimates. This conclusion is also supported by the empirical and fitted CDF plots in Figure 2. Furthermore, in Table 6, we compare the suitability of the CA distribution against eight commonly used probability distributions to modeling wind speed data. The assessment of the goodness-of-fit is based on the log-likelihood values, since all distributions have the same number of parameters. The results are reported in Table 6 and we can see that CA distribution is the best model among the others. The superscripts indicates the rank obtained by the estimation method (the smaller the better). The line named as rank total (TR) shows the sum of the ranks.

Table 5. Point estimates (standard errors) for all weather stations.

	Estimator of α			Estimator of λ		
Station	ML	BC	PB	ML	BC	PB
82659	0.1081 (0.0262)	0.0986 (0.0239)	0.0978 (0.0237)	2.4341 (0.1844)	2.4132 (0.1931)	2.4121 (0.1939)
82863	0.0559 (0.0125)	0.0517 (0.0116)	0.0515 (0.0115)	4.1773 (0.2365)	4.1572 (0.2459)	4.1627 (0.2464)
83033	0.1148 (0.0354)	0.0984 (0.0304)	0.0962 (0.0297)	5.1707 (0.2277)	5.1556 (0.2459)	5.1539 (0.2487)
83064	0.1359 (0.0257)	0.1286 (0.0243)	0.1279 (0.0242)	3.6335 (0.1282)	3.6268 (0.1317)	3.6250 (0.1321)
83228	0.3323 (0.0725)	0.3086 (0.0673)	0.3075 (0.0671)	4.2329 (0.0946)	4.2297 (0.0982)	4.2315 (0.0984)
83235	0.1639 (0.0310)	0.1551 (0.0293)	0.1540 (0.0291)	3.7170 (0.1167)	3.7115 (0.1200)	3.7098 (0.1204)

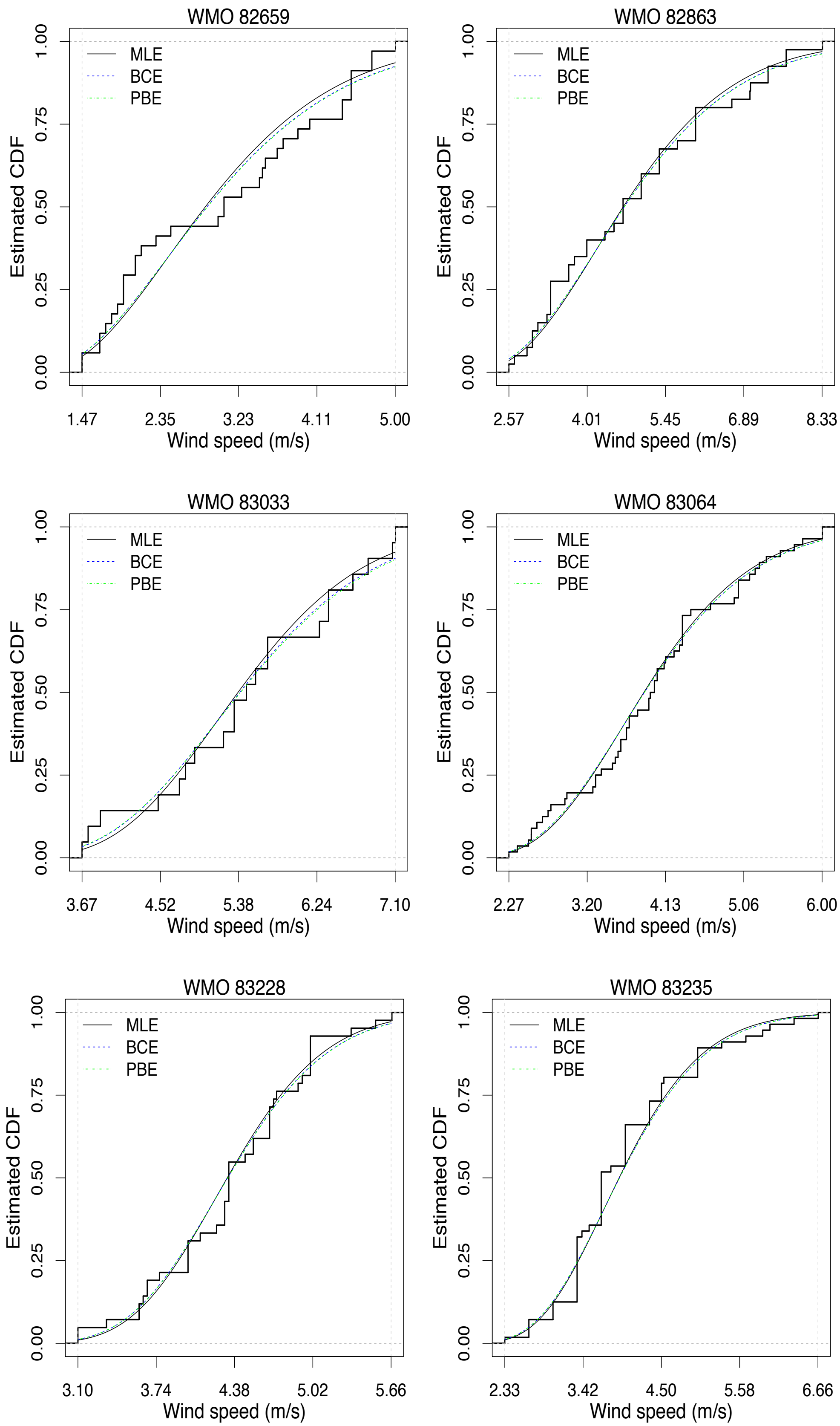


Figure 2. Empirical and fitted CDFs for all examined stations.

Table 6. Negative of the log likelihood values, $-2\log(L)$, of the competing distributions.

Station	CA	Weibull	Gamma	Log-normal	Log-logistic	IG	Gumbel	BS	Nakagami
82659	101.4117^{1}	103.0848^{5}	103.1692^{6}	103.6138^{7}	107.4095^{9}	103.0738^{4}	104.3379^{8}	102.9892^{2}	103.0680^{3}
82863	145.7202^{1}	149.6565^{8}	147.2638^{5}	146.9742^{4}	150.4022^{9}	146.5940^{3}	147.4252^{6}	146.5772^{2}	148.0746^{7}
83033	61.3816^{4}	60.0757^{1}	61.0546^{3}	61.6314^{7}	62.5727^{8}	61.6019^{6}	63.3180^{9}	61.5889^{5}	60.6320^{2}
83064	154.2422^{3}	155.7540^{7}	154.1821^{2}	155.1365^{6}	157.5565^{9}	155.0057^{5}	157.1036^{8}	154.9604^{4}	153.9516^{1}
83228	78.1260^{4}	77.5426^{3}	77.4511^{2}	78.2403^{7}	79.0705^{8}	78.2372^{6}	83.3099^{9}	78.2280^{5}	76.8940^{1}
83235	143.7640^{5}	156.2722^{9}	145.2559^{7}	143.5298^{4}	144.2417^{6}	143.4797^{2}	142.3888^{1}	143.5114^{3}	147.7691^{8}
RT	18^{1}	33^{6}	25^{4}	35^{7}	49^{9}	26^{5}	41^{8}	21^{2}	22^{3}

Table 7. Voung test (p-values) comparing CA distribution with others.

Weibull	Gamma	Log-normal	Log-logistic	IG	Gumbel	BS	Nakagami
0.595 (0.276)	2.146 (0.016)	3.782 (0.000)	8.895 (0.000)	2.538 (0.006)	5.876 (0.000)	2.773 (0.003)	0.882 (0.189)
1.522 (0.064)	2.171 (0.015)	2.460 (0.007)	4.628 (0.000)	1.806 (0.035)	2.156 (0.016)	1.959 (0.025)	1.479 (0.070)
-0.650 (0.742)	-0.788 (0.785)	1.378 (0.084)	1.218 (0.112)	1.564 (0.059)	2.545 (0.005)	1.482 (0.069)	-1.014 (0.845)
0.497 (0.310)	-0.072 (0.529)	2.107 (0.018)	1.929 (0.027)	2.267 (0.012)	3.379 (0.000)	2.203 (0.014)	-0.192 (0.576)
-0.168 (0.567)	-1.282 (0.900)	0.872 (0.192)	0.654 (0.256)	1.090 (0.138)	2.854 (0.002)	1.008 (0.157)	-1.233 (0.891)
3.457 (0.000)	1.943 (0.026)	-0.555 (0.711)	0.244 (0.404)	-0.799 (0.788)	-0.686 (0.754)	-0.750 (0.773)	2.375 (0.009)

Table 8. p-values associated to goodness-of-fit measures for all weather stations.

		KS			CM			AD	
Station	ML	BC	PB	ML	BC	PB	ML	BC	PB
82659	0.4697	0.5185	0.5205	0.3129	0.3970	0.4039	0.3247	0.4056	0.4112
82863	0.6574	0.7285	0.7192	0.7725	0.8539	0.8515	0.7825	0.8570	0.8547
83033	0.9390	0.9763	0.9775	0.8936	0.9252	0.9262	0.8599	0.9042	0.9047
83064	0.8883	0.8736	0.8684	0.7438	0.7689	0.7677	0.7557	0.7962	0.7976
83228	0.4790	0.4961	0.5090	0.6607	0.6748	0.6836	0.7368	0.7534	0.7578
83235	0.2996	0.3054	0.3097	0.4005	0.3801	0.3795	0.4824	0.4735	0.4731

Therefore, using the interpretation for λ given in Section 1, the estimates given in Table 5 can be interpreted as follows. The most frequent wind speed at: station 82659 is around 2.4; station 82863 is around 4.2; station 83033 is around 5.2; station 83064 is around 3.6; station 83228 is around 4.2; station 83235 is around 3.7.

5. Conclusions

In this paper, we have adopted a corrective approach to derive analytical expressions for the second order biases of the maximum likelihood estimators of the parameters of the Chaudhry-Ahmad distribution. Furthermore, we have also considered an alternative bias-correction mechanism through bootstrap resampling. The biases of the proposed estimators are of order $\mathcal{O}\left(n^{-2}\right)$, whereas for the maximum likelihood estimators they are of order $\mathcal{O}\left(n^{-1}\right)$, indicating that the proposed estimates converge to their true value considerably faster than those of the maximum likelihood estimates.

The numerical evidence shows that the proposed bias corrected estimators are quite attractive because they outperform the maximum likelihood estimates in terms of biases, integrated bias squared norm and root mean-squared error. Further, our analytic bias correction is found to be superior to the alternative of bias-correction via the bootstrap in terms of bias reduction. The proposed bias-corrected estimators are strongly recommended over maximum likelihood estimator, especially when the sample size is small or moderate since it has smaller bias and root mean-squared error

Acknowledgements

The authors are thankful to the referees and to the Editors-in-Chiefs for many valuable suggestions. Josmar Mazucheli gratefully acknowledge the partial financial support from Fundação Araucária (Grant 064/2019 - UEM/Fundação Araucária).

References

Abramowitz, M. and Stegun, I.A., 1974. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover, New York.

Chaudhry, M.A. and Ahmad, M., 1993. On a probability function useful in size modelling. Canadian Journal of Forest Research, , 23, 1679–1683.

Cox, D.R. and Hinkley, D.V., 1979. Theoretical Statistics. CRC Press, Boca Raton, FL, US.

Cox, D.R. and Reid, N., 1987. Parameter orthogonality and approximate conditional inference. Journal of the Royal Statistical Society B, 49, 1–39.

Cox, D.R. and Snell, E.J., 1968. A general definition of residuals. Journal of the Royal Statistical Society B, 30, 248–275.

Cribari-Neto, F. and Vasconcellos, K.L.P., 2002. Nearly unbiased maximum likelihood estimation for the beta distribution. Journal of Statistical Computation and Simulation, 72, 107–118.

Davison, A.C. and Hinkley, D.V., 1997. Bootstrap Methods and Their Applications. Cambridge University Press, Cambridge.

Delignette-Muller, M.L. and Dutang, C., 2015. fitdistrplus: An R package for fitting distributions. Journal of Statistical Software, 64, 1–34.

Doornik, J.A., 2007. Object-Oriented Matrix Programming Using Ox Timberlake Consultants Press and Oxford, London.

Edwards, A.W.F., 1992. Likelihood. Johns Hopkins University Press, Baltimore.

Efron, B., 1982. The Jackknife, the Bootstrap and other Resampling Plans. SIAM.

Efron, B. and Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall, New York.

Faraway, J., Marsaglia, G., Marsaglia, J., and Baddeley, A., 2014. goftest: Classical goodness-of-fit tests for univariate distributions. R package version 1.0-2.

Giles, D.E., 2012a. Bias reduction for the maximum likelihood estimators of the parameters in the Half-Logistic distribution. Communication in Statistics: Theory and Methods, 41, 212–222.

Giles, D.E., 2012b. A note on improved estimation for the Topp-Leone distribution. Technical Report. Department of Economics, University of Victoria.

Giles, D.E. and Feng, H., 2009. Bias of the maximum likelihood estimators of the two-parameter Gamma distribution revisited. Technical Report. Department of Economics, University of Victoria.

Giles, D.E., Feng, H., and Godwin, R.T., 2013. On the bias of the maximum likelihood estimator for the two-parameter Lomax distribution. Communications in Statistics: Theory and Methods, 42, 1934–1950.

Koutras, V.P., 2011. Two-level software rejuvenation model with increasing failure rate degradation. Dependable Computer Systems, 97, 101–115.

Lagos-Àlvarez, B., Jiménez-Gamero, M.D., and Alba-Fernández, V., 2011. Bias correction in the type I generalized Logistic distribution. Communications in Statistics: Simulation and Computation, 40, 511–531.

Lai, M.T., 2013. Optimum number of minimal repairs for a system under increasing failure rate shock model with cumulative repair-cost limit. International Journal of Reliability and Safety, 7, 95–107.

Lehmann, E.J. and Casella, G., 1998. Theory of Point Estimation. Springer, New York.

Lemonte, A.J., Cribari-Neto, F., and Vasconcellos, K.L., 2007. Improved statistical inference for the two-parameter Birnbaum-Saunders distribution. Computational Statistics and Data Analysis, 51, 4656 – 4681.

Ling, X. and Giles, D.E., 2014. Bias reduction for the maximum likelihood estimator of the parameters of the generalized Rayleigh family of distributions. Communications in Statistics: Theory and Methods, 43, 1778–1792.

Mazucheli, J., Bapat, S.R., and Menezes, A.F.B., 2020. A new one-parameter unit-Lindley distribution. Chilean Journal of Statistics, 11, 53–67.

Mazucheli, J., and Dey, S., 2018. Bias-corrected maximum likelihood estimation of the parameters of the generalized half-normal distribution. Journal of Statistical Computation and Simulation, 88, 1027–1038.

Mazucheli, J., Menezes, A.F.B., and Dey, S., 2018. Improved maximum likelihood estimators for the parameters of the unit-gamma distribution. Communications in Statistics: Theory and Methods, 47, 3767–3778.

Nanos, N. and Montero, G., 2001. Spatial Prediction of Diameter Distribution Models in Forestry. Springer, Dordrecht.

Nanos, N. Tadesse, W., Montero, G., Gil, L., and Alia, R., 2000. Modelling resin production distributions for pinus pinaster by using two probability functions. Annals of Forest Science, 57, 369–377.

Pearson, K., 1893. Contributions to the mathematical theory of evolution. Proceedings of the Royal Society of London, 54, 329–333.

Pearson, K., 1895. Contributions to the mathematical theory of evolution, II: Skew variation in homogeneous material. Philosophical Transactions of the Royal Society of London A, A186, 343–414.

Pearson, K., 1901. Mathematical contributions to the theory of evolution, X: Supplement to a memoir on skew of variation. Philosophical Transactions of the Royal Society of London A, 197, 443–459.

Pearson, K., 1916. Mathematical contributions to the theory of evolution, XIX: Second supplement to a memoir on skew of variation. Philosophical Transactions of the Royal Society of London A, 216, 429–457.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Reath, J., 2016. Improved parameter estimation of the Log-Logistic distribution with applications. Ph.D. thesis. Michigan Technological University.

Saha, K. and Paul, S., 2005. Bias-corrected maximum likelihood estimator of the negative Binomial dispersion parameter. Biometrics, 61, 179–185.

Sankaran, P.G., Nair, N.U., and Sindhu, T.K., 2003. A generalized Pearson system useful in reliability analysis. Statistical Papers, 44, 125–130.

Schwartz, J. and Giles, D.E., 2016. Bias-reduced maximum likelihood estimation of the zero-inflated Poisson distribution. Communications in Statistics: Theory and Methods, 45, 465–478.

Schwartz, J., Godwin, R.T., and Giles, D.E., 2013. Improved maximum-likelihood estimation of the shape parameter in the Nakagami distribution. Journal of Statistical Computation and Simulation, 83, 434–445.

Shakil, M., Kibria, B.M.G., and Singh, J.N., 2010. A new family of distributions based on the generalized Pearson differential equation with some applications. Austrian Journal of Statistics, 39, 259–278.

Shakil, M., Kibria, B.M.G., and Singh, J.N., 2016. Review on generalized Pearson system of probability distributions. Journal of Mathematical Sciences and Mathematics Education, 11, 13–33.

Stavroyiannis, S., 2014. On the generalised Pearson distribution for application in financial time series modelling. Global Business and Economics Review, 16, 1–14.

Teimouri, M. and Nadarajah, S., 2013. Bias corrected MLEs for the Weibull distribution based on records. Statistical Methodology, 13, 12–24.

Teimouri, M. and Nadarajah, S., 2016. Bias corrected MLEs under progressive type-II censoring scheme. Journal of Statistical Computation and Simulation, 86, 2714–2726.

Tsarouhas, P.H. and Arvanitoyannis, I.S., 2010. Reliability and maintainability analysis of bread production line. Critical Reviews in Food Science and Nutrition, 50, 327–343.

Wang, M. and Wang, W., 2017. Bias-corrected maximum likelihood estimation of the parameters of the weighted Lindley distribution. Communications in Statistics: Theory and Methods, 46, 530–545.

Woosley, R.L., Cossman, J., 2007. Drug development and the FDA's critical path initiative. Public Policy, 81, 129–133.

Zhang, G. and Liu, R., 2015. Bias-corrected estimators of scalar skew Normal. Communications in Statistics: Simulation and Computation, 46, 831–839.

Chilean Journal of Statistics
Vol. 11, No. 2, December 2020, 151–162

Data Science and Education
Research Paper

A timetabling system for scheduling courses of statistics and data science: Methodology and case study

André Leite[1], Abel Borges[1], Geiza Silva[1] and Raydonal Ospina[1,⋆]

[1]Departamento de Estatística, CASTLab,
Universidade Federal de Pernambuco, Recife, PE, Brazil

(Received: 25 August 2020 · Accepted in final form: 25 October 2020)

Abstract

The purpose of this work is to present a decision support system for scheduling courses of statistics and data science to help educational institutions. Currently, an increasing demand of statisticians and data scientists around the world of businesses and organizations is observed. By distributing resources, such as the available time for teachers to form those human personnel, is challenging because of the many dependencies that can exist, which must be taken into account. We describe an integer programming formulation to handle a real instance of a courses-to-lecturers timetabling problem based on a case study. The proposed system is successfully applied by experimental runs using course offerings and classroom data from past semesters.

Keywords: Integer programming· Mathematical programming · Timetabling problem.

Mathematics Subject Classification: Primary 90C10 · Secondary 6207.

1. Introduction

Integer programming tries to allocate finite resources, in an optimal way. In usual applications, the problem description leads to obtaining an optimal value, either a maximum or minimum, for an objective function based on the number of units of resources allocated to each competing entity and constraints on the allocation of the resources. The parameters that describe the number of units are referred to as decision variables. This potential has served to solve relevant questions in areas such as industrial research and economy, administration or scientific issues. Hence, within a good theoretical basis, it is possible to automatize general processes or even promote their optimization, as well as keep managers informed to support decision making.

A classic problem which appears often in the literature of combinatorial optimization is the matching problem (Papadimitriou and Steiglitz, 1982; Even et al., 1975). In a particular situation, it is also known as the timetabling problem and can be formulated as a binary integer programming problem (MirHassani, 2006; de Werra, 1985; Selim, 1983).

Periodically, universities, schools or human resources departments face the challenge of assigning tasks to their staff. There exist legal and institutional constraints which must be satisfied besides, eventually, the aim to optimize some aspect, according to a given criteria.

∗Corresponding author. Email: raydonal@de.ufpe.br

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

Usually each instance has very specific features. A huge part of the papers in this area are motivated by scheduling issues in universities or schools (Burke and De Causmaecker, 2002). When handling educational institutions – schools and universities –, it is common to talk about educational timetabling. At this point, it is possible to distinguish three main classes of timetabling problems (Burke et al., 1997; Schaerf, 1999):

- School timetabling problem (STP): In this ranking are found the problems of the weekly schedule between teacher/class in schools/colleges. Here it is considered that the subjects are fixed for each class and the main objective is to avoid one teacher being allocated to two classes simultaneously or two classes having lessons with one particular teacher in one timetable.
- University course timetabling problem (UCTP): It has in view the weekly schedule of all subjects for all periods of the university courses by determining the relation (teacher/classroom/classroom). It differs from the school timetabling, because in this classification the students can choose the subjects (electives) they wish to enroll. The objective is also to minimize the overlap of any of the variables involved.
- Examination timetabling problem (ETP): It deals with the scheduling of exams for university courses, avoiding overlapping of exams of subjects that have students in common and keeping the exam dates as far away from students as possible.

Now it is more clear why models built for specific problems do not claim to serve for all instances. The reader is referred to de Werra (1985), in which the author considers general formulations for the timetabling problem, starting from the simpler or less specific one and then including common constraints in practical applications. Also, he solves this problem using an approach based on graph coloring methods. Despite these generalization difficulties (or maybe because of them), there is a wide scientific production on timetabling which deals with its theory and applications and several approaches are proposed. The massive use of computers to solve timetabling problems probably started with the construction of class-teacher timetables in 1963 (Gotlieb, 1963). A survey conducted by the Automated Scheduling and Planning Group at the University of Nottingham in the year of 1995 obtained feedback from 56 British universities on the use of computers to build timetables (Burke et al., 1997). Then, 42% of the British universities were used to schedule manually, 37% were assisted by computers, and 21% totally automated. The timetabling problem is very popular because it promotes competitions such as the International Timetabling Competition (ITC), which have had three versions (2002, 2007 and 2011), Post et al. (2016). These events had a positive impact in the research community in the sense of stating common instances and so enabling comparisons between the models and algorithms proposed. The binary integer linear programming model is an useful approach to this problem (Bakir and Aksop, 2008; Ferreira et al., 2011; Havas et al., 2013; Sánchez-Partida et al., 2014; Eledum, 2017)).

Data science is currently getting attention because of the Big Data trend. Different sectors of society are now able to continuously gather data from websites, mobile devices, social media tools and legacy systems, not to mention the burgeoning Internet of Things (IoT). Institutions, governments and businesses today are drowning in data. This is information that, when examined with discerning eyes, ostensibly reveals the trends that could help better serve customers, increase sales, keep their businesses growing and help save lives. In global emergencies like the coronavirus disease (COVID-19) pandemic (Ghebreyesus, 2020), open science policies can remove obstacles to the free flow of research data and ideas, and thus accelerate the pace of research critical to combating the disease (Zastrow, 2020). In this sense, Statisticians and Data scientists are in demand because there is a shortage of qualified data science professionals on the market today.

In this work, we design a decision support system for scheduling courses of statistics and data science. We focus to help educational institutions to distributing resources efficiently such as the available time spent for teachers to form those human resources, once dependencies that can exist in the process of distribution need to be taken into account to reduce tensions that confront teachers of statistics and data science in practice (Cobb, 2011; Batanero and Díaz, 2012).

The objective of this paper is to provide a case study of the university courses timetabling class. We examine the Department of Statistics from the Federal University of Pernambuco in Brazil (`https://www.ufpe.br/dep-estatistica`) because it has a similar structure to other undergraduate and graduate programs in statistics and data sciences in Latin America and other countries. Educational questions must be satisfied and we try to take into account and answer the lecturers' subjective preferences for the courses and schedules (weekly, in this case). So far, this process has been realized manually, taking some weeks until a conclusion. We attempt to promote its automation by implementing an optimization model that looks for the best schedules.

The rest of this paper is organized as follows. Section 2 has brief comments on common and specific scheduling rules of the present instance regarding matching lecturers to courses and times that must be considered and describes the integer linear programming model proposed in this case. Section 3 introduce a coefficient for multiple solutions to choose the optimal allocation based on satisfaction of lecturers. Section 4 presents the results obtained after using this model in several semesters. Section 5 concludes the paper.

2. Scheduling rules and mathematical model

Firstly, the model is based on the satisfiability approach of optimizing the lecturer's preferences by courses and schedules. This paradigm guides the objective function. Secondly, as usual when building courses-timetables, we consider common scheduling rules such as (there may be more):

(a) One, and only one, lecturer teaches each class.
(b) Lecturers just may be in one place at a time.
(c) Each course must be taught by the same lecturer.
(d) Lecturers have a maximum load of courses to teach.
(e) Only assign adjacent shifts.
(f) Try to maximize number of graduating students.

Additionally, we consider specific features. For instance, we need to deal with basic and external courses, which have a prefigured timetable by other departments, and manage the choice of courses to offer by semester. Also, classes of the same course should be properly spaced over the weekdays and so on.

Next, we describe the structure of the proposed model. The particular contents and details of the computational implementations are not exhaustive. The notations present in Table 1 is considered, calling this set and explaining as it is convenient in the text. Furthermore, variables and parameters are defined as follows.

In our approach we consider the following decision and auxiliary variables and parameters:

(a) $u(t, c)$ (parameter): Ordinal utility of relation professor-course. It represents the preference of the lecturer t about course c. Each lecturer informs a ordered list of preferred courses and the first one has the greater u value, the second one the second greater u value and so on.

Table 1. Description of sets and indexes used in the model.

Set	Index	Description
$\mathcal{T}, \mathcal{T}_d, \mathcal{T}_s$	t	Lecturers, department lecturers and assistants ones
$\mathcal{C}, \mathcal{C}_{\text{und}}, \mathcal{C}_{\text{ext}}$	c	All courses, undergraduate courses and external ones
$\mathcal{D}$	d	Weekdays
$\mathcal{S}$	s	Shifts: morning, afternoon, night
$\mathcal{B} \equiv \{1, 2\}$	b	First and second time blocks (or time slots) in a given shift
$\mathcal{P}, \mathcal{P}_b$	p	All semesters and semesters in which are offered basic courses
$\mathcal{C}_p$	c	Courses distinguished by semesters
$\mathcal{N}$	n	Students near graduation
$\mathcal{F}_n$	c	Courses required by undergraduating student n
$\mathcal{D}_{\text{grad}}$	(t, d)	Days in which lecturer t teaches some graduate course
$\mathcal{H}$	(t, d, s, b)	Graduate schedules that must to be avoided
$\mathcal{L}$	(t, d, s, b)	Locked schedules of lecturer t
$\mathcal{A}_p$	(d, s, b)	External undergraduate courses schedules by semester
$\mathcal{E}$	(c, d, s, b)	External courses (offered to others departments) schedules

(b) load(t) (parameter): This parameter regards the classes load that lecturer t must satisfy with undergrad courses (or external courses). In the model, its value provides an upper limit to how many courses of this kind he must teach. This depends, for instance, on lecturer being assistant or not, teaching courses or having administrative responsibilities.

(c) $x(t, c, d, s, b)$ (decision variable): Indicator variable of event "lecturer t is allocated to teach course c in day d, shift s and time slot b".

(d) $y(t, c)$ (auxiliary variable): Binary variable which informs whether lecturer t is matched to course c.

(e) $z(t, d)$ (auxiliary variable): Binary variable which indicates whether lecturer t teaches some class in day d.

Here, by building a timetable to lecturer's educational tasks is guided, first of all, by the following criteria: answer as much as possible the preferences of the lecturers for courses and schedules. Then, the objective function is defined. It is intended to maximize the quantity (objective function)

$$Q = \sum_{\mathcal{T}} \sum_{\mathcal{C}} u(t, c) y(t, c) - M \sum_{\mathcal{T}} \sum_{\mathcal{D}} z(t, d).$$

The constant M is a positive large number, and it promotes a penalization on Q when increasing z values, that is an adjacent purpose is to concentrate the teachings of a given lecturer at the minimum feasible number of days. We remark that, by definition, $z(t, d)$ is the indicator variable of the event "lecturer t teaches some class on the day d".

Next, we introduce the following notations and settings to determine the constraints of the problem:

(a) Definition of the auxiliary variable $y(t, c)$: For each pair lecturer-course scheduled, the variable y equals one if, and only if, summing x over all triples day-shift-block equals two, once each course considered here must have two time blocks of classes per week, which is formulated as

$$\sum_{\mathcal{D}}\sum_{\mathcal{S}}\sum_{\mathcal{B}} x(t,c,d,s,b) = 2y(t,c), \quad \forall\ t \in \mathcal{T}, c \in \mathcal{C}.$$

(b) Definition of the auxiliary variable $z(t,d)$: It is intended to concentrate the lecturers' teachings at minimum feasible number of days, which is done by means of the proposed penalization in Q and an inequality stated as

$$\sum_{\mathcal{C}}\sum_{\mathcal{S}}\sum_{\mathcal{B}} x(t,c,d,s,b) \leqslant 6z(t,d), \quad \forall\ t \in \mathcal{T}, d \in \mathcal{D}.$$

Notice that the maximum hypothetical value assumed by the triple sum is six. This would be a reality if the considered lecturer teaches in the two time blocks of all three shifts. Combination of constraints regarding to control the intervals between classes of each course and the lecturers' loads (to be described) avoid this result.

Indeed, though we talk in definition of z, only this constraint does not guarantee that, if lecturer t_0 is scheduled for no classes on day d_0, $z(t_0,d_0) = 0$. But, once we are handling with an integer linear programming model, in the optimal solution the combined effects of this constraint and the penalization in Q act to make z work according to the interpretation we gave to it.

(c) Each department lecturer t should teach a maximum of load(t) undergraduate or external courses, which is established by

$$\sum_{\mathcal{C}} y(t,c) \leqslant \text{load}(t), \quad \forall\ t \in \mathcal{T}_d.$$

(d) Some lecturers also cooperate with the statistics graduate program. In order to sum in z this information in the objective function, let $\mathcal{D}_{\text{grad}}$ be the set of couples (t,d) such that lecturer t teaches in day d some course on statistics graduate program, which is formulated as

$$z(t,d) = 1, \quad \forall\ (t,d) \in \mathcal{D}_{\text{grad}}.$$

Under the same argument, let $\mathcal{H}$ be the set of 4-tuples (t,d,s,b) such that lecturer t teaches some class on the statistics graduate program on day d, shift s and time block b. Notice that $\mathcal{D}_{\text{grad}} = \{(t,d) : (t,d,s,b) \in \mathcal{H}\}$. Generally, the undergraduate schedule is subordinated to the statistics graduate program. Then, to avoid time conflict between both programs, we set

$$\sum_{\mathcal{C}} x(t,c,d,s,b) = 0, \quad \forall\ (t,d,s,b) \in \mathcal{H}.$$

(e) Given a specific time slot, a lecturer must be teaching not more than one class on it. This is a constraint present in almost all classes timetabling problems and stated as

$$\sum_{\mathcal{C}} x(t,c,d,s,b) \leqslant 1, \quad \forall\ t \in \mathcal{T}, d \in \mathcal{D}, s \in \mathcal{S}, b \in \mathcal{B}.$$

(f) Each course must be delivered by the same lecturer, which is established by

$$\sum_{\mathcal{T}} y(t,c) = 1, \quad \forall\ c \in \mathcal{C}.$$

(g) It is necessary to avoid that instructors teach classes on extreme shifts in a day. Let $\mathcal{T}_d$ be the set of the department lecturers. Once in our case none of them showed interest by courses supposed to be offered at night, we write this shift setting

$$\sum_{\mathcal{C}} \sum_{\mathcal{D}} \sum_{\mathcal{B}} x(t, c, d, \text{night}, b) = 0, \quad \forall\, t \in \mathcal{T}_d.$$

(h) We want to avoid classes of a same course happening two days in a row, as well as in two consecutive time blocks at the same shift and day, that is, each course has classes in different and properly spaced weekdays. Then, this can be formulated as

$$\sum_{\mathcal{T}} \sum_{\mathcal{S}} \sum_{\mathcal{B}} x(t, c, d, s, b) + \sum_{\mathcal{T}} \sum_{\mathcal{S}} \sum_{\mathcal{B}} x(t, c, d+1, s, b) \leqslant 1, \quad \forall\, c \in \mathcal{C}, d \in \mathcal{D}.$$

(i) Basic courses constraints: These courses (for example, calculus, linear algebra, and analytic geometry) are offered by a specific department. They are necessary to many other departments of exact sciences. Then, basic courses schedules are preset and thenceforth the concerned departments look for conforming their timetable. Thus,
  (1) Let $\mathcal{C}_p$ be the set of undergraduate courses of semester p and
  (2) Let $\mathcal{A}_p$ be the set of triples (d, s, b) scheduled for basic courses in semester p.
  Hence, in each semester, we restrict undergraduate courses to be matched to time blocks by means of

$$\sum_{\mathcal{T}} \sum_{\mathcal{C}_p} \sum_{\mathcal{A}_p} x(t, c, d, s, b) = 0, \quad \forall\, p \in \mathcal{P}_b.$$

(j) We must guarantee that, in each semester p, a time block is filled with a maximum of one lecturer teaching one undergraduate course. Therefore, for all $p \in \mathcal{P}, d \in \mathcal{D}, s \in \mathcal{S}, b \in \mathcal{B}$, we have

$$\sum_{\mathcal{T}} \sum_{\mathcal{C}_p} x(t, c, d, s, b) \leqslant 1,$$

considering a term $\mathcal{C}_0$, which is the set of optional courses, with no defined semesters.

(k) External courses must be also considered, that is, courses offered by the Department of Statistics to others departments, also have a preset schedule. Let $\mathcal{E}$ be the set of 4-tuples (c, d, s, b) such that external course c is scheduled to time block b, shift s and day d. Then, we get that

$$\sum_{\mathcal{T}} x(t, c, d, s, b) = 1, \quad \forall\, (c, d, s, b) \in \mathcal{E}.$$

Thus, we guarantee that exactly one lecturer is matched to every external course.

(l) It often happens that students are subscribed to courses which belong to different semesters. Some difficulties arise from this fact when making decisions about which courses offer in each semester. Hence, in a first moment, it is considered the request of courses coming from students near to achieve undergraduate level; secondarily, students with no disapprovals have a preference. Let $\mathcal{N}$ denote the set of potential undergraduating students and let $\mathcal{F}_n$, for $n \in \mathcal{N}$, be the set of courses required by a student n. Then, for all $n \in \mathcal{N}, d \in \mathcal{D}, s \in \mathcal{S}, b \in \mathcal{B}$, we set

$$\sum_{\mathcal{T}} \sum_{\mathcal{F}_n} x(t, c, d, s, b) \leqslant 1,$$

which is equivalent to avoid time conflicts between every pair of distinct courses in $\mathcal{F}_n$.

(m) We must avoid also time blocks in which lecturers prefer not teach. Let $\mathcal{L}$ be the set of 4-tuples (t, d, s, b) such that lecturer t prefers do not be scheduled on (d, s, b). Therefore, we have that

$$\sum_{\mathcal{C}} x(t, c, d, s, b) = 0, \quad \forall\ (t, d, s, b) \in \mathcal{L}.$$

If all the constraints are considered, we set it is a constraint satisfaction problem.

3. An alternative coefficient for multiple optimal solutions

It is possible that distinct solutions yield the same optimal value of the objective function. This would mean that the problem has multiple optimal solutions, which is not a bad picture. In order to introduce a model-independent criterion to judge whether some solution is better than another, let $\mathcal{I}$ be the set of optimal solutions of some instance.

We propose a coefficient for each solution i, called here G_i, calculated by the following algorithm. To optimize the understanding, the reader may want to take a look at Table 4 to visualize the process. For particular solutions i and lecturer t, compute for the j-th allocation the expression given by

$$g_j \triangleq g_j(i, t) = \frac{l_{j,t} - k_{j,t} + \delta_j}{l_{j,t} - 1 + \delta_j},$$

where δ_j is the indicator variable of the event "the j-th course is the last one of the list". Here, $k_{j,t}$ is the ranking of the j-th course, in order of preference, matched to lecturer t in the list without all courses up to the $(j-1)$-th assigned course, whose length is $l_{j,t} = l_{1,t} - (j-1)$. Let lecturer t is matched to $m_t \leqslant \text{load}(t)$ disciplines. Note that each $g_j \in (0, 1]$, for $j \in \{1, 2, \ldots, m_t\}$, measures the meeting of the j-th preference given that the previous allocations have already been considered.

Thus, setting

$$G_{i,t} \triangleq \frac{1}{m_t} \sum_{j=1}^{m_t} g_j, \qquad \forall\ i \in \mathcal{I}, t \in \mathcal{T},$$

we may define a satisfaction index

$$G_i \triangleq |\mathcal{T}|^{-1} \sum_{\mathcal{T}} G_{i,t}, \qquad \forall\ i \in \mathcal{I},$$

as a coefficient of how well the solution i meet the lecturers' preferences by courses, where $|\cdot|$ denotes "cardinality of the set". For instance, one may compute $G_{1,13}$ for lecturer $t = 13$ to the unique optimal solution $i = 1$ presented in Table 4 and thus obtain

$$G_{1,13} = \frac{g_1 + g_2}{2} = \frac{1}{2}\left(\frac{5-1}{5-1} + \frac{4-2}{4-1}\right) = \frac{5}{6}.$$

Then, given the distinct optimal solutions in $\mathcal{I}$ of some instance, we will say that $\max_i\{G_i : i \in \mathcal{I}\}$ indicates which is the best solution.

4. Application and results

The model was implemented in AMPL (a mathematical programming language) an algebraic modeling language to describe and solve high-complexity problems for large-scale (Fourer et al., 1987) and solved by means of the Gurobi (https://www.gurobi.com) an optimizer software (Bixby, 2007) linked with the R statistical software https://www.r-project.org. A computer with Linux Ubuntu 19.04 - Disco Dingo with AMD quad-core processor with a frequency of 2.80 GHz and 4GB RAM was used to carry out the tests.

The proposed model was applied in three consecutive semesters. In the first semester (Instance 4), there were 903 binary variables and 1,760 linear constraints. For the second semester that the model is applied (Instance 5), the formulation had 2,514 binary variables and 1,055 constraints. In both cases, a unique solution was found after less than 2 seconds. Instance 6 presents a bigger problem with 3,750 binary variables. Therefore, we observe an increase in the size of the instances over the semesters.

Table 4 (Instance 5) shows the preference list of each lecturer, in which the courses and department lecturers are labeled by integers from 1 to 32 and 1 to 18, respectively, besides one assistant lecturer. Courses with square shape labels were later attached to the original lists by a commission to handle feasibility.

For Instance 4, about 70% of total lecturers had first preference met and about 10% of them (lecturers 9 and 10) were not answered in the first three preferences. Tables 2 and 3 compare over the semesters the number of days in which lecturers were scheduled to teach, the secondary criteria. As we said before, the scheduling process was manually executed up to Instance 3. Also, for Instance 5, as high as 80% of total lecturers had first preference and, as Instance 4, 10% were not answered in the first three preferences. Therefore, for Instance 6, about 60% of total lecturers had first preference met and about only 8% of them were not attended in the first three preferences.

Table 2. Distribution of how many days were allocated to each lecturer.

	Institution					
Days	1	2	3	4	5	6
2	75%	48%	75%	89%	75%	83%
3	17%	33%	17%	11%	25%	17%
4	8%	19%	8%	0%	0%	0%

Table 3. Distribution of how many days were allocated to each lecturer whose load is at least two courses.

	Institution					
Days	1	2	3	4	5	6
2	64%	21%	50%	80%	87%	83%
3	24%	50%	33%	20%	13%	17%
4	12%	29%	17%	0%	0%	0%

Table 4. Solution for Instance 4.

Lecturer	Days	Preferences list
1	2	**8**, 4, 3, 1, 7, 9, 5, 10, 6, 14
2	2	**18**, 21
3	2	**11**, 5, 6, 13, 23
4	2	**7**, 5, 1
5	2	2, 9, **14**, 22, 19, 24, **25**
6	2	**2**, 11, 12, 20, 16, 17, 18, 24, 32, **25**
7	2	**25**, **26**
8	2	**21**, 16, 17, **25**
9	2	2, 8, 5, 4, 10, 1, 18, **20**,23
10	3	4, 5, 8, 9, 10, **22**, 23, **12**, **25**
11	2	**1**, 7, 21, 16, **25**
12	2	12, 5, **25**, **26**
13	2	**5**, 8, **10**, 21, 19
14	2	**9**, 11, 15, 20, 18, 21, 23, **24**
15	2	**16**, **17**, **25**
16	2	**6**, **15**, 5, 10, 11, 12, 13, 19, 20, 23, 24
17	2	**4**, **10**, 2, 23, 8
18	3	**19**, 1, **3**, 7, 21, 32, **25**
Assistant	2	**31**, **32**, **23**

where the notations used indicate that:

0 allocated course;

0 allocated course which were not in original list;

0 course added to the original list;

0 course allocated to assistants.

5. Conclusions and perspectives

In this paper we present a scheduling problem at the Department of Statistics of the Federal University of Pernambuco. The case study developed aims to allocate teachers to subjects according to their teaching preferences and reduce the number of days of classes. An entire programming model was proposed and three real instances were tested, relative to three consecutive semesters of the course. The optimal solutions obtained were applied in the department (with good acceptance among teachers).

For future works, it is intended to study, for this model, the influence of the parameters $u(t, c)$ in solutions and even in computational complexity. Also, a parallel approach of the problem based on the coefficients G_i, in the sense of being specific about the required quality of the solutions, may has interesting features. Therefore, during these difficult times of mitigations measures given by the COVID-19 pandemic, we are studying different adaptations of the timetabling formulation present here to include Social Distancing in the re-opening processes of courses in institutions of safe form and include time-space structures as a preferences of teachers, information asymmetries on real-time updates, and travel times.

Acknowledgements

The authors thank Editors and Referees for their valuable suggestions. We gratefully acknowledge partial financial support from CNPq (grant 305305/2019-0) and CAPES, Brazil.

References

Bakir, M. and Aksop, C., 2008. A 0-1 integer programming approach to a university timetabling problem. Hacettepe Journal of Mathematics and Statistics, 37, 41–55.

Batanero, C. and Díaz, C., 2012. Training school teachers to teach probability: reflections and challenges. Chilean Journal of Statistics, 3, 3–13.

Bixby, B., 2007. The gurobi optimizer. Transportation Research Part B, 41, 159–178.

Burke, E. and De Causmaecker, P., 2002. Practice and theory of automated timetabling. Proceeding of the IV 4th International Conference, Gent, Belgium.

Burke, E., Jackson, K., Kingston, J., and Weare, R., 1997. Automated university timetabling: The state of the art. The Computer Journal, 40, 565–571.

Cobb, G.W., 2011. Teaching statistics: Some important tensions. Chilean Journal of Statistics, 2, 31–62.

de Werra, D., 1985. An introduction to timetabling. European journal of operational research, 19, 151–162.

Eledum, H., 2017. The university timetabling problem: Modeling and solution using binary integer programming with penalty functions. International Journal of Applied Mathematics and Statistics, 56, 164–178.

Even, S., Itai, A., and Shamir, A., 1975. On the complexity of time table and multicommodity flow problems. Proceeding of the 16th Annual Symposium on Foundations of Computer Science, pp. 184–193.

Ferreira, P., Karas, E., Palucoski, F., Ribeiro, A., and Silva, A. (2011). Application of integer programming in the distribution of teaching charges in educational institutions. Trends in Applied and Computational Mathematics, 12, 135–144.

Fourer, R., Gay, D.M., and Kernighan, B.W., 1987. AMPL: A mathematical programming language. AT & T Bell Laboratories Murray Hill, New Jersey.

Ghebreyesus, T.A., 2020. WHO Director-General opening remarks at the Member State Briefing on the COVID-19 pandemic evaluation. Available at https://bit.ly/36nRIvG. Retrieval date: July 9, 2020.

Gotlieb, C., 1963. The construction of class-teacher timetables. Proceeding IFIP, Vol. 62, pp. 73–77, Amsterdam.

Havas, J., Olsson, A., Persson, J., and Schierscher, M. (2013). Modeling and optimization of university timetabling: A case study in integer programming. Gothenburg University Publisher, Gothenburg.

MirHassani, S., 2006. A computational approach to enhancing course timetabling with integer programming. Applied Mathematics and Computation, 175, 814–822.

Papadimitriou, C.H. and Steiglitz, K., 1982. Combinatorial Optimization: Algorithms and Complexity. Dover, New York.

Post, G., Di Gaspero, L., Kingston, J., McCollum, B., and Schaerf, A., 2016. The third international timetabling competition. Annals of Operations Research, 239, 69–75.

Schaerf, A., 1999. A survey of automated timetabling. Artificial Intelligence Review, 13, 87–127.

Selim, S., 1983. An algorithm for producing course and lecturer timetables. Computers and Education, 7, 101–108.

Sánchez-Partida, D., Flores, J.L., and Olivares-Benitez, E., 2014. An integer linear programming model for a university timetabling problem considering time windows and consecutive periods. Mathematical Methods of Operational Research, 6, 158–173.

Zastrow, M., 2020. Open science takes on the coronavirus pandemic. Nature, 581, 109–110

Chilean Journal of Statistics
Vol. 11, No. 2, December 2020, 163–176

Distribution Theory
Research Paper

Modeling bounded data with the trapezoidal Kumaraswamy distribution and applications to education and engineering

Jorge Figueroa-Zúñiga[1,*], Rodrigo Sanhueza-Parkes[1], Bernardo Lagos-Álvarez[1], and Germán Ibacache-Pulgar[2,3]

[1]Department of Statistics, Universidad de Concepción, Concepción, Chile,
[2]Department of Statistics, Universidad de Valparaíso, Valparaíso, Chile.
[3]Centro Interdisciplinario de Estudios Atmosféricos y Astroestadística
Universidad de Valparaíso, Valparaíso, Chile.

(Received: 13 October 2020 · Accepted in final form: 30 November 2020)

Abstract

The Kumaraswamy distribution has been a very studied tool in the analysis and modeling of limited-range continuous random variables. Several variants of this distribution have been studied, but they do not have the possibility of lifting the tails of this distribution. However, in many situations, scenarios where the data are bounded and tail-area events occur at one or both tails independently. In order to model these scenarios, we propose the trapezoidal Kumaraswamy distribution. This paper is centered on the trapezoidal Kumaraswamy distribution, which has two intuitive additional parameters with respect to the Kumaraswamy distribution and generalizes this. We study its probability density function and derive some fundamental properties, such as the moments, moment generating function, and characteristic function. Then, the trapezoidal Kumaraswamy distribution is rewritten conveniently as a finite mixture showing that its parameters can be easily estimated using the expectation-maximization algorithm. We report results of a simulation and an application to a real data set. Comparison with several competing distributions indicates that the trapezoidal Kumaraswamy distribution presents a better fit and so it can be quite useful in empirical applications.

Keywords: EM algorithm · Maximum likelihood · Mixture distributions.

Mathematics Subject Classification: 62E15 · 62F10.

1. Introduction

A good alternative for modeling continuous data restricted to a bounded interval is the double bounded distribution (Kumaraswamy, 1980), named after as the Kumaraswamy distribution (Jones, 2009). This distribution provides a wide variety of shapes for its probability density function (PDF) allowing different type of data to be accommodated.

The Kumaraswamy distribution is very flexible. However, it does not consider tail-area events nor high flexibility in the variance specification. In order to add flexibility into the model, other distributions derived from the Kumaraswamy distribution have been pro-

*Jorge Figueroa-Zúñiga. Email: jifiguer@gmail.com

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

posed. For example, the Kumaraswamy Weibull (Cordeiro et al., 2010) and Kumaraswamy-G (Cordeiro and de Castro, 2011) distributions have been derived including two additional positive parameters. The authors studied some of their mathematical properties by presenting special submodels such as: the Kumaraswamy generalized gamma distribution (de Pascoa et al., 2011), which is able to model bathtub-shaped hazard rate functions. The importance of Kumaraswamy generalized gamma distribution is in its capacity to model functions of monotonous failure frequency and non-monotone, which are fairly common in life-time data analysis and reliability. Another case is the Kumaraswamy Gumbel distribution (Cordeiro et al., 2012), which is probably the most widely applied statistical distribution to problems in engineering. Similarly, the Kumaraswamy-log-logistic (De Santana et al., 2012), Kumaraswamy-geometric (Akinsete et al., 2014), and Kumaraswamy Fréchet (Mead and Abd-Eltawab, 2014) distributions, among others of the same family have been proposed. Furthermore, in the same direction, in order to make some existing distributions flexible, other models have been proposed as in Liang et al. (2014), Nadarajah and Kotz (2004), Nadarajah and Kotz. (2006), Akinsete and Famoye (2008), Eugene et al (2002), Cordeiro and dos Santos Brito (2012), among others. Note that the Kumaraswamy distribution, and its extensions, are unable to fit data which are concentrated at both tails. The main objective of this work is to propose a new bounded distribution which is able to model data which are concentrated at both tails.

The reminder of this article is organized as follows. In Section 2, the trapezoidal Kumaraswamy (TK) distribution is proposed and its basic properties are discussed. In Section 3, we estimate parameters through a convenient reparametrization of the TK distribution given in Section 2. Section 4 conducts a Monte Carlo simulation study for both the TK and Kumaraswamy distributions, comparing them. In Section 5, two empirical illustrations are provided corresponding to (i) percent slacks for reduction in pollutant emissions/discharges for carbon dioxide (CO_2) and water (H_2O) in Angolan thermal power plants, and (ii) scores of a university admission test in 1295 school establishments in Metropolitan region of Chile. The results are compared with the classical Kumaraswamy distribution. Finally, discussions, conclusions and further research of the proposed distribution appear in Section 6.

2. The new distribution

In this section, we discuss some properties of the Kumaraswamy distribution and we present the TK distribution as well as its properties.

2.1 Background

The PDF of a random variable Y following a Kumaraswamy distribution is given by

$$f_{\mathrm{K}}(y;\alpha,\beta)=\alpha\beta y^{\alpha-1}(1-y^{\alpha})^{\beta-1},\quad y\in(0,1), \tag{1}$$

where $\alpha>0$ and $\beta>0$. Then, note that

$$\mathrm{E}(Y)=m_1,\quad \mathrm{Var}(Y)=m_2-m_1^2,$$

with m_k denoting the k-th moment of the Kumaraswamy distribution stated as

$$m_k=\frac{\beta\Gamma(1+\frac{k}{\alpha})\Gamma(\beta)}{\Gamma(1+\frac{k}{\alpha}+\beta)}=\beta B\left(1+\frac{k}{\alpha},\beta\right),$$

where B is the beta function.

In practice, the Kumaraswamy distribution has been a useful tool for modeling bounded data. However, it is common in many cases to have data concentrated at both tails independently. Hence, we propose the TK distribution as an extension which allows to model this situation and that it conserve the flexibility of the Kumaraswamy distribution.

2.2 The trapezoidal Kumaraswamy distribution

Let Y follow a TK distribution of parameters a, b, α, β which we denote by $Y \sim \text{TK}(a, b, \alpha, \beta)$. Then, the PDF of Y is established as

$$f_{\text{TK}}(y; a, b, \alpha, \beta) = a + (b - a)y + \left(1 - \frac{a+b}{2}\right) f_{\text{K}}(y; \alpha, \beta), \tag{2}$$

with $0 < y < 1$, $0 \leq a, b \leq 2$, $0 \leq a + b \leq 2$ and $f_{\text{K}}(y; \alpha, \beta)$ being the Kumaraswamy PDF of parameters α and β given in Equation (1). The parameters a and b can be intuitively interpreted as the lift at the left and right tails of the PDF respectively; see Figure 1. As a particular case, we have that, when $a = b = 0$, the standard Kumaraswamy distribution is recovered –see Equation (1)– and we propose the rectangular Kumaraswamy distribution when $a = b = \theta$.

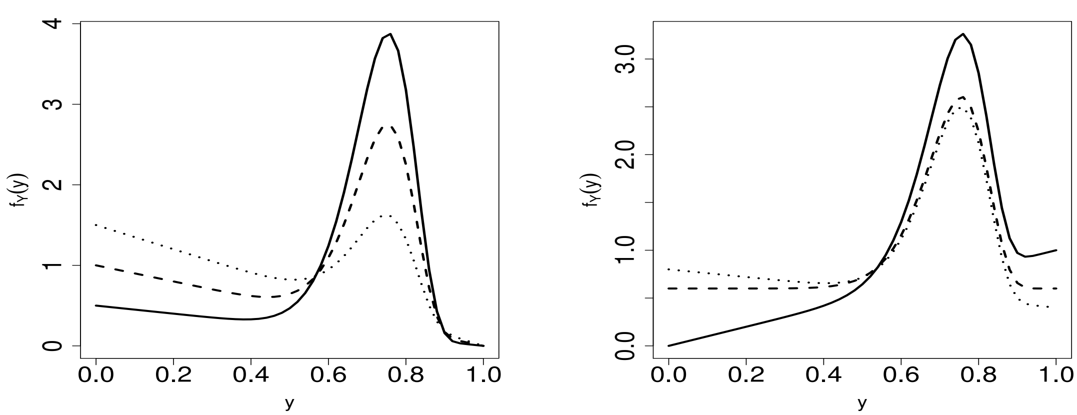


Figure 1. Examples of TK PDF with $\alpha = 10, \beta = 15$ and different values of the parameters (a, b). Left: $(a, b) = (0.5, 0)$ (solid line), $(a, b) = (1, 0)$ (dashed line) and $(a, b) = (1.5, 0)$ (dotted line); right: $(a, b) = (0, 1)$ (solid line), $(a, b) = (0.6, 0.6)$ (dashed line) and $(a, b) = (0.8, 0.4)$ (dotted line).

We now present some properties of the TK distribution. Let $Y \sim \text{TK}(a, b, \alpha, \beta)$. Then, the k-th moment of Y is given by

$$m_k = \text{E}(Y^k) = \frac{a}{k+1} + \frac{b-a}{k+2} + \left(1 - \frac{a+b}{2}\right) m_k^*, \tag{3}$$

where m_k^* is the k-th moment of the Kumaraswamy distribution of parameters α, β. Then, Equation (3) can be written as

$$\begin{aligned} m_k &= \frac{a}{k+1} + \frac{b-a}{k+2} + \left(1 - \frac{a+b}{2}\right) \frac{\beta\Gamma\left(1 + k/\alpha\right)\Gamma\left(\beta\right)}{\Gamma\left(1 + \beta + k/\alpha\right)} \\ &= \frac{a}{k+1} + \frac{b-a}{k+2} + \left(1 - \frac{a+b}{2}\right) \beta B\left(1 + k/\alpha, \beta\right). \end{aligned} \tag{4}$$

With the expression defined in Equation (4), it is easy to deduce that

$$\begin{aligned}
\mathrm{E}(Y) &= \frac{a+2b}{6} + \left(1 - \frac{a+b}{2}\right)\beta B\left(\frac{\alpha+1}{\alpha}, \beta\right), \\
\mathrm{Var}(Y) &= \frac{3a+9b-(a+2b)^2}{36} \\
&+ \left(1 - \frac{a+b}{2}\right)\beta\left(B\left(\frac{\alpha+2}{\alpha}, \beta\right) - \frac{(a+2b)}{3}B\left(\frac{\alpha+1}{\alpha}, \beta\right)\right. \\
&- \left.\left(1 - \frac{a+b}{2}\right)\beta B^2\left(\frac{\alpha+1}{\alpha}, \beta\right)\right).
\end{aligned}$$

The moment generating function of the random variable Y is given by

$$M_Y(t) = \mathrm{E}\left(\mathrm{e}^{tY}\right) = 1 + \sum_{k=1}^{\infty} m_k \frac{t^k}{k!}, \quad t \in \mathrm{R},$$

and its characteristic function is stated as

$$\varphi_Y(t) = \mathrm{E}\left(\mathrm{e}^{itY}\right) = 1 + \sum_{k=1}^{\infty} m_k \frac{(it)^k}{k!}, \quad t \in \mathrm{R}.$$

3. Estimation of trapezoidal Kumaraswamy distribution parameters

In this section, we discuss how to estimate the parameters of the TK distribution efficiently.

3.1 Log-likelihood function

The likelihood function for a sample of n observations from the TK distribution is specified as

$$\mathcal{L}(a,b,\alpha,\beta) = \prod_{i=1}^{n}\left(a + (b-a)y_i + \left(1 - \frac{a+b}{2}\right) f_{\mathrm{K}}(y_i;\alpha,\beta)\right). \tag{5}$$

Then, one strategy to build estimators for its parameters is to maximize the corresponding log-likelihood given by

$$\ell(a,b,\alpha,\beta) = \sum_{i=1}^{n}\log\left(a + (b-a)y_i + \left(1 - \frac{a+b}{2}\right) f_{\mathrm{K}}(y_i;\alpha,\beta)\right). \tag{6}$$

The maximum likelihood estimators of a, b, α and β are obtained from the differentiation of Equation (6) with respect to the mentioned parameters and equating to zero. However, in this case, the obtained equations do not have closed-form. Hence, they need to be obtained by numerically maximizing the log-likelihood function using a nonlinear optimization algorithm, such as the Newton algorithm or the quasi-Newton algorithm, such the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Nocedal and Wright, 1999).

An efficiently strategy to estimate the parameters of the TK distribution is solving this problem as a missing data problem, specifying the likelihood function defined in Equation (5) conveniently, as described in next subsection.

3.2 The EM algorithm

First, we can observe that Equation (2) can be rewrite as a mixture of beta distributions and a Kumaraswamy distribution, that is, by means of

$$f_{\mathrm{TK}}(y; a, b, \alpha, \beta) = \frac{a}{2}(2-2y) + \frac{b}{2}2y + \left(1 - \frac{a+b}{2}\right) f_{\mathrm{K}}(y; \alpha, \beta), \tag{7}$$

where $f_1(y) = f_B(y; 1, 2) = 2 - 2y$ and $f_2(y) = f_B(y; 2, 1) = 2y$ are particular cases of the beta PDF defined as $f_B(y; \alpha^*, \beta^*)$, whereas $f_3(y) = f_{\mathrm{K}}(y; \alpha, \beta)$ corresponds to Kumaraswamy PDF described in Equation (1). In addition, here $w_1 = a/2$, $w_2 = b/2$ and $w_3 = (1-(a+b)/2)$ are the weights such that $w_1+w_2+w_3 = 1$ and $0 \leq w_1, w_2, w_3 \leq 1$. Then, this problem can be solved as a finite mixture of distributions by using the expectation-maximization (EM) algorithm (McLachlan and Peel, 2004). The EM algorithm is a general method for finding maximum likelihood estimates when there are missing values or latent variables. The idea behind the EM algorithm applied to mixture models is to assume that the mixture is generated by missing observations of a discrete random variable Z, where $z_i \in \{1, 2, 3\}$ indicates which mixture component generated the observation y_i. The likelihood function of the complete data formed by the observed data ($\boldsymbol{y}$) and the unobserved data ($\boldsymbol{z}$), for a sample of n, is established by

$$\begin{aligned} p_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y}, \boldsymbol{z}; \Theta) = \prod_{i=1}^{n} p_{\boldsymbol{Y},\boldsymbol{Z}}(y_i, z_i; \Theta) = \prod_{i=1}^{n} & \left(\frac{a}{2}(2-2y_i)\right)^{\mathbb{1}_{z_i=1}} \left(\frac{b}{2}(2y_i)\right)^{\mathbb{1}_{z_i=2}} \\ & \times \left(\left(1 - \frac{a+b}{2}\right) f_{\mathrm{K}}(y_i; \alpha, \beta)\right)^{\mathbb{1}_{z_i=3}}, \end{aligned}$$

where $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are the random vectors associated with ($\boldsymbol{y}$) and ($\boldsymbol{z}$), respectively. In addition, $\Theta = (a, b, \alpha, \beta)$ is the parameter vector and $\mathbb{1}$ is the indicator function, that is $\mathbb{1}_{z_i=j} = 1$ if $z_i = j$ (with $j \in \{1, 2, 3\}$) holds, and $\mathbb{1}_{z_i=j} = 0$, otherwise. Note that, in the EM algorithm, it is necessary to specify an auxiliary function Q, corresponding to the conditional expectation of the log-likelihood function with complete data $(\boldsymbol{y}, \boldsymbol{z})$ given the observed data $Y = y$, and a parameterization $\Theta^{(p-1)}$, that is, we have that

$$\begin{aligned} Q\left(\Theta, \Theta^{(p-1)}\right) &= \mathrm{E}_{\boldsymbol{Y},\boldsymbol{Z},\Theta^{(p-1)}}(\log(p_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{Y}, \boldsymbol{Z}; \Theta))) \\ &= \sum_{i=1}^{n} \mathrm{E}_{\boldsymbol{Y},\boldsymbol{Z},\Theta^{(p-1)}}(\log(p_{\boldsymbol{Y},\boldsymbol{Z}}(Y_i, Z_i; \Theta))) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{3} r_{ij}^{(p-1)} \log(p_{\boldsymbol{Y},\boldsymbol{Z}}(y_i, z_i; \Theta)) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{3} r_{ij}^{(p-1)} (\log(w_j f_j(y_i; \Theta))), \end{aligned}$$

where $w_1 = a/2$, $w_2 = b/2$, $w_3 = (1-(a+b)/2)$, $f_1(y_i; \Theta) = 2 - 2y_i$, $f_2(y_i; \Theta) = 2y_i$, $f_3(y_i; \Theta) = f_{\mathrm{K}}(y_i; \alpha, \beta)$ as in Equation (7), and

$$r_{ij}^{(p-1)} = \mathrm{P}(Z_i = j; Y_i = y_i, \Theta^{(p-1)}) = \frac{w_j^{(p-1)} f_j(y_i; \Theta^{(p-1)})}{\sum_{l=1}^{3} w_l^{(p-1)} f_l(y_i; \Theta^{(p-1)})}.$$

In the E-Step, we need to find the expected value of $\mathbb{1}_{z_i=j}$ for $j = 1, 2, 3$ given y_i and the current parameterization $\Theta^{(p-1)}$, stated as

$$\mathrm{E}\left[\mathbb{1}_{z_i=j}; y_i, \Theta^{(p-1)}\right] = r_{ij}^{(p-1)}.$$

In the M-Step, we find $\Theta^{(p)}$ which maximizes $Q(\Theta, \Theta^{(p-1)})$. Calculating the derivates of Q with respect to w_1, w_2, w_3 under the restriction $w_1 + w_2 + w_3 = 1$, is possible obtain the estimators

$$w_j^{(p)} = \frac{\sum_{i=1}^n r_{ij}^{(p-1)}}{\sum_{i=1}^n \sum_{j=1}^3 r_{ij}^{(p-1)}} = \frac{n_j^{(p-1)}}{n}.$$

Additionally, the derivates with respect to α and β lead to the usual maximum likelihood estimators of the Kumaraswamy distribution, which solve the equations expressed as

$$(\beta - 1)\, \frac{\sum_{i=1}^n r_{i3}^{(p-1)} y_i^{\alpha} \log(y_i)}{1 - y_i^{\alpha}} - \frac{n_3^{(p-1)}}{\alpha} - \sum_{i=1}^n r_{i3}^{(p-1)} \log(y_i) = 0 \qquad (8)$$

$$\frac{n_3^{(p-1)}}{\beta} + \sum_{i=1}^n r_{i3}^{(p-1)} \log(1 - y_i^{\alpha}) = 0. \qquad (9)$$

The corresponding estimates generated from Equations (8) and (9) can be obtained using the quasi-Newton algorithm. Once we update the parameters, we must repeat both the E and M steps, iteratively. In our case, in the M-step of the algorithm, we use the BFGS method to iteratively solve the non-linear maximization problem associated. The BFGS method is implemented in the `R` software by the functions `optim` and `optimx`; see `www.R-project.org` and R Core Team (2018).

4. Simulation study

In this section, we conduct a simulation study to compare the performance of the TK distribution with the Kumaraswamy distribution for samples generated from each of them.

4.1 Scenario of the simulations

In order to capture the particular tail behavior of each one, we use a sample size of 1000 and generate 100 sample sets to calculate the mean log-likelihood function and the Akaike information criterion (AIC). First, we simulate from the TK distribution with parameters given by $\Theta = (0.2, 0.5, 7, 10)$, that is, we simulate an asymmetric distribution with independent lifting in both tails to capture the essense of the proposed TK distribution. Second, we collect a sample from the Kumaraswamy distribution with parameters stated as $\Theta_B = (7, 10)$, that is, an asymmetric distribution but without lifted tails in its PDF.

4.2 Results of the simulations

In our first simulation from the TK distribution, we can observe in Table 1 that the TK distribution achieves a better fit than the Kumaraswamy distribution. In Table 2, we can appreciate that the Kumaraswamy distribution tries to fit the model by increasing the variance, that is, finding small values for α and β to overcome the inability of this distribution to raise the tails.

Table 1. Comparison between the mean log-likelihood and mean AIC of the TK and Kumaraswamy distributions for 100 samples of size 1000 drawn from a TK distribution with parameters $(0.2, 0.5, 7, 10)$

Distribution	Log-likelihood	AIC
TK	363.26	−718.53
Kumaraswamy	237.38	−470.75

Table 2. Comparison between the mean of the estimated parameters of the TK and Kumaraswamy distributions for 100 samples of size 1000 drawn from a TK distribution with parameters $(0.2, 0.5, 7, 10)$

	Estimated parameter			
Distribution	a	b	α	β
True	0.20	0.50	7.00	10.00
TK	0.20	0.49	7.03	10.28
Kumaraswamy	-	-	2.72	1.94

In Figure 2, we can see the histogram for simulated data from the TK distribution and the adjusted PDFs for the TK and Kumaraswamy distributions. The interpretation of the estimated parameters a, b is straightforward and corresponds exactly to the lifting of the tails of PDF in left and right tails respectively. In addition, note that the Kumaraswamy distribution is unable to capture this lifting.

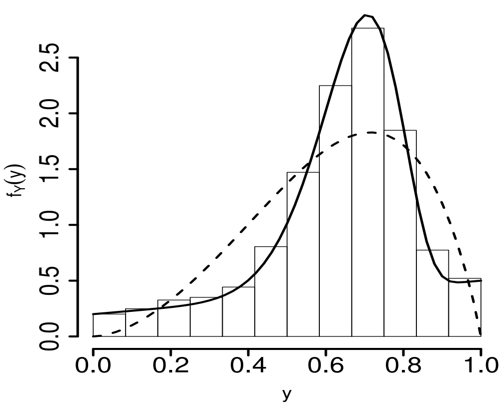


Figure 2. Histogram for simulate data set from TKD and adjusted PDFs for two different models: In solid line, the TK model; In dashed line the Kumaraswamy model.

Table 3 reports the relative bias (RB) and the root-mean-squared error (RMSE) for each parameter estimator over the 100 simulated samples under the TK distribution. They are defined as

$$\mathrm{RB}(\theta) = \frac{1}{100}\sum_{i=1}^{100}\left(\frac{\widehat{\theta}^{(i)} - \theta}{\theta}\right), \quad \mathrm{MSE}(\theta) = \frac{1}{100}\sum_{i=1}^{100}(\widehat{\theta}^{(i)} - \theta)^2,$$

where θ represents any particular parameter, and $\widehat{\theta}^{(i)}$ is the estimate of θ for the i-th sample. Table 3 reports that the estimate of each parameter in each data set is reasonable when fitting the TK distribution.

Table 3. RB and RMSE of each parameter under 100 samples of size 1000 drawn from a TK distribution with parameters $(0.2, 0.5, 7, 10)$.

	Parameter			
Indicator	a	b	α	β
RB	0.00088	−0.00287	0.00038	0.00276
RMSE	0.00554	0.04537	0.08497	0.87242

In our second simulation from the Kumaraswamy distribution, we can observe in Table 4 that the TK distribution achieve an equally good fit than the Kumaraswamy distribution. In Table 5, note that the TK distribution gives similar estimates for the parameters, compared to the Kumaraswamy distribution.

Table 4. Log-likelihood and AIC for simulated data

Distribution	Log-likelihood	AIC
TK	843.52	−1679.03
Kumaraswamy	843.29	−1682.58

Table 5. Comparison between the mean of the estimated parameters of the TK and Kumaraswamy distributions for 100 samples of size 1000 drawn from a Kumaraswamy distribution with parameters $(7, 10)$

	Estimated parameter			
Distribution	a	b	α	β
True	0.00	0.00	7.00	10.00
TK	2.85e-04	1.12e-03	7.07	10.29
Kumaraswamy	-	-	7.05	10.22

Unsurprisingly, when the sample is generated from the Kumaraswamy distribution, we do not see significant differences on the mean log-likelihood and AIC achieved by the two adjusted Kumaraswamy and TK distributions. When the sample is drawn from the TK distribution with a difference between the its two tails, $a = 0.2$ and $b = 0.5$, the best fit in terms of the mean log-likelihood and AIC is achieved by the TK distribution. This can be explained by the fact that the data generated from the tails of the distribution cannot be captured only by using a Kumaraswamy distribution.

5. Empirical illustrations with real data

In this section, in order to illustrate the TK distribution in practice, we apply the proposed results to two real data sets. We compare the goodness of fit between the TK and Kumaraswamy distributions.

5.1 Pollutant emissions in Angolan thermal power plants

Data on Angolan thermal power plants span the period 2010 to 2014 were obtained from a enterprise named ENE-EP. They are based on the plants balance sheets and income statements, which are gathered and organized by ENE-EP as part of regular reporting. The variables of interest for our study are the percent slacks for reduction in pollutant emissions/discharges for CO2 and H2O. This scalar measure deals directly with the input excesses and the output shortfalls of the decision making unit concerned and is typically

Table 7. Log-likelihood and AIC values or H2O data

Indicator	Distribution TK	Kumaraswamy
Log-likelihood	82.21	24.40
AIC	−156.43	−44.81

used as efficiency measure for modeling environmental performance (Barros and Wanke, 2017).

Efficiency scores computed from the slacks based model with undesirable (bad) outputs (SBM-Undesirable) range between 0 and 1, where 1 denotes a maximum or 100 % of efficiency. This suggests that a given thermal plant is operating at the frontier of the productive technology. In fact, efficiency is a productivity ratio between two DMUs: in data envelopment analysis (DEA) based models, all plants are assessed against a convex frontier of best practices formed by the most productive DMUs that can deliver higher outputs consuming lower inputs or benchmarks. In DEA, each production unit is known as a decision making unit (DMU).

Before proceeding, it is worth noting that if the variable assumes the extreme values of zero and one ($Y^* \in [0,1]$), then a practical transformation must be applied (Smithson and Verkuilen, 2006) by

$$y = \frac{(n-1)}{n} y^* + \frac{1}{2n}, \quad y^* \in [0,1],$$

where n is the sample size.

In our study, we consider 160 efficiency scores ($n = 160$) for the 32 Angolan thermal power plants from 2010 to 2014. This efficiency scores has been measures for CO2 and H2O. From Figures 3 and **??**, note that the data distribution have a lifted left tail. Then, it is justified to fit the TK distribution to model these data. The model under consideration is defined by

$$Y_i \overset{\text{IND}}{\sim} \text{TK}(a, b, \alpha, \beta), \quad i = 1, \ldots, 160,$$

where IND stands for independent. Note in Tables 6 and 7 that the TK distribution achieves a best fit compared to the Kumaraswamy distribution. In Tables 8 and 9, we report the estimated parameters. It is clear that the distribution in this example is lifted in the left tail, since for CO2 data we have $\widehat{a} = 0.3806$ and $\widehat{b} = 0$, whereas for H2O data, $\widehat{a} = 0.3303$ and $\widehat{b} = 0$, and then we can see that these estimates have a very intuitive interpretation since the tails of the PDF are lifted visually in these quantities. This fact is attempted to be compensated in the Kumaraswamy distribution by increasing the variance (decreasing $\widehat{\alpha}$ and $\widehat{\beta}$).

Table 6. Log-likelihood and AIC values for CO2 data

Indicator	Distribution TK	Kumaraswamy
Log-likelihood	66.86	14.00
AIC	−125.73	−23.99

In Figure 3, we can see the adjusted PDFs for the two different models, with the TK distribution being the model that better captures the distribution of the data.

Table 8. Estimated parameters for CO2 data

	Estimated parameter			
Distribution	a	b	α	β
TK	0.3806	2.50e-45	7.0541	5.1930
Kumaraswamy	-	-	1.7546	1.2278

Table 9. Estimated parameters for H2O data

	Estimated parameter			
Distribution	a	b	α	β
TK	0.3303	1.12e-43	8.2015	5.5768
Kumaraswamy	-	-	2.1070	1.2778

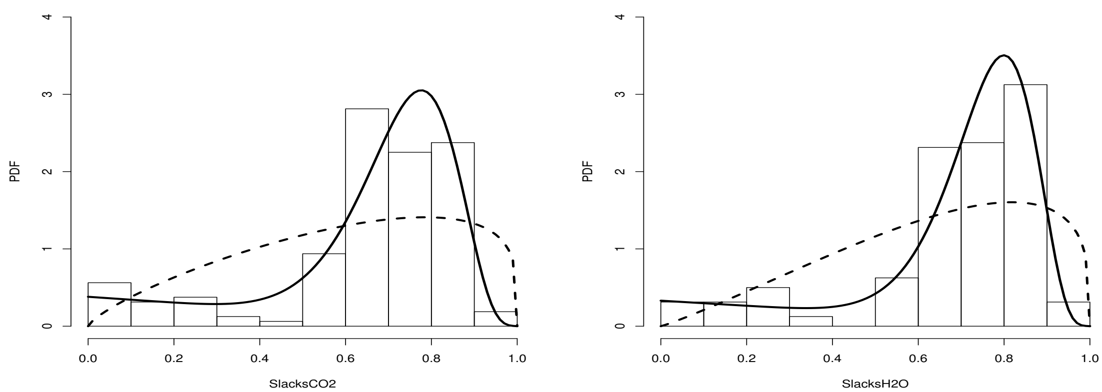


Figure 3. Adjusted PDFs for two different models: in solid line, the TK distribution; and in dotted line the Kumaraswamy distribution for CO2 (left) and H2O (right) data.

5.2 University admission score

We analyze the average score of university admission test in 1295 school establishments in Metropolitan region of Chile, 2016. This test is applied to students who have graduated from school in Chile, which is carried out at a national level and covers different areas of knowledge. In Chile, this test is named "prueba de selección universitaria (PSU)" and allows the student's admission to the different universities of the country, depending on the result obtained in this test. The data set is available in the website https://es.datachile.io.

We are interested in the performance of the students who have applied to the PSU. To measure performance, a total of 1295 average scores per establishment have been collected in the Metropolitan region of Chile and scored in the interval $(0, 1)$ through the transformation proposed by Smithson and Verkuilen (2006) formulated as

$$y = \frac{n-1}{n}\frac{y^* - a_1}{a_2 - a_1} + \frac{1}{2n}, \quad y^* \in [a_1, a_2].$$

Then, $y \in (0, 1)$ and in our case $a_1 = 293.5$, $a_2 = 715.5$ and $n = 1295$. We can see in Figure 4 that the data distribution have a lifted right tail and slightly lifted left tail. Thus, it is justified to fit the TK distribution to model these data. The model under consideration is

Table 10. Log-likelihood and AIC values for PSU data

	Distribution	
Indicator	TK	Kumaraswamy
Log-likelihood	393.68	352.95
AIC	−779.35	−701.90

Table 11. Estimated parameters for PSU data

	Estimated parameter			
Distribution	a	b	α	β
TK	0.0066	0.3072	2.9844	6.6608
Kumaraswamy	-	-	2.3976	3.3506

defined by

$$Y_i \overset{\text{IND}}{\sim} \text{TK}(a, b, \alpha, \beta), \quad i = 1, \ldots, 1295.$$

We can see in Table 10 that the TK distribution achieves a best fit compared to the Kumaraswamy distribution. In Table 11 we report the estimated parameters. It is clear that the distribution in this example is lifted in the tails ($\widehat{a} = 0.0066$ and $\widehat{b} = 0.3072$) and we can see that these estimates have a very intuitive interpretation since the tails of the PDF are lifted visually in these quantities. This fact is once again attempted to be compensated in the Kumaraswamy distribution by increasing the variance (decreasing $\widehat{\alpha}$ and $\widehat{\beta}$).

In Figure 4, we can see the adjusted PDFs for the two different models, with the TK distribution being the model that better captures the distribution of the data.

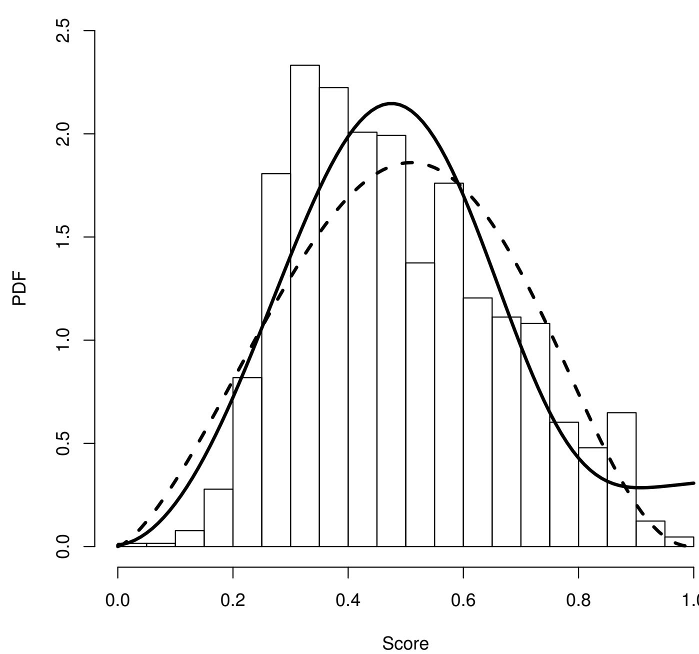


Figure 4. Adjusted PDFs for two different models: in solid line, the TK distribution; and in dotted line the Kumaraswamy distribution for PSU data.

6. Concluding remarks and future research

The Kumaraswamy distribution and other distributions derived from this have been very used in practice. However, until now, it has not been proposed a distribution that allows us to raise the tails of the probability density function in the case of having data accumulated in one or both ends. In this work, we introduced a new four-parameter model called the trapezoidal Kumaraswamy distribution, that is a generalization of the Kumaraswamy distribution which has the rectangular Kumaraswamy distribution as a particular case. The trapezoidal Kumaraswamy distribution comes to solve the problem of adjusting data with some concentration in the extremes. The trapezoidal Kumaraswamy distribution can be represented as a finite mixture model generated by two specific beta distributions and the Kumaraswamy distribution. The trapezoidal Kumaraswamy distribution presented two additional parameters with respect to the Kumaraswamy distribution and they have the advantage of being very intuitive, because they represent the lifting of the probability density function in the tails. The estimation procedure for their parameters is straightforward and in this paper was presented a methodology of estimation achieving good results both with the simulated and real data. In the simulation studies, we observed marked differences in favor of the trapezoidal Kumaraswamy distribution when the samples have some concentration in the tails. In the empirical illustration, the trapezoidal Kumaraswamy distribution turned out to be the model that best adjusted the data and that attended to the essence of the data distribution with some accumulation at the ends. Then, we can conclude that the trapezoidal Kumaraswamy distribution seems to be a new robust alternative for modeling data bounded on the unit interval.

Some open problems that arose from the present investigation are the following:

- An extension of this work that is under development is to propose the reparametrized trapezoidal Kumaraswamy distribution in terms of its mean and connect to it a regression structure, then we will propose a trapezoidal Kumaraswamy regression model.
- The development of a bayesian methodology can be of interest for an alternative implementation.
- The benefits of the distribution will be extended to any bounded distribution.
- A re-parametrization of the trapezoidal Kumaraswamy distribution in terms of its mode is of interest, as this will allow us to connect its mean to a regression structure in a similar manner to that as in generalized linear models.
- A quantile regression model with a trapezoidal Kumaraswamy distributed response will be studied.

Therefore, the proposed results in this study opens opportunities to explore other theoretical and numerical issues.

Acknowledgements

The authors thank Editors and Referees for their suggestions which allowed us to improve the presentation of this work. J.I. Figueroa-Zúñiga acknowledges funding support by grant VRID 217.014.027-1.0, from the Universidad de Concepción, Chile. G. Ibacache-Pulgar acknowledges funding support by grant FONDECYT 11130704, Chile. B. Lagos-Alvarez acknowledges funding support by grant VRID 216.014.026-1.0, from University of Concepción, Chile.

References

Akinsete, A., Famoye, F. and Lee, C., 2014. The Kumaraswamy-geometric distribution. Journal of Statistical Distributions and Applications, 1, 1–17.

Akinsete, A. and Famoye, F., 2008. The beta-Pareto distribution. Statistics, 42, 547–563

Barros, C. and Wanke, P., 2017. Efficiency in Angolan thermal power plants: Evidence from cost structure and pollutant emissions. Energy, 130, 129–143.

Cordeiro, G. and de Castro, M., 2011. A new family of generalized distributions. Journal of Statistical Computation and Simulation, 81, 883–898.

Cordeiro, G. and dos Santos, B., 2012. The beta power distribution. Brazilian Journal of Probability and Statistics, 26, 88–112

Cordeiro, G., Nadarajah, S., and Ortega, E., 2012. The Kumaraswamy Gumbel distribution. Statistical Methods and Applications, 21, 139–168.

Cordeiro, G., Ortega, M. and Nadarajah, S., 2010. The Kumaraswamy Weibull distribution with application to failure data. Journal of the Franklin Institute, 347, 1399–1429.

de Pascoa, M., Ortega, E. and Cordeiro G., 2011. The Kumaraswamy generalized gamma distribution with application in survival analysis. Statistical Methodology, 8, 411–433.

De Santana, T., Ortega, E., Cordeiro, G. and Silva, G., 2012. The Kumaraswamy-log-logistic distribution. Journal of Statistical Theory and Applications, 11, 265–291.

Eugene, N., Lee, C. and Famoye, F., 2002. Beta-normal distribution and its applications. Communications in Statistics: Theory and Methods, 3, 497–512.

Jones, M.C., 2009. Kumaraswamy distribution: A beta-type distribution with some tractability advantages. Statistical Methodology, 6, 70–81.

Kumaraswamy, P., 1980. A generalized probability density function for double-bounded random processes. Journal of Hydrology, 46, 79–88.

Liang, Y., Sun, D., He, C. and Schootman, M., 2014. Modeling bounded outcome scores using the binomial-logit-normal distribution. Chilean Journal of Statistics, 5, 3–14.

McLachlan, G. and Peel, D., 2004. Finite Mixture Models. Wiley, New York.

Mead, M. and Abd-Eltawab, A., 2014. A note on Kumaraswamy Fréchet distribution. Australian Journal of Basic and Applied Sciences, 8, 294–300.

Nadarajah, S. and Kotz, S., 2004. The beta-Gumbel distribution. Mathematical Problems in Engineering, 10, 323–332

Nadarajah, S. and Kotz, S., 2006. The beta exponential distribution. Reliability Engineering and System Safety, 91, 689–697

Nocedal, J. and Wright, S., 1999. Numerical Optimization. Springer, New York.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. Available at http://www.r-project.org

Smithson, M. and Verkuilen, J., 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychological Methods, 11, 54–71.

Appendix

This appendix presents one piece of R codes used for fitting the trapezoidal Kumaraswamy distribution.

```
library(extraDistr)
# For evaluation of Kumaraswamy probability density function (dkumar)
## Trapezoidal Kumaraswamy probability density function ##
dtrapkum<-function(data,w1,w2,alpha,beta){ # w1 and w2 are the weights
                                             # described in the paper
eval<-w1*dbeta(data,1,2)+w2*dbeta(data,2,1)+(1-w1-w2)*dkumar(data,alfa,beta)
return(eval)
}

# Function used in Algorithm to estimate the Kumaraswamy parameters
model<-function(x,data){
alfa0<-(sum(tau3)/x[1])+sum(tau3*log(data))
-sum(tau3*(x[2]-1)*data^x[1]*log(data)/(1-data^x[1]))
beta0<-(sum(tau3)/x[2])+sum(tau3*log(1-data^x[1]))
c(alfa0=alfa0,beta0=beta0)
}

# Initial values
a<-0.1
b<-0.2
alfa<-2
beta<-2
w1<-a/2
w2<-b/2
w3<-1-w1-w2

# EM algorithm #
for(k in 1:1000){
# E step
tau1<-w1*dbeta(data,1,2)/(dtrapkum(data,w1,w2,alpha,beta))
tau2<-w2*dbeta(data,2,1)/(dtrapkum(data,w1,w2,alpha,beta))
tau3<-(1-w1-w2)*dkumar(data,alfa,beta)/(dtrapkum(data,w1,w2,alpha,beta))
# M step
pi1<-sum(tau1)/length(data)
pi2<-sum(tau2)/length(data)
solution<-multiroot(f=model,start = c(alfa,beta),maxiter=5000,data=data)
solution
alfa<-solution$root[1]
beta<-solution$root[2]
}
```

Information for authors

The editorial board of the Chilean Journal of Statistics (ChJS) is seeking papers, which will be refereed. We encourage the authors to submit a PDF electronic version of the manuscript in a free format to the Editors-in-Chief of the ChJS (E-mail: chilean.journal.of.statistics@gmail.com). Submitted manuscripts must be written in English and contain the name and affiliation of each author followed by a leading abstract and keywords. The authors must include a"cover letter" presenting their manuscript and mentioning: "We confirm that this manuscript has been read and approved by all named authors. In addition, we declare that the manuscript is original and it is not being published or submitted for publication elsewhere".

Preparation of accepted manuscripts

Manuscripts accepted in the ChJS must be prepared in Latex using the ChJS format. The Latex template and ChJS class files for preparation of accepted manuscripts are available at http://chjs.mat.utfsm.cl/files/ChJS.zip. Such as its submitted version, manuscripts accepted in the ChJS must be written in English and contain the name and affiliation of each author, followed by a leading abstract and keywords, but now mathematics subject classification (primary and secondary) are required. AMS classification is available at http://www.ams.org/mathscinet/msc/. Sections must be numbered 1, 2, etc., where Section 1 is the introduction part. References must be collected at the end of the manuscript in alphabetical order as in the following examples:

Arellano-Valle, R., 1994. Elliptical Distributions: Properties, Inference and Applications in Regression Models. Unpublished Ph.D. Thesis. Department of Statistics, University of São Paulo, Brazil.

Cook, R.D., 1997. Local influence. In Kotz, S., Read, C.B., and Banks, D.L. (Eds.), Encyclopedia of Statistical Sciences, Vol. 1., Wiley, New York, pp. 380-385.

Rukhin, A.L., 2009. Identities for negative moments of quadratic forms in normal variables. Statistics and Probability Letters, 79, 1004-1007.

Stein, M.L., 1999. Statistical Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Tsay, R.S., Peña, D., and Pankratz, A.E., 2000. Outliers in multivariate time series. Biometrika, 87, 789-804.

References in the text must be given by the author's name and year of publication, e.g., Gelfand and Smith (1990). In the case of more than two authors, the citation must be written as Tsay et al. (2000).

Copyright