CHILEAN JOURNAL OF **STATISTICS** Volume 13 Number 1 - April 2022

Volume 13 Number 1 - April 2022

CHILEAN JOURNAL OF **STATISTICS**

CHILEAN JOURNAL OF STATISTICS

Edited by Víctor Leiva and Carolina Marchant

A free open-access journal indexed by

Volume 13 Number 1
April 2022
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 13, Number 1
April 2022

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

http://www.soche.cl/chjs

Chilean Journal of Statistics

Volume 13, Number 1
April 2022

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

http://www.soche.cl/chjs

Contents

Chilean Journal of Statistics
Volume 13, Number 1 (April 2022), Pages 1–2
DOI: 10.32372/ChJS.13-01-00

Thirteenth Volume – First Issue
Editorial Paper

Chilean Journal of Statistics: Thirty eight years generating quality knowledge

Carolina Marchant[1] and Víctor Leiva[2]

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

Welcome to the first issue of the thirteenth volume of the Chilean Journal of Statistics (ChJS), published on 28 April 2022. The ChJS celebrates 38 years of life during very difficult times due, on the one hand, to the COVID-19 pandemic, which is still present, and, on the other hand, to diverse conflicts around the world. Statistical scientific research provides relevant information in decision making in different phenomena and national and international events that affect us. For example, in the COVID-19 pandemic has permitted the governments to establish regulations stopping its spread.

The scientific and editorial production of this volume would not have been achieved without the valuable contributions of many people. We are pleased to inform the international community that outstanding researchers have honored us by publishing their interesting work in our journal. We are publishing articles written by colleagues from Algeria, Australia, Brazil, Chile, Colombia, France, India, and Spain. We also thank all the anonymous reviewers who have contributed to maintaining ChJS' high-quality standards. Furthermore, we feel obliged and pleased to thank our prestigious editorial board listed in http://soche.cl/chjs/board.html. 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 on us, the Editors-In-Chief of the ChJS, their confidence in our work.

The first issue of the thirteenth volume of the ChJS comprises six articles as follows:

(i) In our first paper, María Dueñas and Ramón Giraldo, from Colombia, explored ordinary kriging for functional data based on Andrews curves as an alternative to the classical multivariate approach.
(ii) The second paper is authored by Ernesto San Martín and Eduardo Alarcón-Bustamante from Chile, which carry out a dissection of three Chilean surveys.
(iii) In the third paper, an analysis about Bayesian detection of change in the parameters of an autoregressive process of known order was proposed by Abdeldjalil Slama from Algeria.
(iv) The fourth paper is authored by Christophe Chesneau, Muhammed Rasheed Irshad, Damodaran Santhamani Shibu, Soman Latha Nitin and Radhakumari Maya from France and India, who proposed a new version of the two-parameter log-normal distribution with applications to astronomy and cancer data.
(v) In the fifth paper, Lucas D. Ribeiro-Reis, Gauss M. Cordeiro, and José J. de Santana e Silva, from Brazil, derived a new bimodal distribution named the Mc-Donald Chen model.

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

http://www.soche.cl/chjs

(vi) The sixth and last paper is authored by Emilio Gómez-Déniz, Enrique Calderín-Ojeda, and José María Sarabia, from Australia and Spain, which explored the arctan family of distributions and provided three numerical applications related to insurance.

As the Chilean Statistics Society, we are proud because we continue to provide, by means of the ChJS, an open-access forum, publishing high-quality works free of any article processing charges (APC). In addition, we are indexed to the Elsevier Scopus and Clarivate ISI WoS systems. We are very motivated because, at the beginning of 2022, we received 30 submissions from different countries.

Finally, we would like the international statistical and data-science communities, our editorial board, and our collaborators, to champion the ChJS as a long-lived, international, free of charges, and open-access forum, with fair and high-quality reviews. We encourage the international scientific community to submit their works to the ChJS.

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

Chilean Journal of Statistics
Volume 13, Number 1 (April 2022), Pages 3–16
DOI: 10.32372/ChJS.13-01-01

Spatial Statistics
Research Paper

Multivariate spatial prediction based on Andrews curves and functional geostatistics

María Dueñas[1] and Ramón Giraldo[1,*]

[1]Department of Statistics, Universidad Nacional de Colombia, Bogotá, Colombia
(Received: 03 March 2022 · Accepted in final form: 06 April 2022)

Abstract

There are two usual ways for modeling the realizations of multivariate random fields: Applying kriging individually on each variable or using cokriging, which considers the spatial cross-dependence between the variables. It has been shown that the second way, in general, allows a prediction variance reduction. The use of cokriging may be limited in practice when the number of variables increases because estimating the linear model of coregionalization (the cross-dependence between the variables) becomes complex. This work explores ordinary kriging for functional data based on Andrews curves as an alternative to the classical multivariate approach. Employing a simulation study, we compare the predictor proposed with kriging and cokriging. The methodology is applied to an environmental dataset.

Keywords: Andrews curves · Cokriging · Functional data · Geostatistics · Kriging

Mathematics Subject Classification: Primary 60G10 · Secondary 60G25.

1. Introduction

In many fields of applied science, it is required to simultaneously model data of several variables. Several statistical tools have been adapted to deal challenging multivariate problems. Among other areas, regression analysis (Bilodeau and Brenner, 1999), ANOVA (Smith et al., 1962), longitudinal data (Verbeke et al., 2014), and generalized linear models (Fahrmeir et al., 1994) have been tailored to this challenge. When the number of characteristics increases, the modeling becomes more complex. Also, the analysis of multivariate data is a big problem if there are inherent temporal and spatial dependence structures. One example is the multivariate spatial statistics (Gelfand et al., 2010), where it is necessary to consider auto and cross-correlations. The problem is solved using cokriging (assuming stationarity)(Giraldo et al., 2021). An advantage of this method is that it does not require that the variables are measured at the same sites. Its use has demonstrated to reduce uncertainty concerning ordinary kriging (spatial prediction of each variable separately). In its simplest form, cokriging assumes that the joint spatial correlation of the multivariate random field is generated from combinations of basic spatial covariance models and coregionalization matrices. If there are p variables, it is then necessary to estimate $p(p+1)/2$ variograms (including simple and cross-variograms). This makes this technique difficult to implement when p increases.

*Corresponding author. Email: rgiraldoh@unal.edu.co

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

http://www.soche.cl/chjs

Andrews curves (Andrews, 1972) are generally utilized in multivariate analysis to detect outliers (Embrechts et al., 1986), carry out clustering (Moustafa, 2011) and discriminant analysis. In this work, we propose its usage in multivariate geostatistics (Genton and Kleiber, 2015) as a tool for solving the high dimensionality problem. When the number of variables increases, it is not easy to estimate the coregionalization model and, therefore, to make predictions using cokriging. Employing Andrews curves combined with functional geostatistics (Giraldo et al., 2011) can simplify the problem because it only requires to fit a single variogram model. Once an Andrews curve is predicted on an unsampled site, implicitly all the variables of the multivariate random field of interest are predicted too.

Classical tools for spatial data analysis can be extended to functional data. Particularly in geostatistics, several alternatives for this purpose have been proposed. Ordinary, residual, and universal kriging for functional data (Mateu and Giraldo, 2022) are some approaches to solve the problem of spatial prediction when we have a realization of a functional random field (when a curve or, in general, a function is recorded at several sites of a region with spatial continuity). Here we propose an alternative for carrying out spatial prediction in multivariate geostatistics using ordinary kriging for functional data (Giraldo et al., 2011) based on Andrews curves. This alternative does not require to estimate a linear coregionalization model (Wackernagel, 2003), and consequently reducing the complexity of the problem.

The work is organized as follows. Section 2 gives a review on Andrews curves, multivariate geostatistics, and functional geostatistics. Section 3 presents the methodology proposed. An illustration with simulated data and an application to real data are shown in Section 4. The article ends with some conclusions, limitations and ideas for further research in Section 5.

2. Background

In this section, we present a short overview about Andrews curves (Andrews, 1972; Moustafa, 2011), multivariate geostatistics (Wackernagel, 2003), and ordinary kriging for functional data (Giraldo et al., 2011).

2.1 Andrews curves

A statistical multivariate analysis is considered when we have data of a p-dimensional random vector ($p > 1$). Given a realization of size n of a random vector $X = (X_1, \ldots, X_p)^\top$, we obtain the data matrix

$$\boldsymbol{x} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}. \tag{2.1}$$

The underlying idea of Andrews curves is that each multivariate data point (observation) can be represented by a curve using a Fourier interpolation function where the coefficients are the observation's components (Moustafa, 2011). Andrews curves are used as a descriptive tool for summarizing a multivariate data set as represented in Matrix given in expression (2.1) or employed to identify atypical values or clustering the individuals (Moustafa, 2011). These are built as linear combinations of the observations (Andrews, 1972). Specifically, for all i, for $i = 1, \ldots, n$, the i-th Andrews curve is given by

$$x_i(t) = \frac{1}{\sqrt{2\pi}} x_{i1} + \sin(t)x_{i2} + \cos(t)x_{i3} + \sin(2t)x_{i4} + \cdots, \tag{2.2}$$

with $t \in [-\pi, \pi]$. The order of the variables plays an important role in obtaining the curve: when there are many variables, the last ones have a low contribution to the shape of the

curve. For this reason, they are usually ordered previously according to the amount of information that each of them provides. Generally, for this, a principal component analysis is initially carried out.

2.2 Multivariate geostatistics

This subsection is based on Giraldo et al. (2017). Let $\{\boldsymbol{X}(s) = (X_1(s), \ldots, X_m(s)) : s \in D\}$ be a multivariate spatial process defined over a domain $D \subset \mathbb{R}^2$. Assume $\boldsymbol{X}(s) = \boldsymbol{\mu}(s) + \boldsymbol{\epsilon}(s)$ is a stationary process with $\boldsymbol{\mu}(s)$ the mean vector and $\boldsymbol{\epsilon}(s)$ a stationary noise process with $\mathrm{E}(\boldsymbol{\epsilon}(s)) = \mathbf{0}$. We use the following notation: (i) $2\gamma_{lq}(s_i, s_j) = \mathrm{V}(X_l(s_i) - X_q(s_j))$, for $l, q = 1, \ldots, m$, $i, j = 1, \ldots, n$; (ii) $\boldsymbol{\gamma}_{lk}^{\top} = (\gamma_{lk}(s_1, s_0), \ldots, \gamma_{lk}(s_n, s_0))$; and (iii)

$$\boldsymbol{\Gamma}_{lq} = \begin{pmatrix} \gamma_{lq}(s_1, s_1) & \cdots & \gamma_{lq}(s_1, s_n) \\ \vdots & \ddots & \vdots \\ \gamma_{lq}(s_n, s_1) & \cdots & \gamma_{lq}(s_n, s_n) \end{pmatrix}.$$

The cokriging predictor of the random variable $X_k(s_0)$ based on the realization $\boldsymbol{X}(s_i)$, for $i = 1, \ldots, n$, is defined as

$$\widehat{X}_k(s_0) = \sum_{j=1}^{m} \lambda_{1j}^k X_j(s_1) + \cdots + \sum_{j=1}^{m} \lambda_{nj}^k X_j(s_n) = \sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_{ij}^k X_j(s_i). \tag{2.3}$$

The predictor given in Equation (2.3) is unbiased if $\sum_{i=1}^{n} \lambda_{ik}^k = 1$ and $\sum_{i=1}^{n} \lambda_{ij}^k = 0$ for $j \neq k$, $j = 1, \ldots, m$. Using the Lagrange method to minimize the mean squared prediction error, $\mathrm{E}(\widehat{X}_k(s_0) - X_k(s_0))^2$, subject to the unbiasedness constraints gives the cokriging system of equations, which in matrix notation can be expressed by $\boldsymbol{C}\boldsymbol{\lambda}^k = \boldsymbol{c}^k$, with

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{\Gamma}_{11} & \cdots & \boldsymbol{\Gamma}_{1k} & \cdots & \boldsymbol{\Gamma}_{1m} & \mathbf{1} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & & & \vdots & \vdots & \ddots & & & \vdots \\ \boldsymbol{\Gamma}_{k1} & & \boldsymbol{\Gamma}_{kk} & & \boldsymbol{\Gamma}_{km} & \mathbf{0} & & \mathbf{1} & & \mathbf{0} \\ \vdots & & & \ddots & \vdots & \vdots & & & \ddots & \vdots \\ \boldsymbol{\Gamma}_{m1} & \cdots & \boldsymbol{\Gamma}_{m2} & \cdots & \boldsymbol{\Gamma}_{mm} & \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{1} \\ \mathbf{1}^{\top} & \cdots & \mathbf{0}^{\top} & \cdots & \mathbf{0}^{\top} & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & & & \vdots & \vdots & \ddots & & & \vdots \\ \mathbf{0}^{\top} & & \mathbf{1}^{\top} & & \mathbf{0}^{\top} & 0 & & 0 & & 0 \\ \vdots & & \vdots & \ddots & \vdots & \vdots & & & \ddots & \vdots \\ \mathbf{0}^{\top} & \cdots & \mathbf{0}^{\top} & \cdots & \mathbf{1}^{\top} & 0 & \cdots & 0 & \cdots & 0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Gamma} & \boldsymbol{Z} \\ \boldsymbol{Z}^{\top} & \mathbf{0}^{*} \end{pmatrix},$$

$$\boldsymbol{\lambda}^k = \begin{pmatrix} \boldsymbol{\lambda}_1^k \\ \vdots \\ \boldsymbol{\lambda}_k^k \\ \vdots \\ \boldsymbol{\lambda}_m^k \\ \delta_1 \\ \vdots \\ \delta_k \\ \vdots \\ \delta_m \end{pmatrix}, \boldsymbol{c}^k = \begin{pmatrix} \boldsymbol{\gamma}_1^k \\ \vdots \\ \boldsymbol{\gamma}_k^k \\ \vdots \\ \boldsymbol{\gamma}_m^k \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix},$$

where $(\boldsymbol{\Gamma}_{ij})_{(n\times n)}$, $\mathbf{1} = (1,\ldots,1)^\top_{(n\times 1)}$, $\mathbf{0} = (0,\ldots,0)^\top_{(n\times 1)}$, $(\boldsymbol{\Gamma})_{(m\times n)\times(n\times m)}$, $(\boldsymbol{Z})_{(n\times m)\times m}$, $(\mathbf{0}^*)_{(m\times m)}$, $\boldsymbol{\lambda}_j^k = (\lambda_{1j}^k,\ldots,\lambda_{nj}^k)^\top$, and $\boldsymbol{\gamma}_j^k = (\gamma_{1j}^k,\ldots,\gamma_{nj}^k)^\top$, for all $i,j = 1,\ldots,m$. Cokriging could be used for predicting simultaneously all m variables instead of predicting a variable, one at a time.

2.3 Functional geostatistics

Let $\{X_t(s), t \in \mathbb{R}, s \in D \subseteq \mathbb{R}^2\}$ be a second-order stationary and isotropic functional random field (Giraldo et al., 2011) whose realizations are functions defined in the real interval T with $X_t(s) \in L_2(T)$ the space of square integrable functions. From the stationarity conditions and taking $h = \|s_i - s_j\|$ we have

- $\mathrm{E}(X_t(s)) = \mu_t$.
- $\mathrm{V}(X_t(s)) = \sigma_t^2$.
- $\mathrm{C}(X_t(s_i), X_t(s_j)) = C(||s_i - s_j||; t) = C(h;t)$.
- $\frac{1}{2}\mathrm{V}(X_t(s_i) - X_t(s_j)) = \gamma(||s_i - s_j||; t) = \gamma(h;t)$.

The ordinary kriging predictor of the function on a site s_0 is defined as (Giraldo et al., 2011)

$$\widehat{X}_t(s_0) = \sum_{i=1}^{n} \lambda_i X_t(s_i),\ \lambda_1,\ldots,\lambda_n \in \mathbb{R}. \tag{2.4}$$

Optimal λ in Equation (2.4) that guarantee $\mathrm{E}(\widehat{X}_t(s_0)) = X_t(s_0)$ are obtained by solving the system

$$\begin{pmatrix} \int_T \gamma\left(||s_1 - s_1||, t\right) \mathrm{d}t & \cdots & \int_T \gamma\left(||s_1 - s_n||, t\right) \mathrm{d}t & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \int_T \gamma\left(||s_1 - s_n||, t\right) \mathrm{d}t & \cdots & \int_T \gamma\left(||s_n - s_n||, t\right) \mathrm{d}t & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \nu \end{pmatrix} = \begin{pmatrix} \int_T \gamma\left(||s_0 - s_1||, t\right) \mathrm{d}t \\ \vdots \\ \int_T \gamma\left(||s_0 - s_n||, t\right) \mathrm{d}t \\ 1 \end{pmatrix}. \tag{2.5}$$

The function $\gamma(h) = \int_T \gamma(h,t)\,\mathrm{d}t = (1/2)\mathrm{E}(\int_T (X_t(s_i) - X_t(s_j))^2 \mathrm{d}t)$ is called the trace-variogram. A review on its estimation based on the observed data is provided in Giraldo et al. (2011). Note that ν in Equation (2.5) is the Lagrange multiplier used to consider the unbiasedness constraint.

3. Multivariate geostatistics based on Andrews curves

We show how ordinary kriging based on Andrews curves is an alternative to perform multivariate spatial prediction. We assume isotropy and that all variables are recorded in the same sites.

3.1 From multivariate to functional Kriging

Let $\{\boldsymbol{X}(s), s \in D \subset \mathbb{R}^d\}$ be a p-dimensional random field and $[\boldsymbol{X}(s_1), \boldsymbol{X}(s_2), ..., \boldsymbol{X}(s_n)]^\top$ a sample of the process with

$$\boldsymbol{X}(s_i) = \begin{bmatrix} X_1(s_i) \\ X_2(s_i) \\ \vdots \\ X_p(s_i) \end{bmatrix}, \quad i = 1,\ldots,n.$$

Suppose we want to predict the random field at a site s_0. Employing Andrews curves given in Equation (2.2), the sample of the multivariate random field can be used to define a sample of a functional random field of Andrews curves $\{X_t(s), s \in D \subset \mathbb{R}^d, t \in [-\pi, \pi] \subset \mathbb{R}\}$ with the transformation

$$X_t(s_i) = \sum_{k=1}^{p} X_k(s_i)\phi_k(t), \tag{3.6}$$

with $\phi_k(t)$ the k-th coefficient of a Fourier series as defined in Equation (2.2). Likewise, from the multivariate observed sample of the random process $[\boldsymbol{x}(s_1), \boldsymbol{x}(s_2), ..., \boldsymbol{x}(s_n)]$, we have that

$$x_t(s_i) = \sum_{k=1}^{p} x_k(s_i)\phi_k(t). \tag{3.7}$$

Assuming that the curves defined in Equation (3.6) are a sample of a functional random field, we can use functional geostatistical methods (Giraldo et al., 2011, 2017) for carrying spatial prediction of all variables. Particularly, using ordinary kriging for functional data given in Equation (2.4) and taking as input the observed curves in Equation (3.7) we can predict the Andrews curves on unsampled sites. Note that the coefficients in Equation (2.2) are known and correspond to the data recorded from the p variables in the n sites $s_1, \ldots, s_n$.

3.2 Functional random field of Andrews curves

Assume the multivariate random field of interest is second order stationary. Consequently, we have the following properties for the random field of Andrews curves $\{X_t(s), s \in D \subset \mathbb{R}^d, t \in [-\pi, \pi] \subset \mathbb{R}\}$:

(i)

$$\begin{aligned}
\mu(t) = \mathrm{E}\Big[X_t(s)\Big] &= \mathrm{E}\left[\sum_{k=1}^{p} X_k(s)\phi_k(t)\right] \\
&= \sum_{k=1}^{p} \mathrm{E}\left[X_k(s)\phi_k(t)\right] \\
&= \sum_{k=1}^{p} \phi_k(t)\mathrm{E}\left[X_k(s_i)\right] \\
&= \sum_{k=1}^{p} \phi_k(t)\mu_k,
\end{aligned}$$

with $\mu_k = \mathrm{E}\left[X_k(s_i)\right]$ being the mean of the k-th random field.

(ii)

$$\begin{aligned}\mathrm{V}\left[X_t(s)\right] &= \sigma_t^2 \\ &= \mathrm{V}\left[\sum_{k=1}^{p} X_k(s)\phi_k(t)\right] \\ &= \sum_{k=1}^{p}\sum_{l=1}^{p}\phi_k(t)\phi_l(t)\mathrm{Cov}\left[X_k(s), X_l(s)\right] \\ &= \sum_{k=1}^{p}\sum_{l=1}^{p}\phi_k(t)\phi_l(t)C_{kl}(0),\end{aligned}$$

with $C_{kl}(0)$ being the covariance between the variables k and l.

(iii)

$$\begin{aligned}\mathrm{C}\Big[X_t(s_i), X_t(s_j)\Big] &= \mathrm{C}(h,t) \\ &= \mathrm{C}\left[\sum_{k=1}^{p} X_k(s_i)\phi_k(t), \sum_{k=1}^{k} X_k(s_j)\phi_k(t)\right] \\ &= \sum_{k=1}^{p}\sum_{l=1}^{p}\phi_k(t)\phi_l(t)\mathrm{C}\left[X_k(s_i), X_l(s_j)\right] \\ &= \sum_{k=1}^{p}\sum_{l=1}^{p}\phi_k(t)\phi_l(t)\mathrm{C}_{kl}(||s_i - s_j||) \\ &= \sum_{k=1}^{p}\sum_{l=1}^{p}\phi_k(t)\phi_l(t)\mathrm{C}_{kl}(h).\end{aligned}$$

Note that the functional covariance depends only on the distance between sites s_i and s_j.

3.3 Spatial prediction of Andrews curves

Let $X_t(s_i)$, for $i = 1, \ldots, n$, be the sample of a functional random field of Andrews curves. Then the ordinary kriging predictor of an Andrews curve on a site s_0 is given by

$$\begin{aligned}\widehat{X}_t(s_0) &= \sum_{i=1}^{n}\lambda_i X_t(s_i) \\ &= \sum_{i=1}^{n}\lambda_i \sum_{k=1}^{p} X_k(s_i)\phi_k(t) \\ &= \sum_{k=1}^{p}\sum_{i=1}^{n}\lambda_i X_k(s_i)\phi_k(t). \qquad (3.8)\end{aligned}$$

In Equation (3.8), each term $\sum_{i=1}^{n}\lambda_i X_k(s_i)$ is an scalar corresponding to the predictor $\widehat{X}_k(s_0)$. This is an unbiased and minimum variance predictor if $\lambda_1, \ldots, \lambda_n$ are such that

$$\int_T \mathrm{V}\left(\widehat{X}_t(s_0) - X_t(s_0)\right)\mathrm{d}t,$$

is minimum subject to $\sum_{i=1}^{n}\lambda_i = 1$.

3.4 Relationship between the trace-variogram function and univariate variograms

Note that

$$\int_t \Big(X_t(s_i) - X_t(s_j)\Big)^2 \mathrm{d}t = \int_t \left(\sum_{k=1}^{p} X_k(s_i)\phi_k(t) - \sum_{k=1}^{p} X_k(s_j)\phi_k(t)\right)^2 \mathrm{d}t$$
$$= \int_t \left[\sum_{k=1}^{p} \left(X_k(s_i) - X_k(s_j)\right)\phi_k(t)\right]^2 \mathrm{d}t,$$

with $T = [-\pi, \pi]$. In matrix notation, we get

$$\begin{aligned}\int_t \left(X_t(s_i) - X_t(s_j)\right)^2 \mathrm{d}t &= \int_t \Big[\left(\boldsymbol{X}(s_i) - \boldsymbol{X}(s_j)\right)^\top \Phi(t)\Big]^2 \mathrm{d}t \\ &= \int_t \left(\boldsymbol{X}(s_i) - \boldsymbol{X}(s_j)\right)^\top \left(\Phi(t)\Phi(t)^\top\right) \left(\boldsymbol{X}(s_i) - \boldsymbol{X}(s_j)\right) \mathrm{d}t \\ &= \left(\boldsymbol{X}(s_i) - \boldsymbol{X}(s_j)\right)^\top \int_t \left(\Phi(t)\Phi(t)^\top\right) dt \left(\boldsymbol{X}(s_i) - \boldsymbol{X}(s_j)\right) \\ &= \left(\boldsymbol{X}(s_i) - \boldsymbol{X}(s_j)\right)^\top W \left(\boldsymbol{X}(s_i) - \boldsymbol{X}(s_j)\right),\end{aligned}$$

with W being the matrix of inner products of $\Phi(t)$. Taking into account that $\Phi(t)$ is an orthonormal basis, we have that $W = I_n$. Thus, we reach

$$\gamma(h) = \frac{1}{2}\mathrm{E}\left[\sum_{k=1}^{p} \left(X_k(s_i) - X_k(s_j)\right)^2\right].$$

Under second order stationarity, we have that

$$\gamma_k(h) = \frac{1}{2} E\left[\left(X_k(s_i) - X_k(s_j)\right)^2\right]. \tag{3.9}$$

From Equation (3.9), the trace-variogram can be expressed as

$$\begin{aligned}\gamma(h) &= \frac{1}{2}\mathrm{E}\left[\sum_{k=1}^{p} \left(X_k(s_i) - X_k(s_j)\right)^2\right] \\ &= \frac{1}{2}\sum_{k=1}^{p} \mathrm{E}\left[\left(X_k(s_i) - X_k(s_j)\right)^2\right] \\ &= \sum_{k=1}^{p} \gamma_k(h). \end{aligned} \tag{3.10}$$

Therefore, the theoretical trace-variogram corresponds to the sum of the univariate semivariograms associated to the variables used to define the Andrews curves. This sum can be modeled with a single model once the empirical trace-variogram has been calculated. To carry out the spatial prediction we need to estimate the trace-variogram function $\int \gamma_t\left(||s_i - s_j||\right) \mathrm{d}t$,

for all $i = 1, \ldots, n$. The corresponding estimator is given by

$$\widehat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \left(\boldsymbol{X}(s_i) - \boldsymbol{X}(s_j)\right)^{\top} \boldsymbol{W} \left(\boldsymbol{X}(s_i) - \boldsymbol{X}(s_j)\right),$$

where $N(h)$ is the number of pairs (s_i, s_j) such that $h = ||s_i - s_j||$ and $|N(h)|$ is the number of sites separated by a distance h. Hence, the moment estimator of the trace-variogram function is stated as

$$\widehat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \sum_{k=1}^{p} \left(X_k(s_i) - X_k(s_j)\right)^2. \tag{3.11}$$

From Equation (3.10), the total prediction variance can be defined as

$$\sigma^2(s_0) = \sum_{i=1}^{n} \lambda_i \gamma(\|s_i - s_0\|) + \mu = \sum_{i=1}^{n} \lambda_i \sum_{k=1}^{p} \gamma_k(h) + \mu,$$

and its estimation is formulated by

$$\begin{aligned} \widehat{\sigma}^2(s_0) &= \sum_{i=1}^{n} \lambda_i \widehat{\gamma}(\|s_i - s_0\|) + \mu \\ &= \sum_{i=1}^{n} \lambda_i \left(\frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \sum_{k=1}^{p} \left(x_k(s_i) - x_k(s_j)\right)^2 \right) + \mu. \end{aligned}$$

4. Numerical Applications

This section initially compares kriging, cokriging and functional kriging using a small simulated dataset. Posteriorly, an application to a real dataset is presented. The computational routines were developed using the R software (R, 2022) version 4.1.3 for Windows platform.

4.1 Simulated Data

Suppose we have data of a stationary bivariate Gaussian random field $\{\boldsymbol{X}(s) = (X_1(s), X_2(s)) : s \in [0,1] \times [0,1]\}$ with means $\mu_1(s) = 2$ and $\mu_2(s) = 90$ and spatial dependence defined by the following variogram models:

$$\begin{aligned} \gamma_{X_1}(h) &= 0.30\gamma_1(h) + 0.26\gamma_2(h) \\ \gamma_{X_2}(h) &= 11\gamma_1(h) + 71\gamma_2(h) \\ \gamma_{X_1X_2}(h) &= 1.2\gamma_1(h) + 3.8\gamma_2(h), \end{aligned}$$

with $\gamma_1(h) = (1 - \exp((-h/0.7))$ and $\gamma_2(h) = (1.5(h/0.95) - 0.5(h/0.95)^2)$. In both models, the parameter ϕ is relatively high ($\phi = 0.7$ for the exponential model and $\phi = 0.95$ in the case of the spherical model), which is an indicator of high spatial simple and cross correlation. Note that ϕ is the parameter that defines the spatial correlation. The values assigned to this parameter correspond respectively to 70% and 95% of the maximum distance between sites of the simulation region.

Table 1. Four simulated data sets of a bivariate Gaussian random field defined on the square $[0,1] \times [0,1]$.

s	Coordinates	$X_1(s)$	$X_2(s)$
s_1	(1.00, 0.22)	1.51	80.83
s_2	(0.00, 0.33)	2.88	80.49
s_3	(0.67, 0.00)	2.94	102.29
s_4	(0.22, 0.78)	1.84	79.22

The corresponding covariance matrix is given by

$$\Sigma = \begin{bmatrix} 0.56 & 0.07 & 0.27 & 0.08 & 5.00 & 0.29 & 2.22 & 0.31 \\ 0.07 & 0.56 & 0.12 & 0.22 & 0.29 & 5.00 & 0.66 & 1.68 \\ 0.27 & 0.12 & 0.56 & 0.08 & 2.22 & 0.66 & 5.00 & 0.35 \\ 0.08 & 0.22 & 0.08 & 0.56 & 0.31 & 1.68 & 0.35 & 5.00 \\ 5.00 & 0.29 & 2.22 & 0.31 & 82.00 & 2.61 & 34.96 & 2.81 \\ 0.29 & 5.00 & 0.66 & 1.68 & 2.61 & 82.00 & 8.38 & 25.78 \\ 2.22 & 0.66 & 5.00 & 0.35 & 34.96 & 8.38 & 82.00 & 3.40 \\ 0.31 & 1.68 & 0.35 & 5.00 & 2.81 & 25.78 & 3.40 & 82.00 \end{bmatrix}.$$

Assume that we want to predict the variables $X_1(s_0)$ and $X_2(s_0)$, $s_0 = (0.22, 0.00)$, using four observations of the process; see Table 1. Based on the covariance matrix and employing univariate ordinary Kriging, ordinary Cokriging, and Functional Kriging predictions are obtained for $X_1(s_0)$ and $X_2(s_0)$.

Table 2. Predictions using the three methods. $\widehat{\sigma}_T^2$ correspond to the total prediction variance (sum of the prediction variances).

Method	$\widehat{X}_1(s_0)$	$\widehat{X}_2(s_0)$	$\widehat{\sigma}_T^2$
Kriging	2.199	93.708	68.269
Cokriging	2.707	93.602	68.163
Functional kriging	2.819	93.789	68.278

Table 2 shows that we obtain reasonable predictions with the three methods (values around the means $\mu_1(s)$ and $\mu_2(s)$ of the processes) with variances of the predictions that only differ slightly. A more intensive simulation study was conducted posteriorly. Considering the same spatial dependence structure defined above by $\gamma_{X_1}(h)$, $\gamma_{X_2}(h)$, and $\gamma_{X_1X_2}(h)$, a realization of size 100 of the bivariate process was generated. A cross-validation analysis was carried out with these data, that is, each simulated datum was partially deleted and predicted based on the remaining 99 observations through the three methods (Kriging, Cokriging, and Functional Kriging). We do not present the results in detail. The means of the prediction errors were in all cases (three methods) close to zero and the means prediction variances were also very similar (around 11.71). A detailed review of the results can be seen in Dueñas (2017). Here, we consider only two processes to show that even when the number of variables is small the methodology based on functional kriging can be applied. If the number of processes increases, there is more significant differences between the methods, but cokriging also is more complex. In these cases, the approach based on functional kriging may be more appropriate.

4.2 Real Data

The lagoon-estuarine system Ciénaga Grande de Santa Marta (CGSM) located at the north coast of Colombia (Figure 1) is of interest for its ecological and hydrological characteristics

and its richness in fish, mollusks, and crustaceans (Rodríguez-Rodríguez et al., 2021). Monitoring its physicochemical and biological conditions is essential due to its environmental and economic impact on the region.

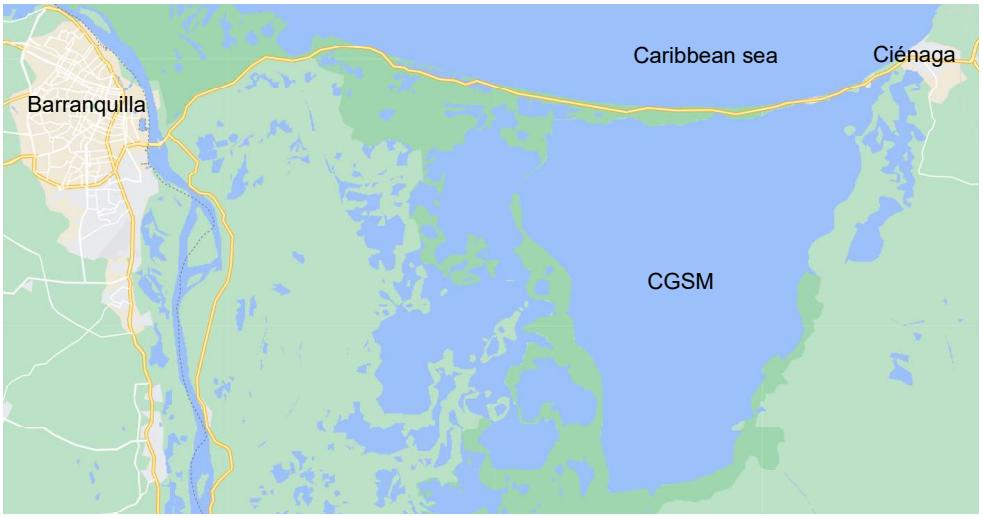


Figure 1. The lagoon-estuarine ecosystem Ciénaga Grande de Santa Marta (CGSM) is located at the north coast of Colombia between the cities of Barranquilla and Ciénaga. A narrow, continuous sandbar borders the entire CGSM complex to the north. Source: Google Maps 2021.

This work shows how to use functional kriging based on Andrews curves to jointly predict the spatial distribution of some of these variables. Specifically, we analyze data of six variables (salinity, dissolved oxygen (mg O_2/L), temperature (°C), chlorophyll-a (μg/l), total suspended solids (mg/l), and depth (cm)) collected in 95 sampling sites of the system. The spatial distribution of these variables according to the quartiles of the recorded values is shown in Figure 2. These plots suggest that is reasonable assuming stationarity, because there is not a defined spatial trend in any case. There are three alternatives for doing prediction in this case. We can apply ordinary kriging (without considering the dependency between the variables), ordinary cokriging which require the estimation of a LMC (a complex procedure in this scenario because we must to take into account data of six random processes simultaneously), or ordinary kriging based on Andrews curves. Below, we show the results considering this last option. We also do a comparison with the results obtained using ordinary kriging.

In Table 3, we report the variation coefficients calculated with the 95 observations from each one of the six variables considered in the study. These values are ordered from highest to lowest. Following Andrews (1972), we employ this order to define the coefficients x_{ij} from Equation (2.2) of the Andrews curves for the dataset of interest (top panel of Figure 3). We note that the curves have a similar behavior. Only two curves have a different pattern (see curves with the lowest values for $t \in (0, 1.7)$). These correspond to places in the north of the Ciénaga that have different conditions of salinity and depth.

Using Equation (3.11), we calculate the empirical trace-variogram function (white circles in bottom panel of Figure 3). An exponential semivariogram model with $\widehat{\phi} = 6460$m y $\widehat{\sigma}^2 = 19224.08$ was fitted to this scatterplot (red curve in bottom panel of Figure 3). The value of $\widehat{\phi}$ indicates that the Andrews curves are correlated up to a distance of about 6.4 km. Using this model, we can estimate the weights λ_i, for $i = 1, \ldots, 95$ in Equation (3.8) and predict the six variables on unsampled sites of the region utilizing functional kriging based on Andres curves. To evaluate the performance of the predictor we do a cross-validation analysis comparing the results with the ones obtained with ordinary kriging. In Table 4 we show the sum of squares of prediction errors for each one of the six variables based on

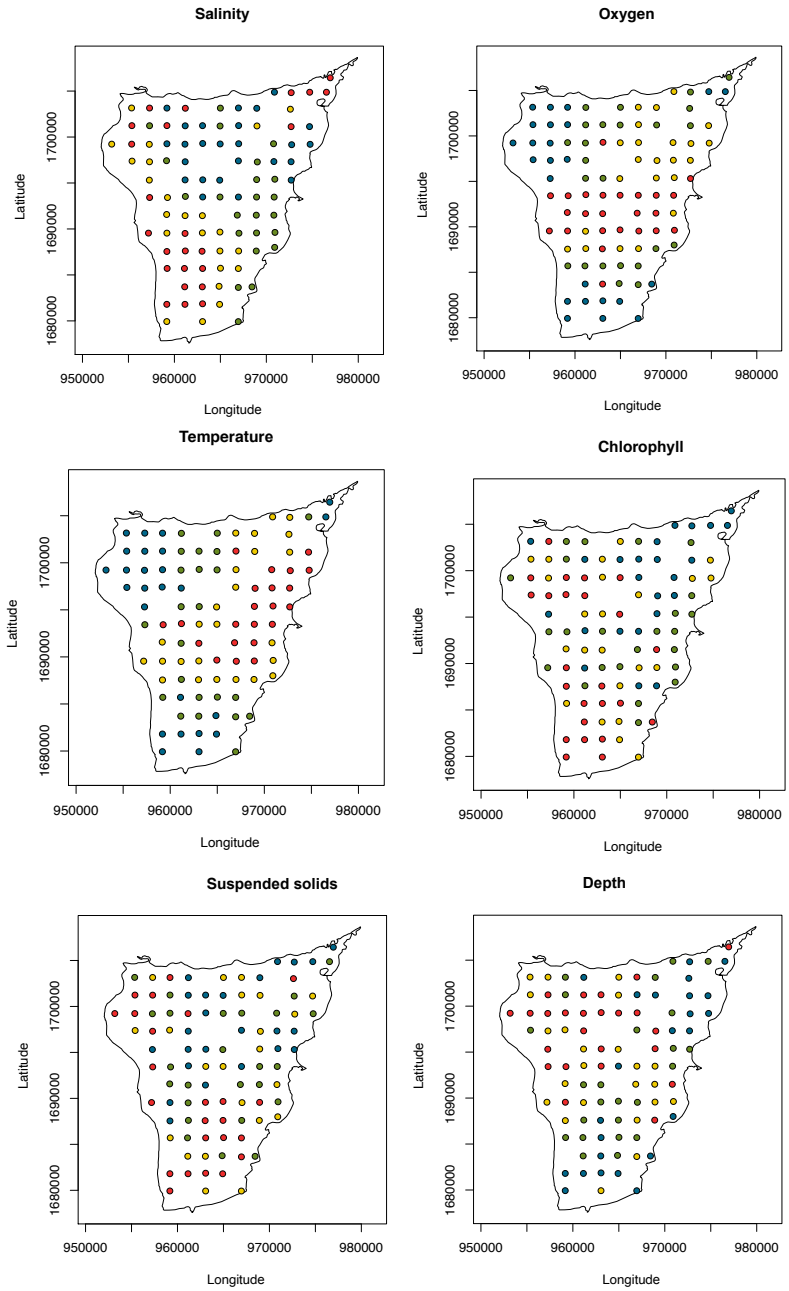


Figure 2. Spatial distribution of data for each one of the six variables considered. The values are, in each case, divided according to the quartiles.

Table 3. Coefficients of variation calculated with data recorded in 95 sites of the lagoon-estuarine system Ciénaga Grande de Santa Marta.

Variable	Coefficient of variation (%)
Oxygen	36.5
Depth	24.5
Chlorophyll	23.8
Suspended solids	19.3
Salinity	17.1
Temperature	7.2

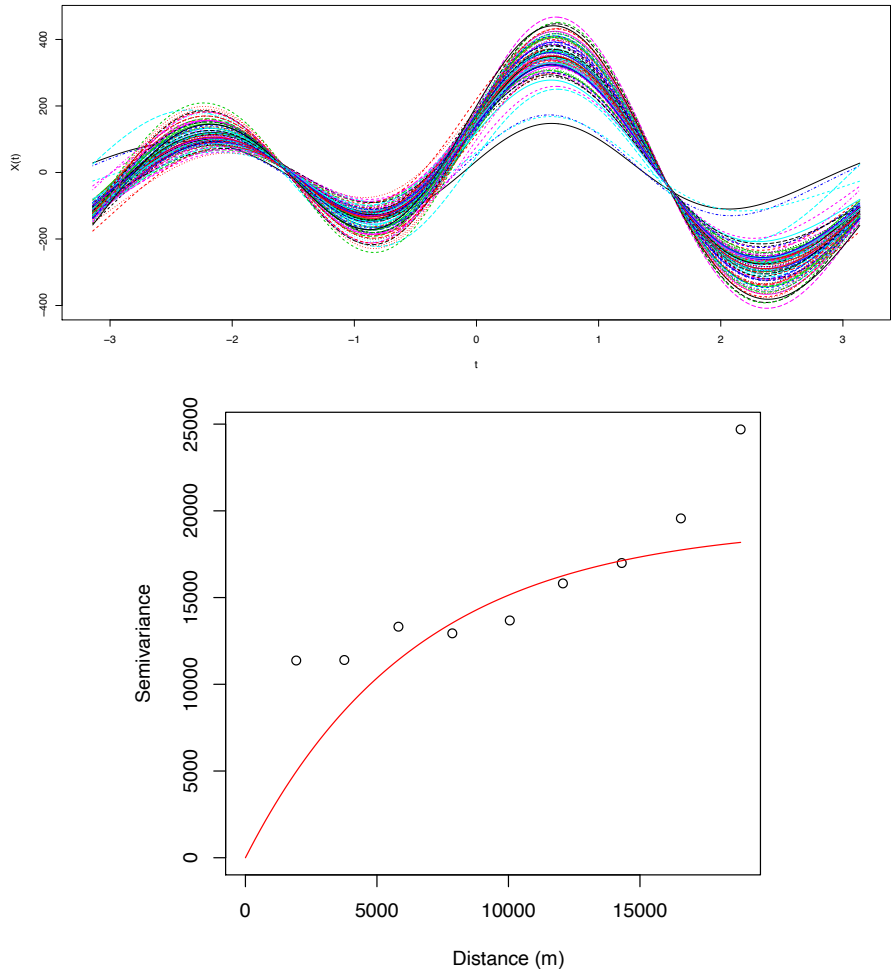


Figure 3. Andrews curve calculated for each one of 95 sites of the lagoon-estuarine system Ciénaga Grande de Santa Marta, based on the values of six physicochemical variables (top); and variogram model (red line) fitted to the empirical trace-variogram function (bottom).

the two approaches considered. In general the results look similar. To test for significant differences we use Wilcoxon tests based on the cross-validation residuals. These indicate that the method based on functional kriging using Andrews curves is better than ordinary kriging in the case of the variables depth and suspended solids. In the other cases there are not significant differences between the two strategies.

Table 4. Sum of squares errors of cross-validation (using the data of 95 sites) obtained by functional kriging based on Andrews curves and ordinary univariate kriging.

	Functional kriging	Univariate kriging
Oxygen	218.8	216.9
Depth	12.3	12.5
Chlorophyll	46335.9	46457.2
Suspended solids	169397.1	170136.2
Salinity	223.2	229.9
Temperature	46.9	46.4

5. Conclusions, limitations, and future research

In the paper, we have proposed the predictor ordinary kriging for functional data (Giraldo et al., 2011) based on Andrews curves (Andrews, 1972) as a method for making spatial prediction in multivariate geostatistics (Smith et al., 1962). The results based on a simulation study and an analysis of real-world data have indicated that this strategy has a good performance. Obviously, if the geostatistical analysis is carried out with two or three variables, it is more convenient to use cokriging, since the prediction variance is reduced. However, when the number of variables increases, this option is limited and the proposed technique emerges as a very appropriate alternative, because it only requires the estimation of just one variogram and does not have the limitations of the linear coregionalization model.

The proposed methodology could be adapted to the case of optimal sampling (Bohorquez et al., 2016), regression, and analysis of variance of multivariate spatial data. Other research alternatives are the extension to the case of non-stationary processes and the treatment of outliers (Borssoi et al., 2011).

Author contributions Conceptualization, M.D., R.G.; methodology, M.D., R.G.; software, M.M.; investigation, M.D., R.G; writing original draft preparation, M.D., R.G., writing review and editing, R.G. Both authors have read and agreed to the published version of the paper.

Acknowledgements The authors would like to thank the Editors and the reviewers for their valuable comments and suggestions which improved the quality of this paper.

Funding Not applicable.

Conflicts of interest The authors declare no conflict of interest.

References

Andrews, D., 1972. Plots of high-dimensional data. Biometrics, 28, 125–136.

Bilodeau, M. and Brenner, D., 1999. Multivariate Regression. Springer, New York, USA.

Bohorquez, M., Giraldo, R., and Mateu, J., 2016. Optimal sampling for spatial prediction of functional data. Statistical Methods Applications, 25, 39–54.

Borssoi, J.A., De Bastiani, F., Uribe-Opazo, M.A., and Galea, M., 2011. Local influence of explanatory variables in Gaussian spatial linear models. Chilean Journal of Statistics, 2(2), 29–38.

Dueñas, M., 2017. Análisis geoestadístico multivariado a través de métodos funcionales y curvas de Andrews. Master thesis. Department of Statistics, Universidad Nacional de Colombia, Colombia.

Embrechts, P., Herzberg, A., and Allen, C., 1986. An investigation of Andrews's plots to detect period and outliers in time series data. Communications in Statistics: Simulation and Computation, 15, 1027–1051.

Fahrmeir, L., Tutz, G., Hennevogl, W. and Salem, E., 1994. Multivariate Statistical Modelling Based on Generalized Linear Models. Springer, New York, USA.

Gelfand, A., Diggle, P. Guttorp, P., and Fuentes, M., 2010. Handbook of Spatial Statistics. CRC Press, New York, USA.

Genton, M. and Kleiber, W., 2015. Cross-covariance functions for multivariate geostatistics. Statistical Science, 30, 147–163.

Giraldo, R., Delicado, P., and Mateu, J., 2011. Ordinary kriging for function-valued spatial data. Environmental Ecological Statistics, 18, 411–426.

Giraldo, R., Delicado, P., and Mateu, J., 2017. Spatial prediction of a scalar variable based on data of a functional random field. Comunicaciones en Estadística, 10, 315–344.

Giraldo, R., Herrera, L., and Leiva, V., 2020. Cokriging prediction using as secondary variable a functional random field with application in environmental pollution. Mathematics, 8, 1305.

Mateu, J. and Giraldo, R., 2021. Geostatistical Functional Data Analysis. Wiley, New York, USA.

Mustafa, R., 2011. Andrews curves. Computational Statistics, 3, 373–382.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rodríguez-Rodríguez, J.A., Mancera-Pineda, J.E., and Tavera, H., 2021. Mangrove restoration in Colombia: Trends and lessons learned. Forest Ecology and Management, 496, 119414.

Smith, H., Gnanadesikan, R., and Hughes, J., 1962. Multivariate analysis of variance (MANOVA). Biometrics, 18, 22–41.

Verbeke, G., Fieuws, S., Molenberghs, G., and Davidian, M., 2014. The analysis of multivariate longitudinal data: a review. Statistical Methods in Medical Research, 23, 42–59.

Wackernagel, H., 2003. Multivariate Geostatistics: An Introduction with Applications. Springer, New York, USA.

Chilean Journal of Statistics
Volume 13, Number 1 (April 2022), Pages 17–46
DOI: 10.32372/ChJS.13-01-02

PSYCHOMETRICS
RESEARCH PAPER

Dissecting Chilean surveys: The case of missing outcomes

ERNESTO SAN MARTÍN [1,2,3,4,*] and EDUARDO ALARCÓN-BUSTAMANTE[1,2,3]

[1]Faculty of Mathematics, Department of Statistics, Pontificia Universidad Católica de Chile, Santiago, Chile.
[2]Millenium Nucleus on Intergenerational Mobility: From Modelling to Policy, MOVI.
[3]Interdisciplinary Laboratory of Social Statistics, Santiago, Chile.
[4]The Economics School of Louvain, Université Catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium.

(Received: 19 January 2022 · Accepted in final form: 29 March 2022)

Abstract

In this paper, the strengths and weaknesses of two Chilean political polls and the National Socioeconomic Characterisation Survey are analyzed from a statistical modelling point of view. The rationale of the analytical strategy is based on a distinction between identified parameters and parameters of interest. This is equivalent to make a distinction between what we can learn from the data provided by a survey and what we want to learn from those data. Using partial identification techniques, each survey is analyzed at different levels according to specific subpopulations. Based on these analyses, we emphasize not only the way in which the results should be reported, but also the necessity to make explicit the uncertainty induced by the non-response rates at the survey report.

Keywords: Ignorability condition · Missing data · Non-response · Partial identifiability · Quantile function

Mathematics Subject Classification: 46N30 · 78M31.

1. INTRODUCTION

Broadly speaking, public surveys are applied either to get a better gauge of citizens' political opinions (Berinsky, 2017) or to collect information that is useful for policy makers. These surveys are perceived as reliable tools as it is argued that they are applied to "representative samples". If this were the case, the analysis of the strength of a survey would be reduced to indicating how a sample design ensures access to a "representative sample". However, it is necessary to emphasize that the expression "representative sample" is not a statistical concept because it is logically contradictory. As a matter fact, a survey is applied to know the behavior of a population in relation to an outcome of interest. Doing so means that we have no idea about this outcome: how then can we ensure the representativeness of the survey? In addition, if we know this outcome at the population level, why do we need to conduct a survey?

*Corresponding author. Email: esanmart@mat.uc.cl

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

http://www.soche.cl/chjs

A question then arises: how can we assess a survey? This paper intends to answer this question in a specific but quite typical case, namely when some surveyed individuals do not answer a specific question. Our approach is based on two questions: what can be learned from the data provided by a survey? And what do we want to learn from those data? The difference between these two question relies on the statistical concept of identifiablity.

As a matter of fact, a statistical model is a family of probability distributions indexed by a parameter and defined on a sample space. From a modelling point of view, a set of data is fully represented by a probability distribution that generates them. Consequently, a parameter of this distribution represents a specific characteristic of the set of data under analysis; see Fisher (1922). Technically speaking, these correspond to the identified parameter. However, if we attribute a characteristic to a set of data that cannot be represented by a parameter (it is not a functional of the probability distribution generating the data), then we face an identification problem. Technically speaking, these correspond to a parameter of interest. Thus, the identified parameters summarize what can be learned from the data, whereas the parameters of interest represent what we want to learn from the data. When an injective relationship is established between them, the identification problem is solved. For details and references, see Koopmans and Reiersol (1950); San Martín (2018), San Martín et al. (2015) and San Martín and González (2022).

In this paper, we use this conceptual distinction to assess both the strengths and weaknesses of three Chilean surveys: two of political opinion (CADEM survey and the Araucanía citizen consultation), and one related to the income distribution of employees (National Socioeconomic Characterization Survey, CASEN in Spanish). We analyze the identification problem raised by missing outcomes. To do that, we employ Manski's technique of partial identification, which allows us to evaluate how strong are the ignorability conditions (also known as missing at random condition) typically used to impute missing data. Based on this discussion, we emphasize the way in which these survey should report their results.

Let us remark the type of conclusion that can be done from a partial identification analysis. Typically, an identification analysis allows a parameter of interest to be point identified. For instance, in a fixed effect ANOVA model, the mean of the observations nested into a same group (for example, scores of students of a specific school) is parameterized as an addition of two parameters, namely $\mathrm{E}(Y_{ij}) = \alpha + \theta_j$, where j labels the groups and i labels the statistical units. Let us call θ_j, parameter of group j, and α, global parameter. The group parameters are point identified if, for instance, the parameter of the first group is assumed to be equal to 0. In this case, the parameter of a specific group is equal to the difference between the means of that group and of the first group (this explains why this identification constraint is known as deviation from the mean). However, a partial identification analysis provides an identification region to which the parameter of interest belongs, rather than identifying it pointwise. This is due to the fact that an identification analysis makes explicit certain assumptions (identification restrictions) under which the parameter of interest is point identified, but, in the context of application, such a restriction is incredible (Manski, 2011, 2020). Therefore, the analysis strategy consists of relaxing such assumptions to establish a region to which this parameter belongs. The reader may ask where is the disadvantage of accepting incredible identification constrains to point identify the parameters of interest. The drawback lies in the fact that scientific conclusions and/or policy recommendations depend more on such constrains than on the data and, consequently, an illusion of scientific certainty is created based only on incredible certainty.

These considerations are illustrated through the dissection of three Chilean surveys. This paper is accordingly organized as follows. In Section 2, the political opinion survey CADEM is analyzed. Section 3 focuses its attention on the National Socioeconomic Characterisation Survey CASEN. Section 4 analyzes a recent citizen consultation applied in the Araucanía region in the south of Chile. In each of these sections, we provide the corresponding method-

ological information of each survey and also the political and/or economical context in which the survey is used. The paper ends with concluding remarks in Section 5.

2. CADEM survey

We begin by dissecting the CADEM political opinion survey. After describing the purpose of the survey and summarizing the methodology used to deal with missing data, we perform a conditional identification analysis of different sub-populations of interest.

2.1 General objective and methodological information

According to the information provided on its website, the CADEM survey is one of the many services offered by the market research company CADEM Research & Estrategia. Specifically, it is related to the service called Plaza Pública, which describes itself as "the first and only polling platform that measures public opinion on a weekly basis to provide data and analysis on a wide range of topics of interest"[1]. This particular aspect is related to one of the general objectives of this marketing company: "We want to connect people with decision makers, through data and not from intuition, providing strategies and action plans to achieve the expected results based on a deep knowledge of the new consumer/citizen"[2].

CADEM survey delivers "reliable, timely and contingent information on the political, economic and social debate in Chile on a weekly basis". The study published by CADEM "contemplates a probabilistic survey of 700 weekly cases (with a monthly consolidation that goes from 2,800 surveys to 3,500 depending on whether the month has 4 or 5 weeks), applied 100% through cell phones, using CADEM's own database that contains more than 18 million cell phones considering both prepaid and postpaid numbers, all obtained through Random Digit Dialing and consolidated during the last four years". Its target group is, therefore, all individuals living in the national territory, Chileans and immigrants, men and women over 18 years old, inhabitants of the 15 regions of the country. This led to perform a previous stratification of the total population based on the population projections made by the National Institute of Statistics (NIS) of the Chilean Government for the year 2017 at the national level. Table 1 presents the estimated population aged 18 and over for each region of the country as of 2017 and the number of surveys proposed for each region to comply with the national proportionality. In addition to the distribution by region, the previous stratification considers, only as a control, the combination of sex and age variables; for more details, see CADEM (2018).

It is important to emphasize that this general information is not published week by week, except for the total number of people selected and the total number of people who agreed to answer the survey.

2.2 How are the missing responses treated?

Taking into account that the survey is conducted by telephone, the main issue is the non-response rate. CADEM is not only aware of this problem, but distinguishes three cases of non-response: cases of no contact, namely no one answers the call either because the phone is busy or out of service; cases of a non-eligible person, namely a person who answers the call, but does not satisfy the requirements of the target group; and a person who is correctly selected but refuses to answer the survey. The impact of the non-response rate is assessed in the following terms:

[1]Retrieved from https://cadem.cl/sobre-cadem/ on December 30, 2021.
[2]Retrieved from https://cadem.cl/plaza-publica/ on December 30, 2021.

Table 1. NIS population projections for 2017 and number of surveyed

Region	Population over 18 years old	Theoretical sample
XV	182,301	9
I	252,814	13
II	471,980	24
III	234,933	12
IV	595,594	30
V	1,430,182	72
VI	706,014	35
VII	804,214	40
VIII	1,634,325	82
IX	756,349	38
XIV	313,112	16
X	636,432	32
XI	80,797	4
XII	126,772	6
RM	5,713,842	287
Total	13,939,661	700

> Estimating the magnitude of non-response is critical because of the direct relationship it may have with self-selection biases in public opinion polls. The calculation of the non-response rate is also used as a measure of validation of the results. Under the assumption that those who rejects to answer the survey are equal to those who answers it, the magnitude of the non-response rate does not offer major disadvantages, but when there is evidence that the two groups are not equivalent, the non-response can introduce serious distortions in the results (CADEM, 2018).

CADEM accordingly reports the rate of non-response. Three types of results are reported by the survey: those that make explicit the number of cases surveyed, which is approximately equal to 700; those that use a subset of these cases; and trends over time, using previous survey results. However, the impact of the non-response rate on both the results of the survey and their report are not discussed. For an example, see the survey published on the fourth week of December 2021 (CADEM, 2022).

2.3 Dissecting the CADEM survey

The objective of this section is to answer the following questions: What can be learned from the data collected by the CADEM survey? How reliable is the CADEM survey? To be consistent with a certain degree of reliability, how should its results be communicated?

Example Let us consider the collected results during the fifth week of December 2021 (CADEM, 2022). As mentioned above, each study contains a methodological sheet, which indicates that the sampling is a probability sample with random selection of individuals and previously stratified by region; that the sample consists of 705 cases, which required making 6,401 telephone calls, so the response rate is equal to 11%. Let us focus our attention on the first question of the study:

> Do you have a very positive, positive, negative or very negative image of Gabriel Boric?

The results are the following: 63% have a very positive or positive (denoted by a) image of Gabriel Boric; 27% have a negative or very negative (denoted by b) image; and 10% do not know or non-response (denoted by c).

WHAT WE CAN LEARN FROM THE DATA? Let M be the sample space whose components are the numbers of cellular phones. On this space we define the vector of random variables (E, R, S, C, G): $M \to \{0,1\}^4 \times \{1,\ldots,15\}$, where for each $m \in M$

- $E(m) = 1$ if the person associated with cell phone m is eligible, and $E(m) = 0$ if not.
- $R(m) = 1$ if the cell phone m answers the call, and $R(m) = 0$ if not.
- $S(m) = 1$ if the person associated with cell phone m is selected, and $S(m) = 0$ if not.
- $C(m) = 1$ if the person associated with cell phone m answers the survey, and $C(m) = 0$ if not.
- $G(m) = g$ with $g \in \{1,\ldots,15\}$ if the person associated with cell phone m belongs to region g.

From these definitions, it follows that

$$\begin{aligned} &\{m \in M\colon S(m) = 1\} \subset \{m \in M\colon E(m) = 1\} \cap \{m \in M\colon R(m) = 1\}; \\ &\{m \in M\colon S(m) = 1\} = \{m \in M\colon C(m) = 0\} \cup \{m \in M\colon C(m) = 1\}. \end{aligned} \tag{2.1}$$

Let Y be the outcome of interest, taking values in the set $\{a, b, c\}$. The data inform about the conditional distribution of Y given $(E = 1, R = 1, S = 1, C = 1)$; that is,

$$\begin{aligned} &P(Y = a \mid E = 1, R = 1, S = 1, C = 1) = 0.63; \\ &P(Y = b \mid E = 1, R = 1, S = 1, C = 1) = 0.27; \\ &P(Y = c \mid E = 1, R = 1, S = 1, C = 1) = 0.10; \\ &P(C = 1 \mid E = 1, R = 1, S = 1) = 0.11. \end{aligned}$$

Both $P(Y = y \mid E = 1, R = 1, S = 1, C = 1)$ for $y \in \{a, b, c\}$, and $P(C = c \mid E = 1, R = 1, S = 1)$ for $c \in \{0, 1\}$ correspond to the identified parameter, and therefore they represent all that can be learned from the data.

WHAT WE WANT TO LEARN FROM THE DATA? The results of the CADEM survey can be interpreted conditionally to different sub-populations.

First level of analysis The first level corresponds to what we can learn from the data and it is captured by the identified parameter $P(Y = y \mid E = 1, R = 1, S = 1, C = 1)$ for $y \in \{a, b, c\}$.

Second level of analysis A second level corresponds to focus the attention on the surveyed persons, namely $\{m \in M\colon S(m) = 1\}$, which by Equation (2.1) is equivalent to $\{m \in M\colon E(m) = 1, S(m) = 1, R(m) = 1\}$. In this case, it is not longer possible to identified $P(Y = y \mid E = 1, S = 1, R = 1)$. As a matter of fact, by the law of total probability (Kolmogorov, 1950),

$$\begin{aligned} &P(Y = y \mid E = 1, S = 1, R = 1) = P(Y = y \mid S = 1) \quad \text{by Equation (2.1)} \\ &= P(Y = y \mid S = 1, C = 1)P(C = 1 \mid S = 1) + P(Y = y \mid S = 1, C = 0)P(C = 0 \mid S = 1) \end{aligned} \tag{2.2}$$

for each $y \in \{a, b, c\}$. In this decomposition, $P(Y = y \mid S = 1, C = 1)$ and $P(C = 1 \mid S = 1)$ are identified, whereas $P(Y = y \mid S = 1, C = 0)$ is not identified because it depends of those

persons who refuse to answer the survey. Taking into account that such a probability takes values between 0 and 1, we can provide an interval of all plausible values for $P(Y = y \mid E = 1, S = 1, R = 1)$ which are compatible with the observed information: for each $y \in \{a, b, c\}$,

$$\begin{aligned} P(Y = y \mid S = 1, C = 1)P(C = 1 \mid S = 1) &\leq P(Y = y \mid S = 1) \\ &\leq P(Y = y \mid S = 1, C = 1)P(C = 1 \mid S = 1) + P(C = 0 \mid S = 1). \end{aligned} \tag{2.3}$$

Following Manski (2007), this interval corresponds to the region where $P(Y = y \mid S = 1)$ is partially identified. Such an interval deserves the comments:

(i) Considering the example of Subsection 2.3, we have that $P(C = 1 \mid S = 1) = 0.11$ and $P(Y = a \mid S = 1) = 0.63$. Therefore,

$$0.0693 \leq P(Y = a \mid S = 1) \leq 0.9593. \tag{2.4}$$

Thus, the survey report should be phrased in the following terms: at least 6.93% of the surveyed people have a positive or very positive image of Gabriel Boric, and at most 95.93% of the surveyed people have such positive or very positive image.

(ii) This interval provides information about the uncertainty inherent to the non-response rate. In fact, the width of Equation (2.4) is equal to $P(C = 0 \mid S = 1)$, which in this example is equal to 89%. This means that the interval is close to be uninformative.

(iii) Different scenarios should be considered when reporting $P(Y = a \mid S = 1)$, $P(Y = b \mid S = 1)$ and $P(Y = c \mid S = 1)$ because these three probabilities belongs to the 2-dimensional simplex $S_3 = \{(p_1, p_2, p_3) \in [0, 1]^3 \colon p_1 + p_2 + p_3 = 1\}$. Thus, for instance, it can be said that 95.93% of surveyed people have a positive or very positive image of Gabriel Boric and, consequently, a 4.07% have a poor or very poor image or Gabriel Boric, or do not known or non-response, that is,

$$\begin{aligned} 1 - [P(Y = a, C = 1 \mid S = 1) + P(C = 0 \mid S = 1)] & \\ = & P(C = 1 \mid S = 1) + P(C = 0 \mid S = 1) \\ & - P(Y = a, C = 1 \mid S = 1) - P(C = 0 \mid S = 1) \\ = & P(C = 1 \mid S = 1) - P(Y = a, C = 1 \mid S = 1) \\ = & P(Y \neq a, C = 1 \mid S = 1) \\ = & P(Y \in \{b, c\}, C = 1 \mid S = 1) \\ = & P(Y = b, C = 1 \mid S = 1) + P(Y = c, C = 1 \mid S = 1), \end{aligned}$$

which is the lower bound of $P(Y \in \{b, c\} \mid S = 1)$. In the example, $P(Y = b, C = 1 \mid S = 1) = 0.0297$ and $P(Y = c, C = 1 \mid S = 1) = 0.011$.

Once the partial identification of $P(Y = y \mid S = 1)$ $(y \in \{a, b, c\})$ is established, it is possible to qualify CADEM's claims about non-responses. As it was mentioned in Subsection 2.2, CADEM considers that, "under the assumption that those who rejects to answer the survey are equal to those who answers it, the magnitude of the non-response rate does not offer major disadvantages, but when there is evidence that the two groups are not equivalent, the non-response can introduce serious distortions in the results". If we consider the decomposition of Equation (2.2), the assumption advanced by CADEM corresponds to the equality $P(Y = y \mid S = 1, C = 1) = P(Y = y \mid S = 1, C = 0)$, for all $y \in \{a, b, c\}$, which, by definition of conditional independence, is equivalent to $Y \perp\!\!\!\perp C \mid \{S = 1\}$; where

$V \perp\!\!\!\perp W \mid Z$ corresponds to the conditional independence between V and W given Z; for details and properties on conditional independence, see Florens et al. (1990, Ch. 2). This condition, typically known as missing at random (Rubin, 1976; Little and Rubin, 2019), is not empirically refutable because it depends on the component $P(Y = y \mid S = 1, C = 0)$ which in turn is not based on actual observations. Consequently, it is impossible to find out evidence establishing that "the two groups are not equivalent".

Correctly stated, condition in Equation (2.6) is an identification restriction (San Martín and González, 2022) under which $P(Y = y \mid S = 1)$ is point identified in the sense that $P(Y = y \mid S = 1) = P(Y = y \mid S = 1, C = 1)$, for all $y \in \{a, b, c\}$.

In other words, under assumption of Equation (2.6), the uncertainty induced by the non-response decreases from an interval of width $P(C = 0 \mid S = 1)$ to the singleton $\{P(Y = y \mid S = 1, C = 1)\}$. Thus, what we want to learn from the data coincides with what we can learn from the data. In passing, let us mention that condition in Equation (2.6) should be viewed as a characterization of absence of (self-)biased and, consequently, the identification problem induced by the non-response is exactly the same as the identification problem induced by self-selection.

Third level of analysis A third level of analysis corresponds to focus the attention on the eligible persons, namely $\{m \in E(m) = 1\}$. In this case, the parameter of interest is given by $P(Y = y \mid E = 1)$ for $y \in \{a, b, c\}$. Let us analyze its identifiability using only the information available at the CADEM survey as published.

Using the law of total probability, we have

$$\begin{aligned} P(Y = y \mid E = 1) = {} & P(Y = y \mid E = 1, R = 1)P(R = 1 \mid E = 1) \\ & + P(Y = y \mid E = 1, R = 0)P(R = 0 \mid E = 1), \end{aligned}$$

for $y \in \{a, b, c\}$. In this decomposition, $\gamma \doteq P(Y = y \mid E = 1, R = 0)$ is not identified because it is impossible to know whether a person associated with a cell phone that does not answer a call is eligible or not. Also, $P(Y = y \mid E = 1, R = 1)$ can be decomposed as

$$\begin{aligned} P(Y = y \mid E = 1, R = 1) = {} & P(Y = y \mid E = 1, R = 1, S = 1)P(S = 1 \mid R = 1, E = 1) \\ & + P(Y = y \mid E = 1, R = 1, S = 0)P(S = 0 \mid R = 1, E = 1) \\ = {} & P(Y = y \mid S = 1)P(S = 1 \mid R = 1, E = 1) \\ & + P(Y = y \mid E = 1, R = 1, S = 0)P(S = 0 \mid R = 1, E = 1), \end{aligned}$$

where the last equality follows from Equation (2.1).

Note that $\{m \in M\colon E(m) = 1, R(m) = 1, S(m) = 0\} = \emptyset$, because there are no eligible persons associated with a cell phone that answered the call who are not selected. Consequently, $P(S = 0 \mid R = 1, E = 1) = P(S = 0, R = 1, E = 1)/P(R = 1, E = 1) = 0$. Moreover, $P(Y = y \mid E = 1, R = 1, S = 0)$ is a probability conditional on an event of probability 0 and, therefore, takes an arbitrary value in $[0, 1]$ (see Remark 2.1). It follows that $P(Y = y \mid E = 1, R = 1) = P(Y = y \mid S = 1)P(S = 1 \mid R = 1, E = 1)$. Thus, for each $y \in \{a, b, c\}$,

$$\begin{aligned} P(Y = y \mid E = 1) = {} & P(Y = y \mid S = 1)P(S = 1 \mid R = 1, E = 1)P(R = 1 \mid E = 1) \\ & + \gamma\, P(R = 0 \mid E = 1) \\ = {} & P(Y = y \mid S = 1)P(S = 1 \mid E = 1) + \gamma\, P(R = 0 \mid E = 1), \end{aligned}$$

for all $\gamma \in [0, 1]$. In this decomposition, $P(Y = y \mid S = 1)$ is partially identified by

the interval in Equation (2.3); by Equation (2.1), $P(S = 1 \mid E = 1)$ corresponds to the ratio $\sharp\{\text{selected persons}\}/\sharp\{\text{eligible persons}\}$, which is identified; and $P(R = 0 \mid E = 1)$ corresponds to the proportion of eligible persons who did not respond to the telephone call. Taking into account that a person can be classified as eligible once he/she has answered the telephone call (see Section 2.1), then it is impossible to identify this parameter. Nevertheless, Equation (2.1) implies that $\{m \in M\colon R(m) = 0\} \subset \{m \in M\colon S(m) = 0\}$ and, therefore,

$$P(R = 0 \mid E = 1) \leq P(S = 0 \mid E = 1) = 1 - P(S = 1 \mid E = 1) = \frac{\sharp\{\text{non-selected persons}\}}{\sharp\{\text{eligible persons}\}},$$

which is identified. Hence, $P(Y = y \mid E = 1)$ is partially identified, where the lower bound of the identification region is given by

$$P(Y = y \mid S = 1, C = 1)P(C = 1 \mid S = 1)P(S = 1 \mid E = 1),$$

which by Equation (2.1) reduces to $P(Y = y, S = 1, C = 1 \mid E = 1)$; and its upper bound is expressed as

$$\begin{aligned}&[\, P(Y = y \mid S = 1, C = 1)P(C = 1 \mid S = 1) + P(C = 0 \mid S = 1)\,]\times\\ &P(S = 1 \mid E = 1) + P(S = 0 \mid E = 1),\end{aligned}$$

which by Equation (2.1) reduces to

$$P(Y = y, S = 1, C = 1 \mid E = 1) + P(C = 0, S = 1 \mid E = 1) + P(S = 0 \mid E = 1).$$

Using the data of the example, $P(S = 1 \mid E = 1) \sim 6,401/14 \times 10^6$ and $0.00003198 \leq P(Y = a \mid E = 1) \leq 0.9999814$, clearly this interval is non-informative.

Remark 2.1 Let $(M, \mathcal{M}, P)$ be a finite probability space. Let $\mathcal{C} = (C_1, \ldots, C_n\} \subset \mathcal{M}$ be a partition of M such that $P(C_1) = 0$ and $P(C_j) > 0$ for $j = 2, \ldots, n$. Therefore, let $A \in \mathcal{M}$. In this case, the conditional probability $P(A \mid \mathcal{C})$ is a random variable defined as $P(A \mid \mathcal{C}) = \sum_{j=1}^{n} P(A \mid C_j)\mathbb{1}_{C_j}$, where $\mathbb{1}_{C_j}$ is the indicator function of the event C_j (Kolmogorov, 1950, §6). Here, the numbers $P(A \mid C_j)$ are computed using a rule stated as

$$P(A \mid C_j) = \begin{cases} \dfrac{P(A \cap C_j)}{P(C_j)}, \text{ if } P(C_j) > 0; \\ \eta \in [0, 1], \text{ if } P(C_j) = 0; \end{cases} \tag{2.5}$$

with η arbitrary. This rule is a correct rule (that is, it avoids paradoxes) because it satisfies the equality $P(A) = E[P(A \mid \mathcal{C})]$, which ensures the existence of the conditional probability. As a matter of fact, under rule in Equation (2.5), this equality reduces to the law of total probability –in the general case, it corresponds to the Radon-Nikodym theorem. Moreover, the number $P(A \mid C_a)$ can be arbitrarily chosen because the random variable $P(A \mid \mathcal{C})$ does not change since $P(C_1) = 0$. For more details, see Rao (2005, Ch. 2).

Fourth level of analysis The non-informativity of the above identification region is primarily due to the fact that $P(S = 1 \mid E = 1)$ is extremely small, so $P(S = 0 \mid E = 1)$ is extremely large. This undesired effect could be counteracted by taking into account the information provided by the CADEM survey regarding how persons are selected: "Probabilistic sampling with random selection of individuals and previously stratified by region" (CADEM, 2022).

By the CADEM sampling design, the reasoning should be done conditionally on $\{m \in M\colon E(m) = 1, R(m) = 1\}$: it is impossible to know whether a person is eligible if he/she has not answered the phone call. Thus, the statement "random selection of individuals and previously stratified by region" corresponds to the condition

$$P(Y = y \mid E = 1, R = 1, G, S = 1) = P(Y = y \mid E = 1, R = 1, G, S = 0),$$

which, by definition of conditional independence, is equivalent to

$$Y \perp\!\!\!\perp C \mid \{S = 1\}; \tag{2.6}$$

By the law of total probability, this condition implies by Equation (2.1) that

$$\begin{aligned} P(Y = y \mid E = 1, R = 1, G) &= P(Y = y \mid S = 1, E = 1, R = 1, G) \\ &= P(Y = y \mid S = 1, G). \end{aligned} \tag{2.7}$$

Thus, to identify $P(Y = y \mid E = 1, R = 1)$, we marginalize with respect to G, namely

$$\begin{aligned} P(Y = y \mid E = 1, R = 1) &= \sum_{g=1}^{15} P(Y = y \mid E = 1, R = 1, G = g) P(G = g \mid E = 1, R = 1) \\ &= \sum_{g=1}^{15} P(Y = y \mid S = 1, G = g) P(G = g \mid E = 1, R = 1), \end{aligned}$$

where the last equality follows from Equation (2.7).

In this decomposition, the conditional probability $P(G = g \mid E = 1, R = 1)$ is in principle identified, although the current information provided by CADEM does not allow to identify it. Moreover, the conditional probability $P(Y = y \mid S = 1, G = g)$ has the same identification problem that was discussed in the second level of analysis and, therefore, it is partially identified: for each $y \in \{a, b, c\}$ and $g \in \{1, \ldots, 15\}$,

$$\begin{aligned} &P(Y = y \mid S = 1, C = 1, G = g) P(C = 1 \mid S = 1, G = g) \\ &\leq P(Y = y \mid S = 1, G = g) \\ &\leq P(Y = y \mid S = 1, C = 1, G = g) P(C = 1 \mid S = 1, G = g) + P(C = 0 \mid S = 1, G = g). \end{aligned}$$

Therefore, the random selection of each individual in each stratum is far from helping to identify $P(Y = y \mid E = 1, R = 1)$. Furthermore, it does not help to identify $P(Y = y \mid E = 1)$ either, since $P(Y = y \mid E = 1, R = 0)$ is still unidentified.

2.4 Discussion

CADEM research & estrategia offers services that "connect people with decision makers, through data and not from intuition". Nevertheless, after dissecting the CADEM survey, we can say that this motto is far from being fulfilled. In fact, the dissection of the CADEM survey shows how weak its reliability is whatever the level of analysis.

The first level of analysis corresponds to a description of the collected data. For the sake of transparency, CADEM must not only remember for each question of the survey the total number of people who answered it, but also indicate, together with the percentages of preference for each option, the absolute frequencies. This warn the readers and especially the

press that the results reflect the opinion of a very small number of people. The second level of analysis makes explicit the uncertainty induced by the non-response. CADEM should be made explicit such uncertainty by reporting both the lower and the upper bound of the identification region of $P(Y = y \mid S = 1)$. In the example, the impact of the non-response rate is dramatic, which prevents the reader from a false illusion of certainty. It should be emphasized that condition in Equation (2.6) is a plausible way to treat the non-responses. A transparent treatment of non-response should show the impact of such a condition on the conclusions of the study. As we have seen in the example, the conclusion depends much more on Equation (2.6) than on the data itself. The third level of analysis focuses on the eligible population. Once again, for the sake of transparency, it is necessary to report both the lower and the upper bound of the identification region. The example we have used shows how uninformative the survey results are. This information is more than relevant, showing the intrinsic limits of this type of public opinion instruments.

3. CASEN survey

The National Socioeconomic Characterisation Survey (CASEN, for their initials in Spanish) is a Chilean household survey that has been applied since 1987. It is used to assess the impact of social programs on the living conditions of the population[1]. According to the Technical data sheet, the target population is the population residing in private households throughout the national territory. The units of analysis are families and individuals living in a household. A suitable respondent is the head of household or, alternatively, a man or woman over 18 years old.

The sampling process of the CASEN survey consists on two steps. First, blocks are chosen that correspond to sets of households; second, a household is chosen in which individuals are surveyed. Due to the pandemic by COVID19, the last version of the survey, called 2020 CASEN survey in pandemic, was carried out in two steps: first, from the households selected in the previously mentioned sampling process, a face-to-face pre-contact was applied to obtain a contact telephone number. Second, the survey was administered by telephone.

In the 2020 CASEN in pandemic survey, 97,848 households were pre-contacted. Of these, only 86,189 households provided at least a telephone number to be contacted. Of these, 62,540 households had individuals who answered the survey, which amounted to 185,437 individuals[2]. It should be remarked that the available CASEN data set contains information of these individuals[3].

3.1 Treatment of missing outcomes in the CASEN survey

One of the objectives of the CASEN survey is to obtain an overview of the income distribution in Chile, and in particular to have an overview of poverty in the country in terms of income. However, some of the selected individuals did not answer the question on income. CASEN considers appropriate to impute these missing data, so that researchers and policy makers can use a database without missing data. The chosen imputation procedure is called conditional mean imputation. The rationale of this technique can be summarized as follows: first, observed covariates are used to define classes. Second, individuals who did not report their income and individuals who reported it are classified in the same class if they share the

[1]Retrieved from http://casenpandemia2020.cl/ on December 30, 2021.
[2]For details, see Nota técnica N7: Desempeño del Trabajo de Campo, Casen en Pandemia en sección Notas Técnicas 2020: http://observatorio.ministeriodesarrollosocial.gob.cl/encuesta-casen-en-pandemia-2020.
[3]The data base can be downloaded from http://observatorio.ministeriodesarrollosocial.gob.cl/encuesta-casen-en-pandemia-2020.

characteristics of that class. For example, those people from city A, with an age range 30-35 years old who do not report the income, are classified in the same class as those people from the same city in the same age range that report the income. Third, it is computed the mean of the observed incomes conditionally on a class: the missing incomes are imputed through this mean (Little and Rubin, 2019).

More precisely, let Y be an outcome of interest, and let $\boldsymbol{X}$ be a set of fully observed covariates which are used to define the classes. Let Z be a binary random variable such that $Z = 1$ if the outcome is observed, and $Z = 0$ if not. The conditional mean of both respondents and non-respondents in the same class are given by $\mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}, Z = 1)$ and $\mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}, Z = 0)$, respectively. The conditional mean imputation assumes that, for each $\boldsymbol{x}$,

$$\mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}, Z = 0) = \mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}, Z = 1). \tag{3.1}$$

This assumption is also known as Mean Missing at Random (Manski, 2007), Weak Ignorability (Imbens, 2000; Hirano and Imbens, 2004), and is equivalent to the conditional orthogonality between Y and Z given $\boldsymbol{X}$.

Remark 3.1 Equation (3.1) is equivalent to $\mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}, Z) = \mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$ for all x, which in turn is equivalent to the conditional orthogonality of Y and Z given $\boldsymbol{X}$. In fact, in the Hilbert space $L^2(M, \mathcal{M}, P)$, Y and Z are conditionally orthogonal given $\boldsymbol{X}$ if and only if

$$Y - \mathrm{E}(Y \mid \boldsymbol{X}) \perp Z - \mathrm{E}(Z \mid \boldsymbol{X});$$

that is, if the correlation between both residual is equal to 0. Florens and Mouchart (1982) prove that this last condition is equivalent to $\mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}, Z) = \mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$. It should be remarked that this condition is implied by $Y \perp\!\!\!\perp Z \mid \boldsymbol{X}$.

3.2 Dissecting the CASEN survey

Example Let us focus our attention on the incomes of the salaried employees. According to the technical report Measuring income and poverty in Chile, 2020 Casen Survey in Pandemic[1], 45,642 individuals were considered in this category. These individuals were exposed to the following question:

> The last month, what was your net income at your main job?

The non-response rate was approximately 11.4% (40,418 valid responses); only 5,062 responses were imputed; the remaining responses (namely, 162) were kept as missing. The following covariates were used to define the classes to impute the missing incomes: X_1 = geographic location, X_2 = range age, X_3 = sex, X_4 = educational level, X_5 = category of the occupation, X_6 = class of activity of the company where the individual works, and X_7 = type of occupation into the company[2].

If we consider the original data (that is, the people who reported their income), the average income is equal to 653,891.6 Chilean pesos, while the average income considering the imputed data also was equal to 653,327 Chilean pesos. The quantiles of the income distributions for both data sets are given in Table 2. Considering the original data, it can be seen that the 5% of the surveyed individuals have an income at most equal to 150,000

[1]Retrieved from http://observatorio.ministeriodesarrollosocial.gob.cl on January 11, 2022.
[2]For details on the imputation procedure, see the technical report: Measuring income and poverty in Chile, Casen Survey in Pandemic 2020.

Table 2. Quantiles of the income distribution for both original and imputed incomes

Percentage	Quantile of the original data	Quantile of the imputed data
5%	150,000	160,000
10%	230,000	242,000
25%	320,000	320,000
50%	400,000	420,000
75%	750,000	750,000
90%	1,300,000	1,300,000
95%	1,800,000	1,800,000
99%	3,500,000	3,500,000

Chilean pesos, while the 10% of the salaried surveyed people have an income at most equal to 230,000 Chilean pesos. When the imputed incomes are considered, these values change.

Remark 3.2 Let Y be a real random variable. The quantile function is defined as

$$q_X(\alpha) = \inf\{t \in \mathrm{R} : P(Y \leq t) \geq \alpha\}, \quad \alpha \in [0, 1].$$

This corresponds to the generalized inverse of the cumulative distribution function of Y; see Embrechts and Hofert (2013). The quantiles reported in Table 2, as in other part of this paper, were calculated using this definition (for a code, see Alarcón-Bustamante, 2022), which respects the nature of the data (the income is a discrete random variable), and not using the Hyndman and Fan (1996)'s recommendations which is used, for instance, in R Core Team (2020).

Table 2 shows the impact of the imputation procedure on the quantiles of the income distribution. How relevant is this impact on a global view of income distribution and poverty? Could we say that it is negligible? These questions can be answered by addressing the following one: what can we learn about the income by using the empirical evidence only? The remaining of this section is devoted to answer this question.

What can we learn from the data? It was previously mentioned that the CASEN data set contains information of 185,437 individuals, that is, those individuals who answered the survey in the application step. For this reason, we consider the sample space M as the set of these individuals. Let us define the coordinates of following random vector $(C, S, Z, Y) : M \to \{0, 1\}^3 \times R^+ \cup \{0\}$: for each $m \in M$:

- $C(m) = 1$ if the individual m answers the survey at the application step, and $C(m) = 0$ if not.
- $S(m) = 1$ if the individual is classified as a salaried employee in the application step, and $S(m) = 0$ if not.
- $Z(m) = 1$ if the individual m reports the income, and $Z(m) = 0$ if not.
- Let $Y(m)$ be the income of individual m.

From these definitions it follows that

(i) $\{m \in M\colon S(m) = 1\} \subset \{m \in M\colon C(m) = 1\}$;

(ii) $\{m \in M\colon Z(m) = 1\} \subset \{m \in M\colon S(m) = 1\} \cap \{m \in M\colon C(m) = 1\}$.

From the CASEN survey, the information summarized in Table 3 is available. This shows that the following conditional probabilities are identified:

$$P(S = 1 \mid C = 1) = 0.246; \quad P(Z = 1 \mid S = 1, C = 1) = 0.885544.$$

Furthermore, the conditional distribution of the income $P(Y \leq y \mid Z = 1, C = 1, S = 1)$ is identified, which is depicted in Figure 1. In particular, the average income $\mathrm{E}(Y \mid Z = 1, C = 1, S = 1)$ is identified, and it is equal to 653,891.6 Chilean pesos.

Table 3. Total of individuals by random variable – 2020 CASEN survey

Event	Cardinality
$\{m \in M\colon C(m) = 1\}$	185,437
$\{m \in M\colon S(m) = 1\}$	45,642
$\{m \in M\colon Z(m) = 1\}$	40,418

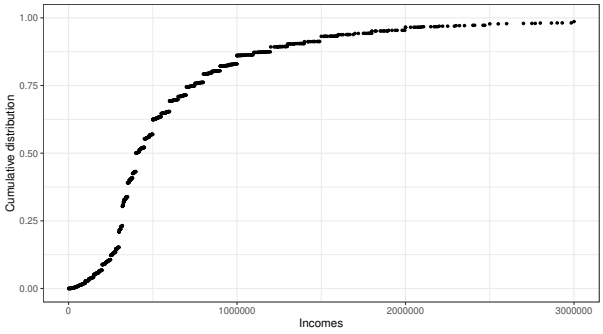


Figure 1. The observed income distribution $P(Y \leq y \mid Z = 1, C = 1, S = 1)$

What we want to learn from the data Analogous to the analysis of the CADEM survey, the results of the CASEN survey can be interpreted conditionally to different subpopulations. This is the content of this section.

First level of analysis The first level corresponds to what we can learn from the data. This level is accordingly captured by the identified parameters above described. Regarding the distribution of the reported incomes, Figure 1 shows that the slope of the curve rapidly increases for lower incomes. As a matter of fact, until 75% of the salaried employees, there are non-dramatic changes in the income, so there is a low variability. In contrast, in the 25% of employees with highest incomes this slope increase slowly, which means that there is a great variability among the incomes.

Second level of analysis: Surveyed salaried employees The second level of analysis is focused on the parameter of interest $P(Y \leq y \mid C = 1, S = 1)$, that is, the income distribution of the salaried employees who answered the survey. The objective of this section is to make explicit the impact of the non-response rate on the income distribution, the average income and the corresponding quantiles. By doing so, it is appreciated how strong is the conditional mean imputation implemented by the CASEN survey.

Income distribution: Let us start by the income distribution. Using the law of total probability, we have that

$$P(Y \leq y \mid C=1, S=1) = P(Y \leq y \mid C=1, S=1, Z=1)P(Z=1 \mid C=1, S=1) + P(Y \leq y \mid C=1, S=1, Z=0)P(Z=0 \mid C=1, S=1).$$

In this decomposition, both $P(Y \leq y \mid C=1, S=1, Z=1)$ and $P(Z=z \mid C=1, S=1)$, $z \in \{0,1\}$, are identified, whereas $P(Y \leq y \mid C=1, S=1, Z=0)$ is not identified because it depends on the employees who did not report their income. Instead of using an ignorability condition (as the conditional mean imputation), the relevant question is what can be learned about $P(Y \leq y \mid C=1, S=1)$ without introducing additional assumptions. Taking into account that $P(Y \leq y \mid C=1, S=1, Z=0) \in [0,1]$, it is possible to bound $P(Y \leq y \mid C=1, S=1)$ as

$$\begin{aligned} &P(Y \leq y \mid C=1, S=1, Z=1)P(Z=1 \mid C=1, S=1) \\ &\leq P(Y \leq y \mid C=1, S=1) \\ &\leq P(Y \leq y \mid C=1, S=1, Z=1)P(Z=1 \mid C=1, S=1) + P(Z=0 \mid C=1, S=1), \end{aligned} \tag{3.2}$$

where $P(Z=1 \mid C=1, S=1) = 0.866$. This identification region, depicted in Figure 2, includes an infinite number of income distributions that are compatible with the observations. Moreover, it reflects the uncertainty induced by the non-response rate: in fact, the width of this interval is equal to the non-response rate, namely $P(Z=0 \mid C=1, S=1)$.

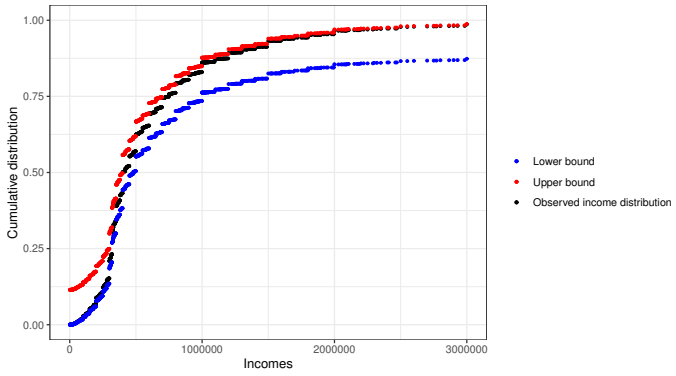


Figure 2. Identification region for $P(Y \leq y \mid C=1, S=1)$

Average income At the second level, the average income corresponds to the conditional expectation $\mathrm{E}(Y \mid C=1, S=1)$, which is decomposed as

$$\mathrm{E}(Y \mid C=1, S=1) = \mathrm{E}(Y \mid C=1, S=1, Z=1)P(Z=1 \mid C=1, S=1) + \mathrm{E}(Y \mid C=1, S=1, Z=0)P(Z=0 \mid C=1, S=1).$$

In this decomposition, $\mathrm{E}(Y \mid C=1, S=1, Z=1)$ and $P(Z=z \mid C=1, S=1)$, for $z \in \{0,1\}$, are identified, whereas $\mathrm{E}(Y \mid C=1, S=1, Z=0)$ is not identified because it depends on the employees who did not report their income. However, this last conditional expectation could be partially identified provided the support of Y is bounded. Although theoretically the support of Y is bounded, in practice the lower bound is known, whereas

the upper bound is finite but unknown: how large is it? 5,000,000 Chilean pesos? 25,000,000 Chilean pesos? There is no way to answer this question and, therefore, there is no way to provide a partial identification region for $\mathrm{E}(Y \mid C = 1, S = 1)$. For additional discussion on partial identifiability of a conditional expectation, see Alarcón-Bustamante et al. (2020).

Quantiles of $P(Y \leq y \mid C = 1, S = 1)$: Although the first moment of the income distribution $P(Y \leq y \mid C = 1, S = 1)$ is not even partially identified, it is possible to learn from the respective quantiles, and to appreciate the impact of the non-response rate on them. The quantiles of the income distribution $P(Y \leq y \mid C = 1, S = 1)$ are given by

$$q_{Y|C=1,S=1}(\alpha) = \inf\{t \in \mathrm{R}\colon P(Y \leq t \mid S = 1, C = 1) \geq \alpha\} \quad \text{for } \alpha \in [0, 1].$$

This quantile function is non identified because it is defined in terms of a non identified probability distribution, namely $P(Y \leq t \mid S = 1, C = 1)$. However, using the bounds in Equation (3.2), it is possible to partially identified the quantile function $q_{Y|C=1,S=1}$ by using the quantiles of the income distribution $P(Y \leq y \mid S = 1, C = 1, Z = 1)$: for $\alpha \in (0, 1)$,

$$\begin{aligned} & q_{Y|C=1,S=1,Z=1}\left(\frac{\alpha - P(Z = 0 \mid C = 1, S = 1)}{P(Z = 1 \mid C = 1, S = 1)}\right) \leq \\ & \leq q_{Y|C=1,S=1}(\alpha) \leq \\ & \leq q_{Y|C=1,S=1,Z=1}\left(\frac{\alpha}{P(Z = 1 \mid C = 1, S = 1)}\right). \end{aligned} \tag{3.3}$$

For a proof, details and reference, see San Martín and González (2022, Section 4).

The identification region given in Equation (3.3) shows the impact of the non-response rate on the quantile function of $P(Y \leq y \mid C = 1, S = 1)$ in the sense that one of the bounds of the quantile function is non-informative for some values of α. As a matter of fact,

- If $\alpha \leq P(Z = 0 \mid C = 1, S = 1)$, then the lower bound in Equation (3.3) is equal to the minimum of the support of the conditional distribution $P(Y \leq y \mid C = 1, S = 1, Z = 1)$ and, therefore, it is non-informative.
- If $\alpha \geq P(Z = 1 \mid C = 1, S = 1)$, then the upper bound in Equation (3.3) is equal to the maximum of the support of the conditional distribution $P(Y \leq y \mid C = 1, S = 1, Z = 1)$ and, therefore, it is non-informative.

Therefore, the quantile function of $P(Y \leq y \mid C = 1, S = 1)$ is informative (that is, provides values in the interior of the support of $P(Y \leq y \mid S = 1, C = 1, Z = 1)$) in the following two cases:

(i) If $P(Z = 0 \mid S = 1, C = 1) < P(Z = 1 \mid S = 1, C = 1)$ or, equivalently, the non-response rate among the employees individuals is smaller than 50%, then the quantile function $q_{Y|C=1,S=1}$ is informative for all

$$\alpha \in [\, P(Z = 0 \mid S = 1, C = 1),\ P(Z = 1 \mid S = 1, C = 1)\,].$$

(ii) If $P(Z = 0 \mid S = 1, C = 1) > P(Z = 1 \mid S = 1, C = 1)$ or, equivalently, the non-response rate among the employees individuals is greater than 50%, then the quantile function $q_{Y|C=1,S=1}$ is informative for all

$$\alpha \in [0, P(Z = 1 \mid C = 1, S = 1)\,] \cup [\, P(Z = 0 \mid C = 1, S = 1),\ 1].$$

Let us illustrate this result with the data of the Example. In this case, $P(Z = 0 \mid C = 1, S = 1) = 0.114456$; the corresponding identification regions of the quantile $q_{Y|C=1,S=1}(\alpha)$ for some values of α are summarized in Table 4. We also summarize the quantiles of the income distribution with imputations, thereafter called CASEN income distribution and denoted as $\widetilde{q}_{Y|C=1,S=1}(\alpha)$. It should be noted that the CASEN income distribution almost overlapped with the distribution of observed incomes. Furthermore, the CASEN income distribution is in the interior of the identification region in Equation (3.2), as theoretically expected; see Figure 3.

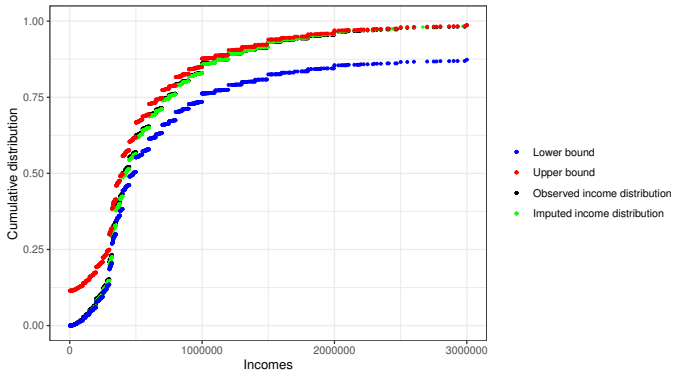


Figure 3. Identification region for $P(Y \leq y \mid C = 1, S = 1)$ and CASEN income distribution

Table 4 deserves the following comments:

(i) For α smaller than the non-response rate, the income of employees can be much lower than the income that can be deduced from the CASEN income distribution. In other words, the non-response rate has such an impact that it is not possible to know how poor the "poorest of the income of employees" are.

(ii) For α greater that the response rate, the income of employees can be much higher than the income that can be deduced from the CASEN income distribution. In other words, the response rate has such an impact that it is not possible to know how rich the "richer of the income of employees" are.

(iii) It can be remarked that for (some) $\alpha_1 \leq \alpha_2$, the identification region of $q_{Y|C=1,S=1}(\alpha_1)$ at least intersects the identification region of $q_{Y|C=1,S=1}(\alpha_2)$. This clearly increases the uncertainty of the conclusions that can be drawn using the partially identified income distribution and which is rendered invisible when using the CASEN income distribution.

The previous conclusions allow us to understand the meaning of ignorability conditions, such as the conditional mean imputation technique or, more generally, Missing at Random conditions. These conditions come from the identification restriction

$$Y \perp\!\!\!\perp Z \mid C = 1, S = 1, \boldsymbol{X}$$

which, by definition of conditional independence, is equivalent to

$$\begin{aligned} P(Y \leq y \mid C = 1, S = 1, \boldsymbol{X}) &= P(Y \leq y \mid C = 1, S = 1, Z = 1, \boldsymbol{X}) \\ &= P(Y \leq y \mid C = 1, S = 1, Z = 0, \boldsymbol{X}). \end{aligned}$$

These equalities means that the missing observations do not provide more relevant information about the output Y, being the only "statistical job" to carefully estimate $P(Y \leq y \mid C = 1, S = 1, Z = 1, \boldsymbol{X})$ –this is the standard procedure.

Table 4. Quantiles of both the partial identified income distribution and the CASEN income distribution

	$q_{Y\|C=1,S=1}(\alpha)$		
α	LB	UB	$\widetilde{q}_{Y\|C=1,S=1}(\alpha)$
0.05	1,200	170,000	160,000
0.10	1,200	250,000	242,000
$P(Z=0 \mid C=1,S=1)$	1,200	265,000	250,000
0.25	300,000	320,000	320,000
0.50	400,000	480,000	420,000
0.75	700,000	1,000,000	750,000
0.80	800,000	1,300,000	865,172
$P(Z=1 \mid C=1,S=1)$	1,100,000	25,000,000	1,200,000
0.90	1,200,000	25,000,000	1,300,000
0.95	1,800,000	25,000,000	1,800,000
0.99	3,500,000	25,000,000	3,500,000
1.00	25,000,000	25,000,000	25,000,000

3.3 Discussion

One of the objectives of the CASEN survey is to obtain an overview of the income distribution of employees and, in particular, to have a look at the incomes of the lowest paid employees as well as those of the highest paid. For this purpose, the self-reported income of survey respondents who fall into the category of salaried employees is used. However, individuals who are exposed to the survey are not required to report their income. As a consequence, the survey includes a non-response rate which, for the 2020 CASEN survey in pandemic, is equal to 11.4456%. Before providing an overview of the distribution of incomes, CASEN applies statistical techniques designed to impute missing income, specifically the conditional mean imputation technique.

Our dissection of the CASEN survey aims to make explicit the policy meaning of this imputation technique. To this end, a partial identification analysis was developed to show the impact of the non-response rate on both the mean of the distribution of the income distribution of employees and on the respective quantiles. One of the main conclusions is that "the poor may be poorer" than what can be asserted from the CASEN income distribution, and that "the rich may be richer" than what can be stated from it.

With this conclusion in mind, it is possible to assess the sense of the imputation technique used by CASEN: the conditional mean imputation technique corresponds to an assumption of income homogeneity. As a matter of fact, it is assumed that, among individuals with characteristics $\boldsymbol{X} = \boldsymbol{x}$ who did not report their income, there is no relevant income information that was not accessed: all the effectively relevant information has already been observed in those who did report their income. Consequently, the income of an employee who did not report it should be related to the average income of all employees sharing the same characteristics $\boldsymbol{X} = \boldsymbol{x}$. The partial identification shows how heterogenous could be the income distributions of employees. Policy decisions should be aware on this uncertainties.

4. The Araucanía citizen consultation

The Araucanía citizen consultation is of special political interest given the ongoing violent conflicts in the region. This is the main motivation for having chosen to analyze it. But there is also a relevant methodological aspect: the information provided by the consultation can be related to the national referendum held in 2020. We study how plausible this relationship is, and how it affects the conclusions that can be drawn.

4.1 Historical and economical context

The capital of the Araucanía region, Temuco, is located 620 kilometers to south of Santiago, the capital of Chile. The Araucanía Region is known for being the original area of the Mapuche People (in the 16th century called "Araucanos"), possibly the only indigenous people with whom the Crown of Spain made a Capitulation of Peace, known as the Paces de Quilín, made on January 5 and 6, 1614. This treaty established the Biobío River as the border, south of which "the Mapuches lived independently for two hundred and forty years, until 1881" (Bengoa, 2007). In 1881, "Manuel Recabarren, Minister of the Interior [at the time], led Chilean troops to the south and, together with General Gregorio Urrutia, advanced hundreds of kilometers along the border and militarily occupied the area" (Bengoa, 2016). This completed the occupation of Araucanía by the Chilean government.

The Araucanía Region, in addition to the Biobío, Los Ríos and Maule regions, develop the country's forestry industry: "the forestry sector represents 1.9% of the domestic GDP, reaching in 2017 USD 5,196 million (3,373 billion of Chilean pesos). Biobío region represents 60.0% of the forestry GDP, followed by La Araucanía region with 10.5%, and Los Ríos, and Maule regions with 10.1% each. Regarding the participation of the three forestry subsectors included in the sectorial GDP, the paper, and pulp industry, as well as products derived from paper represents 44.3%, forestry participates with 29.4%, and the wood products, and wood industry represent 26.3%" (Instituto Forestal, 2021).

Many of the conflicts in the area are due to the presence of forestry companies, whose worldview on nature and its resources is not entirely shared by the Mapuche people's worldview. In addition, part of the forestry exploitation takes place on what were once Mapuche lands, which has triggered a series of territorial claims (Andrade, 2019).

4.2 Recent political context

On October 12, 2021, the President of the Chilean Republic declared a state of emergency for the provinces of Biobío and Arauco, in the Biobío region, and in the provinces of Cautín and Malleco, in the Araucanía Region, for a 15 days period (Diario Oficial de la República de Chile, 2021). According to the Chilean Constitution, this is one of its prerogatives, and it may declare such state of emergency for no more than 15 days. Once a state of emergency is declared, the respective zones are under the immediate dependence of the Chief of National Defense appointed by the President of the Republic, who assume the direction and supervision of his jurisdiction with the powers and obligations established by law (Constitución de la República de Chile, 2005, Art.42). By declaring a state of emergency, the President of the Republic may restrict the freedom of locomotion and assembly (Constitución de la República de Chile, 2005, Art.43).

Among the reasons that led to this decision, the Diario Oficial de la República de Chile (2021) mentions the following ones:

(i) An increase of violence acts linked to drug trafficking, terrorism and organized crime, committed by armed groups that have not only made attempts on the lives of members of the Law Enforcement and Security Forces, but have also attacked people and destroyed facilities and machinery used in industrial, agricultural and commercial activities.
(ii) Since 2018, there has been an increase in crimes and offenses against persons and against property; against public order, including attacks against authority, attacks and threats against prosecutors of the Public Prosecutor's Office and the Judiciary.
(iii) There has been a 116% increase in reported incidents related to crimes contemplated in Law No. 17,798 on Arms Control, including the seizure of weapons and ammunition.

(iv) The number, magnitude and seriousness of the crimes and facts indicated, committed in the provinces of the regions of Biobío and Araucanía, imply a serious alteration of public order –understood as the "situation that allows the peaceful exercise of rights and the fulfillment of obligations, ensuring peaceful coexistence"– in the terms established in Article 42 of the Constitution of the Chilean Republic, which allows the enactment of the state of emergency constitutional exception with respect to such areas of the national territory, provided for in said article.

As it was mentioned above, the state of emergency may not be extended for more than fifteen days, notwithstanding that the President of the Republic may extend it for the same period. However, for successive extensions, the President always requires the consent of the National Congress, specifically the Senate (Constitución de la República de Chile, 2005, Art.42). Until January 2022, the National Congress has approved the extension of the state of emergency for 6 consecutive times[1]. It should be mentioned that the official account of the recent conflicts in La Araucanía does not relate these conflicts to the territorial claims of the Mapuche people.

4.3 Organization of the Araucanía consultation and results

To know the opinion of the citizens of the 32 communes of La Araucanía regarding the renewal of the state of emergency in the region, the Regional Intendancy and the Association of Municipalities of La Araucanía organized a citizen consultation, which took place on November 5, 6 and 7, 2021. The consultation was carried out electronically, and all persons over 18 years old registered in the electoral registry in any of the 32 municipalities may participate from a computer, cell phone or another device connected to the internet[2].

The citizen consultation was limited to the following question:

Do you agree with Congress extending the state of emergency in the Araucanía Region?

The results of the consultation are summarized in Table 5.

Table 5. Results of the Araucanía consultation

Option	Votes	% wrt the consultation	% wrt electoral roll
Yes	118,258	81.56	13.34
No	26,655	18.38	3.01
Blank votes	54	0.04	0.01
Null votes	27	0.02	0.00
Total	144,994	100	16.36

where "wrt" denotes "with respect to".

4.4 How these results were used?

Subsections 4.1 and 4.2 attempt to illustrate the complexity of the political situation in the Araucanía region. This complex context may explain why successive extensions of the state of emergency have been subject to lively debate. In fact, those extensions did not achieve unanimity in the Senate: they were approved not more than 2 or 3 votes in favor. Let us mention the Senate session of November 24, 2021, where the extension of the state

[1]For details, see https://www.senado.cl/senado/site/cache/search/pags/search164185188127928.html on January 10, 2021.
[2]Retrieved from https://www.consultaaraucania.cl/ on January 10, 2022.

of emergency was approved by 16 votes in favor, 14 against, and one abstention. Among the reasons that were mentioned for approving the extension, the Araucanía consultation was explicitly mentioned as an important factor. This was stated by Senator Francisco Chahuán, from the right coalition Chile Vamos, who affirmed that "the state of exception has generated greater tranquility. Attacks against property and arson crimes have decreased. We must listen actively and in La Araucanía there was a citizen consultation that supported this measure"[3]. These expressions are in line with the assessment made by the Governor of La Araucanía, Luciano Rivas, independent, near to the Chile Vamos coalition, at the end of the consultation: "With great respect, but also with great strength, we ask politicians, especially all the deputies and senators of Chile, that our voice be heard, do not turn a deaf ear"[4].

As mentioned by Governor Rivas[1], the Araucanía citizen consultation was one of the first, if not the first, non-binding consultations to be held in Chile. This, added to the complex political situation in the Araucanía region, could explain the interest that this consultation aroused, especially in the relationship that its results have with recent elections, namely the 2020 national referendum on the possibility of a new constitution and the 2021 governor elections. One of these studies is the one conducted by Cayul et al. (2021), which was initially published in the electronic journal CIPER[2]. This study analyzes the representativeness at the municipality level of the Araucanía Consultation on three axes: Mapuche population, rurality, and population that voted for the non-approval option in the 2020 national referendum. According to the authors, "these axes are fundamental to establish whether or not there is a bias in the results, since it analyzes the cultural, socioeconomic and political dimension". To achieve this objective, the authors analyze, on the one hand, the participation in the second round of the election of Regional Governors in Araucanía with the percentage of Mapuche population, the percentage of rural population and the percentage of non-approval in the 2020 referendum; and, on the other hand, the participation in the citizen consultation in Araucanía with the same percentages already mentioned. The choice of the regional governors is due to the fact that in that election "a similar universe of approximately 125,000 people participated". We are able to reproduce the third analysis by considering the data summarized at Table 6.

Figures 4 (a)-(b) reproduce their analysis. Cayul et al. (2021) conclude that "those municipalities with a higher percentage of votes for the non-approval to a new Constitution also had a higher participation in both the citizen consultation and in the second round of governors' elections, but the effect is significantly lower in the latter. That is, there would be a political bias of those who participate in the consultation".

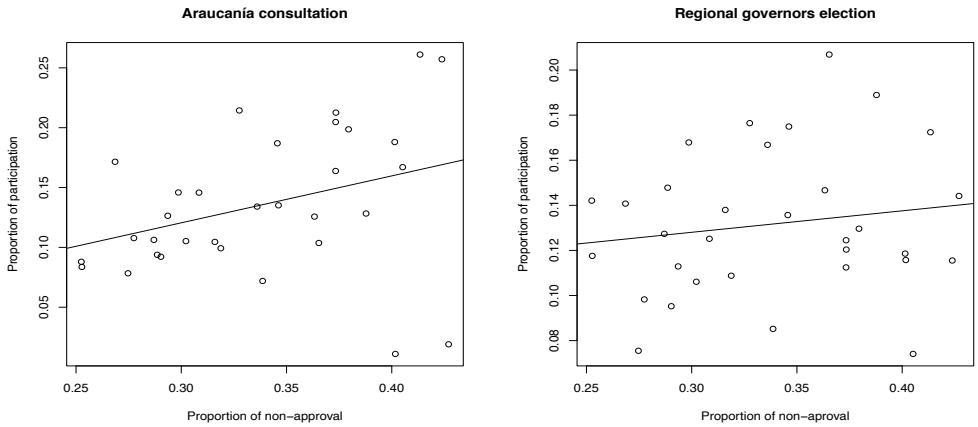
The final conclusion of the study is the following:

> We observe then that, when comparing two elections with a similar participation rate, the people who participate in them are very different. While participation in the consultation was higher in urban, non-Mapuche municipalities that voted for non-approval, these same biases are not observed in the second round of governors election.
>
> Electors, then, are not representative at the municipal level, and this suggests that the consultation is not necessarily representative of the population of Araucanía. This implies that the interpretation of the results should be done with caution, and without extrapolating conclusions for the entire region, especially given the relevance that has been sought to give to the consultation.

[3]Retrieved from https://www.senado.cl/estado-de-excepcion-constitucional on January 10, 2021.
[4]Retrieved from https://assets.eldesconcierto.cl/2021/11/Copia-de-Copia-de-Discurso-Consulta-Araucani%CC%81a.pdf on January 11, 2022.
[1]See his speech of November 7, 2021 in https://assets.eldesconcierto.cl/2021/11/Copia-de-Copia-de-Discurso-Consulta-Araucani%CC%81a.pdf.
[2]At https://www.ciperchile.cl/2021/11/10/consulta-ciudadana-en-la-araucania.

Table 6. 2020-2022 elections in the Araucanía Region by municipality

		2021 electoral roll	2020 referendum for a new constitution				2021 regional governors election				2022 Araucanía consultation
c	Municipality	Total	Approve	Non-approve	Null votes	Blank votes	Tuma	Rivas	Null Votes	Blank votes	Total of votes
1	Angol	46,976	12,017	7,349	97	57	2,050	3,924	85	30	9,331
2	Carahue	24,663	4,978	2,220	39	25	1,552	1,481	45	9	3,595
3	Cholchol	11,149	2,951	1,197	51	19	815	804	24	5	1,046
4	Collipulli	22,172	5,159	3,075	65	41	1,120	1,500	32	17	4,714
5	Cunco	20,105	4,090	1,914	47	31	1,041	1,108	24	14	1,996
6	Curacautín	21,097	3,491	2,378	33	20	599	911	39	14	3,524
7	Curarrehue	8,521	1,740	891	12	12	337	371	11	7	613
8	Ercilla	8,036	1,521	1,020	26	26	433	508	8	4	1,511
9	Freire	22,892	5,198	2,401	51	20	1,578	1,535	31	14	2,394
10	Galvaino	12,092	2,472	995	46	24	877	629	24	10	1,285
11	Gorbea	15,651	3,290	2,084	25	20	1,170	1,743	34	10	2,008
12	Lautaro	34,520	8,037	4,788	68	22	1,615	2,186	66	15	5,656
13	Loncoche	23,234	6,125	2,072	27	13	1,411	1,280	32	9	1,944
14	Lonquimay	12,115	1,963	1,318	42	26	852	523	11	16	133
15	Los Sauces	7,464	1,409	1,050	24	12	346	712	12	6	142
16	Lumaco	9,187	1,711	1,206	31	31	478	1,080	18	8	2,399
17	Melipeuco	7,723	1,542	584	15	18	335	236	10	2	605
18	Nueva Imperia	30,850	8,765	3,218	83	34	2,407	1,849	64	22	5,292
19	Padres las Casas	59,516	16,892	7,020	142	67	3,063	3,522	107	28	7,523
20	Perquenco	6,814	1,778	757	11	7	721	400	14	9	994
21	Pitrufquén	23,226	5,566	2,947	44	10	1,529	2,494	28	12	3,139
22	Pucón	29,594	8,176	3,140	36	23	1,054	1,776	59	19	3,189
23	Purén	12,240	2,657	1,583	30	19	648	824	38	14	2,506
24	Renaico	9,806	2,640	1,080	17	8	430	477	20	7	904
25	Saavedra	13,239	2,960	1,000	34	27	831	983	42	25	1,165
26	Temuco	238,028	78,332	38,159	360	169	15,375	25,684	740	200	51,039
27	Teodoro Scmidt	14,052	3,026	1,742	28	23	1,304	1,559	32	12	1,457
28	Toltén	10,554	2,126	1,213	30	29	526	1,001	15	6	1,327
29	Traiguén	18,595	4,144	2,189	34	35	1,012	1,445	48	18	3,477
30	Victoria	32,607	7,060	5,193	76	44	1,418	2,264	60	25	8,385
31	Vilcún	24,718	5,971	3,022	70	56	1,614	2,455	41	14	3,315
32	Villarrica	56,170	15,087	6,534	89	40	2,301	3,537	95	27	5,911



(a) (b)

Figure 4. Relationship between 2020 referendum and 2021 (a); and between 2021 governor election and 2021 (b) Araucanía consultation participation

4.5 Dissecting the use of Araucanía citizen consultation

Statement of the problem The previous analysis consists in comparing two or more elections that share a common electoral roll. Now, for each election, there are specific probability distributions that are identified, namely (i) the distributions of participation and non-participation, and (ii) the distribution of preferences conditionally on the electors participating in the election. More precisely, let M be the sample space composed of the electors, and define the following random variables on M:

- $V_1(m) = 1$ if the elector m participated at the 2020 referendum, and $V_1(m) =$ if not.
- $V_2(m) = 1$ if the elector m participated at the 2021 governor election, and $V_2(m) =$ if not.
- $V_3(m) = 1$ if the elector m participated at the 2021 citizen consultation, and $V_3(m) =$ if not.
- Let Y_1 be the preference at the 2020 referendum, namely $\mathcal{Y}_1 = \{\text{approve}, \text{non-approve}, \text{blank vote}, \text{null vote}\}$.
- Let Y_2 be the preference at the 2021 governor election, namely $\mathcal{Y}_2 = \{\text{Tuma}, \text{Rivas}, \text{blank vote}, \text{null vote}\}$.
- Let Y_3 be the preference at the 2021 citizen consultation, namely $\mathcal{Y}_3 = \{\text{yes}, \text{no}, \text{blank vote}, \text{null vote}\}$.
- Let C be the municipality in which each elector is registered. C takes 32 different values because there are 32 municipalities; see Table 6.

If we consider each election separately, then the identified parameters are the following:

$$P(Y_i = y_i \mid V_i = 1, C = c) \quad (y_i, c) \in \mathcal{Y}_i \times \{1, \ldots, 32\}; \quad P(V_i = 1 \mid C = c); \quad c \in \{1, \ldots, 32\},$$

for $i \in \{1, 2, 3\}$.

If these elections are jointly used, it should be verified if the set of electors is the same, that is, if the following equality holds:

$$\begin{aligned} &\{m \in M \colon V_1(m) = 1\} \cup \{m \in M \colon V_1(m) = 0\} \\ &= \{m \in M \colon V_2(m) = 1\} \cup \{m \in M \colon V_2(m) = 0\} \\ &= \{m \in M \colon V_3(m) = 1\} \cup \{m \in M \colon V_3(m) = 0\}. \end{aligned}$$

Certainly, the Chilean Electoral Service (SERVEL in Spanish) has access to this information and it can verify such equality. In what follows, we assume that this equality is fulfilled.

The analysis described in Subsection 4.4 consists in comparing

$$\{P(Y_1 = \text{non-aproval} \mid V_1 = 1, C = c)\colon c \in \{1, \ldots, 32\}\},$$

with

$$\{P(V_3 = 1 \mid C = c)\colon c \in \{1, \ldots, 32\}\}.$$

However, for each $c \in \{1, \ldots, 32\}$, $\{m \in M\colon V_1(m) = 1, C(m) = v\}$ is not necessarily equal to $\{m \in M\colon V_3(m) = 1, C(m) = c\}$.

A fair comparison needs to use the same electors, which in turn lead to consider

$$\{P(Y_1 = \text{non-aproval} \mid V_1 = 1, V_3 = 1, C = c)\colon c \in \{1, \ldots, 32\}\}$$

and

$$\{P(V_3 = 1 \mid V_1 = 1, C = c)\colon c \in \{1, \ldots, 32\}\}.$$

This is due to the fact that the political behavior of those who participate in both elections is not necessarily the same as the political behavior of those who participate in one, or the other, or both. It is, therefore, necessary to identify $P(Y_1 = \text{non-aproval} \mid V_1 = 1, V_3 = 1, C = c)$ and $P(V_3 = 1 \mid V_1 = 1, C = c)$ for each c.

Partial identification of $P(V_1 = v_1, V_3 = v_3 \mid C = c)$: Both $P(Y_1 = \text{non-aproval} \mid V_1 = 1, V_3 = 1, C = c)$ and $P(V_3 = 1 \mid V_1 = 1, C = c)$ require the identifiability of $P(V_1 = v_1, V_3 = v_3 \mid C = c)$ for $(v_1, v_3) \in \{0, 1\}^2$. Taking into account that $P(V_1 = v_1 \mid C = c)$ and $P(V_3 = v_3 \mid C = c)$ are identified, the way to relate them to the joint distribution $P(V_1 = v_1, V_3 = v_3 \mid C = c)$ is through the Fréchet inequality (Fréchet, 1960a,b), namely for each $c \in \{1, \ldots, 32\}$, by means of

$$\begin{aligned}
&\max\{1, P(V_1 = v_1 \mid C = c) + P(V_3 = v_3 \mid C = c) - 1\} \leq \\
&\leq P(V_1 = v_1, V_3 = v_3 \mid C = c) \leq \\
&\leq \min\{P(V_1 = v_1 \mid C = c), P(V_3 = v_3 \mid C = c)\}, \quad \forall\, (v_1, v_3) \in \{0, 1\}^2.
\end{aligned}$$

It should be emphasized that these bounds are the best ones; see the constructive proof in Fréchet (1960a). Thus, for $(v_1, v_3) = (1, 1)$, it follows that

$$\begin{aligned}
&\max\{0, P(V_1 = 1 \mid C = c) - P(V_3 = 0 \mid C = c)\} \leq \\
&\leq P(V_1 = 1, V_3 = 1 \mid C = c) \leq \min\{P(V_1 = 1 \mid C = c), P(V_3 = 1 \mid C = c)\}.
\end{aligned} \tag{4.1}$$

For municipality c the lower bound is informative (that is, greater than 0) if $P(V_1 = 1 \mid C = c) > P(V_3 = 0 \mid C = c)$, that is, if the rate of participation at the 2020 referendum is greater than the rate of non-participation at the 2021 citizen consultation; or, equivalently, if the rate of non-participation at the 2020 referendum is smaller than the rate of participation at the 2021 citizen participation. Table 7 summarizes the results, where LB_{13} is the lower bound of Equation (4.1) and UB_{13} is the corresponding upper bound. It can be seen that, for each municipality, the lower bound is always 0, which means that a plausible assumption

is that none of those who participated in one election participated in the other. Another plausible assumption is that

$$P(V_1 = 1, V_3 = 1 \mid C = c) = P(V_3 = 1 \mid C = c), \tag{4.2}$$

that is, the rate of joint participation is equal to the rate of participation at the 2021 citizen consultation. In this case, $P(V_1 = 0, V_3 = 1 \mid C = c) = 0$, that is, no elector did not participate at the 2020 referendum and participated at the 2021 citizen consultation. Certainly this conclusion may seem implausible, which in turn would imply that Equation (4.2) is implausible as an assumption.

Table 7. Bounds of joint participation ratios

c	Municipality	$P(V_1 = 1 \mid C = c)$	$P(V_2 = 1 \mid C = c)$	$P(V_3 = 1 \mid C = c)$	LB_{13}	UB_{13}	LB_{23}	UB_{23}
1	Angol	0.42	0.13	0.20	0.00	0.20	0.00	0.13
2	Carahue	0.29	0.13	0.15	0.00	0.15	0.00	0.13
3	Cholchol	0.38	0.15	0.09	0.00	0.09	0.00	0.09
4	Collipulli	0.38	0.12	0.21	0.00	0.21	0.00	0.12
5	Cunco	0.30	0.11	0.10	0.00	0.10	0.00	0.10
6	Curacautín	0.28	0.07	0.17	0.00	0.17	0.00	0.07
7	Curarrehue	0.31	0.09	0.07	0.00	0.07	0.00	0.07
8	Ercilla	0.32	0.12	0.19	0.00	0.19	0.00	0.12
9	Freire	0.34	0.14	0.10	0.00	0.10	0.00	0.10
10	Galvaino	0.29	0.13	0.11	0.00	0.11	0.00	0.11
11	Gorbea	0.35	0.19	0.13	0.00	0.13	0.00	0.13
12	Lautaro	0.37	0.11	0.16	0.00	0.16	0.00	0.11
13	Loncoche	0.35	0.12	0.08	0.00	0.08	0.00	0.08
14	Lonquimay	0.28	0.12	0.01	0.00	0.01	0.00	0.01
15	Los Sauces	0.33	0.14	0.02	0.00	0.02	0.00	0.02
16	Lumaco	0.32	0.17	0.26	0.00	0.26	0.00	0.17
17	Melipeuco	0.28	0.08	0.08	0.00	0.08	0.00	0.08
18	Nueva Imperia	0.39	0.14	0.17	0.00	0.17	0.00	0.14
19	Padres las Casas	0.41	0.11	0.13	0.00	0.13	0.00	0.11
20	Perquenco	0.37	0.17	0.15	0.00	0.15	0.00	0.15
21	Pitrufquén	0.37	0.17	0.14	0.00	0.14	0.00	0.14
22	Pucón	0.38	0.10	0.11	0.00	0.11	0.00	0.10
23	Purén	0.35	0.12	0.20	0.00	0.20	0.00	0.12
24	Renaico	0.38	0.10	0.09	0.00	0.09	0.00	0.09
25	Saavedra	0.30	0.14	0.09	0.00	0.09	0.00	0.09
26	Temuco	0.49	0.18	0.21	0.00	0.21	0.00	0.18
27	Teodoro Scmidt	0.34	0.21	0.10	0.00	0.10	0.00	0.10
28	Toltén	0.32	0.15	0.13	0.00	0.13	0.00	0.13
29	Traiguén	0.34	0.14	0.19	0.00	0.19	0.00	0.14
30	Victoria	0.38	0.12	0.26	0.00	0.26	0.00	0.12
31	Vilcún	0.37	0.17	0.13	0.00	0.13	0.00	0.13
32	Villarrica	0.39	0.11	0.11	0.00	0.11	0.00	0.11

Partial identification of $P(V_3 = 1 \mid V_1 = 1, C = c)$: From Equation (4.1) it can be deduced the identification region for $P(V_3 = 1 \mid V_1 = 1, C = c)$, namely

$$\max\left\{0, \frac{P(V_1 = 1 \mid C = c) - P(V_3 = 0 \mid C = c)}{P(V_1 = 1 \mid C = c)}\right\} \leq \tag{4.3}$$
$$\leq P(V_3 = 1 \mid V_1 = 1, C = c) \leq \min\left\{1, \frac{P(V_3 = 1 \mid C = c)}{P(V_1 = 1 \mid C = c)}\right\}.$$

For each municipality c, the lower bound is informative if $P(V_1 = 1 \mid C = c) > P(V_3 = 0 \mid C = c)$, whereas the upper bound is informative (that is, smaller than 1) if $P(V_3 = 1 \mid C = c) < P(V_1 = 1 \mid C = c)$, that is, if the rate of participation at the citizen consultation is smaller than the rate of participation at the 2020 referendum. Table 8 shows the corresponding lower and upper bound. It can be seen that the lower bound is uninformative, whereas the upper bound is informative: it corresponds to the ratio of participation at the citizen consultation given that electors participated at the 2020 referendum.

Table 8. Partial identification of $P(V_3 = 1 \mid V_1 = 1, C = c)$ and $P(Y_1 = \text{non-approve} \mid V_1 = 1, V_3 = 1, C = c)$

		$P(V_3 = 1 \mid V_1 = 1, C = c)$		$P(Y_1 = \text{non-approve} \mid V_1 = 1, V_3 = 1, C = c)$	
c	Municipality	LB	UB	LB	UB
1	Angol	0.00	0.48	0.00	0.72
2	Carahue	0.00	0.50	0.00	0.61
3	Cholchol	0.00	0.25	0.05	0.38
4	Collipulli	0.00	0.57	0.00	0.85
5	Cunco	0.00	0.33	0.00	0.47
6	Curacautín	0.00	0.60	0.00	0.99
7	Curarrehue	0.00	0.23	0.14	0.44
8	Ercilla	0.00	0.58	0.00	0.94
9	Freire	0.00	0.31	0.00	0.46
10	Galvaino	0.00	0.36	0.00	0.44
11	Gorbea	0.00	0.37	0.02	0.61
12	Lautaro	0.00	0.44	0.00	0.66
13	Loncoche	0.00	0.24	0.02	0.33
14	Lonquimay	0.00	0.04	0.37	0.41
15	Los Sauces	0.00	0.06	0.39	0.45
16	Lumaco	0.00	0.81	0.00	1.00
17	Melipeuco	0.00	0.28	0.00	0.38
18	Nueva Imperia	0.00	0.44	0.00	0.47
19	Padres las Casas	0.00	0.31	0.00	0.42
20	Perquenco	0.00	0.39	0.00	0.49
21	Pitrufquén	0.00	0.37	0.00	0.54
22	Pucón	0.00	0.28	0.00	0.38
23	Purén	0.00	0.58	0.00	0.89
24	Renaico	0.00	0.24	0.06	0.38
25	Saavedra	0.00	0.29	0.00	0.35
26	Temuco	0.00	0.44	0.00	0.58
27	Teodoro Scmidt	0.00	0.30	0.08	0.52
28	Toltén	0.00	0.39	0.00	0.59
29	Traiguén	0.00	0.54	0.00	0.75
30	Victoria	0.00	0.68	0.00	1.00
31	Vilcún	0.00	0.36	0.00	0.52
32	Villarrica	0.00	0.27	0.04	0.41

Partial identification of $P(V_2 = 1, V_3 = 1 \mid C = c)$: Following the arguments developed in Subsection 4.5, it follows that

$$\begin{aligned} &\max\{0, P(V_2 = 1 \mid C = c) - P(V_3 = 0 \mid C = c)\} \leq \\ &\leq P(V_2 = 1, V_3 = 1 \mid C = c) \leq \min\{P(V_2 = 1 \mid C = c), P(V_3 = v_3 \mid C = c)\}. \end{aligned} \tag{4.4}$$

Table 7 shows the corresponding lower and upper bounds. Lower bounds are always uninformative because, for each municipality, the rate of participation at the 2021 governor election is smaller than the rate of non-participation at the 2021 citizen consultation. This means that, although the overall participation rates in both elections are very similar (16% for the citizen consultation, 14% for the governor election), a plausible assumption is that there are no electors who participated in both elections. In addition, sometimes the upper bound is equal to $P(V_2 = 1 \mid C = c)$, sometimes to $P(V_3 = 1 \mid C = c)$. In the first case, namely when it is assumed that $P(V_2 = 1, V_3 = 1 \mid C = c) = P(V_2 = 1 \mid C = c)$, then there are no electors who participated in the 2021 governors election and did not participate in the 2021 citizen consultation. In the second case, namely $P(V_2 = 1, V_3 = 1 \mid C = c) = P(V_3 = 1 \mid C = c)$, then there are no electors who did not participate in the 2021 governors election and who participated in the 2021 citizen consultation. Again, it can be stated that these assumptions may not seem entirely plausible, which in turn shows that it is possible to have turnout rates in both elections lower than the upper bound. By passing, this jeopardizes the argument according to which the governor election and the citizen consultation can be compared because their rate of participation are similar.

Partial identification of $P(Y_1 = y \mid V_1 = 1, V_2 = 1, C = c)$: For each municipality c, the conditional probability $P(Y_1 = y \mid V_1 = 1, V_2 = 1, C = c)$ can not vary arbitrarily because it

is related to the identified conditional probability $P(Y_1 = y \mid V_1 = 1, C = c)$ through the following decomposition:

$$P(Y_1 = y \mid V_1 = 1, C = c) = P(Y_1 = y \mid V_1 = 1, V_3 = 1, C = c)P(V_3 = 1 \mid V_1 = 1, C = c) + P(Y_1 = y \mid V_1 = 1, V_3 = 0, C = c)P(V_3 = 0 \mid V_1 = 1, C = c).$$

In this decomposition, $\gamma_c \doteq P(Y_1 = y \mid V_1 = 1, V_3 = 0, C = c)$ is non identified, whereas $P(V_3 = 1 \mid V_1 = 1, C = c)$ and $P(V_3 = 0 \mid V_1 = 1, C = c)$ are partially identified by Equation (4.4).

Let $C \in \{1, \ldots, 32\}$ and $p_c \doteq P(V_3 = 0 \mid V_1 = 1, C = c)$ be fixed. It follows that $P(Y_1 = y \mid V_1 = 1, V_3 = 1, C = c)$ belongs to the set

$$\left\{ \frac{P(Y_1 = y \mid V_1 = 1, C = c) - \gamma_c p_c}{1 - p_c} \colon \gamma_c \in [0, 1] \right\},$$

which reduces to the interval

$$A_{p_c} \doteq \left[\frac{P(Y_1 = y \mid V_1 = 1, C = c) - p_c}{1 - p_c}, \ \frac{P(Y_1 = y \mid V_1 = 1, C = c)}{1 - p_c} \right].$$

Now, if $p_{1,c} < p_{2,c}$, then

$$A_{p_{1,c}} \subset A_{p_{2,c}}.$$

Therefore, for each $c \in \{1, \ldots, 32\}$, we have that

$$P(Y_1 = y \mid V_1 = 1, V_3 = 1, C = c) \in \bigcup_{p_c \in [l_c, u_c]} A_{p_c} = A_{u_c},$$

where $[l_c, u_c]$ is given by Equation (4.3). It follows that, for each $c \in \{1, \ldots, 32\}$, $P(Y_1 = y \mid V_1 = 1, V_3 = 1, C = c) \in$ belongs to an identification region where the lower bound is given by

$$\max \left\{ 0, \frac{P(Y_1 = y \mid V_1 = 1, C = c) - \min \left\{ 1, \frac{P(V_3 = 1 \mid C = c)}{P(V_1 = 1 \mid C = c)} \right\}}{1 - \min \left\{ 1, \frac{P(V_3 = 1 \mid C = c)}{P(V_1 = 1 \mid C = c)} \right\}} \right\},$$

and the upper bound is given by

$$\min \left\{ 1, \frac{P(Y_1 = y \mid V_1 = 1, C = c)}{1 - \min \left\{ 1, \frac{P(V_3 = 1 \mid C = c)}{P(V_1 = 1 \mid C = c)} \right\}} \right\}.$$

Table 8 shows the corresponding lower and upper bound of $P(V_3 = 1 \mid V_1 = 1, C = c)$ and $P(Y_1 = \text{non-approve} \mid V_1 = 1, V_3 = 1, C = c)$. It can be observed the uncertainty induced by the joint participation in both elections. In particular, four municipalities have an extreme uncertainty because the width of their identification regions is at least equal to 0.9: Curacautín, Ercilla, Lumaco and Victoria. In these municipalities, the proportion of non-approval conditionally on the participation at both the 2020 referendum and the 2021

consultation is any value. Moreover, the conditional probability to participate at the citizen referendum given that electors participated at the 2020 referendum is, respectively, 0.6 0.58, 0.81 and 0.68. Also, two municipalities, Lonquimay and Los Sauces, have the smaller uncertainty and, consequently, the rate of approval conditionally on the joint participation is less uncertainty: between 0.37 and 0.41 for Lonquimay; and between 0.39 and 0.45 for Los Sauces. Nevertheless, the rate of participation at the 2021 consultation given participation at the 2020 referendum are quite small: 0.04 and 0.06, respectively.

4.6 Discussion

The partial identification analysis shows the impact of the uncertainty due to the joint participation in both elections, namely 2020 referendum and 2021 consultation, on the proportion of electors who chose the non-approve option at the 2020 referendum. This impact can be diminished if, for each municipality, the joint distribution $P(V_1 = v_1, V_3 = v_3 \mid C = c)$ with $(v_1, v_3) \in \{0,1\}^2$ were known. This seems to be feasible for the Chilean Electoral Service, without having to transgress elector identity protection. If this were the case, then $P(V_3 = 1 \mid V_1 = 1, C = c)$ would be identified. However, this fact does not ensure that $P(Y_1 = y \mid V_1, = 1, V_3 = 1, C = c)$ is point identified because $P(Y_1 = y \mid V_1 = 1, V_3 = 0, C = c)$ is not identified given the secrecy of the vote. Consequently, following the arguments developed in Subsection 4.5, $P(Y_1 = y \mid V_1, = 1, V_3 = 1, C = c)$ belongs to an identification interval with a lower bound given by

$$\max\left\{0, \frac{P(Y_1 = y \mid V_1 = 1, C = c) - P(V_3 = 1 \mid V_1 = 1, C = c)}{P(V_3 = 0 \mid V_1 = 1, C = c)}\right\}$$

and an upper one given by

$$\min\left\{1, \frac{P(Y_1 = y \mid V_1 = 1, C = c)}{P(V_3 = 0 \mid V_1 = 1, C = c)}\right\}.$$

It can be deduced that this interval is informative (that is, strictly included in $[0,1]$) if

$$P(V_3 = 1 \mid V_1 = 1, C = c) < P(V_3 = 0 \mid V_1 = 1, C = c),$$

which is a surprising result.

It could be argued that, under "mild conditions", it is possible to ignore joint participation, and thus argue for the reliability of studies such as the one reported in Subsection 4.4. The partial identification analysis developed in Subsection 4.5 shows that there are two possible assumptions that could be made: the first one would be to assume that $P(Y_1 =$ non-approve $\mid V_1 = 1, V_3 = 0) = 0$, that is, that no elector who participated in the 2020 referendum and did not participate in the 2021 citizen consultation chose the option non-approve. It should be mentioned that this assumption is quite strong and hard to believe. A second assumption would be $Y_1 \perp\!\!\!\perp V_3 \mid \{V_1 = 1\}, C$, which is equivalent to the following two equivalent conditions:

$$\begin{aligned} &P(Y_1 = y \mid V_1 = 1, V_3 = 1, C = c) = P(Y_1 = y \mid V_1 = 1, C = c); \\ &P(V_3 = v_3 \mid Y = y, V_1 = 1, C = c) = P(V_3 = v_3 \mid V_1 = 1, C = c) \quad v_3 \in \{0,1\}. \end{aligned}$$

The last condition means that, once an elector of a specific municipality participated at the 2020 referendum, the participation at the 2021 consultation does not depend on the preference at the 2020 referendum: again hard to believe.

Therefore, it should be emphasized that the previous analysis of partial identification is applicable to critically assess the comparisons over time of political surveys as they would be correctly done if made conditionally on joint participation.

5. Concluding remarks

This paper illustrates a traditional service that Applied Statistics can render to society. In fact, during the XIX century, statistics was considered as "the science of social facts, expressed in numerical terms", as indicated by Moreau de Jones (1847), or as the prospectus of the Statistical Society of London stated, "Statistics [...] may be said [...] to be ascertaining and bringing together of those ≪facts which are calculated to illustrate the condition and prospects of society;≫and the objective of Statistical Science is to consider the results which they produce, with the view to determine those principles upon which the well-being of society depends" (Journal of the Statistical Society of London, 1838). As it is well known, these considerations go back to Süßmilch (1998) and his idea of seeking order in the figures that summarize the profile of a state – hence the term Statistics.

These original ideas show clearly the need of every statesman for statistics to "illustrate, with new or more accurate data, a multitude of issues that arise every day, stimulating public opinion, being the subject of parliamentary discussions, and forming problems whose solution can only be offered by Statistics" Moreau de Jones (1847). The two surveys analyzed in this paper, as well as the citizen consultation, are examples of the scenario described by Moreau de Jones. As a matter of fact, the socioeconomic survey CASEN is used by policy makers either to assess social policies or to have a global view of poverty or income distribution. Stake-holders, as the press or politicians, use the two political opinion polls (CADEM and Araucanía citizen consultation) either to influence citizens' political opinion or to justify political arguments at the parliament.

We complement Moreau de Jones' scenario by making explicit new frontiers of what statisticians and social scientists call data of good quality. From a statistical point of view, we focus our assessment of the surveys on the correct way of communicating their results, so that the uncertainty induced by non-responses is made explicit. The results can be reported at different levels depending on the population of interest to which the results are to be generalized. The advantage of this strategy is that it makes explicit how this uncertainty could be reduced, which part of it can not be reduced unless very strong assumptions are introduced. The price to be paid in the face of these strong assumptions is the drawing of non-credible conclusions –that is, Law of Decreasing Credibility (Manski, 2013). For instance, in the Araucanía citizen consultation, the uncertainty of the option at the 2020 referendum conditionally on the joint participation at both the 2020 referendum and 2021 consultation can decreased whether the Chilean Electoral Service provides information on such joint participation. However, the uncertainty can not decrease to a point value because of the secrecy of the vote.

At the methodological level, the assessment or dissection of the Chilean surveys was performed making a distinction between identified parameters and parameters of interest: what we can learn from the data is represented by the identified parameters, while what we want to learn from the data is represented by the parameters of interest. In (almost) all empirical research there is a gap between both types of parameters; it is quite relevant to highlight the difference and to study their possible relationships –which is equivalent to solving an identification problem. The way Clifford (1982) expresses himself is illuminating and perhaps summarizes the perspective developed in this paper:

> Anyone who has tried to make sense to real data will, sooner or later, have come across the problem of non-identifiability. Broadly speaking this means that their first explanation of the data is not the only one. The existence of alternative explanations becomes important when decisions have to be made and particularly so when different explanations suggest completely different courses of action.

The identification regions we established for each of the Chilean survey contain such different substantive explanations.

Supplementary material

The computational routine implemented in `R` is available online at https://github.com/edalarconb?tab=repositories.

Author contributions Conceptualization, E.S.M., E.A-B.; methodology, E.S.M., E.A-B.; software, E.S.M., E.A-B.; validation, E.S.M., E.A-B.; formal analysis, E.S.M., E.A-B.; investigation, E.S.M., E.A-B.; data curation, E.S.M., E.A-B.; writing—original draft preparation, E.S.M., E.A-B.; writing—review and editing, E.S.M., E.A-B.; visualization, E.S.M., E.A-B. supervision, E.S.M., E.A-B. All authors have read and agreed to the published version of the paper.

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.

Funding This research was supported by the Agencia Nacional de Investigación y Desarrollo (ANID) Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI) NCS2021072. The first author was also partially funded by the FONDECYT Project No. 1181261, whereas the second author was partially funded by the Postdoctoral FONDECYT grant No. 3220422.

Conflicts of interest The authors declare no conflict of interest.

References

Alarcón-Bustamante, E., 2022. EmpiricalEvidence: an R package for empirical research. Available from https://github.com/edalarconb.

Alarcón-Bustamante, E., San Martín, E., and González, J., 2020. Predictive validity under partial observability. In Wiberg, M., Molenaar, D., González, J., Böckenholt, U., and Kim, J.S. (Eds.) Quantitative Psychology, pp. 135–145. Springer, Cham, Switzerland.

Andrade, M.J., 2019. La lucha por el territorio mapuche en Chile: una cuestión de pobreza y medio ambiente. L'Ordinaire des Amériques, 2019, 225..

Bengoa, J., 2007. El Tratado de Quilín. Catalonia, Santiago, Chile.

Bengoa, J., 2016. Sarmiento y sarmientadas. In Sarmiento, D.F. (Ed.). Conflicto y Armonía de las Razas en América, pp. 5–14. Akal-Inter Pares, Santiago, Chile.

Berinsky, A.J., 2017. Measuring public opinion with surveys. Annual Review of Political Science, 20, 309–329.

CADEM, 2018. Diseño Metodológico de Plaza Pública CADEM 2018. Santiago, Chile.

CADEM, 2022. Encuesta Plaza Pública. Santiago, Chile.

Cayul, P., Durán, E., and Jaimovich, D., 2021. ¿Es representativa la consulta ciudadana en la Araucanía? (Is the citizen consultation in Araucanía representative?). Santiago, Chile.

Clifford, P., 1982. Some General Comments on Nonidentifiability. In LeCam, L. and Neyman, J. (Eds.) Probability Models and Cancer, pp. 81–83. North-Holland Publishing Company, Amsterdam, Netherlands.

Constitución de la República de Chile, 2005. Decreto 100. fija el texto refundido, coordinado y sistematizado de la Constitución Política de la República de Chile. Santiago, Chile.

Diario Oficial de la República de Chile, 2021. Declara estado de excepción constitucional de emergencia en las zonas del territorio nacional que indica. Santiago, Chile.

Embrechts, P. and Hofert, M., 2013. A note on generalized inverses. Mathematical Methods of Operations Research, 77, 423–432.

Fisher, R.A., 1922. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London A, 222, 309–368.

Florens, J.P. and Mouchart, M., 1982. A note on noncausality. Econometrica, 50, 583–592.

Florens, J.P., Mouchart, M., and Rolin, J.M., 1990. Elements of Bayesian Statistics. Marcel Dekker, New York, USA.

Fréchet, M., 1960a. Les tableaux dont les marges sont données. Trabajos de Estadística, 11, 1–18.

Fréchet, M., 1960b. Sur les tableaux dont les marges et des bornes sont données. Revue de l'Institut International de Statistique, 28, 10–32.

Hirano, K. and Imbens, G. W., 2004. The propensity score with continuous treatments. In Shewhart, W.A. and Wilks, S.S. (Eds.) Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family, pp. 73–84. Wiley, New York, USA.

Hyndman, R. J. and Fan, Y., 1996. Sample Quantiles in Statistical Packages. The American Statistician, 50, 361–365.

Imbens, G.W., 2000. The role of the propensity score in estimating dose-response functions. Biometrika, 87, 706–710.

Instituto Forestal, 2021. Anuario Forestal. Boletín Estadístico Número 174. Instituto Forestal, Santiago, Chile.

Journal of the Statistical Society of London, 1838. Introduction. Journal of the Statistical Society of London, 1, 1–5.

Kolmogorov, A N., 1950. Foundations of the Theory of Probability. Chelsea Publishing Company, New Yok.

Koopmans, T. and Reiersol, O., 1950. The identification of structural characteristics. The Annals of Mathematical Statistics, 21, 165–181.

Little, R.J. and Rubin, D.B., 2019. Statistical Analysis with Missing Data. Wiley, New York.

Manski, C.F., 2007. Identification for Prediction and Decision. Harvard University Press, Cambridge, MA, USA.

Manski, C.F., 2011. Policy analysis with incredible certitude. The Economic Journal, 121:F261–F289.

Manski, C.F., 2013. Public Policy in an Uncertain World. Harvard University Press, Cambridge, MA, USA.

Manski, C.F., 2020. The lure of incredible certitude. Economics and Philosophy, 36, 216–245.

Moreau de Jones, A., 1847. Éléments de Statistique Comprenant les Principes Généraux de cette Science, et un Aperu Historiques de ses Progrès. Guillaumn et Cie, Paris.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Rao, M., 2005. Conditional Measures and Applications. Second Edition. Chapman Hall, New Yok.

Rubin, D.B., 1976. Inference and missing data. Biometrika, 63, 581–592.

San Martín, E., 2018. Identifiability of structural characteristics: How relevant is it for the Bayesian approach? Brazilian Journal of Probability and Statistics, 32, 346–373.

San Martín, E. and González, J, 2022. A critical view on the NEAT equating design: Statistical modelling and identifiability problems. Journal of Educational and Behavioral Statistics, pages in press.

San Martín, E., González, J., and Tuerlinckx, F., 2015. On the unidentifiability of the fixed-effects 3PL model. Psychometrika, 80, 450–467.

Süßmilch, J.P., 1998. L'Ordre Divin dans les changements de l'espèce humaine, démontré par la naissance, la mort et la propagation de celle-ci.. À l'Institut National D'Études Démographiques, Paris.

Chilean Journal of Statistics
Volume 13, Number 1 (April 2022), Pages 47–66
DOI: 10.32372/ChJS.13-01-03

Time Series
Research Paper

A Bayesian detection of structural changes in autoregressive time series models

Abdeldjalil Slama[1,2,*]
[1]Department of Mathematics and Computer Sciences,
[2]Laboratory of Mathematics, Modeling and Applications
University of Adrar, Adrar, Algérie
(Received: 01 February 2022 · Accepted in final form: 10 March 2022)

Abstract

This study investigates a Bayesian detection of a change in any parameter, or in any collection of parameters of the autoregressive time series model of known order p. An unconditional Bayesian test based on the highest posterior density credible sets is determined. Using the Gibbs sampler algorithm, some simulated results are given to approximate the posterior densities of the change point and other parameters of the model. The performance of our proposed method has been investigated on simulated and real data sets.

Keywords: Bayesian analysis· Change point· Gibbs sampler· HPD credible set· p-value.

Mathematics Subject Classification: Primary 62M10 · Secondary 62F15.

1. Introduction

Change point detection is an important element in time series analysis that arises in many fields such as quality control procedures (Basseville and Nikiforov (1993)), anomaly detection in internet traffic data (Lévy-Leduc and Roueff, 2009; Tartakovsky et al., 2006), metrology (Jandhyala et al., 2014), economics and financial analysis (Georgescu, 2012), and biology (Fan et al., 2015), among others. Change point detection is the problem of detecting abrupt changes in the parameters of temporal or other sequential data. Since the papers of Page (1954) and Page (1955), who proposed a sequential scheme for identifying changes in the mean of a sequence of independent random variables, the problem of detecting changes has been an important issue between statisticians and considerable attention has been given to this problem in a variety of settings. For example, changes in a sequence of random variables have been considered by Eastwood (1993), Gombay and Horvath (1999) and Guo and Modarres (2020) from the nonparametric viewpoint. Montoril and da Silva Ferreira (2018) proposed a method based on the coefficient of determination, to estimate the change points in the Beer-Lambert law problems. Among the approach based on likelihood ratio, Worsley (1983, 1986) proposed a numerical method for computing the p-value of the generalized likelihood ratio test to detect a change in the binomial probability and in the location of an

*Corresponding author. Email: aslama@univ-adrar.edu.dz, slama_dj@yahoo.fr

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

http://www.soche.cl/chjs

exponential family distribution. Kim (1996) considered a likelihood ratio test for a change in the mean when observations are correlated. Kim and Siegmund (1989) considered likelihood ratio tests to detect a change-point in simple linear regression. Wang et al. (2020) used the likelihood ratio test to detect changes in the parameters of the skew slash distribution.

From a Bayesian point of view, the problem of detecting a change has received much attention and has been studied by many authors like Chernoff and Zacks (1964), Kander and Zacks (1966), Sen and Srivastava (1975), Jani and Pandya (1999), Fan and Chen (2005) and Shah and Patel (2007). Ming Ng (1990) analyzed a linear model in which both the mean and the precision change once at an unknown time point, the posterior distributions of the change point, and the ratio of the precisions are derived.

Kim (1991) proposed a Bayesian significance test for the stationarity of a regression equation using the highest posterior density (HPD) credible set. From a Monte Carlo simulation study, he showed that the Bayesian significance test has a stronger power than the Cusum and the Cusum of squares tests suggested by Brown et al. (1975). Sáfadi and Morettin (2000) considered a Bayesian analysis for threshold autoregressive moving average models. Pan et al. (2017) considered a Bayesian analysis of threshold autoregressive (TAR) model with various possible thresholds. Recently, Hahn et al. (2020) introduced a computationally inexpensive Bayesian approach (BayesProject) for detecting changes in mean within multivariate data sequences.

For autoregressive time series models, many papers about detecting and estimating changes in autoregressive time series of known order p (AR(p)) processes have been published. For example, Davis et al. (1995) studied the asymptotic behavior of a Gaussian-type likelihood ratio statistic for testing a change in the parameters of an AR(p) model. Husková et al. (2007, 2008) used an approach based on partial sums of weighted residuals (asymptotic and bootstrapping methods). Venkatesan and Arumugam (2007) considered the problem of gradual changes in the parameters of an autoregressive time series model. Gombay (2008) used the efficient score vector to detect change in the parameter(s) of autoregressive time series. Berkes et al. (2011) developed the likelihood ratio test for the structural change of an AR model to a threshold AR model. Slama (2014) examined the effect of correlation on the performance of the Bayesian significance test derived under the assumption of no correlation. By numerical studies, he showed that the Bayesian significance test based on the HPD region is sensitive to the correlation in the data. Kezim and Abdelli (2004) proposed a Bayesian analysis of a first order autoregressive process subject to one change in both the variance of the error terms and the autocorrelation coefficients at an unknown time point. The detection of possible changes in the parameters of autoregressive models for binary time series can be found in Hudecová (2013). Cheon and Kim (2014) proposed a general solution to detect the Bayesian estimation in Bayesian autoregressive structural-change time series models. A Bayesian approach to estimate the multiple structural change-points in a level and the trend when the number of change-points is unknown was proposed. Slama and Saggou (2017) investigated the Bayesian approach using HPD credibles sets and p-values for detecting an abrupt change in the parameters of an AR(p). In a recent work, Gamage and Ning (2021) proposed a nonparametric method based on the empirical likelihood is proposed to detect the structural changes in the autoregressive parameters of autoregressive models. In the last three works, the mean is assumed constant and equal to 0.

Bauwens et al. (2014) solved the problem of the computation of the marginal likelihood for a Markov-switching GARCH or change-point GARCH models by applying a particle Markov chain Monte Carlo (PMCMC) method. Recently, Romano et al. (2021) proposed a principled approach to detect abrupt changes in mean in univariate time-series that models local fluctuations as a random walk process and autocorrelated noise via an AR(1) process. For a review of methods of inference for single and multiple change-points in time series, we refer the reader to Jandhyala et al. (2013) and Truong et al. (2020).

In this paper, we investigate a Bayesian detection of a change in any parameter, or in any collection of parameters of an AR(p). We consider a Bayesian significance test for an abrupt change at an unknown time point in the mean, the autocorrelation coefficients and the variance of the error terms of an AR(p). This work is an extension of the paper by Slama and Saggou (2017) to the case where the mean is unknown and changes at an unknown time.

The rest of the paper is organized as follows. Section 2 presents the model AR(p) with change in the parameters at an unknown time point and some notations used along this paper. In Section 3 we give the conditional posterior distributions of the parameters of change and Bayesian significance test of change in AR(p) model. In Section 4 we present a simulation results with the application of the Gibbs sampler algorithm. A real data analysis is provided in Section 5. Finally, our conclusion is presented in Section 6.

2. Definition of the Model and notations

Assume that we observe a real time series, $y_1, \ldots, y_n$ namely, generated from an AR(p) model, with a change in the mean μ, the autocorrelation coefficients ϕ_i and in the variance σ^2 at an unknown time point m. The AR(p) model with structural change is given by

$$
\begin{aligned}
Y_t - \mu_1 &= \sum_{i=1}^{p} \phi_i \left(Y_{t-i} - \mu_1\right) + \epsilon_t, \quad t = 1, \ldots, m, \\
Y_t - \mu_2 &= \sum_{i=1}^{p} \psi_i \left(Y_{t-i} - \gamma_{t-i}\mu_1 - (1-\gamma_{t-i})\mu_2\right) + \epsilon_t, \quad t = m+1, \ldots, m+p, \\
Y_t - \mu_2 &= \sum_{i=1}^{p} \psi_i (Y_{t-1} - \mu_2) + \epsilon_t, \quad t = m+p+1, \ldots, n,
\end{aligned}
\tag{1}
$$

where γ_t is the indicator function such that $\gamma_{t-i} = 1$ if $t - i \leq m$ and $\gamma_{t-i} = 0$ if $t - i > m$. $\epsilon_t \sim \mathrm{N}(0, \sigma_1^2)$, for $t = 1, \ldots, m$ and $\epsilon_t \sim \mathrm{N}(0, \sigma_2^2)$, for $t = m+1, \ldots, n$. The parameters $\mu_i \in \mathrm{R}$, $\sigma_i > 0$), for $i = 1, 2$, and ϕ_i, ψ_i, for $i = 1, \ldots, p$, are assumed to be unknown, and $m = 1, \ldots, n-2$ is the change point assumed also unknown. If $\phi_i \neq \psi_i$ for some $i = 1, \ldots, p$, the structure of the series has changed from an AR(p) model with coefficient ϕ_i to another AR(p) model with coefficient ψ_i. We assume that the autoregressive parameters correspond to stationary processes in the sense that the parameter vector $\phi^{(p)} = (\phi_1, \phi_2, \ldots, \phi_p)$ lies in the stationary region $\Phi_1^{(p)} = \{z/1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0\}$, which implies $|z| > 1$, and likewise $\psi^{(p)} = (\psi_1, \psi_2, \ldots, \psi_p)$ lies in the stationary region $\Phi_2^{(p)} = \{z/1 - \psi_1 z - \psi_2 z^2 - \cdots - \psi_p z^p = 0\}$, which implies $|z| > 1$. The quantities $y_{1-p}, \ldots, y_{-1}, y_0$ are the initial observations assumed to be stated.

The model given in Equation (1) is more general than the model considered in Slama and Saggou (2017). In Slama and Saggou (2017), the mean μ is assumed to be constant and equal to 0. Whereas, in Equation (1) the mean is assumed unknown and changes at an unknown time point m, which increases the size of the parameter space. The parameter space for the model given in Equation (1) is $\Theta = \{\theta = (m, \mu_1, \mu_2, \phi_1, \phi_2, \ldots, \phi_p, \psi_1, \ldots, \psi_p, r_1, r_2)\}$, where $r_i = 1/\sigma_i^2, i = 1, 2$, with $m = 1, \ldots, n-2, \mu_1, \mu_2 \in \mathrm{R}, r_1, r_2 \in \mathrm{R}_+^*$, and $\phi_i, \psi_i \in \Phi^{(p)}$, for $i = 1, \ldots, p$.

We want to test whether or not a change-point occurs in the autoregressive parameters. Thus, we build an inference about testing the hypotheses: H_0: $\delta = \mu_2 - \mu_1 = 0$ and $\rho_j = \psi_j - \phi_j = 0, \forall j = 1, \ldots, p$ and $\tau = \sigma_2^2/\sigma_1^2 = 1$, against H_1: $\delta = \mu_2 - \mu_1 \neq 0$ or for at least one $\rho_j = \psi_j - \phi_j \neq 0, j = 1, \ldots, p$, or $\tau = \sigma_2^2/\sigma_1^2 \neq 1$. Hence, under the alternative hypothesis, there is a change in at least one of the $2p + 5$ parameters at an unknown time point. The proposed test is based on the posterior distribution of the shift $\delta = \mu_2 - \mu_1$, $\rho_j = \psi_j - \phi_j$ and of the ratio $\tau = \sigma_2^2/\sigma_1^2$. The hypothesis meaning "no change" is equivalent to H_0': $m = n$ and H_1 is equivalent to $H_1' : m \neq n$.

For the rest of the paper, we consider the notations: $\phi^{(p)} = (\phi_1, \ldots, \phi_p)$, $\psi^{(p)} = (\psi_1, \ldots, \psi_p)$, $\rho^{(p)} = (\rho_1, \ldots, \rho_p)$, $\phi^{(-j)} = (\phi_1, \ldots, \phi_{j-1}, \phi_{j+1}, \ldots, \phi_p)$ and $\rho^{(-j)} = (\rho_1, \ldots, \rho_{j-1}, \rho_{j+1}, \ldots, \rho_p)$, where $\rho_j = \psi_j - \phi_j$, $j = 1, \ldots, p$. The functional forms $\pi(\cdot)$ and $\pi(\cdot \mid \cdot)$ represent a prior and a posterior distribution, respectively.

The parameter set $\theta = (m, \mu_1, \mu_2, \phi^{(p)}, \psi^{(p)}, r_1, r_2)$, where $r_i = 1/\sigma_i^2$, for $i = 1, 2$, is a vector of dimension $(2p + 5)$. The conditional likelihood function based on the observations $y = (y_1, \ldots, y_n)$ is given by

$$
\begin{aligned}
l(y|\theta) \propto \quad & r_1^{\frac{m}{2}} r_2^{\frac{n-m}{2}} \exp\left\{ -\frac{r_1}{2} \left[\sum_{t=1}^{m} (y_t - \mu_1 - \sum_{i=1}^{p} \phi_i (y_{t-i} - \mu_1) \right]^2 \right\} \\
& \exp\left\{ -\frac{r_2}{2} \left[\sum_{t=m+1}^{m+p} (y_t - \mu_2 - \sum_{i=1}^{p} \psi_i \Big(y_{t-i} - \gamma_{t-i}\mu_1 - (1 - \gamma_{t-i})\mu_2 \Big) \right]^2 \right\} \\
& \exp\left\{ -\frac{r_2}{2} \left[\sum_{t=m+p+1}^{n} (y_t - \mu_2 - \sum_{i=1}^{p} \psi_i (y_{t-i} - \mu_2) \right]^2 \right\}.
\end{aligned} \tag{2}
$$

The conditional likelihood approach is based on the assumption that the initial observations $y_0, y_{-1}, \ldots, y_{1-p}$ are also available (Reinsel, 1997). Moreover, if the sample size n is sufficiently large, the first observation makes a negligible contribution to the total likelihood (Hamilton, 1994).

3. Bayesian analysis

In this section, the conditional posterior distribution of the shift in the mean δ, in the autocorrelation coefficients ρ_j, $j = 1, \ldots, p$, of the variance ratio τ and of the change point m are derived. These distributions are used to define an unconditional Bayesian significance test of change in the parameters of an AR(p).

Since prior knowledge of $\theta' = (\mu_1, \mu_2, r_1, r_2)$ is often vague or diffuse, we employ a diffuse prior for θ'. Assume that the priors of the change-point m, of $\phi^{(p)}$ and of $\psi^{(p)}$ are given by

$$
\begin{aligned}
\pi(m) &\propto \frac{1}{n-2}; \quad m = 1, \ldots, n-2, \\
\pi(\phi^{(p)}) &\propto \text{constant in} \quad \Phi^{(p)}, \\
\pi(\psi^{(p)}) &\propto \text{constant in} \quad \Phi^{(p)},
\end{aligned}
$$

where $\Phi^{(p)} = \Phi_1^{(p)} \cap \Phi_2^{(p)}$.

The parameters m, $\phi^{(p)}$ and θ' being assumed independent. The prior distribution of θ is, therefore, stated as

$$\pi(\theta) \propto \frac{1}{r_1 r_2}, \tag{3}$$

where $m = 1, \ldots, n-2$, $\mu_1, \mu_2 \in \mathrm{R}$, $r_1, r_2 \in \mathrm{R}_+^*$ and $\phi_i, \psi_i \in \Phi^{(p)}$ for $i = 1, \ldots, p$. The posterior distribution of θ, obtained by combination of Equations (2) and (3) is formulated as

$$\begin{aligned}\pi(\theta|y) \propto r_1^{\frac{m}{2}-1} r_2^{\frac{n-m}{2}-1} \exp\left\{-\frac{r_1}{2}\left[\sum_{t=1}^{m}(y_t - \mu_1 - \sum_{i=1}^{p}\phi_i(y_{t-i} - \mu_1)\right]^2\right\} \\ \exp\left\{-\frac{r_2}{2}\left[\sum_{t=m+1}^{m+p}(y_t - \mu_2 - \sum_{i=1}^{p}\psi_i\Big(y_{t-i} - \gamma_{t-i}\mu_1 - (1-\gamma_{t-i})\mu_2\Big)\right]^2\right\} \\ \exp\left\{-\frac{r_2}{2}\left[\sum_{t=m+p+1}^{n}(y_t - \mu_2 - \sum_{i=1}^{p}\psi_i(y_{t-i} - \mu_2)\right]^2\right\}.\end{aligned}$$

In the following, we give the joint posterior distribution of the parameter $\Phi = (m, \mu_1, \delta, \phi^{(p)}, \rho^{(p)}, \tau)$. By transforming the parameter set $\Theta = (m, \mu_1, \mu_2, \phi^{(p)}, \psi^{(p)}, r_1, r_2)$ into $\Phi = (m, \mu_1, \delta, \phi^{(p)}, \rho^{(p)}, \tau)$, we can form the joint posterior distribution of Φ, that is, we have

$$\begin{aligned}\pi(\Phi \mid y) = \int_{r_2} \pi(m, \mu_1, \delta + \mu_1, \phi_1, \rho^{(p)} + \phi^{(p)}, r_2\tau, r_2/y) \mid r_2 \mid \mathrm{d}r_2, \\ \tau^{\frac{m}{2}-1}\left\{\tau\sum_{t=1}^{m}\left(y_t - \mu_1 - \sum_{i=1}^{p}\phi_i(y_{t-i} - \mu_1)\right)^2\right. \\ + \sum_{t=m+1}^{m+p}\left(y_t - \delta - \mu_1 - \sum_{1}^{p}(\rho_i + \phi_i)(y_{t-i} - \gamma_{t-i}\mu_1 - (1-\gamma_{t-i})(\delta + \mu_1))\right)^2 \\ \left. + \sum_{t=m+p+1}^{n}\left(y_t - \delta - \mu_1 - \sum_{1}^{p}(\rho_i + \phi_i)(y_{t-i} - \delta - \mu_1)\right)^2\right\}^{-\frac{n}{2}}.\end{aligned} \tag{4}$$

The posterior conditional distribution of δ is stated as follows. Equation (4) can be written as

$$\pi(\Phi \mid y) \propto \tau^{\frac{m}{2}-1}\left\{\tau SS_1(m, \mu_1, \phi^{(p)}) + SS_2(m, \mu_1, \phi^{(p)}, \rho^{(p)}) + \Lambda_1\left(\delta - \widehat{\delta}\left(m, \mu_1, \phi^{(p)}, \rho^{(p)}\right)\right)^2\right\}^{-\frac{n}{2}}, \tag{5}$$

where

$$\Lambda_1 = \sum_{m+1}^{m+p}\left(1 - \sum_{1}^{p}(1-\gamma_{t-i})(\rho_i + \phi_i)\right)^2 + (n-m-p)\left(1 - \sum_{1}^{p}(\rho_i + \phi_i)\right)^2,$$

$\widehat{\delta}(m, \mu_1, \phi^{(p)}, \rho^{(p)}) = \Lambda_2/\Lambda_1$, with

$$\Lambda_2 = \sum_{m+1}^{m+p} (1 - (1 - \gamma_{t-i})(\rho_i + \phi_i)) \left(y_t - \mu_1 - \sum_{i=1}^{p} (\rho_i + \phi_i)(y_{t-i} - \mu_1 \right) + \left(1 - \sum_{i=1}^{p} (\rho_i + \phi_i) \right) \left(\sum_{m+p+1}^{n} (y_t - \mu_1 - \sum_{i=1}^{p} (\rho_i + \phi_i)(y_{t-i} - \mu_1) \right),$$

and

$$SS_1(m, \mu_1, \phi^{(p)}) = \sum_{t=1}^{m} \left(y_t - \mu_1 - \sum_{i=1}^{p} \phi_i (y_{t-i} - \mu_1) \right)^2, \tag{6}$$

$$SS_2(m, \mu_1, \phi^{(p)}, \rho^{(p)}) = \sum_{t=m+1}^{n} \left(y_t - \mu_1 - \sum_{1}^{p} (\rho_i + \phi_i)(y_{t-i} - \mu_) \right)^2 - \frac{\Lambda_2^2}{\Lambda_1}. \tag{7}$$

Following the Bayes theorem, the conditional posterior distribution of δ is given by

$$\pi(\delta | m, \mu_1, \phi^{(p)}, \rho^{(p)}, \tau, y) \propto \left\{ 1 + \frac{(\delta - \widehat{\delta}\left(m, \mu_1, \phi^{(p)}, \rho^{(p)}\right))^2}{(n-1) S_1^2(m, \mu_1, \phi^{(p)}, \rho^{(p)}, \tau)} \right\}^{-\frac{n}{2}},$$

where $S_1^2(m, \mu_1, \phi^{(p)}, \rho^{(p)}, \tau) = (\tau SS_1(m, \mu_1, \phi^{(p)}) + SS_2(m, \mu_1, \phi^{(p)}, \rho^{(p)}))/((n-1)\Lambda_1)$. Given m, μ_1, $\phi^{(p)}$, $\rho^{(p)}$ and τ, the conditional posterior distribution of δ is distributed as a Student-t distribution with location parameter $\widehat{\delta}(m, \mu_1, \phi^{(p)}, \rho^{(p)})$, precision $S_1^2(m, \mu_1, \phi^{(p)}, \rho^{(p)}, \tau)$ and $(n-1)$ degrees of freedom. Equivalently, the quantity

$$T(\delta) = \frac{\delta - \widehat{\delta}\left(m, \mu_1, \phi^{(p)}, \rho^{(p)}\right)}{S_1\left(m, \mu_1, \phi^{(p)}, \rho^{(p)}, \tau\right)},$$

is distributed a posteriori as a conditional Student-t distribution with $(n-1)$ degrees of freedom.

The posterior conditional distributions of ρ_j is formulated as follows. Equation (4) can also be written as

$$\pi(\Phi \mid y) \propto \tau^{\frac{m}{2}-1} \left\{ \Lambda_{5j} - \frac{\Lambda_{4j}^2}{\Lambda_{3j}} + \Lambda_{3j} \left(\rho_j - \widehat{\rho}_j \left(m, \mu_1, \delta, \phi^{(p)}, \rho^{(-j)} \right) \right)^2 \right\}^{-\frac{n}{2}},$$

where, $\widehat{\rho}_j\left(m,\mu_1,\delta,\phi^{(p)},\rho^{(-j)}\right)=\Lambda_{4j}/\Lambda_{3j}$, with

$$\Lambda_{3j}=\sum_{t=m+1}^{m+p}\left(y_{t-j}-\gamma_{t-j}\mu_1-(1-\gamma_{t-j})(\delta+\mu_1)\right)^2+\sum_{t=m+p+1}^{n}\left(y_{t-j}-\delta-\mu_1\right)^2;$$

$$\begin{aligned}\Lambda_{4j}=&\sum_{m+1}^{m+p}\Big[y_{t-j}-\gamma_{t-j}\mu_1-(1-\gamma_{t-j})(\delta+\mu_1)\Big]\\&\Big[y_t-\delta-\mu_1-\sum_{i=1}^{p}\phi_i(y_{t-i}-\gamma_{t-i}\mu_1-(1-\gamma_{t-i})(\delta+\mu_1))\\&-\sum_{i\neq j}^{p}\rho_i(y_{t-i}-\gamma_{t-i}\mu_1-(1-\gamma_{t-i})(\delta+\mu_1))\Big]\\&\sum_{m+p+1}^{n}[y_{t-j}-\delta-\mu_1])\Big[y_t-\delta-\mu_1-\sum_{i=1}^{p}\phi_i(y_{t-i}-\delta-\mu_1)-\sum_{i\neq j}^{p}\rho_i(y_{t-i}-\delta-\mu_1)\Big];\end{aligned}$$

$$\begin{aligned}\Lambda_{5j}=&\tau\sum_{t=1}^{m}\Big(y_t-\mu_1-\sum_{i=1}^{p}\phi_i(y_{t-i}-\mu_1)\Big)^2\\&+\sum_{m+1}^{m+p}\Big(y_t-\delta-\mu_1-\sum_{1}^{p}\phi_i(y_{t-i}-\gamma_{t-i}\mu_1-(1-\gamma_{t-i})(\delta+\mu_1))\\&-\sum_{i\neq j}^{p}\rho_i(y_{t-i}-\gamma_{t-i}\mu_1-(1-\gamma_{t-i})(\delta+\mu_1))\Big)^2\\&+\sum_{m+p+1}^{n}\Big(y_t-\delta-\mu_1-\sum_{i=1}^{p}\phi_i(y_{t-i}-\delta-\mu_1)-\sum_{i\neq j}^{p}\rho_i(y_{t-i}-\delta-\mu_1)\Big)^2.\end{aligned}$$

Following the Bayes theorem, the posterior conditional distribution of ρ_j, for $j=1,\ldots,p$, is given by

$$\pi(\rho_j|m,\mu_1,\phi^{(p)},\delta,\rho^{(-j)},\tau,y)\propto\left\{1+\frac{(\rho_j-\widehat{\rho}_j(m,\phi^{(p)},\mu_1,\delta,\rho^{(-j)})^2}{(n-1)S_{2j}^2(m,\phi^{(p)},\mu_1,\delta,\rho^{(-j)},\tau)}\right\}^{-\frac{n}{2}},$$

where

$$S_{2j}^2(m,\phi^{(p)},\mu_1,\delta,\rho^{(-j)},\tau)=\frac{\Lambda_{5j}-\dfrac{\Lambda_{4j}^2}{\Lambda_{3j}}}{(n-1)\Lambda_{3j}}.$$

For $j=1,\ldots,p$, given m, μ_1, $\phi^{(p)}$, δ, $\rho^{(-j)}$ and τ, the conditional posterior distribution of ρ_j is distributed as a Student-t distribution with location parameter $\widehat{\rho}_j(m,\phi^{(p)},\mu_1,\delta,\rho^{(-j)})$, precision $S_{2j}(m,\phi^{(p)},\mu_1,\delta,\rho^{(-j)},\tau)$ and $(n-1)$ degrees of freedom. Thereby, the quantity,

$$S_j(\rho_j)=\frac{\rho_j-\widehat{\rho}_j(m,\phi^{p},\mu_1,\delta,\rho^{(-j)})}{S_2(m,\phi^{(p)},\mu_1,\delta,\rho^{(-j)},\tau)},$$

is distributed a posteriori as a conditional Student-t distribution with $(n-1)$ degrees of freedom.

The posterior conditional distributions of τ is expressed as follows. The integration of Equation (5) with respect to δ gives the joint posterior distribution of m, μ_1, $\phi^{(p)}$, $\rho^{(p)}$ and τ by

$$\pi(m,\mu_1,\phi^{(p)},\rho^{(p)},\tau|y) \propto \tau^{\frac{m}{2}-1}\Lambda_1^{-1/2}\left\{\tau SS_1\left(m,\mu_1,\phi^{(p)}\right)+SS_2\left(m,\mu_1,\phi^{(p)},\rho^{(p)}\right)\right\}^{-\frac{(n-1)}{2}}, \tag{8}$$

by application of the Bayes theorem, the conditional posterior distributions of τ is given by

$$\pi\left(\tau|m,\mu_1,\phi_1,\rho,y\right) \propto \tau^{\frac{m}{2}-1}\left\{\tau SS_1(m,\mu_1,\phi^{(p)})+SS_2(m,\mu_1,\phi^{(p)},\rho^{(p)})\right\}^{-\frac{(n-1)}{2}},$$

where $SS_1(m,\mu_1,\phi^{(p)})$ and $SS_2(m,\mu_1,\phi^{(p),\rho^{(p)}})$ are given in Equations (6) and (7), respectively. Given m, μ_1, $\phi^{(p)}$, $\rho^{(p)}$, the quantity

$$F\left(\tau\right)=\tau\frac{SS_1\left(m,\mu_1,\phi^{(p)}\right)/m}{SS_2(m,\mu_1,\phi^{(p)},\rho^{(p)})/\left(n-m-1\right)},$$

is distributed a posteriori as a conditional F distribution with $(m, n-m-1)$ degrees of freedom.

The posterior conditional distribution of ϕ_j, for $j=1,\ldots,p$, is considered as follows. Still, the formula in Equation (4) can be written as

$$\pi(\Phi \mid y) \propto \tau^{\frac{m}{2}-1}\left\{\Lambda_{8j}-\frac{\Lambda_{7j}^2}{\Lambda_{6j}}+\Lambda_6\left(\phi_j-\widehat{\phi}_j\left(m,\mu_1,\delta,\phi^{(-j)},\rho^{(p)}\right)\right)^2\right\}^{-\frac{n}{2}},$$

where

$$\begin{aligned}
\Lambda_{6j} &= \tau\sum_1^m(y_{t-j}-\mu_1)^2+\sum_{m+1}^{m+p}\left(y_{t-j}-\gamma_{t-j}\mu_1-(1-\gamma_{t-j})(\delta+\mu_1)\right)^2\\
&\quad+\sum_{m+p+1}^{n}(y_{t-j}-\delta-\mu_1)^2;\\
\Lambda_{7j} &= \tau\sum_1^m(y_{t-j}-\mu_1)(y_t-\mu_1-\sum_{i\neq j}\phi_i(y_{t-i}-\mu_1))\\
&\quad+\sum_{m+1}^{m+p}(y_{t-j}-\gamma_{t-j}\mu_1-(1-\gamma_{t-j})(\delta+\mu_1))\\
&\quad\Big(y_t-\delta-\mu_1-\sum_{i=1}^{p}\rho_i(y_{t-i}-\gamma_{t-i}\mu_1-(1-\gamma_{t-i})(\delta+\mu_1))\\
&\quad-\sum_{i\neq j}\phi_i(y_{t-i}-\gamma_{t-i}\mu_1-(1-\gamma_{t-i})(\delta+\mu_1))\Big)\\
&\quad+\sum_{m+p+1}^{n}(y_{t-j}-\delta-\mu_1)(y_t-\delta-\mu_1-\sum_{i=1}^{p}\rho_i(y_{t-i}-\delta-\mu_1)-\sum_{i\neq j}\phi_i(y_{t-i}-\delta-\mu_1));
\end{aligned}$$

$$\Lambda_{8j} = \tau \sum_{1}^{m}(y_t - \mu_1 - \sum_{i\neq j}\phi_i(y_{t-i} - \mu_1)^2$$
$$+ \sum_{m+1}^{m+p}\Big(y_t - \delta - \mu_1 - \sum_{i=1}^{p}\rho_i(y_{t-i} - \gamma_{t-i}\mu_1 - (1-\gamma_{t-i})(\delta+\mu_1))$$
$$- \sum_{i\neq j}\phi_i(y_{t-i} - \gamma_{t-i}\mu_1 - (1-\gamma_{t-i})(\delta+\mu_1))\Big)^2$$
$$+ \sum_{m+p+1}^{n}\Big(y_t - \delta - \mu_1 - \sum_{i=1}^{p}\rho_i(y_{t-i} - \delta - \mu_1) - \sum_{i\neq j}\phi_i(y_{t-i} - \delta - \mu_1)\Big)^2;$$

and $\widehat{\phi}_j(m, \mu_1, \delta, \phi^{(j)}, \rho^{(p)}) = \Lambda_7/\Lambda_6$. Following the Bayes theorem, the posterior conditional distribution of ϕ_j, for $j = 1, \ldots, p$, is given by

$$\pi(\phi_j|m, \rho^{(p)}, \mu_1, \phi^{(-j)}, \delta, \tau, y) \propto \left\{1 + \frac{(\phi_j - \widehat{\phi}_j(m, \rho^{(p)}, \mu_1, \phi^{(-j)}, \delta, \tau))^2}{(n-1)S^2_{3j}(m, \rho^{(p)}, \mu_1, \phi^{(-j)}, \delta, \tau)}\right\}^{-\frac{n}{2}},$$

where $S^2_{3j}(m, \rho^{(p)}, \mu_1, \phi^{(-j)}, \delta, \tau) = (\Lambda_{8j} - \Lambda^2_{7j}/\Lambda_{6j})/((n-1)\Lambda_{6j})$. For $j = 1, \ldots, p$, given m, μ_1, $\phi^{(-j)}$, $\rho^{(p)}$, δ, and τ, the conditional posterior distribution of ϕ_j is distributed as a Student-t distribution with location parameter $\widehat{\phi}_j(m, \rho^{(p)}, \mu_1, \phi^{(-j)}, \delta, \tau)$, precision $S_{3j}(m, \rho^{(p)}, \mu_1, \phi^{(-j)}, \delta, \tau)$ and $(n-1)$ degrees of freedom.

The posterior conditional distribution of μ_1 is given next. We can write Equation (4) as

$$\pi(\Phi \mid y) \propto \tau^{\frac{m}{2}-1}\left\{\Lambda_{11} - \frac{\Lambda^2_{10}}{\Lambda_9} + \Lambda_9\left(\mu_1 - \widehat{\mu}_1\left(m, \phi^{(p)}, \rho^{(p)}, \delta, \tau\right)\right)^2\right\}^{-\frac{n}{2}},$$

where

$$\Lambda_9 = m\tau\left(1 - \sum_{i=1}^{p}\phi_i\right)^2 + (n-m)\left(1 - \sum_{i=1}^{p}(\rho_i + \phi_i)\right)^2;$$
$$\Lambda_{10} = \tau(1 - \sum_{i=1}^{p}\phi_i)\sum_{1}^{m}\left(y_t - \sum_{i=1}^{p}\phi_i y_{t-i}\right)$$
$$+(1 - \sum_{i=1}^{p}(\rho_i + \phi_i)\sum_{m+1}^{m+p}\left(y_t - \delta - \sum_{i=1}^{p}(\rho_i + \phi_i)(y_{t-i} - (1-\gamma_{t-i})\delta\right)$$
$$(1 - \sum_{i=1}^{p}(\rho_i + \phi_i))\sum_{m+p+1}^{n}\left(y_t - \delta - \sum_{i=1}^{p}(\rho_i + \phi_i)(y_{t-i} - \delta)\right);$$
$$\Lambda_{11} = \tau\sum_{1}^{m}\left(y_t - \sum_{i=1}^{p}\phi_i y_{t-i}\right)^2 + \sum_{m+1}^{m+p}\left(y_t - \delta - \sum_{i=1}^{p}(\rho_i + \phi_i)(y_{t-i} - (1-\gamma_{t-i})\delta)\right)^2$$
$$\sum_{m+p+1}^{n}\left(y_t - \delta - \sum_{i=1}^{p}(\rho + \phi_i)(y_{t-i} - \delta)\right)^2;$$

and $\widehat{\mu}_1(m, \phi^{(p)}, \rho^{(p)}, \delta, \tau) = \Lambda_{10}/\Lambda_9$. By the Bayes theorem, the posterior conditional distri-

bution of μ_1 is given by

$$\pi(\mu_1|m,\phi^{(p)},\rho^{(p)},\delta,\tau,y) \propto \left\{1+\frac{(\mu_1-\widehat{\mu}_1(m,\phi^{(p)},\rho^{(p)},\delta,\tau))^2}{(n-1)S_4^2(m,\rho^{(p)},\phi^{(p)},\delta,\tau)}\right\}^{-\frac{n}{2}},$$

where $S_4^2(m,\rho^{(p)},\phi^{(p)},\delta,\tau) = (\Lambda_{11}-\Lambda_{10}^2/\Lambda_9)/((n-1)\Lambda_9)$. Given m, $\phi^{(p)}$, $\rho^{(p)}$, δ and τ, the conditional posterior distribution of μ_1 is distributed as a Student-t distribution with location parameter $\widehat{\mu}_1(m,\rho^{(p)},\phi^{(p)},\delta,\tau)$, precision $S_4(m,\rho^{(p)},\phi^{(p)},\delta,\tau)$ and $(n-1)$ degrees of freedom.

The posterior conditional distributions of m is stated next. From the joint posterior distribution of m, μ_1, $\phi^{(p)}$, $\rho^{(p)}$ and τ given in Equation (8), the conditional posterior distributions of m is given by

$$\pi(m|\mu_1,\phi^{(p)},\rho^{(p)},\tau,y) \propto \tau^{\frac{m}{2}-1}\Lambda_1^{-1/2}\left\{\tau SS_1\left(m,\mu_1,\phi^{(p)}\right)+SS_2\left(m,\mu_1,\phi^{(p)},\rho^{(p)}\right)\right\}^{-\frac{(n-1)}{2}},$$

where $SS_1(m,\mu_1,\phi^{(p)})$ and $SS_2(m,\mu_1,\phi^{(p),\rho^{(p)}})$ are given in Equations (6) and (7), respectively.

Remark 1 As the degrees of freedom m and $n-m-1$ of F distribution are greater or equal to 1, this implies that the change point m belongs to $\{1,n-2\}$.

The unconditional posterior distributions of $T(\delta)$, $S_j(\rho_j)$, for $j=1,\ldots,p$, and $F(\tau)$ are given, respectively, by

$$\begin{aligned}\pi(T(\delta)|y) = \sum_m \int_\tau \int_{\rho^{(p)}} \int_{\phi^{(p)}} \int_{\mu_1} \pi(T(\delta)|m,\mu_1,\phi^{(p)},\rho^{(p)},\tau,y)\pi(\mu_1|m,\phi^{(p)},\rho^{(p)},\tau,y) \\ \pi(\phi^{(p)}|m,\rho,\phi^{(p)},\tau,y)\pi(\rho^{(p)}|m,\rho^{(p-1)},\tau,y)\pi(\tau|m,y)\pi(m|y)\mathrm{d}\mu_1\mathrm{d}\phi^{(p)}\mathrm{d}\rho^{(p)}\mathrm{d}\tau,\end{aligned} \tag{9}$$

$$\begin{aligned}\pi(S_j(\rho_j)|y) = \sum_m \int_\tau \int_\delta \int_{\rho^{(j)}} \int_{\phi^{(p)}} \int_{\mu_1} \pi(S_j(\rho_j)|m,\mu_1,\phi^{(p)},\delta,\tau,y)\pi(\mu_1|m,\phi^{(p)},\delta,\tau,y) \\ \pi(\phi^{(p)}|m,\delta,\tau,y)\pi(\rho^{(-j)}|m,\rho^{(p-j)},\delta,\tau,m)\pi(\delta|m,\tau,y) \\ \pi(\tau|m,y)\pi(m|y)\mathrm{d}\mu_1\mathrm{d}\phi^{(p)}\mathrm{d}\rho^{(-j)}\mathrm{d}\delta\mathrm{d}\tau, j=1,\ldots,p,\end{aligned} \tag{10}$$

$$\begin{aligned}\pi(F(\tau)|y) = \sum_m \int_\delta \int_{\rho^{(p)}} \int_{\phi^{(p)}} \int_{\mu_1} \pi(F(\tau)|m,\mu_1,\phi^{(p)},\rho^{(p)},\delta,y)\pi(\mu_1|m,\phi^{(p)},\rho^{(p)},\delta,y) \\ \pi(\phi^{(p)}|m,\rho^{(p)},\delta,y)\pi(\delta|m,\rho^{(p)},y)\pi(\rho^{(p)}|m,\rho^{(p-j)},y)\pi(m|y)\mathrm{d}\mu_1\mathrm{d}\phi^{(p)}\mathrm{d}\rho^{(p)}\mathrm{d}\delta,\end{aligned} \tag{11}$$

where

$$\pi(\rho^{(p)}|\beta,\rho^{(p-1)},y) = \pi(\rho_1|\beta,\rho_2,\ldots,\rho_p,y)\pi(\rho_2|\beta,\rho_1,\rho_3,\ldots,\rho_p,y),\ldots,\pi(\rho_p|\beta,\rho_1,\ldots,\rho_{p-1},y),$$

and $\pi(\rho^{(-j)}|\beta, \rho^{(p-j)}, y) =$

$$\begin{cases} \pi(\rho_2|\beta,\rho_3,\ldots,\rho_p,y)\pi(\rho_3|\beta,\rho_2,\rho_4,\ldots,\rho_p,y),\ldots,\pi(\rho_p|\beta,\rho_2,\ldots,\rho_{p-1},y), & j=1; \\ \pi(\rho_1|\beta,\rho_3,\ldots,\rho_p,y)\pi(\rho_3|\beta,\rho_1,\rho_4,\ldots,\rho_p,y)\ldots\pi(\rho_p|\beta,\rho_1,\rho_3,\ldots,\rho_{p-1},y), & j=2; \\ \vdots & \vdots \\ \pi(\rho_1|\beta,\rho_2,\ldots,\rho_{p-1},y)\pi(\rho_2|\beta,\rho_1,\rho_3,\ldots,\rho_{p-1},y),\ldots,\pi(\rho_{p-1}|\beta,\rho_1,\rho_2,\ldots,\rho_{p-2},y), & j=p. \end{cases}$$

The null hypothesis H_0 can be divided into $p+2$ sub-hypotheses H_{01}: $\delta = \mu_2 - \mu_1 = 0$, H_{02j}: $\rho_j = \phi_j - \psi_j = 0$, and H_{03}: $\tau = \sigma_2^2/\sigma_1^2 = 1$, and H_0 could be rejected if either of these $p+2$ sub-hypotheses is rejected. The separation of the null into several sub-hypotheses would be helpful to determine which parameters have been changed at time m. One defines separately the HPD credible sets of $T(\delta)$, $S_j(\rho_j)$ and $F(\tau)$ based on conditional distributions. The credible set are used to define the unconditional p-value and thereby an unconditional test, the bayesian significance test of change in the parameters of autoregressive time series.

Given m, μ_1, $\phi^{(p)}$, $\rho^{(p)}$ and τ the $(1-\alpha)$-credible set for $T(\delta)$ is defined as

$$C_\delta = \left\{T(\delta) \mid T(\delta)| < t_{\alpha/2}(n-1)\right\},$$

where $t_{\alpha/2}(n-1))$ is the $100(1-\alpha/2)$th quantile of a Student-t distribution with $(n-1)$ degrees of freedom. Hence, given m, μ_1, $\phi^{(p)}$, $\rho^{(p)}$ and τ the decision rule for H_{01} is to reject if $T(0) \in \overline{C_\delta}$, where $\overline{C_\delta}$ is the complement of C_δ.

The unconditional p-value of the hypothesis H_{01} calculated from Equation (9) yields

$$\begin{aligned} P_{\delta=0|y} &= 2\sum_m \int_\tau \int_{\rho^{(p)}} \int_{\phi^{(p)}} \int_{\mu_1} \{1 - \mathcal{T}_{n-1}(|T(0)|)\} \\ &\quad \pi(m,\mu_1,\phi^{(p)},\delta,\rho^{(p)},\tau|y)\mathrm{d}\mu_1\mathrm{d}\phi^{(p)}\mathrm{d}\rho^{(p)}\mathrm{d}\tau, \\ &= 2E_m E_\tau E_{\rho^{(p)}} E_{\mu_1} E_{\phi^{(p)}} \{1 - \mathcal{T}_{n-1}(|t(0)|)\}, \end{aligned} \tag{12}$$

The sub-hypothesis H_{01} is rejected unconditionally at α significance level if $P_{\delta=0|y} < \alpha$.

The unconditional p-value of the hypothesis H_{02j}, for $j = 1,\ldots,p$, calculated from Equation (10), is given by

$$\begin{aligned} P_{\rho_j=0|y} &= 2\sum_m \int_\tau \int_{\mu_1} \int_\delta \int_{\rho^{(-j)}} \int_{\phi^{(p)}} \int_{\mu_1} \{1 - \mathcal{T}_{n-1}(|S_j(0)|)\} \\ &\quad \pi(m,\mu_1,\phi^{(p)},\delta,\rho^{(-j)},\tau|y)\mathrm{d}\mu_1\mathrm{d}\phi^{(p)}\mathrm{d}\rho^{(-j)}\mathrm{d}\delta\mathrm{d}\tau, \\ &= 2E_m E_\tau E_{\mu_1} E_\delta E_{\rho^{(-j)}} E_{\phi^{(p)}} \{1 - \mathcal{T}_{n-1}(|t(0)|)\}, \end{aligned} \tag{13}$$

where $\mathcal{T}_{n-1}$ is the cumulative distribution function of the Student-t distribution with $(n-1)$ degrees of freedom. For $j = 1,\ldots,p$, the sub-hypothesis H_{02j} is rejected unconditionally at α significance level if $P_{\rho_j=0|y} < \alpha$. Where, the sub-hypothesis H_{02} is rejected unconditionally at α significance level if

$$P_{\rho=0|y} := \min_{1\le j\le p}\{P_{\rho_j=0}|y\} < \alpha.$$

Likewise, the unconditional p-value of H_{03} calculated from Equation (11) is stated as

$$\begin{aligned} P_{\tau=1|y} &= 2\sum_{m}\int_{\rho^{(p)}}\int_{\phi^{(p)}}\int_{\mu_1}\{1-\mathcal{F}_{m,n-m-1}[\max(F(1),1/F(1))]\} \\ &\quad \pi\left(m,\mu_1,\phi^{(p)},\rho^{(p)}|y\right)\mathrm{d}\mu_1\mathrm{d}\phi_1\mathrm{d}\rho, \\ &= 2E_m E_{\rho^{(p)}}E_{\phi^{(p)}}E_{\mu_1}\{1-\mathcal{F}_{m,n-m-1}[\max(F(1),1/F(1))]\}, \end{aligned} \tag{14}$$

where $\mathcal{F}_{m,n-m-1}$ is the cumulative distribution function of the Fisher F distribution with $(m, n-m-1)$ degrees of freedom. The sub-hypothesis H_{03} is rejected unconditionally at α significance level if $P_{\tau=1|y} < \alpha$. Therefore, the null hypothesis H_0 well be rejected unconditionally at α significance level if $\min\{P_{\delta=0|y}, P_{\rho=0|y}, P_{\tau=0|y}\} < \alpha$, and thus define the bayesian significance test of change in the parameters of autoregressive time series AR(p) of known order p. The test allows to test the change in the $p+2$ parameters of the AR(p) model in an individual way.

The notations E_{μ_1}, $E_{\phi^{(p)}}$, $E_{\rho^{(p)}}$, $E_{\rho^{(-j)}}$, E_δ, E_τ and E_m are the expectations taken with respect to μ_1, $\phi^{(p)}$, $\rho^{(p)}$, $\rho^{(-j)}$, δ, τ, and m, respectively.

The quantities given in Equations (12), (13) and (14) are evaluated numerically by the Gibbs sampler algorithm using the conditional posterior distributions given in Section 3.

The Gibbs sampler was introduced by Geman and Geman (1984) as a way of simulating from high-dimensional complex distributions arising in image restoration, is a Markovian updating scheme enabling one to obtain samples from a joint distribution via iterated sampling from full conditional distributions. Although most applications of Gibbs sampler have been in Bayesian models, it is also extremely useful in classical (likelihood) calculations Casella and George (1992). In Bayesian framework, the common objective is to produce posterior densities for, or estimate of, parameters of interest. The algorithm is also very useful for the calculation of high dimensional integrals. Therefore, the use of Gibbs sampler algorithm allows us to reduce in a huge way the calculation of complex high-dimensional integration in Equations (12), (13) and (14). Detailed investigation of the Gibbs sampler applied to general Bayesian calculation is given by Gelfand and Smith (1990), Gelfand et al. (1990) and Gelfand (2000).

4. Simulation results

In this section we conduct a set of controlled simulation studies to evaluate the performance of the proposed test presented in Section 3. We simulated a sample from the model given in Equation (1) with $p=1$, $n=200$, $m=100$, $\mu_1=0.0$, $\mu_2=0.5$, $\phi_1=0.3$, $\phi_2=-0.2$, $\sigma_1^2=1.0$ and $\sigma_2^2=0.5$. The assumed values for y_0 is 1. From these observations, by the application of the Gibbs sampler algorithm with 10,000 repetitions, we approximate the posterior density of the change point m, the posterior density of δ, the posterior density of ρ, of the variance ratio τ and the unconditional p-values for the hypothesis H_{01}: $\delta=0$, H_{012}: $\rho=0$ and H_{03}: $\tau=1$. The results are given in Tables 1-3.

Tables 1 and 2 list the posterior density of the change point at values around the true value of m and the unconditional p-values for the sub-null-hypotheses H_{01}, H_{02} and H_{03}. From Table 1, we can readily see that the posterior mode is equal to the true value of the change point m. Based on the unconditional p-values given in Table 2 the no change in δ, ρ and τ is obviously rejected at 1% significance levels, respectively.

Tables 3 summarize the posterior estimates of the parameters m, δ, ρ and τ. The estimates for the parameters of the series in Figure 1 are generally close to the true values. Also, we

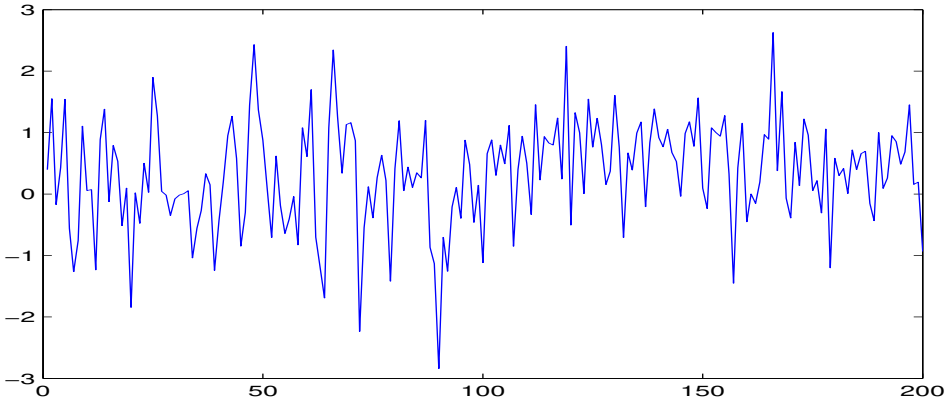


Figure 1. Simulated observations y_t.

Table 1. The posterior density of m.

m	$\pi(m\|y)$	m	$\pi(m\|y)$
89	0.0000	100	0.2321
90	0.0017	101	0.1396
91	0.0040	102	0.0782
92	0.0304	103	0.0541
93	0.0404	104	0.0311
94	0.0368	105	0.0192
95	0.0551	106	0.0120
96	0.0313	107	0.0378
97	0.0214	108	0.0274
98	0.0363	109	0.0154
99	0.0368	110	0.0107

Table 2. The unconditional p-values of the hypothesis H_{01}, H_{02} and H_{03}.

Sub-null-hypothesis	H_{01}	H_{02}	H_{03}
p-values	4.8452×10^{-5}	0.0027	0.0017

clearly see that, all the 95% HPD sets of the parameters contain the true value of all the parameters.

Table 3. Posterior estimates of the parameters m, δ, ρ and τ.

Parameters	True values	Median	Mean (SD)	2.5%	97,5%
m	100	100	100.56(4.9418)	92	111
$\delta = \mu_2 - \mu_1$	0.5	0.4872	0.4869(0.1115)	0.2709	0.7098
$\rho = \phi_2 - \phi_1$	−0.5	−0.4230	−0.4239(0.1368)	−0.6897	−0.1515
$\tau = \sigma_2^2/\sigma_1^2$	0.5	0.5023	0.5138(0.1129)	0.3291	0.7649

Figures 2-5 give the posterior distribution of the parameters m, δ, ρ and τ. They indicate that the posterior mode is around the true values of the parameters. Thus, an estimate of the true values of the parameters is given by the posterior mode of the respective posterior distributions.

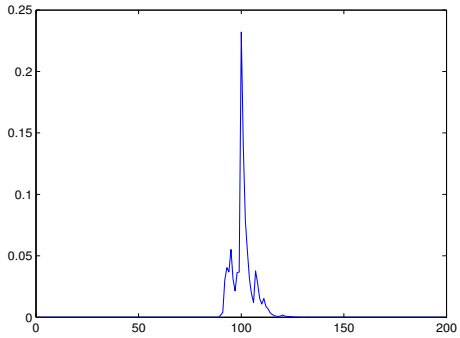


Figure 2. Posterior density function of the change point m.

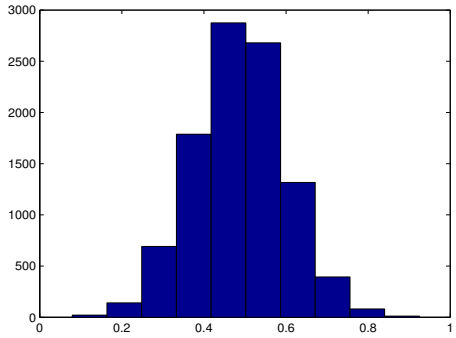


Figure 3. Histogram of posterior distribution of the parameter δ.

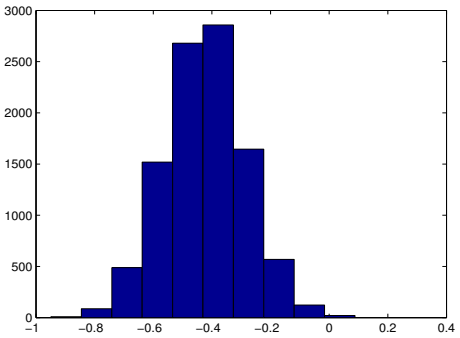


Figure 4. Histogram of posterior distribution of the parameter ρ.

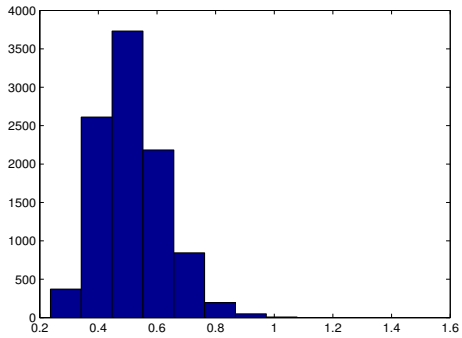


Figure 5. Histogram of posterior distribution of the parameter τ.

Furthermore, Table 4 presents the unconditional p-values of the sub-null-hypotheses H_{01}, H_{02} and H_{03} for $n = 200$, $m = 100$ and different values of the parameters μ_1, μ_2, ϕ_1, ϕ_2, σ_1^2 and σ_2^2. Several cases are considered, stability in one of the parameters and change in the other two and in the last case the three parameters are stable (without change). The results show that the p-values of sub-hypotheses corresponding to the stable parameters do not allow to reject the corresponding sub-hypothesis. While, for the other sub-hypotheses where the parameters exhibiting changes, the corresponding p values make it possible to reject these sub-hypotheses at 5% significance level. For example, with $\mu_1 = 0.0$, $\mu_2 = 0.5$, $\phi_1 = 0.3$, $\phi_2 = 0.3$ and $\sigma_1^2 = 1.0$, $\sigma_2^2 = 0.5$, the p-values $P_{\delta=0|y}$ and $P_{\tau=1|y}$ are respectively 0.0187 and 0.0146. Thus, the sub-hypotheses H_{01} and H_{03} are rejected at 5% significance level. The p-value $P_{\rho=0|y}$ is 0.4134, therefore, the sub-hypothesis H_{02} cannot be rejected. Note that the parameter ϕ is stable, that is, $\rho = \phi_2 - \phi_1 = 0$.

To study the performance of the Bayesian significance test for detecting structural changes in the parameters of autoregressive AR(p), we simulated 1000 samples from the model given in Equation (1) with $p = 1$ and different values of n, m, μ_1, μ_2, ϕ_1, ϕ_2, σ_1^2 and σ_2^2 and we computed the rejection rates (the number of times the hypothesis is rejected divided by the total number of samples) of sub-hypotheses H_{01}, H_{02} and H_{03} at 5% significance level. The results are obtained by Gibbs sampler algorithm with 5000 repetitions and are given in Table 5.

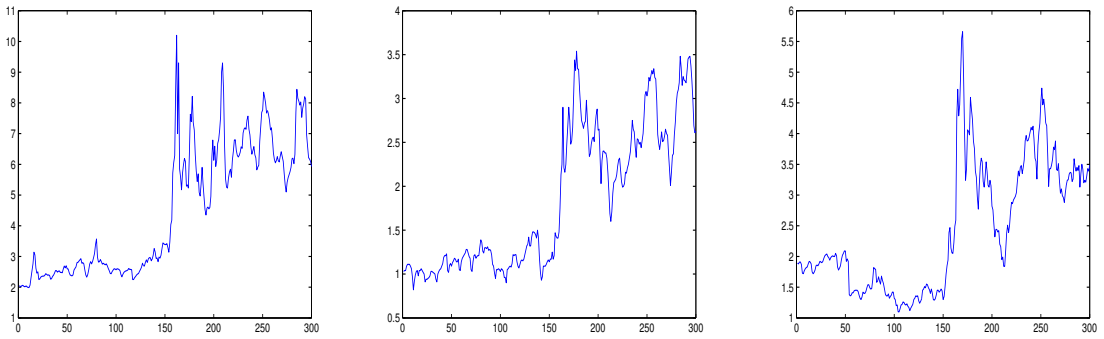
Table 5 illustrates that, for $n = 100$ and $m = 50$, the rejection rates of sub-hypotheses H_{01}, H_{02} and H_{03} are more than 60% at 5% level when the parameter exhibits a change, while it is only 6.6% when the parameter is stable (without change). For example, for the

set of parameters $\mu_1 = 0.0$, $\mu_2 = 0.5$, $\phi_1 = 0.3$, $\phi_2 = 0.3$ and $\sigma_1^2 = 1.0$, $\sigma_2^2 = 0.5$, the rejection rate of the sub hypothesis H_{01} is 0.630, for H_{02} is 0.004 and for H_{03} is 0.711. We note that the parameter ϕ is stable. For the last set of parameters, $\mu_1 = 0.0$, $\mu_2 = 0.0$, $\phi_1 = 0.3$, $\phi_2 = 0.3$ and $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.5$, the three parameters are assumed to be stable, the rejection rate of sub hypotheses H_{01}, H_{02} and H_{03} are respectively 0.006, 0.009 and 0.066. Therefore, the test detects well the autoregressive parameters that are subjects to a change.

It can be seen that the rejection rates of sub-hypotheses H_{01}, H_{02} and H_{03} of AR(1) model increases with the sample size. For $n = 200$, $m = 100$, $\mu_1 = 0.0$, $\mu_2 = 0.5$, $\phi_1 = 0.3$, $\phi_2 = 0.3$ and $\sigma_1^2 = 1.0$, $\sigma_2^2 = 0.5$, the rejection rate of the sub-hypotheses H_{01}, H_{02} and H_{03} of AR(1) are respectively 0.810, 0.018 and 0.877. However, they are respectively only 0.220, 0.007 and 0.516 for $n = 50$ and $m = 25$. Therefore, the sample size has a positive impact on the Bayesian significance test of change in the parameters of autoregressive time series models.

5. Application

In this section, we illustrate our test procedures using three data sets, which are the monthly average soybean, corn and wheat prices achieved by farmers in Illinois from one January 1960 to one December 1984. The prices are given in dollars per bushel. The price y_t is observed each month from one January 1960 until one December 1984 with sample of 300 observations. Data used in this analysis cab be found in (https://farmdoc.illinois.edu/decision-tools/illinois-average-farm-price-received-database). The sample size is 300. The series are plotted in Figures 6 (a)-(c), Berkes et al. (2011) study two real data sets. The first sample consists of monthly average corn prices and the second sample consists of monthly average soybean prices achieved by farmers in Illinois from January 1960 to November 2008. The results of their statistical test indicate that the changes from an AR(1) to a threshold AR(1) occurred around July 1971 (Corn) and October 1974 (Soybeans).



(a) (b) (c)

Figure 6. Monthly average for corn prices (a); Soybean prices (b) and wheat prices (c) from 1960 to 1984.

We are interested of wither there is any evidence for the existence of a change in the parameter of AR(1) model. A visual inspection of this series in Figures 6(a)-(c) seem to suggest that there might be a change in the parameters of the series. By application of the Gibbs sampler algorithm with 10,000 repetitions we approximate the unconditional p-values for the hypothesis $H_{01} : \delta = 0$, $H_{012} : \rho = 0$ and $H_{03} : \tau = 1$ and the posterior estimates of the change point m. The results are given in Tables 6.

Table 4. The unconditional p-value of H_{01}, H_{02} and $H_{0"}$ with different values of $\delta = \mu_2 - \mu_1$, $\rho = \phi_2 - \phi_1$ and $\tau = \sigma_2^2/\sigma_1^2$.

Parameters / p-values	$\mu_1 = 0.0$, $\mu_2 = 0.5$ $\phi_1 = 0.3$, $\phi_2 = -0.2$ $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.5$	$\mu_1 = 0.0$, $\mu_2 = 0.5$ $\phi_1 = 0.3$, $\phi_2 = 0.3$ $\sigma_1^2 = 1.0$, $\sigma_2^2 = 0.5$	$\mu_1 = 0.0$, $\mu_2 = 0.0$ $\phi_1 = 0.3$, $\phi_2 = -0.2$ $\sigma_1^2 = 1.0$, $\sigma_2^2 = 0.5$	$\mu_1 = 0.0$, $\mu_2 = 0.0$ $\phi_1 = 0.3$, $\phi_2 = 0.3$ $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.5$
$P_{\delta=0\|y}$	$4.38.10^{-6}$	0.0187	0.3804	0.3919
$P_{\rho=0\|y}$	0.0029	0.4134	0.0551	0.3728
$P_{\tau=1\|y}$	0.75086	0.0146	0.0211	0.5328

Table 5. Rejection rate of subnull H_{01}, H_{02} and $H_{0"}$ at 5% level for 1000 samples with different values of n, m, $\delta = \mu_2 - \mu_1$, $\rho = \phi_2 - \phi_1$ and $\tau = \sigma_2^2/\sigma_1^2$.

	Subnull / Parameters	$\mu_1 = 0.0$, $\mu_2 = 0.5$ $\phi_1 = 0.3$, $\phi_2 = -0.2$ $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.5$	$\mu_1 = 0.0$, $\mu_2 = 0.5$ $\phi_1 = 0.3$, $\phi_2 = 0.3$ $\sigma_1^2 = 1.0$, $\sigma_2^2 = 0.5$	$\mu_1 = 0.0$, $\mu_2 = 0.0$ $\phi_1 = 0.3$, $\phi_2 = -0.2$ $\sigma_1^2 = 1.0$, $\sigma_2^2 = 0.5$	$\mu_1 = 0.0$, $\mu_2 = 0.5$ $\phi_1 = 0.3$, $\phi_2 = -0.2$ $\sigma_1^2 = 1.0$, $\sigma_2^2 = 0.5$	$\mu_1 = 0.0$, $\mu_2 = 0.0$ $\phi_1 = 0.3$, $\phi_2 = 0.3$ $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.5$
$n = 50$	$H_{01} : \delta = 0$	0.765	0.220	0.000	0.730	0.003
$m = 25$	$H_{02} : \rho = 0$	0.386	0.007	0.393	0.393	0.001
	$H_{03} : \tau = 1$	0.095	0.443	0.516	0.537	0.052
$n = 100$	$H_{01} : \delta = 0$	0.997	0.630	0.000	0.997	0.006
$m = 50$	$H_{02} : \rho = 0$	0.797	0.004	0.770	0.787	0.009
	$H_{03} : \tau = 1$	0.159	0.711	0.825	0.846	0.066
$n = 200$	$H_{01} : \delta = 0$	1	0.810	0.003	0.998	0.009
$n = 100$	$H_{02} : \rho = 0$	0.987	0.018	0.810	0.921	0.000
	$H_{03} : \tau = 1$	0.210	0.865	0.877	0.944	0.016

Table 6. Unconditional p-values of the hypothesis H_{01}, H_{02} and H_{03} and posterior mode and median of change point m for monthly average for soybeans, for corn and for wheat prices.

Dataset	mode	median	$P_{\delta=0\|y}$	$P_{\rho=0\|y}$	$P_{\tau=1\|y}$
Soybean	154	155	0.2001	0.0866	0.0220
Corn	155	155	0.3641	0.1066	0.0390
Wheat	163	163	0.2832	0.2809	0.0014

Table 6 shows some numerical results of the series of monthly average soybeans, corn and wheat prices from one January 1960 to one December 1984. Posterior mode of the change point m indicates that the changes occurred at time $m = 154$, that corresponds to around October 1972 for Soybeans, at time $m = 155$, that corresponds to November 1972 for Corn and at time $m = 163$, that corresponds to around July 1973 for Wheat. As, the smaller the p-value, more the strength of the evidence against H_0 is significant, the values of the p-values indicate that there is evidence against the equality of the variances of the three series of observations. Thus, the unconditional p-values $P_{\delta=0|y}$, $P_{\rho=0|y}$ and $P_{\tau=1|y}$ of the hypotheses H_{01}, H_{02} and H_{03}, respectively, indicate that the no change in the variance of the series of Saybeans, Corn and Wheat is rejected at 5% significance level. While, the no change in the mean cannot be rejected even at 20% significance level for the three crops. For the change in the autocorrelation coefficient it can be rejected at 10% significance level for Saybeans and it can hardly be rejected at 10% significance level for Corn, and cannot be rejected even at 20% significance level for Wheat. Consequently, the results in Table 6 indicate that the prices of Soybeans, Corn and wheat have undergone a significant variation in the variance parameter since October 1972 for Soybeans, since November 1972 for Corn and since July 1973 for Wheat. Period which corresponds to the beginning of the world food crisis of the 1970s (FAO (2009)), a time from mid-1972 to mid-1975 (Gerlach (2015)).

6. Conclusions, limitations, and future research

In this paper, we have investigated a Bayesian detection of change in the parameters of an autoregressive process of known order p. The model is subjected to a change in $p + 2$ parameters, the mean, the variance of the error terms and the p autoregressive parameters at an unknown time point. We derived the conditional posterior distributions of the change point, of the magnitude of the shift in the mean, of the magnitude of the shift in the autocorrelation coefficients and of the variance ratio. An unconditional Bayesian significance test of change based on the calculation of the p-values is determined. The test detects separately the autoregressive parameters which are subject to a change at an unknown time m. The Gibbs sampler algorithm is employed to estimate the model parameters. The performance of the test has been investigated on simulated and real data sets. We showed how inferences can be made readily by using the Bayesian significance test based on the highest posterior density credible sets for detecting a change of an individual parameter of autoregressive models. Also, we have showed the impact of the sample size on the Bayesian significance test of change. We have illustrated the application of the methods using three real datasets available in the literature. The datasets are the monthly average soybean, corn and wheat prices achieved by farmers in Illinois from one January 1960 to one December 1984. Results obtained report the existence of a change point in all three datasets. The change points obtained correspond exactly to the beginning of the food crisis which occurred in the early 1970s. A possible limitation of the adopted approach might be associated with the estimation of all the parameters of the model, a similar approach could be adopted to estimate all the parameters to performance residual analysis. Moreover, it would be interesting to extend the study to examine the problem of multiple structural change points and to study the case where the order of the autoregressive model is unknown.

Author contributions Conceptualization; data curation; formal analysis; investigation; methodology; software; supervision.validation; visualization; writing-original draft preparation; and writing-review and editing: A.S. The author (A.S.) has read and agreed the published version of the paper.

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

Funding This work was partially supported by the Ministry of Higher Education and Scientific Research of Algeria (MESRS) and General Direction of Scientific Research and Technological Development (DGRSDT) through PRFU Project (ref: C00L03UN010120210001).

Conflicts of interest The author declares no conflict of interest.

References

Basseville, M. and Nikiforov, I.V., 1993. Detection of abrupt changes: theory and application). Prentice Hall, New York, USA.

Bauwens, L., Dufays, A., and Rombouts, J.V.K., 2014. Marginal likelihood for Markov-switching and change-point GARCH models. Journal of Econometrics, 178, 508–522.

Berkes, I., Horváth, L., Ling, S., and Schauer, J., 2011. Testing for structural change of AR model to threshold AR model. Journal of Time Series Analysis, 32, 547–565.

Brown, R. L., Durbin, J., and Evans, J.M., 1975. Techniques for testing the constancy of regression relationships over time (with discussion). Journal of the Royal Statistical Society A, 138, 149–163.

Casella, G. and George, E.I., 1992. Explaining the Gibbs sampler. The American Statistician, 46, 167-1-74.

Cheon, S. and Kim, J., 2014. A Bayesian structural-change analysis via the stochastic approximation Monte Carlo and Gibbs sampler. Journal of Statistical Computation and Simulation, 84, 1444–1470.

Chernoff, H. and Zacks, S., 1964. Estimating the current mean distribution which is subjected to change in time, The Annals of Mathematical Statistics, 35, 999–1018.

Davis, R.A., Huang, D., and Yao, Y.C., 1995. Testing for a change in the parameter value and order of an autoregressive model, The Annals of Statistics, 23, 282–304.

Eastwood, V.R., 1993. Some nonparametric methods for changepoint problems. Canadian Journal of Statistics, 2, 209–222.

FAO, 2009. The State of Agricultural Commodity Markets. Food and Agricultural Organisation, Ginebra.

Fan, Z., Dror, R.O., Mildorf, T.J., Piana, S., and Shaw, D.E. 2015. Identifying localized changes in large systems: Change-point detection for biomolecular simulations. Proceedings of the National Academy of Sciences, 112, 7454–7459.

Fan, T.H. and Chen, W.C. (2005). Bayesian change points analysis on the seismic activity in northeastern Taiwan. Journal of Statistical Computation and Simulation, 75, 857–868.

Gamage, R.D.P. and Ning, W., 2021. Empirical likelihood for change point detection in autoregressive models. Journal of the Korean Statistical Society, 50, 69–97.

Geman, S. and Geman, D., 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721–741.

Gelfand, A.E., S.E. Hills, A. Racine-Poon, and A.F.M. Smith, 1990. Illustration of Bayesian inference in normal data models using Gibbs sampling, Journal of the American Statistical Association, 85, 972–985.

Gelfand, A.E. and A.F.M. Smith, 1990. Sampling-based approaches to calculate marginal densities, Journal of the American Statistical Association, 85 398–409.

Gelfand, A.E. Gibbs sampling, 2000. Journal of the American Statistical Association, 95, 1300–1304.

Georgescu, V., 2012. Online change-point detection in financial time series: challenges and experimental evidence with frequentist and Bayesian setups. Methods For Decision Making in an Uncertain Environment, 2012, 131–145.

Gerlach, C., 2015. Famine responses in the world food crisis 1972-5 and the World Food Conference of 1974. European Review of History, 22, 929–939.

Gombay, E., 2008. Change detection in autoregressive time series. Journal of Multivariate Analysis 99, 451–464.

Gombay, E. and Horvath, L., 1999. Change-points and bootstrap. Environmetrics, 10, 725–736.

Guo, L. and Modarres, R., 2020. Nonparametric change point detection for periodic time series. Canadian Journal of Statistics, 48, 518–534.

Hahn, G., Fearnhead, P., and Eckley, I.A., 2020. Bayes Project: Fast computation of a projection direction for multivariate changepoint detection. Statistics and Computing, 30, 1691–1705.

Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, Princeton, NY, USA.

Hudecová, S., 2013. Structural changes in autoregressive models for binary time series, Journal of Statistical Planning and Inference, 143, 1744–1752.

Husková, M., Praskova Z., and Steinebach, J., 2007. On the detection of changes in autoregressive time series. I. Asymptotics. Journal of Statistical Planning and Inference, 137, 1243–1259.

Husková, M., Kirch, C., Praskova, Z., and Steinebach, J., 2008. On the detection of changes in autoregressive time series. II. Resampling procedures. Journal of Statistical Planning and Inference, 138, 1697–1721.

Jandhyala V., Fotopoulos S., MacNeill I. and Liu P., 2013. Inference for single and multiple change-points in time series. Journal of Time Series Analysis, 34, 423–446.

Jandhyala, V.K., Liu, P., Fotopoulos, S.B., and MacNeill, I.B., 2014. Change-point analysis of polar zone radiosonde temperature data. Journal of Applied Meteorology and Climatology, 53, 694–714.

Jani, P.N. and Pandya, M., 1999. Bayes estimation of shift point in left truncated exponential sequence. Communications in Statistics: Theory and Methods, 28, 2623–2639.

Kander, A. and Zacks, S., 1966. Test procedure for possible change in parameters of statistical distributions occurring at unknown time point, The Annals of Mathematical Statistics, 37, 1196–1210.

Kezim, B. and Abdelli, Z., 2004. A Bayesian Analysis of a Structural Change in the Parameters of a Time Series, Communications in Statistics: Theory and Methods, 33, 1863–1876.

Kim D., 1991 A Bayesian significance test of the stationarity of regression parameters. Biometrika, 78, 667–675.

Kim, H.J., 1996. Change-point detection for correlated observations, Statistica Sinica, 6, 275–287.

Kim, H.J. and Siegmund, D., 1989. The likelihood ratio test for a change-point in simple linear regression. Biometrika, 76, 409–423.

Lévy-Leduc, C. and Roueff, F., 2009. Detection and localization of change-points in high-dimensional network traffic data. The Annals of Applied Statistics, 3, 637–662.

Ng, V.M., 1990. A Bayesian analysis of linear models exhibiting changes in mean and precision at an unknown time point. Communications in Statistics: Theory and Methods, 19, 111–120.

Montoril, M.H. and da Silva Ferreira, C., 2018. On the estimation of change points in the Beer-Lambert law problem. Chilean Journal of Statistics, 9(1), 19–32.

Page, E.S., 1954. Continuous inspection schemes. Biometrika, 41, 100–115.

Page E.S., 1955. A test for change in a parameter occurring at an unknown point. Biometrika, 42, 523–527.

Pan, J., Xia, Q., and Liu, J., 2017. Bayesian analysis of multiple thresholds autoregressive model. Computational Statistics, 32, 219–237.

Reinsel, G.C., 1997. Elements of Multivariate Time Series Analysis. Springer, New York, USA.

Romano, G., Rigaill, G., Runge, V., and Fearnhead, P., 2021. Detecting abrupt changes in the presence of local fluctuations and autocorrelated noise. Journal of the American Statistical Association, pages in press, available at https://doi.org/10.1080/01621459.2021.1909598.

Sáfadi, T. and Morettin, P.A., 2000. Bayesian analysis of threshold autoregressive moving average models. Sankhya B, 62, 353–371.

Sen, A. and Srivastava, M.S., 1975. Some one-sided tests for change in level. Technometrics,17, 61–64.

Shah, J.B. and Patel, M.N., 2007. Bayes estimation of shift point in geometric sequence. Communications in Statistics: Theory and Methods, 36, 1139–1151.

Slama, A., 2014. A Bayesian significance test of change for correlated observations, Discussiones Mathematicae, Probability and Statistics, 34, 51–52.

Slama, A. and Saggou, H., 2017. A Bayesian analysis of a change in the parameters of autoregressive time series. Communications in Statistics: Simulation and Computation, 46, 7008–7021.

Tartakovsky, A., Rozovskii, B., Blazek, R., and Kim, H., 2006. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch sequential change-point detection methods. IEEE Transactions on Signal Processing, 54 3372–3382.

Truong, C., Oudre, L., and Vayatis, N., 2020. Selective review of offline change point detection methods. Signal Processing, 167, 107299.

Venkatesan, D. and Arumugam, P., 2007. Bayesian analysis of structural changes in autoregressive models, American Journal of Mathematical and Management Sciences, 27, 153–162.

Wang, T., Tian, W., Ning, W., 2020. Likelihood ratio test change-point detection in the skew slash distribution. Communications in Statistics: Simulation and Computation, pages in press available at https://doi.org/10.1080/03610918.2020.1755869.

Worsley, K.J.,1983. The power of likelihood ratio and cumulative sum tests for a change in a binomial probability. Biometrika, 70, 455–464.

Worsley, K.J., 1986. Confidence regions and tests for a change-point in a sequence of exponential family random variables. Biometrika, 73, 91–104.

Chilean Journal of Statistics
Volume 13, Number 1 (April 2022), Pages 67–90
DOI: 10.32372/ChJS.13-01-04

Statistical Modeling
Research Paper

On the Topp-Leone log-normal distribution: Properties, modeling, and applications in astronomical and cancer data

Christophe Chesneau[1,*], Muhammed Rasheed Irshad[2], Damodaran Santhamani Shibu[3], Soman Latha Nitin[3], and Radhakumari Maya[4]

[1]Department of Mathematics, Université de Caen Basse-Normandie, Caen, France,
[2]Department of Statistics, Cochin University of Science and Technology, Kerala, India,
[3]Department of Statistics, University College, Kerala, India,
[4]Department of Statistics, University of Kerala, Kerala, India

(Received: 03 January 2022 · Accepted in final form: 16 February 2022)

Abstract

In the realm of astronomy, the two-parameter log-normal distribution has ominous implications. In this article, we propose a new version of the two-parameter log-normal distribution with an application to astronomical data. More precisely, a new modulating parameter is added to the two-parameter log-normal distribution through the use of the Topp-Leone generator of distributions. The moments, quantile function, several reliability measures, and other significant aspects of the proposed distribution are investigated. The maximum likelihood approach and a Bayesian technique are both utilized to estimate the unknown parameters. In addition, we present a parametric regression model and a Bayesian regression method. A simulation study is carried out to assess the long-term performance of the estimators of the distribution parameters. Two real datasets are employed to show the applicability of this new distribution. The efficiency of the newly added parameter is tested by utilizing the likelihood ratio test. The parametric bootstrap approach is also utilized to determine the adequacy of the suggested model for the datasets.

Keywords: Bayesian estimation · bootstrapping · maximum likelihood estimation · regression · simulation.

Mathematics Subject Classification: Primary 60E05 · Secondary 62F15.

1. Introduction

In practice, the two-parameter log-normal (LN) distribution can be used to fit empirical data in a variety of ways. This is especially true in the field of astronomy. Studies and research have established evidence of an LN distributional characteristic for very high energy emission of light curves from galaxies; for more information, see Abdalla et al. (2017). It is also worth noting that the distribution of galaxy density contrast, which is a parameter utilized in galaxy formation to indicate where there are local enhancements in matter

*Corresponding author. Email: christophe.chesneau@unicaen.fr, christophe.chesneau@gmail.com

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

http://www.soche.cl/chjs

density, is roughly an LN distribution; whether the distribution of mass fluctuation from the Dark Energy Survey, which is derived from weak lensing convergence in a similar way to convex glass lenses, is an LN distribution is less clear. It was first identified by Hubble (1934) that the distribution of galaxies in angular cells on the celestial sphere is well predicted by an LN distribution. Again, recently, Shah et al. (2018) and Shah et al. (2020) elaborately highlighted the considered LN distributional behavior of the gamma-ray (γ-ray) flux distribution on the brightest blazars, which are observed by the Fermi-LAT, a space observatory's large area telescope (LAT) being used to perform γ-ray. For more applications of the LN distribution in the area of astrophysics and cosmology, one can go through the articles by Bernardeau and Kofman (1994), Blasi et al. (1999) and Parravano et al. (2012).

Fundamental distributions occasionally fail to adequately characterize and anticipate the vast majority of real-world datasets resulting from complicated processes. Because the quality of statistical analysis results is strongly dependent on the assumed model, choosing an adaptive model for data analysis is critical. Therefore, more allied distributions must be found in order to obtain better quality and more accurate results. Since the LN distribution has superior importance in the field of astronomy, it is inevitable to derive new generalized versions of the LN distribution, not only for modeling astronimocal data but also for the variety of datasets from other study areas where the LN distribution has the best fit. Note that the LN distribution has been utilized in a range of domains which includes most of the applied areas such as economics, sociology, biology, and meteorology, to name just a few; for more details, see Jobe et al. (1989).

There has recently been a boom in interest in the art of adding parameters to well-known existing distributions in order to obtain diverse forms of hazard rate functions (HRFs) for use in various real-life circumstances, as well as for evaluating data with a high degree of skewness and kurtosis. Several researchers have started to build families of distributions based on conventional distributions or using different methodologies in order to generalize any baseline distribution; for example, see Afify (2017). In this article, using a flexible generalization technique that includes an additional shape parameter, we investigate a novel lifetime distribution that is also a generalized version of the two-parameter LN distribution. The aim is to uncover some of the suggested model's statistical features and apply them to real-world data. The main motivations for developing this lifetime model are: (i) to propose a new flexible version of the LN distribution that can be used, particularly to analyze astronomical data, because the LN distribution has eminent superiority in the field of astronomy, as well as the ability to be applied to a broader class of reliability problems, (ii) to extend both the LN and Topp-Leone distributions, and (iii) to investigate some additional shapes of the HRF.

The remaining sections of the article are structured as follows. Section 2 reveals our distribution methodology. The specification of the new distribution is presented in Section 3. In Section 4, its moments are calculated. The quantile function (QF) and some of its associated measures are obtained in Section 5. The various functions and moments related to the reliability measures are discussed in Section 6. In Section 7, the maximum likelihood (ML) and Bayesian estimation techniques are employed to estimate the unknown parameters of the new model. Also, a parametric bootstrap method of simulation using the ML estimates is presented in Section 8. A parametric regression model associated with the new distribution is defined in Section 9. Again, a Bayesian regression method is presented in Section 10. A simulation study is proposed in Section 11 to analyze the performance of the ML estimators of the parameters. In Section 12, one univariate uncensored real dataset based on an astronomical study, and one censored real dataset based on a cancer study are evaluated to depict the potential of the new distribution over competing distributions. The final concluding remarks are given in Section 13.

2. Construction of the distribution

A simple bounded J-shaped distribution that has attracted various statisticians as an alternative to uniform(0,1) and beta distributions was proposed by Topp and Leone (1955). It is called the Topp-Leone (TL) distribution. The cumulative distribution function (CDF) and probability density function (PDF) of the TL distribution are respectively stated as

$$F_{\mathrm{TL}}(x;\alpha) = [x(2-x)]^{\alpha},$$

and

$$f_{\mathrm{TL}}(x;\alpha) = 2\alpha x^{\alpha-1}(1-x)(2-x)^{\alpha-1}, \quad \alpha > 0,\ x \in (0,1).$$

It is worth mentioning that the TL distribution has a bathtub shaped HRF for all $\alpha < 1$. Later, Sangsanit and Bodhisuwan (2016) introduced the Topp-Leone generalized exponential distribution, using the TL distribution as a generator distribution with application to the maximum stress per cycle and breaking stress of carbon fiber datasets. Now, we consider the method for generating new distributions, called the TX family, proposed by Alzaatreh et al. (2013). The essence of the TX family is presented below. Let X be a continuous baseline random variable with CDF F_X, and T be a continuous generator random variable of a distribution with support on $[a,b]$ and CDF Ψ. Then, the CDF of the TX family is given by

$$F_{\mathrm{TX}}(x) = \Psi[W(F_X(x))], \tag{1}$$

where $W(F_X(x)) \in [a,b]$ is differentiable and monotonically non-decreasing.

Considering the immense applicability of the TL and LN distributions, we propose to apply both the distributions in Equation (1), in which the LN distribution is the baseline and the TL distribution is a generator distribution, and henceforth, we call the resulting distribution the Topp-Leone log-normal (TLLN) distribution, which provides greater versatility in modeling skewed datasets.

We also propose an entirely different method to derive the new distribution. Sharma (2018) proposed a new three-parameter distribution called the Topp-Leone normal (TLN), which is defined on the entire real line and is ideal for modeling increasing HRF data. The CDF of the TLN distribution is expressed as

$$F_{\mathrm{TLN}}(y) = \left\{\Phi\left(\frac{y-\mu}{\sigma}\right)\left[2-\Phi\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{\alpha}, \quad y,\mu \in \mathbb{R},\ \sigma,\alpha > 0,$$

where Φ is the CDF of the standard normal distribution. Then, the random variable $X = \mathrm{e}^{Y}$ follows the TLLN distribution with parameters α, μ and σ.

3. Definition of the distribution

The definition of the new distribution, as well as several key features, are examined in this section.

Definition 3.1 Let X be a random variable which follows the TLLN distribution with parameters α, μ and σ. Then, its CDF is given by

$$F(x) = \left\{\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\left[2-\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]\right\}^{\alpha}, \tag{2}$$

and its PDF is defined as

$$f(x) = \frac{2\alpha}{\sigma x}\,\phi\left(\frac{\log(x)-\mu}{\sigma}\right)\left[1-\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]\times \left\{\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\left[2-\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]\right\}^{\alpha-1}, \tag{3}$$

where $x > 0$, $\mu \in \mathbb{R}$ and $\sigma, \alpha > 0$. Also, ϕ is the PDF of the standard normal distribution.

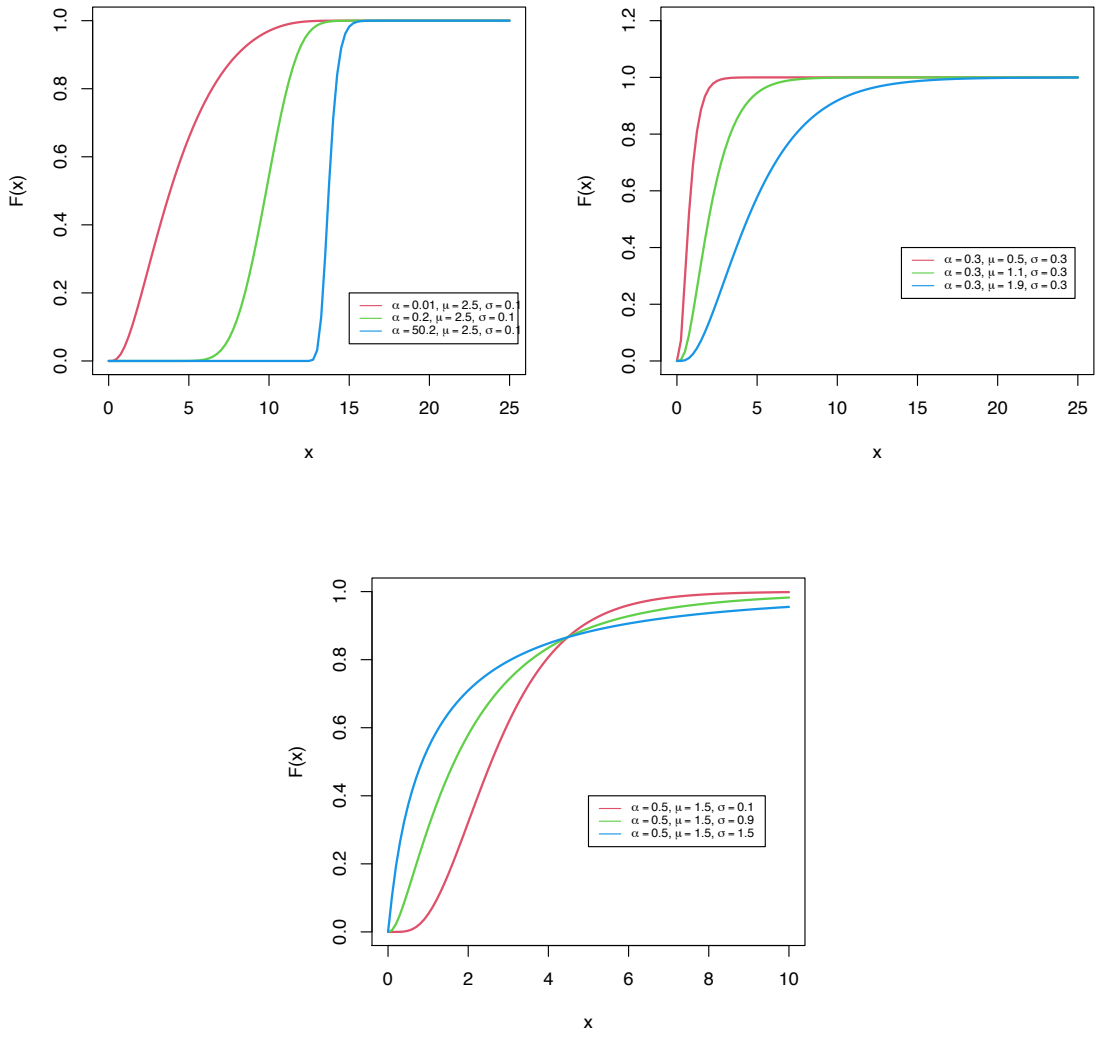


Figure 1. Plots of the CDF of the TLLN distribution.

The plots in Figures 1 and 2 depict the corresponding CDF and PDF of the TLLN distribution. We observe that the PDF may be decreasing and unimodal with a certain flexibility in the mode and tails. It is, however, mainly right-skewed or almost symmetrical. Next, some expansions for the CDF and PDF are provided. It is also interesting to note that the TLLN distribution can be expressed as an infinite sum of exponentiated LN distributions when α is a non-integer or as a finite sum when α is an integer. Indeed, the CDF of the

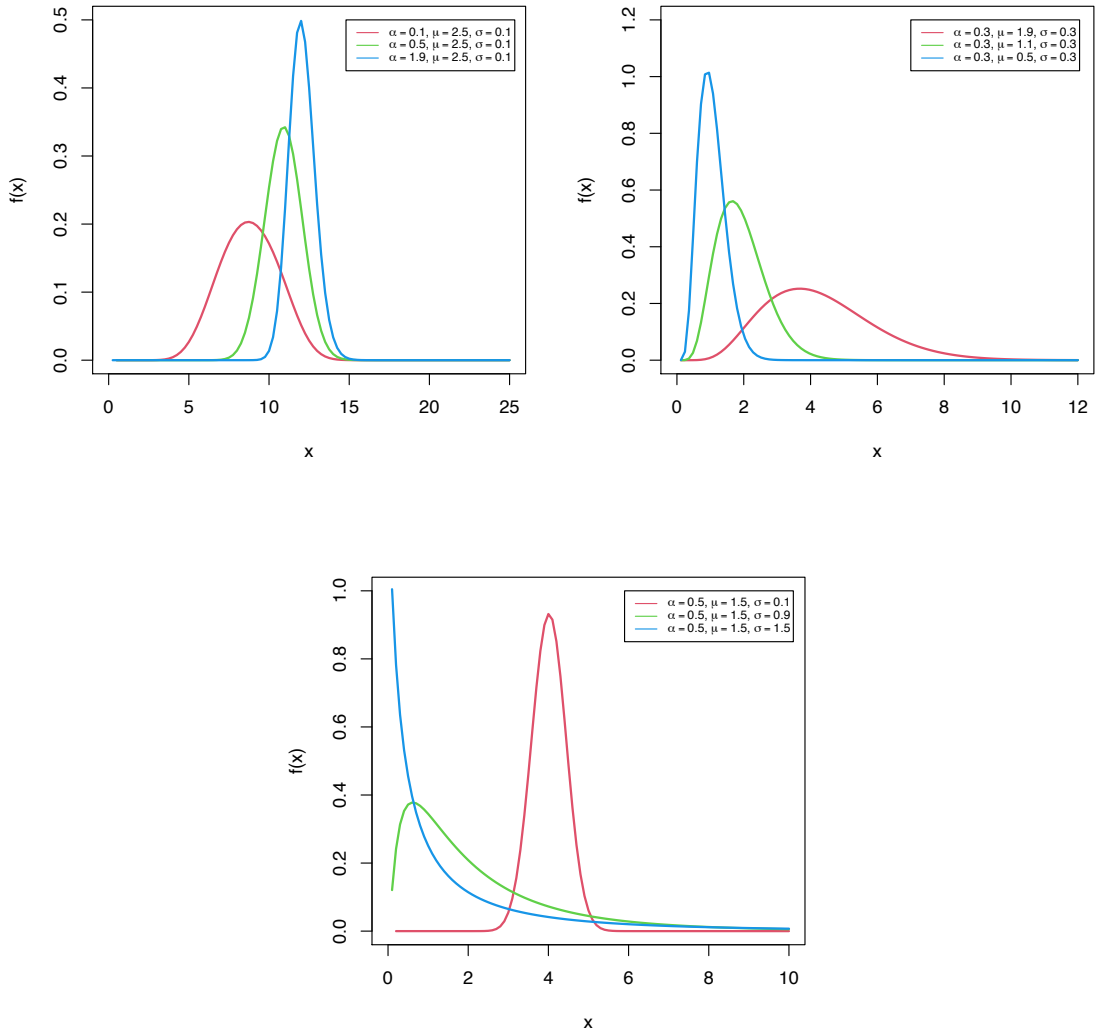


Figure 2. Plots of the PDF of the TLLN distribution.

TLLN distribution in Equation (2) can be simplified as follows:

$$F(x) = \sum_{j=0}^{\infty} \binom{\alpha}{j} (-1)^j 2^{\alpha-j} \left[\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]^{\alpha+j}$$

because of the identity given by

$$(2-b)^{\alpha} = \sum_{j=0}^{\infty} \binom{\alpha}{j} (-1)^j 2^{\alpha-j} b^j,$$

for $|b| < 2$.

Now, note that

$$[\Phi(.)]^{\alpha+j} = [1-(1-\Phi(.))]^{\alpha+j} = \sum_{k=0}^{\infty} \binom{\alpha+j}{k} (-1)^k [1-\Phi(.)]^k$$
$$= \sum_{k=0}^{\infty} \sum_{r=0}^{k} \binom{\alpha+j}{k} \binom{k}{r} (-1)^{k+r} [\Phi(.)]^r.$$

As a result, the CDF of the TLLN distribution takes the form

$$F(x) = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \sum_{r=0}^{k} W_{j,k,r}(\alpha) \left[\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]^r,$$

where

$$W_{j,k,r}(\alpha) = \binom{\alpha}{j} \binom{\alpha+j}{k} \binom{k}{r} (-1)^{j+k+r} 2^{\alpha-j}.$$

Thus, the TLLN distribution can be expressed as the infinite sum of exponentiated LN distributions indexed by the power parameter r. If the parameter α is an integer, the TLLN distribution can be expressed as the finite sum of exponentiated LN distributions given as

$$F(x) = \sum_{j=0}^{\alpha} \sum_{k=0}^{\alpha+j} \sum_{r=0}^{k} W_{j,k,r}(\alpha) \left[\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]^r.$$

Again, applying the series expansion in Equation (3), the PDF of the TLLN distribution can be written as

$$f(x) = \frac{2\alpha}{\sigma x}\, \phi\left(\frac{\log(x)-\mu}{\sigma}\right) \sum_{j=0}^{\infty} \binom{\alpha-1}{j} (-1)^j \left[1-\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]^{2j+1},$$

or

$$f(x) = \sum_{j=0}^{\infty} \sum_{k=0}^{2j+1} b_\alpha(j,k) \psi_k(x), \tag{4}$$

where

$$b_\alpha(j,k) = 2\binom{\alpha-1}{j} \binom{2j+1}{k} \frac{(-1)^{j+k}}{k+1} \tag{5}$$

and

$$\psi_k(x) = \frac{k+1}{\sigma x} \phi\left(\frac{\log(x)-\mu}{\sigma}\right) \left[\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]^k$$

is the PDF of the exponentiated LN distribution with power parameter $k+1$.

Remark 1 If x is fixed, the PDFs of the TLLN and LN distributions correspond when

$$\alpha = \frac{\log\left[\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]}{\log\left\{\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\left[2-\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]\right\}}.$$

The proof is straightforward and omitted for the sake of brevity.

LEMMA 3.2 For $\alpha = 1$, the CDF of the TLLN distribution in Equation (2) becomes the CDF of the transmuted LN distribution with transmuted parameter equals to 1.

Proof: To begin with, a retrospective on the transmuted distributions is necessary. Shaw and Buckley (2009) introduced a new family of distributions called transmuted distributions, and the general expression of its CDF is

$$F_T(x) = (1+\lambda)G(x) - \lambda[G(x)]^2, \quad |\lambda| \leq 1,$$

where G is the baseline CDF and λ is called the transmuted parameter. Thus, the CDF of the TLLN distribution can be written as

$$F(x) = 2\Phi\left(\frac{\log(x)-\mu}{\sigma}\right) - \left[\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]^2,$$

which is also the CDF of the transmuted LN distribution with $\lambda = 1$. It is worth mentioning that the transmuted LN distribution is not discussed in the available literature. Hence, one may study its properties and applications in detail.

4. MOMENTS

In this section, we derive the expression for the raw moments of the TLLN distribution. Let m be a positive integer and X be a random variable following the TLLN distribution. The mth raw moment of the TLLN distribution is then calculated using Equation (4) as

$$\mu'_m = \mathrm{E}(X^m) = \sum_{j=0}^{\infty}\sum_{k=0}^{2j+1}(k+1)b_\alpha(j,k)\mu'_{m,k}, \tag{6}$$

where

$$\mu'_{m,k} = \int_0^\infty \frac{x^{m-1}}{\sigma}\phi\left(\frac{\log(x)-\mu}{\sigma}\right)\left[\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]^k \mathrm{d}x$$

is the probability-weighted moment of the LN distribution. In other words, the raw moments of the TLLN distribution can be written as the weighted sum of the probability-weighted moments of the LN distribution.

Remark 1 If α is an integer, then the mth raw moment of the TLLN distribution is directly derived from Equation (6), and it is stated as

$$\mu'_m = \mathrm{E}(X^m) = \sum_{j=0}^{\alpha-1}\sum_{k=0}^{2j+1}(k+1)b_\alpha(j,k)\mu'_{m,k}.$$

5. Quantile function and associated measures

In this section, we derive an explicit expression for the QF of the TLLN distribution as well as several of its associated measures.

Theorem 5.1 Let $p \in (0,1)$. Then, the pth quantile of the TLLN distribution is given by

$$Q_p = F^{-1}(p) = \exp\left[\mu + \sigma\Phi^{-1}\left(1 - \sqrt{1-p^{1/\alpha}}\right)\right], \tag{7}$$

where Φ^{-1} is the QF of a standard normal distribution.

Proof: For the TLLN distribution, Q_p is the solution of the following equation:

$$\begin{aligned}
&\left\{\Phi\left(\frac{\log(Q_p)-\mu}{\sigma}\right)\left[2-\Phi\left(\frac{\log(Q_p)-\mu}{\sigma}\right)\right]\right\}^{\alpha} = p \\
&\Rightarrow 2\Phi\left(\frac{\log(Q_p)-\mu}{\sigma}\right) - \left[\Phi\left(\frac{\log(Q_p)-\mu}{\sigma}\right)\right]^2 = p^{1/\alpha}.
\end{aligned} \tag{8}$$

On simplifications, since $p \in (0,1)$, Equation (8) reduces to

$$\left[1-\Phi\left(\frac{\log(Q_p)-\mu}{\sigma}\right)\right]^2 = 1-p^{1/\alpha}$$

$$\Rightarrow \frac{\log(Q_p)-\mu}{\sigma} = \Phi^{-1}\left(1-\sqrt{1-p^{1/\alpha}}\right)$$

$$\Rightarrow Q_p = \exp\left[\mu + \sigma\Phi^{-1}\left(1-\sqrt{1-p^{1/\alpha}}\right)\right].$$

Remark 1 Since Φ^{-1} is the QF of the standard normal distribution, Q_p in Equation (7) also gets the form

$$Q_p = \exp\left[\mu + \sigma\sqrt{2}\mathrm{erf}^{-1}\left(1-2\sqrt{1-p^{1/\alpha}}\right)\right], \tag{9}$$

where erf^{-1} is the inverse error function.

By putting $p = 1/2$ in Equation (9), we get the median of the TLLN distribution, and it is given by

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

Equation (9) delivers the first and third quartiles of the TLLN distribution ($Q_{0.25}$ and $Q_{0.75}$) for $p = 1/4$ and $p = 3/4$, respectively.

6. Reliability measures

We derive the expressions for various reliability measures in this section. The HRF of the TLLN distribution is expressed by

$$h(x) = \frac{f(x)}{S(x)},$$

where $S(x) = 1 - F(x)$ is the survival function of the TLLN distribution given by

$$S(x) = 1 - \left\{\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\left[2 - \Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]\right\}^{\alpha}.$$

Thus, the desired HRF gets the following form:

$$h(x) = \frac{2\alpha\phi\left(\frac{\log(x)-\mu}{\sigma}\right)\left[1-\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]\left\{\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\left[2-\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]\right\}^{\alpha-1}}{\sigma x\left[1-\left\{\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\left[2-\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]\right\}^{\alpha}\right]}.$$

Also, plots in Figure 3 refer to the shapes of the HRF and show that the TLLN distribution possesses increasing, decreasing, and upside-down bathtub shapes. Also, as seen in Figure 3, the distribution has a new decreasing-increasing-decreasing shape that we call the inverted N-shaped HRF, as well as a special shape that starts with a flat region and continues with an increasing-decreasing shape that we call the constant-increasing-decreasing shaped HRF.

Let r be a positive integer and X be a random variable following the TLLN distribution. Then, the rth conditional moment of the TLLN distribution is stated as

$$\begin{aligned}\mathrm{E}(X^r|X>t) &= \frac{1}{S(t)}\int_t^{\infty} x^r f(x)\mathrm{d}x \\ &= \frac{1}{S(t)}\sum_{j=0}^{\infty}\sum_{k=0}^{2j+1}\left(\frac{k+1}{\sigma}\right)b_{\alpha}(j,k)I_1(r,k), \end{aligned} \tag{10}$$

where $b_{\alpha}(j,k)$ is given in Equation (5) and $I_1(r,k)$ is formulated as

$$I_1(r,k) = \int_t^{\infty} x^{r-1}\phi\left(\frac{\log(x)-\mu}{\sigma}\right)\left[\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]^k \mathrm{d}x. \tag{11}$$

It is worth mentioning that the rapid aging of a component requires low vitality, whereas high vitality implies relatively slow aging during the given time period.

For $r = 1$, Equation (10) gives the vitality function of the TLLN distribution, which is

$$\begin{aligned} V(t) = \mathrm{E}(X|X>t) &= \frac{1}{S(t)}\int_t^{\infty} x f(x)\mathrm{d}x \\ &= \frac{1}{S(t)}\sum_{j=0}^{\infty}\sum_{k=0}^{2j+1}\left(\frac{k+1}{\sigma}\right)b_{\alpha}(j,k)I_1(1,k), \end{aligned} \tag{12}$$

where $I_1(1,k)$ is obtained by putting $r = 1$ in Equation (11), and is given by

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

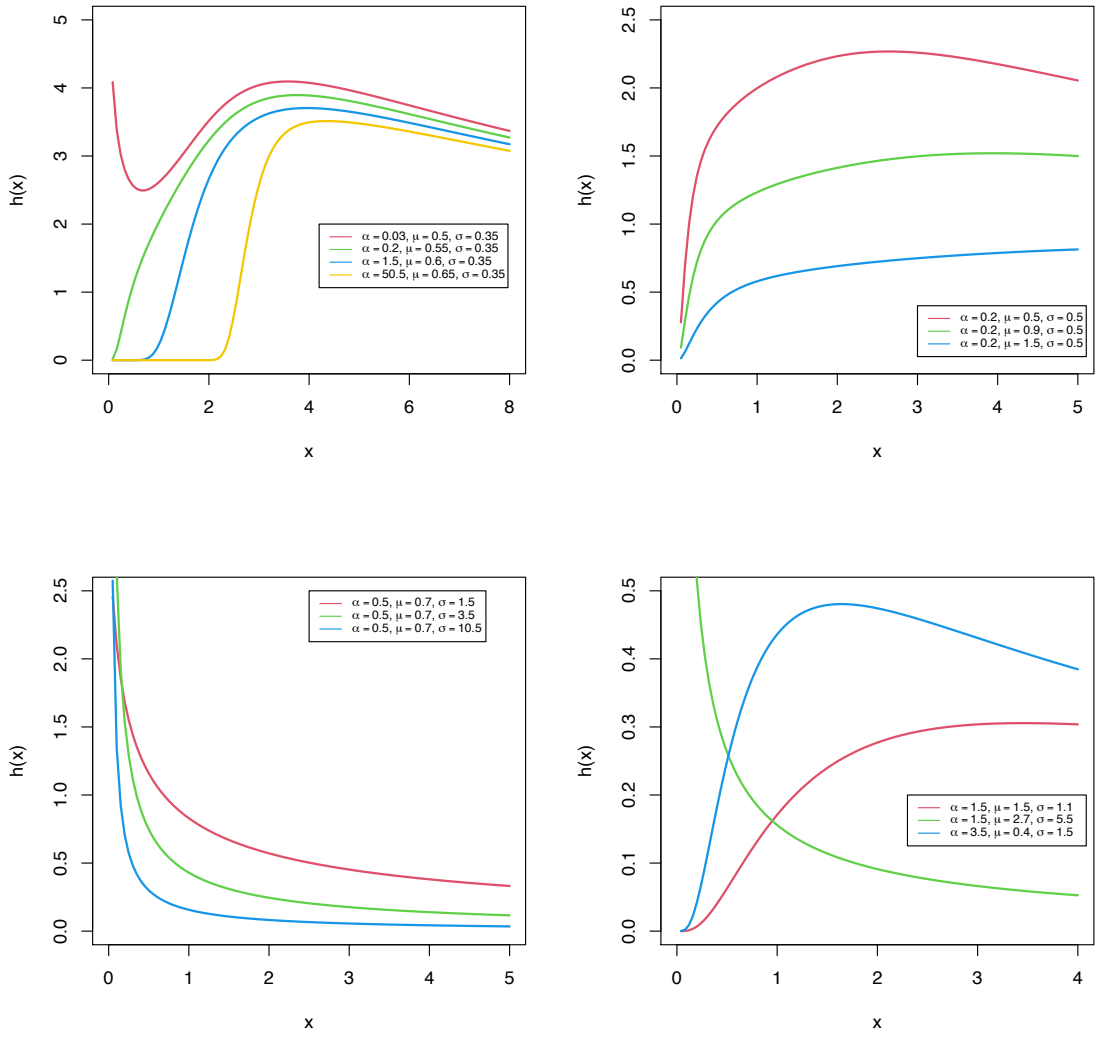


Figure 3. Plots of the HRF of the TLLN distribution.

If X is a random variable representing a component's lifetime, then $\log(G(t)) = \mathrm{E}(\log(X)|X > t)$ represents the ideal geometric mean of the lifetimes of components that have survived up to time t. The geometric vitality function of the TLLN distribution is stated as

$$\log(G(t)) = \frac{1}{S(t)} \sum_{j=0}^{\infty} \sum_{k=0}^{2j+1} (k+1) b_{\alpha}(j,k) I_2(k),$$

where $I_2(k)$ can be expressed as

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

The concept of residual life is of special interest in reliability theory. It measures the amount of time a unit has left after reaching the age of t. The rth order moment of the residual life

of the TLLN distribution is defined as

$$\mu_r(t) = \mathrm{E}\left[(X-t)^r | X > t\right] = \frac{1}{S(t)} \int_t^{\infty} (x-t)^r f(x) \mathrm{d}x$$
$$= \frac{1}{S(t)} \sum_{i=0}^{r} \binom{r}{i} (-1)^{r-i} t^{r-i} I_1(i,k),$$

where $I_1(i,k)$ is given in Equation (11). Now, by taking $r = 1$, we get the expression for the mean residual life (MRL) function, which also gets the form

$$\mu_1(t) = V(t) - t,$$

where $V(t)$ is given in Equation (12). Similarly, the second moment of the residual lifetime of the TLLN distribution is stated as

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

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

$$I_1(2,k) = \int_t^{\infty} x\phi\left(\frac{\log(x)-\mu}{\sigma}\right) \left[\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]^k \mathrm{d}x.$$

Thus, the variance of the residual life function of the TLLN distribution can be obtained using $\mu_1(t)$ and $\mu_2(t)$. The rth order moment of the reversed residual life of the TLLN distribution is formulated as

$$m_r(t) = \mathrm{E}\left[(t-X)^r | X \leq t\right] = \frac{1}{F(t)} \int_0^t (t-x)^r f(x) \mathrm{d}x$$
$$= \frac{1}{F(t)} \sum_{i=0}^{r} \binom{r}{i} (-1)^i t^{r-i} \left[\mu_i' - I_1(i,k)\right], \tag{13}$$

where $I_1(i,k)$ is given in Equation (11). Now, the mean $m_1(t)$ and second moment $m_2(t)$ of the reversed residual life of the TLLN distribution can be obtained by setting $r = 1$ and $r = 2$, respectively, in Equation (13). Again, using $m_1(t)$ and $m_2(t)$, one can obtain the variance of the reversed residual life function of the distribution.

7. Estimation of the parameters

In this section, we discuss how to estimate the parameters of the TLLN distribution utilizing two well-known methods, namely the maximum likelihood (ML) method and the Bayesian method. Next, we consider the ML estimation for the TLLN model parameters α, μ and σ. Let $X_1, \ldots, X_n$ represent a random sample from the TLLN distribution, and $x_1, \ldots, x_n$ represent the observed values. Then, the log-likelihood function can then be written in the

following form:

$$\begin{aligned}\mathcal{L}_n &= \sum_{i=1}^{n} \log[f(x_i)] \\ &= n\log(2) + n\log(\alpha) - n\log(\sigma) - \sum_{i=1}^{n}\log(x_i) + \sum_{i=1}^{n}\log\left[\phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)\right] \\ &+ \sum_{i=1}^{n}\log\left[1-\Phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)\right] + (\alpha-1)\sum_{i=1}^{n}\log\left[\Phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)\right] \\ &+ (\alpha-1)\sum_{i=1}^{n}\log\left[2-\Phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)\right].\end{aligned}$$

The ML estimates of (α, μ, σ) are $(\widehat{\alpha}, \widehat{\mu}, \widehat{\sigma}) = \text{argmax}_{(\alpha,\mu,\sigma)}\mathcal{L}_n$. We can formulate them by using nonlinear log-likelihood equations. First, the score function associated with the log-likelihood function is

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

The associated nonlinear log-likelihood equations are $\boldsymbol{U} = (0,0,0)^{\top}$, that is,

$$\frac{n}{\alpha} + \sum_{i=1}^{n}\log\left[\Phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)\right] + \sum_{i=1}^{n}\log\left[2-\Phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)\right] = 0, \tag{14}$$

$$\begin{aligned}&\sum_{i=1}^{n}\frac{\log(x_i)-\mu}{\sigma^2} + \frac{1}{\sigma}\sum_{i=1}^{n}\frac{\phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)}{1-\Phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)} - \frac{\alpha-1}{\sigma}\sum_{i=1}^{n}\frac{\phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)}{\Phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)} \\ &\qquad + \frac{\alpha-1}{\sigma}\sum_{i=1}^{n}\frac{\phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)}{2-\Phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)} = 0\end{aligned} \tag{15}$$

and

$$\begin{aligned}&-\frac{n}{\sigma} + \sum_{i=1}^{n}\frac{(\log(x_i)-\mu)^2}{\sigma^3} + \sum_{i=1}^{n}\frac{\left(\frac{\log(x_i)-\mu}{\sigma^2}\right)\phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)}{1-\Phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)} \\ &\qquad - \frac{\alpha-1}{\sigma^2}\sum_{i=1}^{n}\frac{(\log(x_i)-\mu)\phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)}{\Phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)} \\ &\qquad + \frac{\alpha-1}{\sigma^2}\sum_{i=1}^{n}\frac{(\log(x_i)-\mu)\phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)}{2-\Phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)} = 0.\end{aligned} \tag{16}$$

Solving the nonlinear Equations (14), (15) and (16) synergistically, one can obtain the ML estimates. For known μ and σ, the ML estimate of α is given by

$$\widehat{\alpha} = -\frac{n}{\sum_{i=1}^{n} \log\left[\Phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)\right] + \sum_{i=1}^{n} \log\left[2-\Phi\left(\frac{\log(x_i)-\mu}{\sigma}\right)\right]}.$$

The asymptotic confidence intervals (CIs) for the parameters α, μ and σ are now executed. On taking the second partial derivatives of Equations (14), (15) and (16) taken at the ML estimates, the observed Hessian matrix of the TLLN distribution can be obtained, and it is given by

$$\widehat{H} = \left.\begin{pmatrix} \frac{\partial^2 \mathcal{L}_n}{\partial \alpha^2} & \frac{\partial^2 \mathcal{L}_n}{\partial \alpha \partial \mu} & \frac{\partial^2 \mathcal{L}_n}{\partial \alpha \partial \sigma} \\ \frac{\partial^2 \mathcal{L}_n}{\partial \mu \partial \alpha} & \frac{\partial^2 \mathcal{L}_n}{\partial \mu^2} & \frac{\partial^2 \mathcal{L}_n}{\partial \mu \partial \sigma} \\ \frac{\partial^2 \mathcal{L}_n}{\partial \sigma \partial \alpha} & \frac{\partial^2 \mathcal{L}_n}{\partial \sigma \partial \mu} & \frac{\partial^2 \mathcal{L}_n}{\partial \sigma^2} \end{pmatrix}\right|_{(\alpha,\mu,\sigma)=(\widehat{\alpha},\widehat{\mu},\widehat{\sigma})}.$$

Now, the observed Fisher's information matrix $\widehat{J}$ is obtained as $\widehat{J} = -\widehat{H}$. The inverse of this matrix provides the variance-covariance matrix of the ML estimators, which can be written as

$$\widehat{\mathbf{\Sigma}} = \widehat{J}^{-1} = \begin{pmatrix} \widehat{\mathbf{\Sigma}}_{11} & \widehat{\mathbf{\Sigma}}_{12} & \widehat{\mathbf{\Sigma}}_{13} \\ \widehat{\mathbf{\Sigma}}_{21} & \widehat{\mathbf{\Sigma}}_{22} & \widehat{\mathbf{\Sigma}}_{23} \\ \widehat{\mathbf{\Sigma}}_{31} & \widehat{\mathbf{\Sigma}}_{32} & \widehat{\mathbf{\Sigma}}_{33} \end{pmatrix},$$

and $\widehat{\mathbf{\Sigma}}_{ij} = \widehat{\mathbf{\Sigma}}_{ji}$ for $i \neq j = 1, 2, 3$. The asymptotically normal distribution of the ML estimators has been thoroughly established. The random version of $\widehat{\Theta} = (\widehat{\alpha}, \widehat{\mu}, \widehat{\sigma})$ follows the multivariate normal distribution $N_3(\Theta, \widehat{\Sigma})$, where $\Theta = (\alpha, \mu, \sigma)$. Thus, we obtain $100 \times (1-\delta)\%$ asymptotic CIs of the parameters using the following formulae:

$$I_\alpha = \left[\widehat{\alpha} \mp \upsilon_{\delta/2}\sqrt{\widehat{\mathbf{\Sigma}}_{11}}\right], \quad I_\mu = \left[\widehat{\mu} \mp \upsilon_{\delta/2}\sqrt{\widehat{\mathbf{\Sigma}}_{22}}\right], \quad I_\sigma = \left[\widehat{\sigma} \mp \upsilon_{\delta/2}\sqrt{\widehat{\mathbf{\Sigma}}_{33}}\right],$$

where υ_δ is the upper δth percentile of the standard normal distribution. Next, we perform the Bayesian analysis for the TLLN model parameters. To do so, each parameter must have a prior distribution. We employ two types of priors for this: the half-Cauchy (HC) and the classical normal (N) priors. The PDF of the HC distribution with scale parameter a is defined as

$$f_{\text{HC}}(x) = \frac{2a}{\pi(x^2+a^2)}, \quad x > 0,\ a > 0. \tag{17}$$

The HC distribution is known to have no mean or variance. Meanwhile, its mode is equal to 0. Since the PDF of the HC distribution is virtually flat but not totally flat at scale value equals 25, which verges on acquiring adequate information for the numerical approximation algorithm to continue looking at the target posterior distribution, the HC distribution with scale parameter equals 25 is recommended as a non-informative prior. According to Gelman and Hill (2006), the uniform distribution, or whether more information is required, is a

superior alternative to the HC distribution. As a result, for the parameters α and σ, the HC distribution with a scale parameter equaling 25 was chosen as a non-informative prior distribution in this study. Thus, we set the prior distributions of the parameters to be

$$\mu \sim \mathrm{N}(0, 1000), \quad \alpha, \sigma \sim \mathrm{HC}(25). \tag{18}$$

Thus, using Equation (18), we obtain the joint posterior PDF as given by

$$\pi(\mu, \alpha, \sigma|x) \propto L_n \times \pi(\mu) \times \pi(\alpha) \times \pi(\sigma), \tag{19}$$

where L_n is the likelihood function of the TLLN distribution. From Equation (19), it is obvious that there is no analytical solution to find out the Bayesian estimates. Thus, we use a remarkable method of simulation, namely the Metropolis-Hastings (MH) algorithm of the Markov Chain Monte Carlo (MCMC) method. Upadhyay et al. (2001) provides a thorough description of the MCMC approach.

8. Bootstrap confidence intervals

In this section, we utilize the parametric bootstrap method to approximate the distribution of the ML estimators of the TLLN model parameters. Then, we can employ the bootstrap distribution to estimate each parameter's CIs for the fitted TLLN distribution. Let $\widehat{\Xi}$ be the ML estimate of Ξ, where $\Xi \in (\alpha, \mu, \sigma)$, using a given dataset $\{x_1, x_2, \ldots, x_n\}$. The bootstrap is a method to estimate the distribution of the statistic $\widehat{\Xi}$ by getting a random sample $\Xi_1^*, \Xi_2^*, \ldots, \Xi_B^*$ for Ξ based on B random samples that are drawn with replacement from the original data $x_1, x_2, \ldots, x_n$. Thus, the bootstrap sample $\Xi_1^*, \Xi_2^*, \ldots, \Xi_B^*$ can be used to construct bootstrap CIs for α, μ and σ.

Thus, using the following formulae, we calculate the $100 \times (1-\delta)\%$ bootstrap CIs for the parameters:

$$\mathfrak{I}_\alpha = \left[\widehat{\alpha} \mp z_{\delta/2}\widehat{\mathrm{SE}}_{\alpha,\mathrm{boot}}\right], \quad \mathfrak{I}_\mu = \left[\widehat{\mu} \mp z_{\delta/2}\widehat{\mathrm{SE}}_{\mu,\mathrm{boot}}\right], \quad \mathfrak{I}_\sigma = \left[\widehat{\sigma} \mp z_{\delta/2}\widehat{\mathrm{SE}}_{\sigma,\mathrm{boot}}\right].$$

where z_δ denotes the δth percentile of the bootstrap sample, SE is the standard error, and, for $\Xi \in \{\alpha, \mu, \sigma\}$,

$$\widehat{\mathrm{SE}}_{\Xi,\mathrm{boot}} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\Xi_b^* - \frac{1}{B}\sum_{b=1}^{B}\Xi_b^*\right)^2}.$$

9. TLLN regression model

In this section, we define a regression model based on the TLLN distribution, called the TLLN regression model.

To begin, consider a random variable X following the TLLN distribution with PDF given in Equation (3), as well as the random variable Y defined by $Y = \log(X)$. Then, Y has the following PDF:

$$f_Y(y) = \frac{2\alpha}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\left[1-\Phi\left(\frac{y-\mu}{\sigma}\right)\right]\left\{\Phi\left(\frac{y-\mu}{\sigma}\right)\left[2-\Phi\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{\alpha-1},$$
$$y \in \mathbb{R},\ \mu \in \mathbb{R},\ \alpha, \sigma > 0. \tag{20}$$

We refer to Equation (20) as the log-Topp-Leone log-normal (log-TLLN) distribution, or otherwise, the Topp-Leone normal (TLN) distribution, which is given by Sharma (2018). In this setting, the standardized random variable $Z = (Y - \mu)/\sigma$ has the PDF given by

$$f_Z(z) = 2\alpha\phi(z)\left[1 - \Phi(z)\right]\left\{\Phi(z)\left[2 - \Phi(z)\right]\right\}^{\alpha-1}. \tag{21}$$

Now, linear location-scale regression model linking the response variable y_i and the explanatory variable vector $\boldsymbol{v}_i^\top = (v_{i1}, \ldots, v_{ip})$, is obtained as

$$y_i = \mu_i + \sigma z_i, \quad i = 1, \ldots, n, \tag{22}$$

where z_i is the random error component that has the PDF in Equation (21), $\mu_i = \mathrm{v}_i^\top \tau$ is the location parameter of y_i, where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)^\top$, α and σ are unknown parameters. The location parameter vector $\mu = (\mu_1, \ldots, \mu_n)^\top$ is represented by a linear model $\mu = \boldsymbol{V}\boldsymbol{\tau}$, where $\boldsymbol{V} = (V_1, \ldots, V_n)^\top$ is a known model matrix. Ultimately, in this study, we propose the TLLN regression model from Equation (22) and it is given by

$$x_i = \mathrm{e}^{y_i} = \mathrm{e}^{\mu_i + \sigma z_i}, \quad i = 1, \ldots, n. \tag{23}$$

Consider a sample $(x_1, \boldsymbol{v}_1), \ldots, (x_n, \boldsymbol{v}_n)$ of n independent observations. Conventional likelihood estimation techniques can be applied here. Now, for the vector of parameters $\boldsymbol{\psi} = (\boldsymbol{\tau}^\top, \alpha, \sigma)^\top$ from model (23), the total log-likelihood function for right censored has the form

$$l(\boldsymbol{\psi}) = \sum_{i=1}^{n} \delta_i \log\left[f(x_i)\right] + \sum_{i=1}^{n} (1 - \delta_i) \log\left[S(x_i)\right],$$

with $\delta_i = 1$, if survival (uncensored) and $\delta_i = 0$, if not (censored). We recall that, for $i = 1, \ldots, n$, $f(x_i)$ and $S(x_i)$ are the PDF and survival function of the TLLN distribution taken at x_i, respectively.

10. Bayesian regression method

The Bayesian technique has been shown to be particularly effective in analyzing survival models in many practical circumstances. Hence, in this section, we look at how the Bayesian approach fits the regression model based on the TLLN distribution when prior pieces of information about the parameters are taken into account. As a result, we use a simulation method in this part for Bayesian analysis of this model. Now, to perform a Bayesian analysis, one should adopt prior distributions for the parameters. Here, as described previously, we utilize the HC and N priors. The PDF of the HC distribution with a as the scale parameter is given in Equation (17). Now, we write the right censored likelihood function as

$$L = \prod_{i=1}^{n} \left[f(x_i)\right]^{\delta_i} \left[S(x_i)\right]^{1-\delta_i}, \tag{24}$$

with $\delta_i = 1$, if survival (uncensored) and $\delta_i = 0$, if not (censored).

$$\mu = V\tau \tag{25}$$

as a linear combination of explanatory variables. Thus, we set the prior distributions of the parameters to be

$$\tau_j \sim \mathrm{N}(0, 1000), \quad j = 1, \ldots, J, \quad \alpha, \sigma \sim \mathrm{HC}(25). \tag{26}$$

Now, using Equations (24), (25) and (26), the joint posterior PDF is obtained as

$$\pi(\tau, \alpha, \sigma | x, V) \propto L(x|V, \tau, \alpha, \sigma) \times \pi(\tau) \times \pi(\alpha) \times \pi(\sigma). \tag{27}$$

From Equation (27), it is clear that the analytical solution is not possible to find out the Bayesian estimates. As a result, we employ the simulation approach, specifically the MH algorithm of the MCMC method.

11. Performance of the estimators using simulations

In this section, we conduct simulation experiments to assess the long-run performance of the ML estimators of the TLLN model parameters for some finite sample sizes. We simulated datasets of sizes $n = 60$, 100, and 250 from the TLLN distribution for the parameter values $\alpha = 0.5, \mu = 0.9, \sigma = 0.6$ and iterated each sample 500 times. The average bias and MSE for all replications in the relevant sample sizes are then computed. That is, the analysis computes the values using the given formulae.

Table 1. Estimates, average bias and MSE values of ML estimators from simulations of the TLLN distribution.

Parameters	Sample Size	Estimates	Bias	MSE
α	60	1.7496	1.2496	67.9030
	100	1.0114	0.5114	6.3232
	250	0.5792	0.0792	0.1677
μ	60	0.8240	-0.0760	0.3754
	100	0.8491	-0.0509	0.2180
	250	0.8873	-0.0127	0.0504
σ	60	0.5520	-0.0480	0.1350
	100	0.5811	-0.0189	0.0849
	250	0.5916	-0.0084	0.0229

- Average bias of the simulated estimates $= \dfrac{1}{500} \sum_{i=1}^{500} (\widehat{\Xi}_i - \Xi)$,
- Average MSE of the simulated estimates $= \dfrac{1}{500} \sum_{i=1}^{500} (\widehat{\Xi}_i - \Xi)^2$,

where $\widehat{\Xi}_i$ represents the estimate of $\Xi \in \{\alpha, \mu, \sigma\}$ at the ith replication. The results are reported in Table 1. It can be concluded that the MSEs of all the estimates decrease with increasing sample size. This shows the consistency of the estimates.

12. Applications and empirical study

This section consists of demonstrating the empirical importance of the TLLN distribution. We use a real dataset from the area of astronomy to compare the data modeling ability of the

TLLN distribution with other competitive distributions. We employ the RStudio software for numerical evaluations of these datasets. The descriptive measures, which include sample size (n), mean (M), median (Md), variance (Var), skewness (Sk), kurtosis (Ku), minimum (min) and maximum (max) values of the dataset, are given in Table 2.

Table 2. Descriptive statistics of the astronomical dataset.

Statistic	n	M	Md	Var	Sk	Ku	min	max
Values	360	14.458	14.54	1.427	-0.395	0.344	10.749	18.052

To show the potential advantage of the TLLN distribution, the following distributions are considered for comparison.

- The two-parameter LN distribution with PDF

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

- The exponentiated LN (ELN) distribution with PDF

$$f(x) = \frac{\alpha}{x\sigma}\phi\left(\frac{\log(x)-\mu}{\sigma}\right)\left[\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]^{\alpha-1}, \ x > 0,\ \mu \in \mathbb{R},\ \alpha, \sigma > 0.$$

- The generalized half-normal (GHN) distribution (see Cooray and Ananda, 2008) with PDF

$$f(x) = \sqrt{\frac{2}{\pi}}\left(\frac{\alpha}{x}\right)\left(\frac{x}{\sigma}\right)^{\alpha} \exp\left\{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^{2\alpha}\right\}, \quad x, \alpha, \sigma > 0.$$

- The exponentiated exponential distribution (EED) with PDF

$$f(x) = \alpha\sigma\left(1 - \mathrm{e}^{-\sigma x}\right)^{\alpha-1}\mathrm{e}^{-\sigma x}, \quad x, \alpha, \sigma > 0.$$

- The new generalized Lindley distribution (NGLD) (see Elbatal et al., 2013) with PDF

$$f(x) = \frac{\mathrm{e}^{-\mu x}}{1+\mu}\left(\frac{\mu^{\alpha+1}x^{\alpha-1}}{\Gamma(\alpha)} + \frac{\mu^{\sigma}x^{\sigma-1}}{\Gamma(\sigma)}\right), \quad x > 0,\ \alpha, \mu, \sigma > 0,$$

  where $\Gamma(\alpha)$ denotes the standard gamma function.
- The gamma distribution.

For the numerical optimization, we maximize the log-likelihood function to find the ML estimates. For fixing a lower and upper bound for each parameter, the numerical optimization technique L-BFGS-B in `fitdistrplus` package of `R` is used. For more information and detailed examples of this package, one should go through the link `https://CRAN.R-project.org/package=fitdistrplus`.

The following statistical tools are utilized in order to compare the competitive models with the proposed models: negative log-likelihood ($-\log(L)$), Kolmogorov-Smirnov (KS), Cramér-von Mises (W^*), Anderson-Darling (A^*) statistics, Akaike information criterion (AIC) and Bayesian information criterion (BIC) values.

We also investigate the empirical HRF for the astronomical dataset using the idea of a total time on test (TTT) plot. It is a graphical representation being used to distinguish

between several types of aging as displayed in the HRF shapes. On the mathematical aspect, the TTT plot is drawn by plotting

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

against i/n, where $i = 1, \ldots, n$ and $x_{1:n}, x_{2:n}, \ldots, x_{n,n}$ are the order statistics of the sample. We also present other important graphs, which consist of the empirical CDF and quantile-quantile (Q-Q) plots for the dataset. We utilize the magnitudes of the near-infrared K-band distribution of 360 globular cluster luminosities in Messier 31 (M31), our nearby Andromeda Galaxy, as an astronomical dataset. The data are from Nantais et al. (2006), and the samples are described in detail in Appendix C.3 of Feigelson and Babu (2012), as well as in the R package `astrodatR`. Note that, the K-band is an atmospheric transmission window in infrared astronomy, referring to an area of the infrared spectrum where atmospheric gases absorb relatively little terrestrial heat radiation. Furthermore, globular clusters are densely packed groups of 10^4 to 10^6 ancient stars packed into a dense, roughly spherical shape that is structurally unique from the general population of stars. Astronomers can use them to determine the age of the universe or to locate the Galactic Center by studying them. The TTT plot in Figure 4 indicates that this dataset has an increasing HRF shape, which is also a characteristic of the TLLN model.

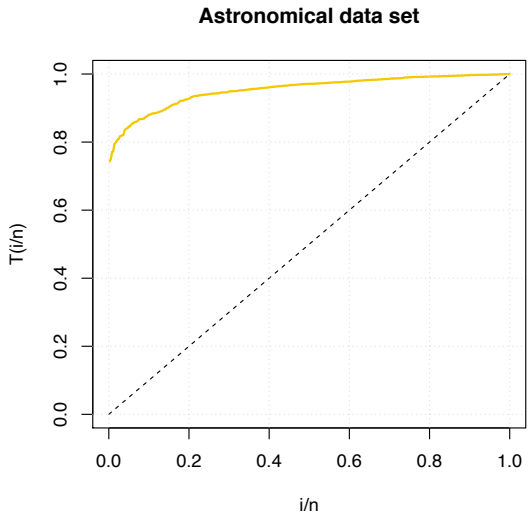


Figure 4. The TTT plot of astronomical dataset.

Next, we present results for the univariate dataset.

Table 3. Astronomical dataset: ML estimates and GOF statistics results.

Model	ML estimate	-log(L)	AIC	BIC	KS	W^*	A^*
TLLN	$\widehat{\alpha} = 0.2694$ $\widehat{\mu} = 2.7796$ $\widehat{\sigma} = 0.0601$	571.7256	1149.451	1161.110	0.0621	0.2083	1.2120
LN	$\widehat{\mu} = 2.6677$ $\widehat{\sigma} = 0.0847$	582.5224	1169.045	1176.817	0.0774	0.5396	3.2814
ELN	$\widehat{\alpha} = 0.1070$ $\widehat{\mu} = 2.7826$ $\widehat{\sigma} = 0.0371$	573.2549	1152.510	1164.168	0.0657	0.2517	1.4640
GHN	$\widehat{\mu} = 9.8248$ $\widehat{\sigma} = 15.2179$	592.761	1189.522	1197.294	0.0753	0.6212	4.1150
EXPPL	$\widehat{\alpha} = 1.8821$ $\widehat{\mu} = 136.4388$ $\widehat{\sigma} = 0.0491$	605.6499	1217.300	1228.958	0.1096	1.2230	7.3301
EED	$\widehat{\alpha} = 44847.58$ $\widehat{\sigma} = 0.7740$	624.776	1253.552	1261.324	0.1293	1.7963	10.8203
NGLD	$\widehat{\alpha} = 137.9676$ $\widehat{\mu} = 9.5441$ $\widehat{\sigma} = 138.2696$	579.4283	1164.857	1176.515	0.0754	0.4665	2.8083
Gamma	$\widehat{\alpha} = 142.2090$ $\widehat{\sigma} = 9.8355$	579.3487	1162.697	1170.470	0.0722	0.4406	2.7088

Table 3 displays the ML estimates and goodness-of-fit (GOF) statistics of the distributions corresponding to the astronomical dataset. The TLLN distribution's GOF statistics values are smaller than those of the other compared distributions. The empirical CDF and Q-Q plots for the dataset are given in Figure 5. The proposed distribution gives acceptable shaped curves for those empirical and fitted functions. As a result, we conclude that the TLLN distribution is superior to the other compared distributions for the astronomical dataset.

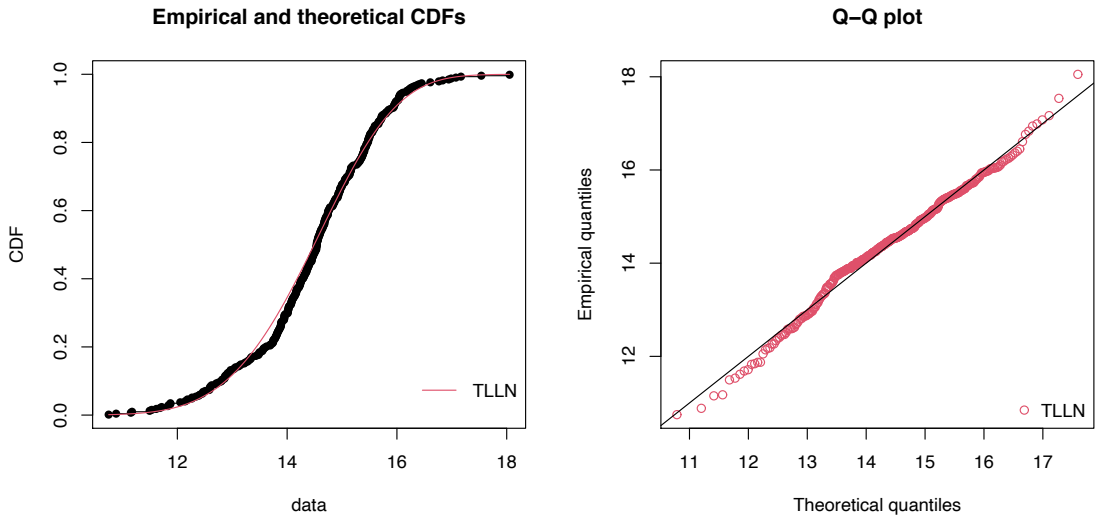


Figure 5. Empirical plots on the astronomical dataset.

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

$$\widehat{\boldsymbol{H}} = \begin{pmatrix} 4960.117 & 21162.32 & -35812.03 \\ 13376.538 & 57566.52 & -72436.86 \\ -35812.031 & -72436.86 & 296223.69 \end{pmatrix},$$

and the corresponding estimated variance-covariance matrix is

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.0105 & -0.0012 & 9.72 \times 10^{-04} \\ -0.0012 & 0.0002 & -1.07 \times 10^{-04} \\ 0.0009 & -0.0001 & 9.49 \times 10^{-05} \end{pmatrix}.$$

Table 4 provides the 95% asymptotic CIs for the TLLN model parameters.

Table 4. The 95% asymptotic CIs of the TLLN model parameters based on the astronomical dataset.

Parameter	Lower	Upper
α	0.0685	0.4703
μ	2.7543	2.8049
σ	0.0410	0.0792

Here, we focus on estimating the parameters of the TLLN distribution using the Bayesian procedure based on the above-discussed univariate astronomical dataset. In the context of Bayesian estimation, the analysis is performed using the MH algorithm of the MCMC method with 1000 iterations. For comparing Bayes estimates with the ML estimates, both of them for the TLLN model parameters for the real dataset are given in Table 5. The numerical computations for Bayesian estimation are done using the `LaplacesDemon` package of the `R` software, which provides a comprehensive environment for Bayesian inference. For more detailed information and examples regarding this package, one should go through the link https://cran.r-project.org/package=LaplacesDemon.

Table 5. ML and Bayes estimates of the TLLN model parameters on the astronomical dataset.

Parameter	ML	Bayes
α	0.2694	0.2811
μ	2.7796	2.7791
σ	0.0601	0.0602

Using the previously discussed astronomical dataset, we construct the 95% bootstrap for the parameters α, μ, and σ using the computed ML estimates. Based on the TLLN distribution, we simulate 1001 samples of the same size as the real dataset, with true values of the parameters chosen as the ML estimate of the respective parameters. For each sample obtained, we compute the ML estimates $\widehat{\alpha}_b^*$, $\widehat{\mu}_b^*$ and $\widehat{\sigma}_b^*$, for $b \in \{1, \ldots, 1001\}$. Table 6 displays the median and 95% bootstrap CI for the parameters α, μ and σ of the dataset. Examining the joint distribution of the bootstrapped values in a matrix of scatter plots to assess the potential structural correlation among the parameters is also noteworthy. Figure 6 displays the matrix scatterplots of the bootstrapped values of the TLLN model parameters, which depict the joint uncertainty distribution of the fitted parameters.

Table 6. The median and 95% bootstrap CI for the TLLN model parameters on the astronomical dataset.

Parameter	Median	Bootstrap CI
α	0.2695	(0.0878, 0.6941)
μ	2.7792	(2.7415, 2.8087)
σ	0.0609	(0.0373, 0.0881)

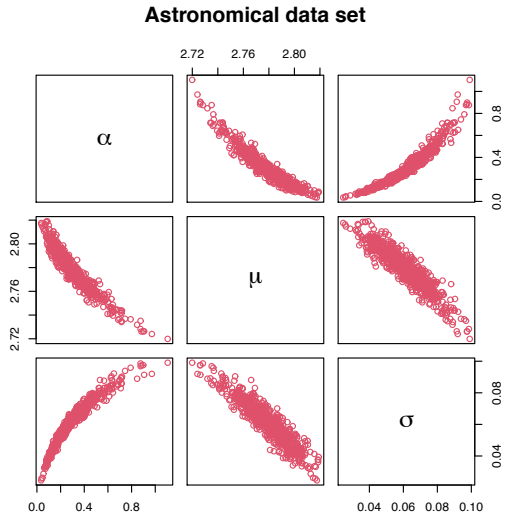


Figure 6. Matrix scatter plot on bootstrapped values of the TLLN model parameters due to the astronomical dataset.

We also utilize the likelihood ratio (LR) test for comparing the TLLN distribution, which has an additional parameter α with the LN and ELN distributions based on the above-discussed univariate astronomical dataset. The LR statistic for comparing the nested models H_0: LN and H_0: ELN against H_1: TLLN is given by

$$\text{LR} = -2\log\left(\frac{\text{likelihood under the null hypothesis}}{\text{likelihood in the whole parameter space}}\right).$$

It is well-known that the random version of this statistic asymptotically follows a chi-square distribution with d degrees of freedom, d being equal to the number of additional parameters in the TLLN model. By using this result and standard statistical tables, we can obtain critical values for the LR test statistics for the given astronomical dataset. Table 7 includes the LR statistics and corresponding p-values for both the datasets. Given the values of test statistics and their associated p-values, we reject the null hypothesis for the above-discussed astronomical dataset and conclude that the TLLN distribution provides a significantly better representation than the LN and ELN distributions.

Table 7. Likelihood ratio statistics and their p-values on the astronomy dataset.

	LR	p-value
TLLN versus LN	21.594	3.37×10^{-06}
TLLN versus ELN	3.0586	2.2×10^{-16}

Now, we use a real, censored dataset based on the prognosis for women with breast cancer. Breast cancer is one of the most common forms of cancer in women. This lifetime dataset

was carried out at the Middlesex Hospital, and documented in Leathem and Brooks (1987) and discussed by Collett (2015) which refers to the survival time in months of women who had received a simple or radical mastectomy to treat a tumor of Grade II, III or IV, between January 1969 and December 1971.

Table 8 summarizes the TLLN regression model as a result of the censored dataset, including estimates of all parameters, negative log-likelihood ($-l(\boldsymbol{\psi})$) and value of AIC. Here, we utilize the `optim` function of the `R` software for the numerical evaluations.

Table 8. Summaries for the TLLN regression model from the breast cancer dataset.

Parameter	τ_0	τ_1	α	σ	$-l(\boldsymbol{\psi})$	AIC
Estimates	0.6372	-1.2392	40.3536	4.3415	154.2923	316.5846

Table 9 represents the summary of 1000 times iterated simulated results, due to the censored dataset using the MH algorithm of the MCMC method, which includes the posterior mean, standard deviation (SD), Monte Carlo standard error (MCSE), effective sample size due to autocorrelation (ESS), 95% CI and the posterior median. Next, we use the `LaplacesDemon` package of `R` for the numerical evaluations.

Table 9. Summaries for the TLLN Bayesian regression model from the breast cancer dataset.

Parameter	Mean	SD	MCSE	ESS	95% CI	Median
τ_0	2.9068	0.6033	0.2791	9.3047	(1.7993, 3.8864)	3.0101
τ_1	-1.1779	0.5730	0.1699	17.7204	(-1.9877, -0.2168)	-1.2055
α	11.4851	2.6373	1.1827	9.2548	(7.8012, 16.7872)	11.3755
σ	3.8347	0.6524	0.2513	8.6492	(2.7314, 5.3484)	3.6765

13. Conclusions, limitations, and future research

We suggest a new distribution, which is a generalized version of the log-normal distribution, mainly to investigate data in the field of astronomy in this research, but it can also be used to match cancer datasets in biological aspects. We call it the Topp-Leone log-normal distribution. We explore several of its mathematical and statistical aspects. On the theoretical aspect, we provide specific expressions for the moments, quantile function, and various reliability measures. The different shapes of the hazard rate function are discussed. In terms of inference, the model parameters are estimated by using Bayesian estimation and the method of maximum likelihood, and also, the observed information matrix is presented. Furthermore, we adopt the parametric bootstrap technique to obtain confidence intervals for the model parameters. More importantly, we introduce a parametric regression model and a Bayesian regression method based on the new distribution. The usefulness of our methodology is illustrated by two applications of real datasets, one related to the astronomical study and the other to censored cancer data, using goodness-of-fit tests. The novel model consistently outperforms previous models in the literature in terms of fitting. We anticipate that the proposed model will find a larger range of applications in the modeling of positive real-world datasets, including not only astronomy but also biology, physics, engineering, survival analysis, hydrology, economics, and other fields.

The possible limitations of the proposed distribution include the impossibility of modeling phenomena with possible negative values or presenting a bimodal nature. The construction of quantile regression models and bivariate variants of the TLLN distribution are two further possible directions for this research. Additional significant improvements and investigations are required for this study, which we will put into future research.

Author contributions Conceptualization, C.C., M.R.I., D.S.S., S.L.N., R.M.; methodology, C.C., M.R.I., D.S.S., S.L.N., R.M.; software, XX; validation, C.C., M.R.I., D.S.S., S.L.N., R.M.; formal analysis, C.C., M.R.I., D.S.S., S.L.N., R.M.; investigation, C.C., M.R.I., D.S.S., S.L.N., R.M.; writing-original draft preparation, C.C., M.R.I., D.S.S., S.L.N., R.M.; writing-review and editing, C.C., M.R.I., D.S.S., S.L.N., R.M. All authors have read and agreed the published version of the paper.

Acknowledgements The authors would express their gratefulness for the constructive criticism of the Editors-in-Chief and the anonymous reviewers which helped to considerably improve the quality of the paper.

Funding The authors received no financial support for the research, authorship, and/or publication of this article.

Conflicts of interest The authors declare no conflict of interest.

References

Abdalla, H. et al., 2017. Characterizing the γ-ray long-term variability of PKS 2155−304 with H.E.S.S. and Fermi-Lat. Astronomy and Astrophysics, 598, A39.

Afify, A.Z., Altun, E., Alizadeh, M., Ozel, G., and Hamedani, G.G., 2017. The odd exponentiated half-logistic-G family: Properties, characterizations and applications. Chilean Journal of Statistics, 8, 65–91.

Alzaatreh, A., Lee, C., and Famoye, F., 2013. A new method for generating families of continuous distributions. Metron, 71, 63–79.

Bernardeau, F. and Kofman, L., 1994. Properties of the cosmological density distribution function. The Astrophysical Journal, 443, 479.

Blasi, P., Burles, S., and Olinto, a., 1999. Cosmological magnetic field limits in an inhomogeneous universe. The Astrophysical Journal Letters, 514, L79–L82.

Collett, D., 2015. Modelling Survival Data in Medical Research. Chapman and Hall, New York, USA.

Cooray, K. and Ananda, M.M.A., 2008. A generalization of the half-normal distribution with applications to lifetime data. Communications in Statistics: Theory and Methods, 37, 1323–1337.

Elbatal, I., Merovci, F., and Elgarhy, M., 2013. A new generalized Lindley distribution. Mathematical Theory and Modeling, 3, 30–47.

Feigelson, E. and Babu, G.J., 2012. In Modern Statistical Methods for Astronomy: With R Applications. Cambridge University Press, Cambridge, UK.

Gelman, A. and Hill, J., 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge, UK.

Hubble, E., 1934. The distribution of extra-galactic nebulae. Astrophysical Journal, 79, 8.

Jobe, J., Crow, E., and Shimizu, K., 1989. Lognormal distributions: Theory and applications. Technometrics, 31, 392.

Leathem, A. and Brooks, S., 1987. Predictive value of lectin binding on breast-cancer recurrence and survival. The Lancet, 329, 1054–1056.

Nantais, J.B., Huchra, J.P., Barmby, P., Olsen, K.A.G., and Jarrett, T.H., 2006. Nearby spiral globular cluster systems. I. Luminosity functions. The Astronomical Journal, 131, 1416–1425.

Parravano, A., Sánchez, N., and Alfaro, E.J., 2012. The dependence of prestellar core mass distributions on the structure of the parental cloud. Astrophysical Journal, 754, 150.

Sangsanit, Y. and Bodhisuwan, W., 2016. The Topp-Leone generator of distributions: Properties and inferences. Songklanakarin Journal of Science and Technology, 38, 537–548.

Shah, Z., Mankuzhiyil, N., Sinha, A., Misra, R., Sahayanathan, S., and Iqbal, N., 2018. Log-normal flux distribution of bright fermi blazars. Research in Astronomy and Astrophysics, 18, 141.

Shah, Z., Misra, R., and Sinha, A., 2020. On the determination of lognormal flux distributions for astrophysical systems. Monthly Notices of the Royal Astronomical Society, 496, 3348–3357.

Sharma, V., 2018. Topp-Leone normal distribution with application to increasing failure rate data. Journal of Statistical Computation and Simulation, 88, 1782–1803.

Shaw, W. and Buckley, I., 2009. The alchemy of probability distributions: beyond Gram-Charlier expansions, and a skew-kurtotic-normal distribution from a rank transmutation map. arXiv:0901.0434.

Topp, C. and Leone, F., 1955. A family of J-shaped frequency functions. Journal of The American Statistical Association, 50, 209–219.

Upadhyay, S. K., Vasishta, N., and Smith, A. F. M., 2001. Bayes inference in life testing and reliability via Markov chain monte Carlo simulation. Sankhya A, 63, 15–40.

Chilean Journal of Statistics
Volume 13, Number 1 (April 2022), Pages 91–111
DOI: 10.32372/ChJS.13-01-05

Distribution Theory
Research Paper

The Mc-Donald Chen distribution: A new bimodal distribution with properties and applications

Lucas D. Ribeiro-Reis[1,*], Gauss M. Cordeiro[1], and José J. de Santana e Silva[1]

[1]Department of Statistics, Universidade Federal de Pernambuco, Recife, Brazil

(Received: 24 June 2021 · Accepted in final form: 02 March 2022)

Abstract

In this paper, the McDonald-Chen distribution is proposed and studied to model different type of data. Its probability density function allows bimodality, thus showing that the model is very flexible. Its failure or hazard rate function may have increasing, decreasing, bathtub, inverted bathtub and increasing-decreasing-increasing shapes depending on the parameter values. The new distribution includes at least five major special cases. Some of its mathematical properties are addressed. The maximum likelihood method is adopted to estimate the model parameters. Monte Carlo simulations evaluate the accuracy of the maximum likelihood estimators. The new distribution is better than three other popular distributions to model two real data sets.

Keywords: Chen distribution · Family of distributions · Maximum likelihood method · Moments · Monte Carlo Simulation.

Mathematics Subject Classification: 46N30 · 78M31.

1. Introduction

Several distributions have been proposed to model data in real applications. Lai (2013) detailed the importance of building new survival distributions and the fact that the failure or hazard rate curves could accommodate different shapes. Thus, there is a need for distributions that are quite flexible to model these shapes. Among the different mechanisms for proposing new continuous distributions, we have: transformation of the random variable; random variable convolution; random variable composition (Cordeiro et al., 2018); mixing distributions between random variables (Nedjar and Zeghdoudi, 2016); distributions that transform the cumulative distribution (Bourguignon et al., 2014). The choice of generated distributions can be carried out using transformation in the cumulative distribution. Some distributions generated using this technique are the beta modified Weibull (Silva et al., 2010), gamma modified Weibull (Cordeiro et al., 2015), transmuted Dagum (Elbatal and Aryal, 2015), Harris extended Lindley (Cordeiro et al., 2019).

Eugene et al. (2002) pioneered the beta-generalized (beta-G) family, which includes nearly all of well-known models as special cases. Further, it can give lighter and heavier tails and be

*Corresponding author. Email: econ.lucasdavid@gmail.com

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

http://www.soche.cl/chjs

applied in several areas such as engineering and biological research, among others. Explicit expressions are reported in several published papers, which facilitate to find its mathematical properties for special models. In the last years, several beta-G models have been proposed; see the list of forty five special models in Table 3 of Tahir et al. (2015). This family has the major benefit for fitting skewed data that can not be fitted by most well-known continuous distributions.

In this paper, a flexible extension of the Chen distribution (Chen, 2000) is proposed, which can be useful in several practical contexts. In particular, adding shape parameters to a baseline distribution can provide better fits to real data in different settings and extended Chen distribution has interesting mathematical properties.

The paper is unfolded as follows. In Section 2, a brief introduction to the McDonald-Chen (MC) distribution is given. In Section 3, the quantile function (QF) of the MC distribution is determined. In Section 4, the new probability density function (PDF) is expressed as a linear combination of Chen PDFs. Moments and moments generating function are obtained in Section 5. In Section 6, its parameters are estimated by the maximum likelihood (ML) method. In Section 7, some simulation results verify the precision of the parameter estimates. In Section 8, the MC distribution is proved to outperform some well-known lifetime models. Finally, Section 9 offers some concluding remarks.

2. Background

Based on the beta-G family, Alexander et al. (2012) defined the cumulative distribution function (CDF) and PDF of the McDonald-generalized (MG) class of distributions as

$$F(x;a,b,c,\boldsymbol{\theta}) = \frac{B_{G(x;\boldsymbol{\theta})^c}(a,b)}{B(a,b)} = \frac{1}{B(a,b)}\int_0^{G(x)^c} w^{a-1}(1-w)^{b-1}\mathrm{d}w \tag{1}$$

and

$$f(x;a,b,c,\boldsymbol{\theta}) = \frac{c}{B(a,b)}g(x;\boldsymbol{\theta})\,G(x;\boldsymbol{\theta})^{ac-1}\left[1-G(x;\boldsymbol{\theta})^c\right]^{b-1}, \tag{2}$$

respectively, where $\boldsymbol{\theta}$ is the parameter vector of the baseline distribution $G(x;\boldsymbol{\theta})$, $g(x;\boldsymbol{\theta}) = \mathrm{d}(x;\boldsymbol{\theta})/\mathrm{d}x$, a, b and c are three positive additional shape parameters, $B(a,b) = \int_0^1 w^{a-1}(1-w)^{b-1}\mathrm{d}w$ denotes the beta function and $B_z(a,b) = \int_0^z w^{a-1}(1-w)^{b-1}\mathrm{d}w$ denotes the lower incomplete beta function.

Let $X \sim \mathrm{MG}(a,b,c,\boldsymbol{\theta})$ be a random variable X having PDF as given in Equation (2). Although this transformation is simple, the MG family is richer than the corresponding baseline $G(x)$. For $G(x) = x$, the MG family reduces to the McDonald distribution pioneered by Mcdonald (2008). For $c = 1$ in Equation (1), it follows the beta-G class defined by Eugene et al. (2002). For $a = 1$, Equation (1) coincides with the Kumaraswamy-generalized (Kw-G) class introduced by Cordeiro and de Castro (2011). The MG family is quite important, since it includes as special cases two of the most well-known classes in the literature, which generated many published distributions in the last twenty years. According to Cordeiro et al. (2012a), the MG family allows greater flexibility in its tails and can be widely used in engineering, biology and other areas.

The hazard rate function (HRF) of X is given by

$$\tau(x;a,b,c,\boldsymbol{\theta}) = \frac{cg(x;\boldsymbol{\theta})G(x;\boldsymbol{\theta})^{ac-1}[1-G(x;\boldsymbol{\theta})^c]^{b-1}}{1-B_{G(x;\boldsymbol{\theta})^c}(a,b)}. \tag{3}$$

The Chen distribution is taken as baseline, since it allows us to model data with bathtub HRF. The CDF and PDF of the Chen distribution are stated as

$$G(y;\lambda,\beta) = 1 - \mathrm{e}^{\lambda(1-\mathrm{e}^{y^\beta})},\ y>0 \tag{4}$$

and

$$g(y;\lambda,\beta) = \lambda\beta y^{\beta-1}\mathrm{e}^{y^\beta+\lambda(1-\mathrm{e}^{y^\beta})},\ y>0, \tag{5}$$

respectively, where $\lambda>0$ is the scale parameter and $\beta>0$ is the shape parameter. Henceforth, $Y\sim \text{Chen}(\lambda,\beta)$ denotes a random variable with PDF as given in Equation (5).

By taking G and g as the CDF and PDF of the Chen distribution, respectively, and substituting in Equations (1), (2) and (3), the CDF, PDF and HRF of the MC distribution are formulated as

$$F(x;a,b,c,\lambda,\beta) = \frac{1}{B(a,b)}\int_0^{\left[1-\mathrm{e}^{\lambda(1-\mathrm{e}^{x^\beta})}\right]^c} w^{a-1}(1-w)^{b-1}\mathrm{d}w, \tag{6}$$

$$f(x;a,b,c,\lambda,\beta) = \frac{c\lambda\beta}{B(a,b)}x^{\beta-1}\mathrm{e}^{x^\beta+\lambda(1-\mathrm{e}^{x^\beta})}\left[1-\mathrm{e}^{\lambda(1-\mathrm{e}^{x^\beta})}\right]^{ac-1}\left\{1-\left[1-\mathrm{e}^{\lambda(1-\mathrm{e}^{x^\beta})}\right]^c\right\}^{b-1} \tag{7}$$

and

$$\tau(x;a,b,c,\lambda,\beta) = \frac{c\lambda\beta x^{\beta-1}\mathrm{e}^{x^\beta+\lambda(1-\mathrm{e}^{x^\beta})}\left[1-\mathrm{e}^{\lambda(1-\mathrm{e}^{x^\beta})}\right]^{ac-1}\left\{1-\left[1-\mathrm{e}^{\lambda(1-\mathrm{e}^{x^\beta})}\right]^c\right\}^{b-1}}{1-\int_0^{\left[1-\mathrm{e}^{\lambda(1-\mathrm{e}^{x^\beta})}\right]^c} w^{a-1}(1-w)^{b-1}\mathrm{d}w},$$

respectively.

Henceforth, let $X\sim \text{MC}(a,b,c,\lambda,\beta)$ have PDF as given in Equation (7). For $c=1$, the MC distribution becomes the beta-Chen (BC), not yet known in the literature. For $a=1$, it follows the (new) Kumaraswamy-Chen (KC). Further, Equation (7) reduces to the exponentiated-Chen ($b=c=1$) (Dey et al., 2017), exponentiated-Chen Lehmann type 2 ($a=c=1$) and the Chen itself ($a=b=c=1$) distributions.

Figure 1 displays plots of the MC PDFs for selected parameter values, where it is shown that this distribution is quite flexible having several forms including bimodality.

The HRF curves for some parameter choices are given in Figure 2. The HRF of X can be increasing, decreasing, unimodal, crescent-descending-crescent and bathtub shape, which shows once again its great flexibility.

Summing up what was said above, we cite six basic motivations for the MC distribution: (i) greater flexibility in the PDF and HRF. In fact, its PDF has bimodality, increasing, decreasing, bathtub and inverted shapes of the HRF, whereas the Chen PDF has only increasing, decreasing and unimodal shapes. In addition, the rate of the MC distribution can be increasing, decreasing, bathtub, inverted bathtub, and increasing-decreasing-increasing shapes. This last form exists for few distributions, but it can be found in many real data sets; (ii) make the skewness and kurtosis more flexible compared to the Chen distribution. The parameter c of the new distribution changes substantially the values of its skewness and kurtosis as shown in the plots of Figures 3 and 4, thus making it very interesting for real applications; (iii) provide consistently better fits than other lifetime models as proved empirically in Section 8; (iv) the proposed distribution includes five others sub-models that

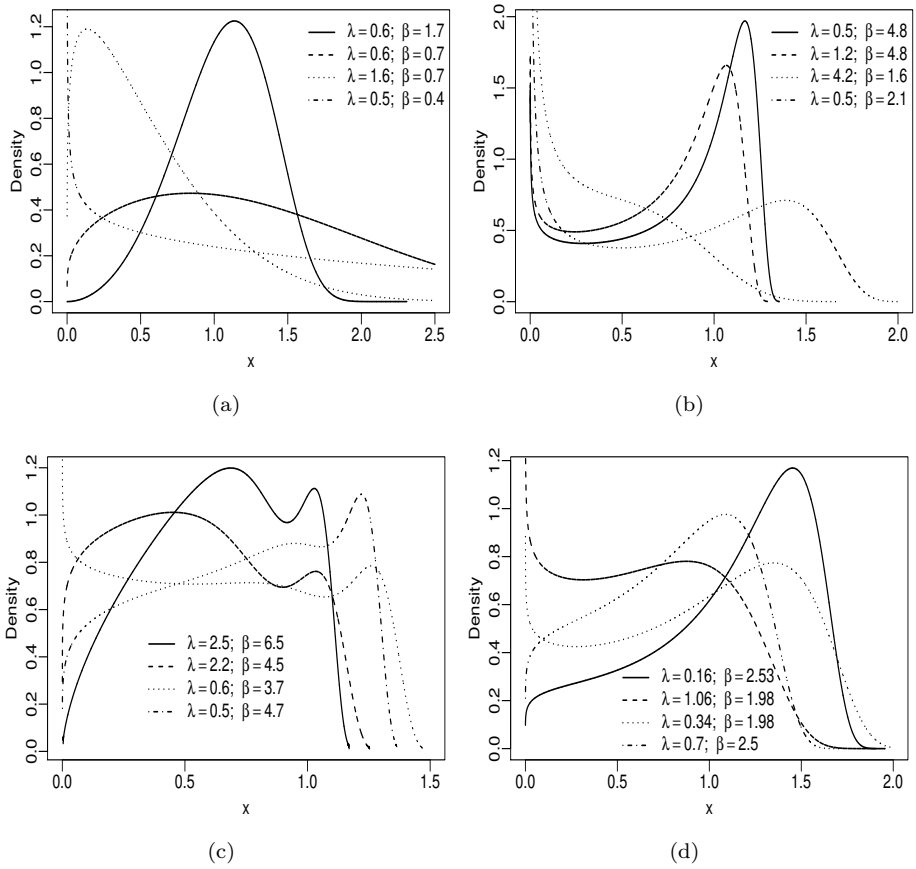


Figure 1. The MC PDF for some parameters values: (a) MC$(0.5, 0.8, 3.5, \lambda, \beta)$, (b) MC$(0.86, 0.32, 0.18, \lambda, \beta)$, (c) MC$(0.05, 0.2, 5, \lambda, \beta)$ and (d) MC$(0.5, 0.5, 0.9, \lambda, \beta)$.

can be compared by using likelihood ratio (LR) tests to choose the best model to explain a data set; (v) the properties of the new distribution are easily obtained from those of Chen due to a linear representation for its PDF; and (vi) construct heavy–tailed special cases that are not longer-tailed for modeling real data.

3. Quantile function

The QF of the MG family, say $Q(u; a, b, c, \boldsymbol{\theta}) = F^{-1}(u; a, b, c, \boldsymbol{\theta})$, can be expressed in terms of the beta QF. Basically, according to Cordeiro et al. (2012b), the QF of the MG distribution (for $0 < u < 1$) has the form

$$Q(u; a, b, c, \boldsymbol{\theta}) = Q_{\mathrm{G}}\{Q_{\beta}(u; a, b)^{\frac{1}{c}}; \boldsymbol{\theta}\},$$

where Q_{G} is the QF of the baseline G and $Q_{\beta}(u; a, b)$ is the beta QF with parameters a and b; see the Wolfram website at http://functions.wolfram.com/06.23.06.0004.01.

Thus, the QF of the MC distribution can be expressed as

$$Q(u; a, b, c, \lambda, \beta) = \left\{\log\left[1 - \frac{1}{\lambda}\log\left(1 - Q_{\beta}(u; a, b)^{\frac{1}{c}}\right)\right]\right\}^{\frac{1}{\beta}}, \ 0 < u < 1.$$

The simulation of X is very easy. If U is a uniform random variable on the unit interval

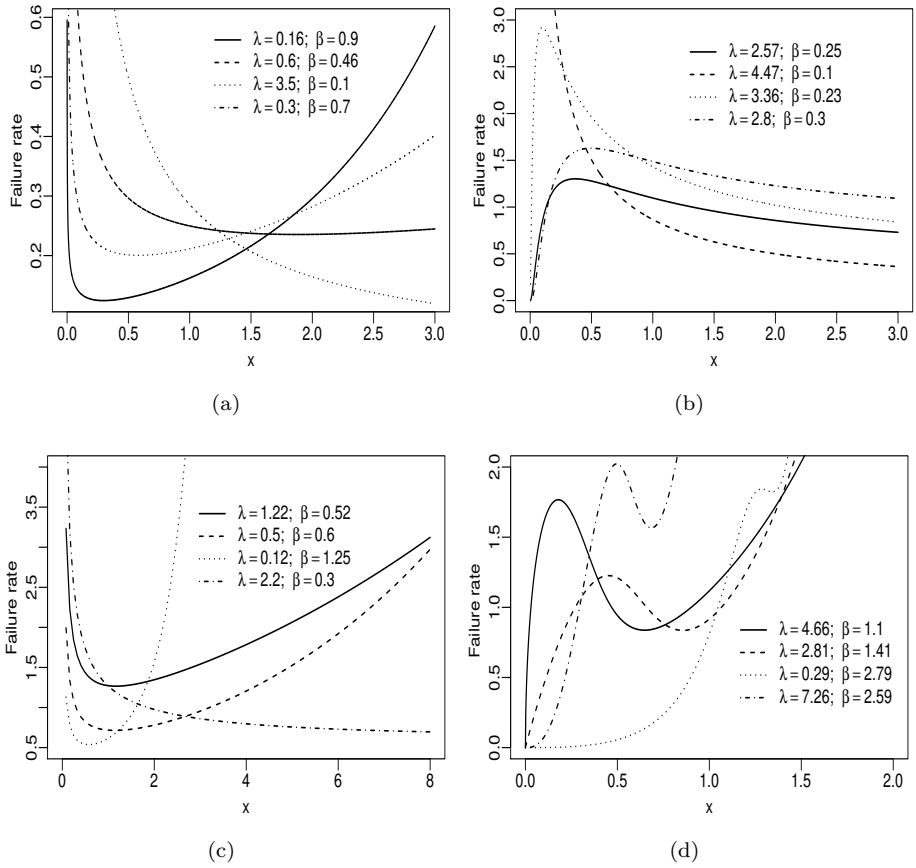


Figure 2. The HRF of the MC model for some parameters values: (a) MC$(0.6, 0.3, 1.35, \lambda, \beta)$, (b) MC$(1.2, 0.7, 15, \lambda, \beta)$, (c) MC$(0.66, 0.7, 0.35, \lambda, \beta)$ and (d) MC$(0.07, 0.08, 20, \lambda, \beta)$.

$(0, 1)$, then

$$X = \left\{ \log \left[1 - \frac{1}{\lambda} \log \left(1 - Q_\beta(U; a, b)^{\frac{1}{c}} \right) \right] \right\}^{\frac{1}{\beta}},$$

wis an MC distributed random variable.

Let $Q(u) = Q(u; a, b, c, \lambda, \beta)$ be the QF of the MC distribution by omitting the arguments. The baseline parameters are $\lambda = 5$ and $\beta = 1.62$ and c varies in $\{0.2, 1, 5, 10\}$ for the scenarios (a)-(d), respectively, to study the influence of the generator parameters a and b on the skewness and kurtosis of the MC distribution. The parameters a and b vary in the interval $(0.1, 1)$. Figure 3 displays the Bowley skewness, as functions of a and b, defined as

$$B = \frac{Q(3/4) + Q(1/4) - 2\,Q(2/4)}{Q(3/4) - Q(1/4)}.$$

The minimum and maximum values for B are then $(-0.1542,\ 1.0000)$, $(-0.1272,\ 0.8406)$, $(-0.0690,\ 0.3101)$ and $(-0.0425,\ 0.2649)$ for the scenarios (a)-(d), respectively. For the selected parameter values, the asymmetry becomes increasingly negative when c increases.

Consider the same parameter values for the Moor kurtosis, as functions of a and b, expressed as

$$M = \frac{Q(7/8) - Q(5/8) - Q(3/8) + Q(1/8)}{Q(6/8) - Q(2/8)},$$

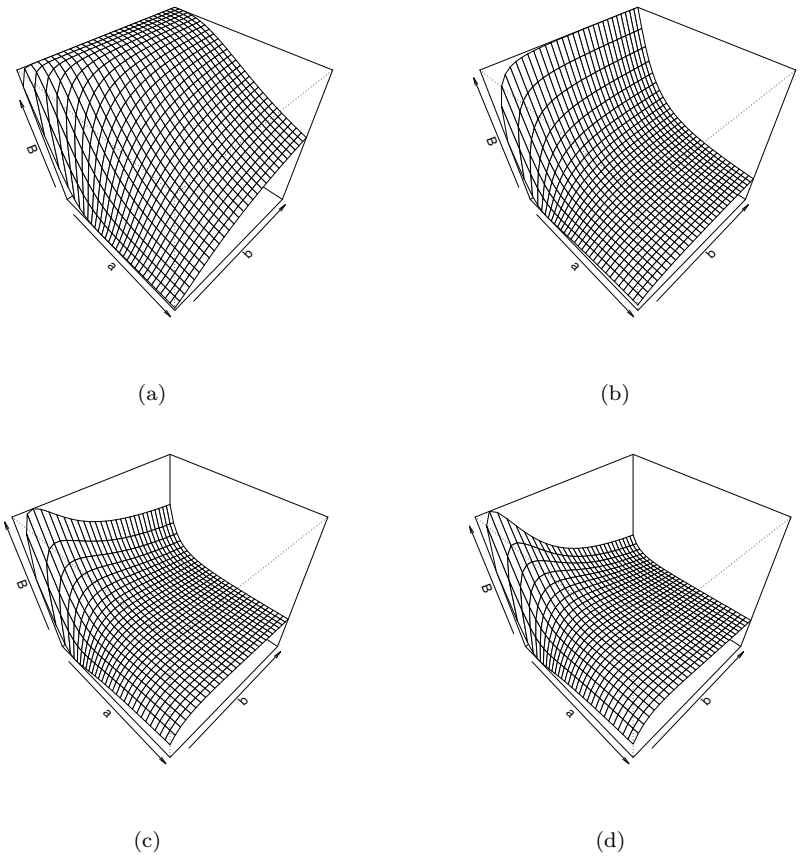


Figure 3. Bowley skewness as function of c: (a) $c = 0.2$, (b) $c = 1$, (c) $c = 5$ and (d) $c = 10$.

Figure 4 displays the Moor kurtosis, where the minimum and maximum values of M are $(-0.1985, 116.5147)$, $(-0.1667, 2.4161)$, $(-0.0744, 0.6191)$ and $(-0.0260, 0.4811)$ for the scenarios (a)-(d), respectively. Small values of c give higher kurtosis. The kurtosis decreases and stabilizes when c increases.

4. Linear representation

Equations (6) and (7) can be expressed in terms of exponentiated distributions. For a given CDF $G(z;\boldsymbol{\theta})$ with parameter vector $\boldsymbol{\theta}$, the random variable Z is exponentiated-G (exp-G) distributed, with power parameter $a > 0$, if its CDF and PDF are

$$H(z; a, \boldsymbol{\theta}) = G(z;\boldsymbol{\theta})^a, \quad h(x) = a\, g(z;\boldsymbol{\theta}) G(z;\boldsymbol{\theta})^{a-1},$$

respectively, where $g(z;\boldsymbol{\theta}) = \mathrm{d}G(z;\boldsymbol{\theta})/\mathrm{d}z$. The exp-G model is also called the Lehmann type I distribution. From now on, we denote it as $Z \sim \text{exp-G}(a, \boldsymbol{\theta})$.

Following Alexander et al. (2012), Equation (2) can be expressed as

$$f(x; a, b, c, \boldsymbol{\theta}) = \sum_{k=0}^{\infty} b_k\, h(x; c(a+k), \boldsymbol{\theta}), \tag{8}$$

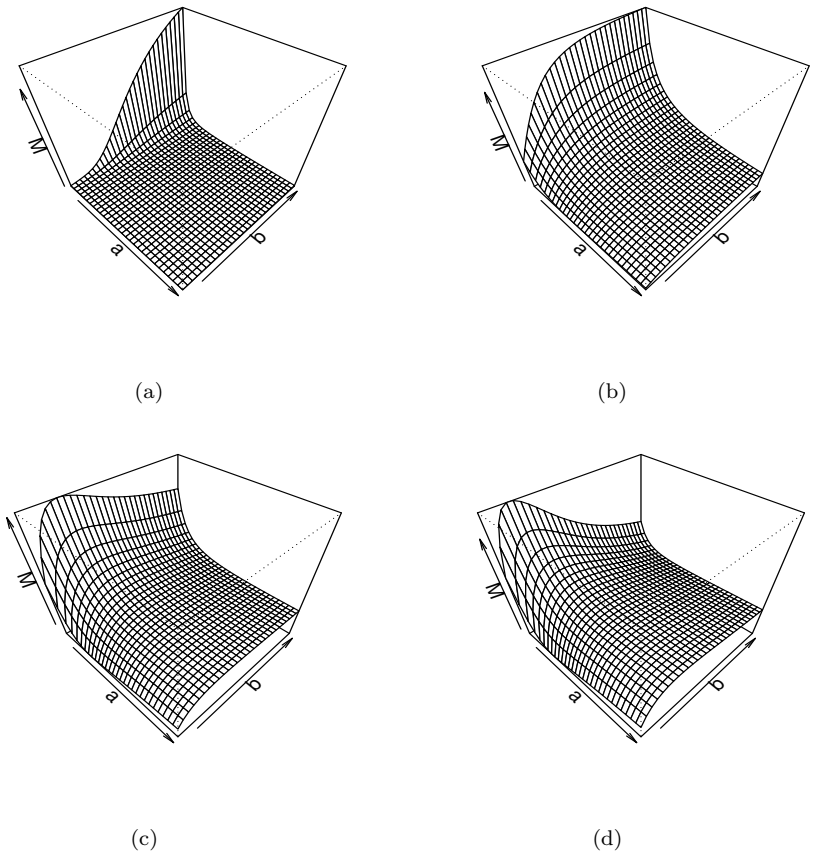


Figure 4. Moor kurtosis as function of c: (a) $c = 0.2$, (b) $c = 1$, (c) $c = 5$ and (d) $c = 10$.

where $h(x; c(a+k), \boldsymbol{\theta})$ is the exp-G$(c(a+k), \boldsymbol{\theta})$ PDF, and the coefficients b_k are

$$b_k = \frac{(-1)^k \, \Gamma(a+b)}{(a+k)\, k! \, \Gamma(a)\, \Gamma(b-k)},$$

where $\Gamma(p) = \int_0^\infty w^{p-1} \mathrm{e}^{-w} \mathrm{d}w$ denotes the gamma function. We can prove that $\sum_{k=0}^{\infty} b_k = 1$.

Equation (8) reveals that MG PDF is a linear combination of exp-G PDFs. Thus, several MG properties can be determined by knowing those corresponding exp-G properties (Cordeiro et al., 2012a). By integrating Equation (8), the MG CDF follows as

$$F(x; a, b, c, \boldsymbol{\theta}) = \sum_{k=0}^{\infty} b_k \, H(x; c(a+k), \boldsymbol{\theta}),$$

where $H(x; c(a+k), \boldsymbol{\theta})$ is the exp-G$(c(a+k), \boldsymbol{\theta})$ CDF.

Theorem 4.1 Let Y be a random variable having a Chen CDF as given in Equation (4). Then, the CDF and PDF of the exp-Chen(a, λ, β) distribution are stated as

$$H(y; a, \lambda, \beta) = 1 + \sum_{m=1}^{\infty} (-1)^m \binom{a}{m} [1 - G(y; m\lambda, \beta)]$$

and

$$h(y;a,\lambda,\beta)=\sum_{m=1}^{\infty} w_m(a)\, g(y;m\lambda,\beta),$$

respectively, where $w_m(a)=(-1)^{m+1}\binom{a}{m}$.

Proof For $|x|<1$ and any real $a\neq 0$, the convergent power series holds by means of

$$(1-x)^a=\sum_{m=0}^{\infty}(-1)^m\binom{a}{m}x^m.$$

Thus, the CDF of the exp-Chen distribution is given by

$$\begin{aligned}H(y;a,\lambda,\beta)&=\left[1-\mathrm{e}^{\lambda(1-\mathrm{e}^{y^\beta})}\right]^a=\sum_{m=0}^{\infty}(-1)^m\binom{a}{m}\mathrm{e}^{m\lambda(1-\mathrm{e}^{y^\beta})}\\&=1+\sum_{m=1}^{\infty}(-1)^m\binom{a}{m}[1-G(y;m\lambda,\beta)].\end{aligned}$$

By differentiating the last equation, we have that

$$h(y;a,\lambda,\beta)=\sum_{m=1}^{\infty}(-1)^{m+1}\binom{a}{m}g(y;m\lambda,\beta),$$

which shows that the exp-Chen PDF is a linear combination of Chen PDFs. ■

Based on Equation (8) and Theorem 4.1, the PDF of X can be expressed as

$$f(x;a,b,c,\lambda,\beta)=\sum_{m=1}^{\infty} d_m\, g(x;m\lambda,\beta), \tag{9}$$

where

$$d_m=d_m(a,b,c)=\sum_{k=0}^{\infty}\frac{(-1)^{k+m+1}\Gamma(a+b)}{(a+k)\,k!\,\Gamma(a)\,\Gamma(b-k)}\binom{c(a+k)}{m},$$

and $g(x;m\lambda,\beta)$ is the Chen PDF with scale parameter $m\lambda$ and shape parameter β. Clearly, the shape parameters of the MC generation are restricted to the coefficients in Equation (9).

Some mathematical properties of the MC distribution can be derived from Equation (9) and those properties of the Chen distribution. For example, the ordinary and incomplete moments and moment generating function (MGF) of X can be determined from the corresponding quantities of the Chen distribution. Consequently, the beta-Chen and Kw-Chen PDFs are also linear combinations of Chen PDFs when $c=1$ and $a=1$, respectively.

By integrating Equation (9), the CDF of the MC distribution is given by

$$F(x;a,b,c,\lambda,\beta)=\sum_{m=1}^{\infty} d_m\, G(x;m\lambda,\beta),$$

where $G(x;m\lambda,\beta)$ is the CDF of the Chen$(m\lambda,\beta)$ distribution.

5. Moments and Moment generating function

Let Y_m be a random variable having the Chen PDF with scale parameter $m\lambda$ and shape parameter β, that is, $Y_m \sim \text{Chen}(m\lambda, \beta)$. By using Equation (9), the rth moment of X can be written as

$$\mathrm{E}[X^r] = \sum_{m=1}^{\infty} d_m \, \mathrm{E}[Y_m^r].$$

Pogany et al. (2017) demonstrated that the rth moment of Y has the form

$$\mathrm{E}[Y^r] = \lambda \, \mathrm{e}^{\lambda} \, \mathrm{D}_t^{r\beta^{-1}} \left[\frac{\Gamma(t+1,\lambda)}{\lambda^{t+1}} \right]_{t=0}. \tag{10}$$

Here, we have that

$$\mathrm{D}_t^p \left[\frac{\Gamma(t+1,\lambda)}{\lambda^{t+1}} \right]_{t=0} = \Gamma(p+1) \sum_{k\geq 0} \frac{(2)_k}{k!} \, \Phi_{\mu,1}^{(0,1)}(-k, p+1, 1) \; {}_1F_1(k+2; 2; -\lambda),$$

where $\Phi_{\mu,1}^{(0,1)}(-a, p+1, 1) = \sum_{n\geq 0} (-a)^n / n!(n+1)^{p+1}$ for $\mu \in \mathrm{C}$, ${}_1F_1(a; b; x) = \sum_{n\geq 0} (a)_n x^n / (b)_n n!$, for $x, a \in \mathrm{C}$ and $b \in \mathrm{C} \setminus Z_0^-$, is the confluent hypergeometric function (Kilbas et al., 2006, p. 29, Eq. 1.6.14) and $(\lambda)_\eta = \Gamma(\lambda+\eta)/\Gamma(\lambda)$, for $\lambda \in \mathrm{C} \setminus \{0\}$, is the generalized Pochhammer symbol, under the convention $(0)_0 = 1$.

The rth ordinary moment of X follows from Equation (10) as

$$\mathrm{E}[X^r] = \lambda \sum_{m=1}^{\infty} m \, d_m \, \mathrm{e}^{m\lambda} \, \mathrm{D}_t^{r\beta^{-1}} \left[\frac{\Gamma(t+1, m\lambda)}{(m\lambda)^{t+1}} \right]_{t=0}.$$

The incomplete moments of a distribution have great applicability to measure inequality. The first incomplete moment is used to construct Lorenz and Bonferroni curves.

For $z > 0$, the rth incomplete moment of Y, say $q_r(z; \lambda, \beta) = \int_0^z y^r \, g(y; \lambda, \beta) \mathrm{d}y$, follows from Pogany et al. (2017) as

$$q_r(z; \lambda, \beta) = \lambda \mathrm{e}^{\lambda} \sum_{n,k\geq 0} \sum_{j=1}^{k} \frac{(2)_{n+k}}{(2)_n} \frac{(-1)^{n+j} \lambda^n \binom{k}{j}}{n!k!(j+1)^{r\beta^{-1}+1}} \gamma(r\beta^{-1}, (j+1)(1-z^{-1})), \tag{11}$$

where $\gamma(p, z) = \int_0^z w^{p-1} \, \mathrm{e}^{-w} \mathrm{d}w$ denotes the lower incomplete gamma function.

The rth incomplete moment of X can be expressed from Equation (9) as

$$m_r(z) = \sum_{m=1}^{\infty} d_m \, q_r(z; m\lambda, \beta),$$

which depends directly on the rth incomplete moment of the Chen$(m\lambda, \beta)$ distribution.

By using Equation (11), the rth incomplete moment de X can be written as

$$m_r(z) = \lambda \sum_{m=1}^{\infty} m \, \mathrm{e}^{m\lambda} \, d_m \sum_{n,k\geq 0} \sum_{j=1}^{k} \frac{(2)_{n+k}}{(2)_n} \frac{(-1)^{n+j} (m\lambda)^n \binom{k}{j}}{n!k!(j+1)^{r\beta^{-1}+1}} \gamma(r\beta^{-1}, (1-z^{-1})(j+1)).$$

The MGF of Y, say $M_Y(t) = \mathrm{E}\left[\mathrm{e}^{-tY}\right]$, for $t > 0$, can be determined from Pogany et al. (2017) as

$$M_Y(t) = \lambda\,\beta\,\mathrm{e}^{\lambda}\,t^{-\beta}\sum_{n\geq 0}\frac{(-\lambda)^n}{n!}\,{}_1\Psi_0\left[(\beta,\beta);-;\frac{n+1}{t^{\beta}}\right], \tag{12}$$

where

$${}_1\Psi_0\left[(a,b);-;z\right] = \sum_{n\geq 0}\frac{\Gamma(a+bn)\,z^n}{n!}, \quad z,a\in\mathrm{C}, b>0,$$

is the generalized Fox-Wright function. Thus, using Equations (9) and (12), the MGF of X is stated as

$$M_X(t) = \lambda\,\beta\,\mathrm{e}^{\lambda}\,t^{-\beta}\sum_{m=1}^{\infty}\sum_{n\geq 0}\frac{(-m\lambda)^n\,d_m}{n!}\,{}_1\Psi_0\left[(\beta,\beta);-;\frac{n+1}{t^{\beta}}\right].$$

6. Estimation

The ML estimators enjoy desirable properties that can be used when constructing confidence intervals for the model parameters. Let $X_1,\ldots,X_n$ be a random sample of size n from $X \sim \mathrm{MC}(a,b,c,\lambda,\beta)$ with observations $x_1,\ldots,x_n$. The log-likelihood function for $\boldsymbol{\theta} = (a,b,c,\lambda,\beta)^{\top}$ from this sample is formulated as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) =& n[\log c\lambda\beta - \log B(a,b) + \lambda] + (\beta-1)\sum_{i=1}^{n}\log x_i + \sum_{i=1}^{n}x_i^{\beta} - \lambda\sum_{i=1}^{n}\mathrm{e}^{x_i^{\beta}}\\ &+ (ac-1)\sum_{i=1}^{n}\log t(x_i) + (b-1)\sum_{i=1}^{n}\log\{1-t(x_i)^c\},\end{aligned} \tag{13}$$

where $t(x_i) = 1-\exp\{\lambda(1-\mathrm{e}^{x_i^{\beta}})\}$.

The function $\mathcal{L}(\boldsymbol{\theta})$ can be maximized either directly by using well-known platforms such as the R (`optim` function), SAS (`PROC NLMIXED`), Ox program (`MaxBFGS` sub-routine) or by solving the nonlinear likelihood equations of the score vector obtained by differentiating Equation (13).

The components of the score vector $U(\boldsymbol{\theta})$ are given by

$$\begin{aligned}U_a(\boldsymbol{\theta}) &= n\psi(a+b) - n\psi(a) + c\sum_{i=1}^{n}\log t(x_i),\\ U_b(\boldsymbol{\theta}) &= n\psi(a+b) - n\psi(b) + \sum_{i=1}^{n}\log\{1-t(x_i)^c\},\\ U_c(\boldsymbol{\theta}) &= \frac{n}{c} + a\sum_{i=1}^{n}\log t(x_i) - (b-1)\sum_{i=1}^{n}\frac{t(x_i)^c\log t(x_i)}{1-t(x_i)^c},\\ U_{\lambda}(\boldsymbol{\theta}) &= \frac{n}{\lambda} + n - \sum_{i=1}^{n}\mathrm{e}^{x_i^{\beta}} - (ac-1)\sum_{i=1}^{n}\frac{r(x_i)}{t(x_i)} + c(b-1)\sum_{i=1}^{n}\frac{r(x_i)t(x_i)^{c-1}}{1-t(x_i)^c},\end{aligned}$$

$$U_\beta(\boldsymbol{\theta}) = \frac{n}{\beta} + \sum_{i=1}^{n} \log x_i + \sum_{i=1}^{n} x_i^\beta \log x_i - \lambda \sum_{i=1}^{n} x_i^\beta \mathrm{e}^{x_i^\beta} \log x_i$$
$$+\lambda(ac-1)\sum_{i=1}^{n} \frac{s(x_i)}{t(x_i)} - c\lambda(b-1)\sum_{i=1}^{n} \frac{s(x_i)t(x_i)^{c-1}}{1-t(x_i)^c},$$

where $\psi(q) = \mathrm{d}\log\Gamma(q)/\mathrm{d}q$ is the digamma function, $r(x_i) = (1-\mathrm{e}^{x_i^\beta})\mathrm{e}^{\lambda(1-\mathrm{e}^{x_i^\beta})}$ and $s(x_i) = x_i^\beta\, \mathrm{e}^{\lambda(1-\mathrm{e}^{x_i^\beta})+x_i^\beta} \log x_i$.

The ML estimate $\widehat{\boldsymbol{\theta}} = (\widehat{a}, \widehat{b}, \widehat{c}, \widehat{\lambda}, \widehat{\beta})^\top$ of $\boldsymbol{\theta} = (a, b, c, \lambda, \beta)^\top$ is determined by the simultaneous solutions of the equations $U(\boldsymbol{\theta}) = \mathbf{0}$. These solutions are those $\widehat{\boldsymbol{\theta}}$ values that maximize Equation (13). The estimates of the unknown parameters can not be obtained analytically, and then interactive methods such as the quasi-Newton BFGS and Newton-Raphson algorithms are required.

The estimated observed information matrix is given by

$$J(\boldsymbol{\theta}) = -\begin{bmatrix} U_{aa}(\boldsymbol{\theta}) & U_{ab}(\boldsymbol{\theta}) & U_{ac}(\boldsymbol{\theta}) & U_{a\lambda}(\boldsymbol{\theta}) & U_{a\beta}(\boldsymbol{\theta}) \\ U_{ba}(\boldsymbol{\theta}) & U_{bb}(\boldsymbol{\theta}) & U_{bc}(\boldsymbol{\theta}) & U_{b\lambda}(\boldsymbol{\theta}) & U_{b\beta}(\boldsymbol{\theta}) \\ U_{ca}(\boldsymbol{\theta}) & U_{cb}(\boldsymbol{\theta}) & U_{cc}(\boldsymbol{\theta}) & U_{c\lambda}(\boldsymbol{\theta}) & U_{c\beta}(\boldsymbol{\theta}) \\ U_{\lambda a}(\boldsymbol{\theta}) & U_{\lambda b}(\boldsymbol{\theta}) & U_{\lambda c}(\boldsymbol{\theta}) & U_{\lambda\lambda}(\boldsymbol{\theta}) & U_{a\beta}(\boldsymbol{\theta}) \\ U_{\beta a}(\boldsymbol{\theta}) & U_{\beta b}(\boldsymbol{\theta}) & U_{\beta c}(\boldsymbol{\theta}) & U_{\beta\lambda}(\boldsymbol{\theta}) & U_{\beta\beta}(\boldsymbol{\theta}) \end{bmatrix}_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}},$$

where $U_{pq}(\boldsymbol{\theta}) = \partial^2\mathcal{L}(\boldsymbol{\theta})/(\partial\phi_p\partial\phi_q)$, and $U_{pq}(\boldsymbol{\theta}) = U_{qp}(\boldsymbol{\theta})$. Thus, we get

$$U_{aa}(\boldsymbol{\theta}) = n\psi'(a+b) - n\psi'(a),\ U_{ab}(\boldsymbol{\theta}) = n\psi'(a+b),\ U_{ac}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log t(x_i),$$

$$U_{a\lambda}(\boldsymbol{\theta}) = -c\sum_{i=1}^{n} \frac{r(x_i)}{t(x_i)},\ U_{a\beta}(\boldsymbol{\theta}) = c\lambda\sum_{i=1}^{n} \frac{s(x_i)}{t(x_i)},\ U_{bb}(\boldsymbol{\theta}) = n\psi'(a+b) - n\psi'(b),$$

$$U_{bc}(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \frac{t(x_i)^c \log t(x_i)}{1-t(x_i)^c},\ U_{b\lambda}(\boldsymbol{\theta}) = c\sum_{i=1}^{n} \frac{r(x_i)t(x_i)^{c-1}}{1-t(x_i)^c},$$

$$U_{b\beta}(\boldsymbol{\theta}) = -c\lambda\sum_{i=1}^{n} \frac{s(x_i)t(x_i)^{c-1}}{1-t(x_i)^c},$$

$$U_{cc}(\boldsymbol{\theta}) = -\frac{n}{c^2} - (b-1)\sum_{i=1}^{n} \frac{t(x_i)^c[\log t(x_i)]^2}{[1-t(x_i)^c]^2},$$

$$U_{c\lambda}(\boldsymbol{\theta}) = -a\sum_{i=1}^{n} \frac{r(x_i)}{t(x_i)} + (b-1)\sum_{i=1}^{n} \frac{r(x_i)t(x_i)^{c-1}}{1-t(x_i)^c} + c(b-1)\sum_{i=1}^{n} \frac{r(x_i)t(x_i)^{c-1}\log t(x_i)}{[1-t(x_i)^c]^2},$$

$$U_{c\beta}(\boldsymbol{\theta}) = a\lambda\sum_{i=1}^{n} \frac{s(x_i)}{t(x_i)} - \lambda(b-1)\sum_{i=1}^{n} \frac{s(x_i)t(x_i)^{c-1}}{1-t(x_i)^c} - c\lambda(b-1)\sum_{i=1}^{n} \frac{s(x_i)t(x_i)^{c-1}\log t(x_i)}{[1-t(x_i)^c]^2},$$

$$U_{\lambda\lambda}(\boldsymbol{\theta}) = -\frac{n}{\lambda^2} - (ac-1)\sum_{i=1}^{n} \frac{(1-\mathrm{e}^{x_i^\beta})r(x_i)t(x_i) + r(x_i)^2}{t(x_i)^2} - c^2(b-1)\sum_{i=1}^{n} \frac{[r(x_i)t(x_i)^{c-1}]^2}{[1-t(x_i)^c]^2}$$
$$+c(b-1)\sum_{i=1}^{n} \frac{(1-\mathrm{e}^{x_i^\beta})r(x_i)t(x_i)^{c-1} - (c-1)r(x_i)^2t(x_i)^{c-2}}{1-t(x_i)^c},$$

$$
\begin{aligned}
U_{\lambda\beta}(\boldsymbol{\theta}) = & -\sum_{i=1}^{n} x_i^{\beta} \mathrm{e}^{x_i^{\beta}} \log x_i + (ac-1)\sum_{i=1}^{n} \frac{s(x_i)}{t(x_i)} \\
& + c^2\lambda(b-1)\sum_{i=1}^{n} \frac{s(x_i)r(x_i)t(x_i)^{2c-2}}{[1-t(x_i)^c]^2} \\
& + \lambda(ac-1)\sum_{i=1}^{n} \frac{(1-\mathrm{e}^{x_i^{\beta}})s(x_i)t(x_i) + s(x_i)r(x_i)}{t(x_i)^2} \\
& - c(b-1)\sum_{i=1}^{n} \frac{s(x_i)t(x_i)^{c-1}}{1-t(x_i)^c} \\
& - c\lambda(b-1)\sum_{i=1}^{n} \frac{(1-\mathrm{e}^{x_i^{\beta}})s(x_i)t(x_i)^{c-1} - (c-1)s(x_i)r(x_i)t(x_i)^{c-2}}{1-t(x_i)^c}, \\
U_{\beta\beta}(\boldsymbol{\theta}) = & -\frac{n}{\beta^2} + \sum_{i=1}^{n} x_i^{\beta}(\log x_i)^2 - \lambda\sum_{i=1}^{n} x_i^{\beta}\mathrm{e}^{x_i^{\beta}}(\log x_i)^2[1+x_i^{\beta}] \\
& + \lambda(ac-1)\sum_{i=1}^{n} \frac{v(x_i)t(x_i) - \lambda s(x_i)^2}{t(x_i)^2} \\
& - c\lambda(b-1)\sum_{i=1}^{n} \frac{v(x_i)t(x_i)^{c-1} + \lambda(c-1)s(x_i)^2 t(x_i)^{c-2}}{1-t(x_i)^c} \\
& - (c\lambda)^2(b-1)\sum_{i=1}^{n} \frac{[s(x_i)t(x_i)^{c-1}]^2}{[1-t(x_i)^c]^2},
\end{aligned}
$$

where $\psi'(q) = \mathrm{d}^2 \log\Gamma(q)/\mathrm{d}q^2$ is the trigamma function and $v(x_i) = s(x_i)[\log x_i - \lambda x_i^{\beta}\mathrm{e}^{x_i^{\beta}} \log x_i + x_i^{\beta} \log x_i]$

The normal approximation for $\widehat{\boldsymbol{\theta}}$ in distribution theory is easily handled numerically. Under general regularity conditions, we have the result $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \overset{a}{\sim} \mathrm{N}_5(\mathbf{0}, \mathbf{K}(\boldsymbol{\theta})^{-\mathbf{1}})$, where $K(\boldsymbol{\theta})$ is the 5×5 expected information matrix and $\overset{a}{\sim}$ denotes asymptotic distribution. For n large, $K(\boldsymbol{\theta})$ can be approximated by the estimated observed information matrix $J(\widehat{\boldsymbol{\theta}})$. This multivariate normal approximation for $\widehat{\boldsymbol{\theta}}$ can be used for construing approximate confidence intervals for the model parameters. The LR statistics can be used for testing hypotheses on these parameters.

7. Simulation study

A Monte Carlo simulation is performed to empirically evaluate some asymptotic properties of the ML estimators for the parameters of the MC distribution. The MC observations are generated from three different combinations of a, b, c, λ and β with samples sizes $n = 25, 50, 75, 100, 200$ and 500 and repeat the simulations $N = 1,000$ times. The subroutine `optim` in `R` (R Core Team, 2020) is used for maximizing the log-likelihood Equation (13). The average estimates (AEs) of the ML estimators and their mean squared errors (MSEs) are reported in Tables 1, 2 and 3. The AEs tend to be closer to the true parameters and the MSEs decrease when the sample size n increases in agreement with first-order asymptotic theory. Note that the parameter β presents the lowest MSE in all scenarios. In addition, the parameter λ is the one which presents the highest MSE.

Table 1. Monte Carlo results under $\boldsymbol{\theta} = (1.3, 1.6, 1.4, 1.2, 0.6)$.

	AE					MSE				
n	$\widehat{a}$	$\widehat{b}$	$\widehat{c}$	$\widehat{\lambda}$	$\widehat{\beta}$	$\widehat{a}$	$\widehat{b}$	$\widehat{c}$	$\widehat{\lambda}$	$\widehat{\beta}$
25	2.219	2.102	2.916	2.958	1.483	7.706	6.801	10.945	11.097	3.319
50	1.942	1.913	2.658	2.689	0.946	5.256	4.602	7.705	8.133	0.728
75	1.834	1.927	2.563	2.423	0.835	4.197	4.062	6.320	5.910	0.348
100	1.908	1.788	2.403	2.399	0.788	4.251	3.314	5.917	5.272	0.218
200	1.726	1.805	2.133	2.089	0.681	2.557	2.654	3.500	3.389	0.053
500	1.630	1.770	1.892	1.717	0.632	1.727	1.712	2.160	1.511	0.013

Table 2. Monte Carlo results under $\boldsymbol{\theta} = (1.4, 2, 0.9, 2.8, 1.1)$.

	AE					MSE				
n	$\widehat{a}$	$\widehat{b}$	$\widehat{c}$	$\widehat{\lambda}$	$\widehat{\beta}$	$\widehat{a}$	$\widehat{b}$	$\widehat{c}$	$\widehat{\lambda}$	$\widehat{\beta}$
25	2.173	3.919	2.110	5.153	2.497	6.493	12.609	8.104	19.458	6.186
50	2.028	3.514	1.822	4.353	2.251	4.719	9.481	6.205	13.103	4.631
75	1.932	3.381	1.612	3.958	2.180	3.297	7.829	4.690	10.588	4.319
100	1.911	3.271	1.614	3.701	1.983	3.407	6.675	4.481	8.492	3.269
200	1.904	3.025	1.443	3.328	1.792	2.845	4.440	2.989	6.200	2.297
500	1.818	2.835	1.347	2.916	1.405	1.972	3.225	2.057	3.555	0.808

Table 3. Monte Carlo results under $\boldsymbol{\theta} = (1.7, 1.9, 1.2, 2.2, 0.7)$.

	AE					MSE				
n	$\widehat{a}$	$\widehat{b}$	$\widehat{c}$	$\widehat{\lambda}$	$\widehat{\beta}$	$\widehat{a}$	$\widehat{b}$	$\widehat{c}$	$\widehat{\lambda}$	$\widehat{\beta}$
25	2.689	3.342	2.302	4.495	2.056	8.880	10.736	8.074	17.959	5.889
50	2.367	3.046	2.132	3.610	1.700	5.293	7.422	6.402	11.657	3.954
75	2.262	2.859	2.094	3.479	1.471	4.589	6.739	5.869	10.489	2.886
100	2.297	2.734	1.897	3.279	1.263	4.167	5.378	4.320	8.920	1.844
200	2.209	2.599	1.925	2.913	0.968	3.824	4.000	3.793	5.655	0.638
500	2.087	2.412	1.694	2.609	0.798	2.235	2.489	2.114	2.963	0.187

8. Applications

Two real data applications prove empirically the adequacy of the MC distribution. The applications are developed using the `R` software (version 3.6.3) (R Core Team, 2020) with the script `AdequacyModel` (Marinho et al., 2019). The criteria for model selection are based on the statistics defined by Chen and Balakrishnan (1995): Anderson Darling (A^*) and Cramér-von Mises (W^*). In addition to these statistics, we consider the Akaike information criterion (AIC), Consistent Akaike information criterion (CAIC), Bayesian information criterion (BIC), Hannan-Quinn information criterion (HQIC) and Kolmogorov-Smirnov (KS) statistic with its p-value for model comparisons. The smaller the value of these statistics evidence we have for a good fit. All these important statistics for selecting the best models are provided in the `AdequacyModel` package. The graphical analysis is also important to identify the best fitted model. We analyze the data histograms, the estimated PDFs and CDFs and the empirical CDF calculated by the Kaplan-Meier (Kaplan and Meier, 1958) method.

The MC distribution is compared with three popular lifetime models. The first one is the beta-modified Weibull (BMW) distribution defined by Silva et al. (2010), whose PDF is given by

$$f_{\text{BMW}}(x;a,b,\alpha,\lambda,\gamma) = \frac{ax^{\gamma-1}(\gamma+\lambda x)\mathrm{e}^{\lambda x}}{B(a,b)}\mathrm{e}^{-b\alpha x^{\gamma}\mathrm{e}^{\lambda x}}\left[1-\mathrm{e}^{-\alpha x^{\gamma}\mathrm{e}^{\lambda x}}\right]^{a-1}, \quad x>0,$$

where a, b, and γ are positive shape parameters, $\alpha > 0$ is a scale parameter and $\lambda > 0$ is an accelerating factor in imperfection time which acts as a fragility factor in the survival of the individual as time increases.

The second one is the three-parameter Burr XII distribution (Zimmer et al., 1998), whose PDF has the form

$$f_{\text{BXII}}(x;s,d,c) = \frac{cd}{s^{c}}x^{c-1}\left[1+\left(\frac{x}{s}\right)^{c}\right]^{-(d+1)}, \quad x>0,$$

where $s > 0$ is a scale parameter and c and d are two positive shape parameters.

The third distribution is the Kumaraswamy-log logistic (KLL) (de Santana et al., 2012) model, whose PDF is stated as

$$f_{\text{KLL}}(x;a,b,\alpha,\gamma) = \frac{ab\gamma}{\alpha^{a\gamma}}x^{a\gamma-1}\left[1+\left(\frac{x}{\alpha}\right)^{\gamma}\right]^{-(a+1)}\left\{1-\left[1-\frac{1}{1+\left(\frac{x}{\alpha}\right)^{\gamma}}\right]^{\alpha}\right\}^{b-1}, \quad x>0,$$

where $\alpha > 0$ is a scale parameter and a, b and γ are positive shape parameters.

8.1 Windshields data

We consider 85 uncensored failure times for a specific windshield model studied by Murthy et al. (2004) and Cordeiro et al. (2015). A problem of interest would be to accurately estimate the probability of failure of this windshield model within a specified period time.

The descriptive statistics for these data are listed in Table 4, including minimum and maximum values, first and third quartile, median (Med), mean, standard deviation (SD), and coefficients of skewness and kurtosis.

Table 4. Descriptive statistics for windshields data.

n	Min	1st quartile	Med	Mean	3rd quartile	Max	SD	Skewness	Kurtosis
85	0.04	1.87	2.38	2.56	3.38	4.66	1.11	0.09	2.37

The ML estimates and their associated standard errors (SEs) in parentheses for the fitted distributions are reported in Table 5. Some estimators have large SEs for the BMW and BXII distributions. In addition, the MC and KLL distribution parameters are significant. Table 6 gives the values of the information criteria described before. The MC distribution has the lowest values for all information criteria. Thus, it is the distribution that yields the best fit to the current data. The p-values of the KS statistic also reveal that the data are described well for all distributions.

Since the MG family includes as special cases the beta-G and Kumaraswamy-G classes, two LR tests are performed: MC versus beta-Chen ($c = 1$) and MC vs Kumasrawamy-Chen ($a = 1$). The LR statistics for these tests are 7.0369 and 7.2441, respectively. The two null hypotheses are rejected, thus indicating that the MC distribution is the most suitable for the current data.

Table 5. ML estimates and their associated standard errors (SEs) in parentheses for the distributions fitted to windshields data.

Distribution	Estimate				
MC$(a, b, c, \lambda, \beta)$	0.0360	0.0994	22.2596	0.0451	1.2201
	(0.0051)	(0.0228)	(0.6653)	(0.0186)	(0.0139)
BMW$(a, b, \alpha, \lambda, \gamma)$	4.9756	0.1824	1.0938	0.6004	0.1290
	(5.2775)	(0.1640)	(0.7173)	(0.1519)	(0.1963)
KLL(a, b, α, γ)	0.3359	3.5033	5.6294	6.2686	
	(0.0374)	(0.6590)	(0.0304)	(0.0306)	
BXII(s, d, c)	13.5330	42.7872	2.4122		
	(8.5003)	(59.8242)	(0.2171)		

Table 6. Statistics for the fitted distributions to windshields data.

Distribution	W^*	A^*	AIC	CAIC	BIC	HQIC	KS	p-value (KS)
MC	0.0576	0.3684	259.0505	259.8100	271.2638	263.9630	0.0810	0.6332
BMW	0.0683	0.4849	264.4261	265.1856	276.6394	269.3387	0.0820	0.6173
KLL	0.0617	0.5597	268.1631	268.6631	277.9337	272.0931	0.0570	0.9454
BXII	0.0590	0.5973	269.0118	269.3081	276.3398	271.9594	0.0538	0.9663

A graphical analysis can show the best choice for a model. First, the estimated PDFs are plotted on the data histogram in Figure 5(a). These plots show that the MC distribution is the most appropriate model for the current data and that its estimated PDFs captures the bimodality of the histogram. Figure 5(b) displays the empirical CDF and the estimated CDFs of the MC, BMW and KLL models, which also reveals the superiority of the MC distribution for these data.

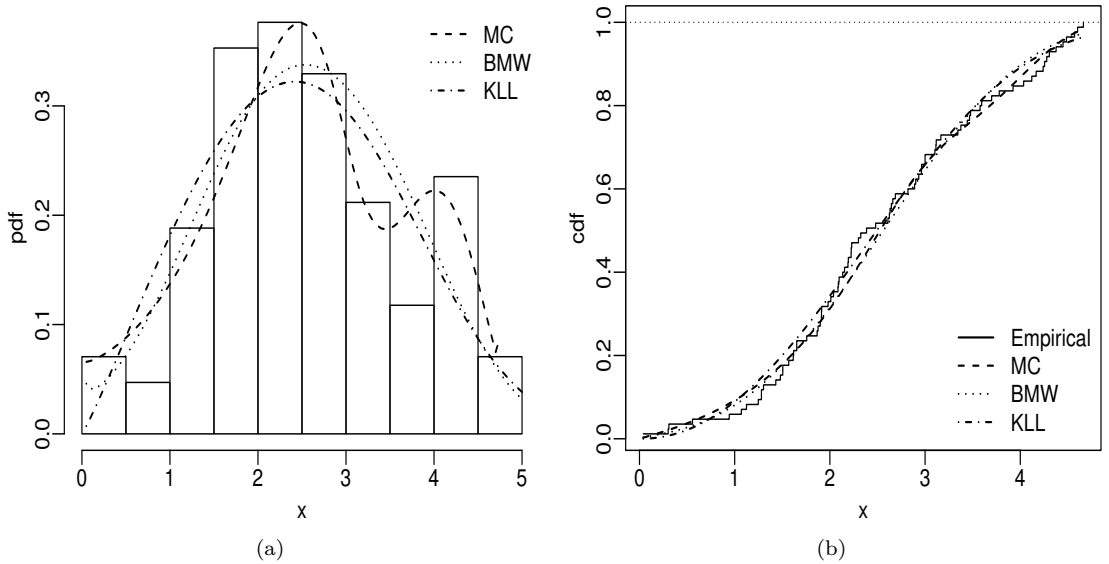


Figure 5. Histogram with estimated PDFs (a) and empirical CDF with estimated CDFs (b) of the MC, BMW and KLL models for windshields data.

8.2 Kevlar/epoxy data

This data set is about the lifetime of spherical pressure vessels under constant pressure until vessel failure, commonly known as static fatigue or stress rupture. NASA space shuttle uses Kevlar/epoxy spherical pressure vessels in a sustained pressure mode for the life of the vessel. The use of this material can be found in air-space breathing apparatus. These data are available in Andrews and Herzberg (1985). The main interest in this application would be to accurately estimate the survival function of these spherical pressure vessels.

The descriptive statistics for these data are given in Table 7. The ML estimates of the parameters for four fitted models are listed in Table 8. Again, the BMW and BXII distributions have large SEs for some estimates. Differently, all the MC and KLL parameters are significant.

Table 7. Descriptive statistics for Kevlar/epoxy data.

n	Min	1st quartile	Med	Mean	3rd quartile	Max	SD	Skewness	Kurtosis
49	1051	5620	8831	8805.69	11745	17568	4553.92	0.10	2.17

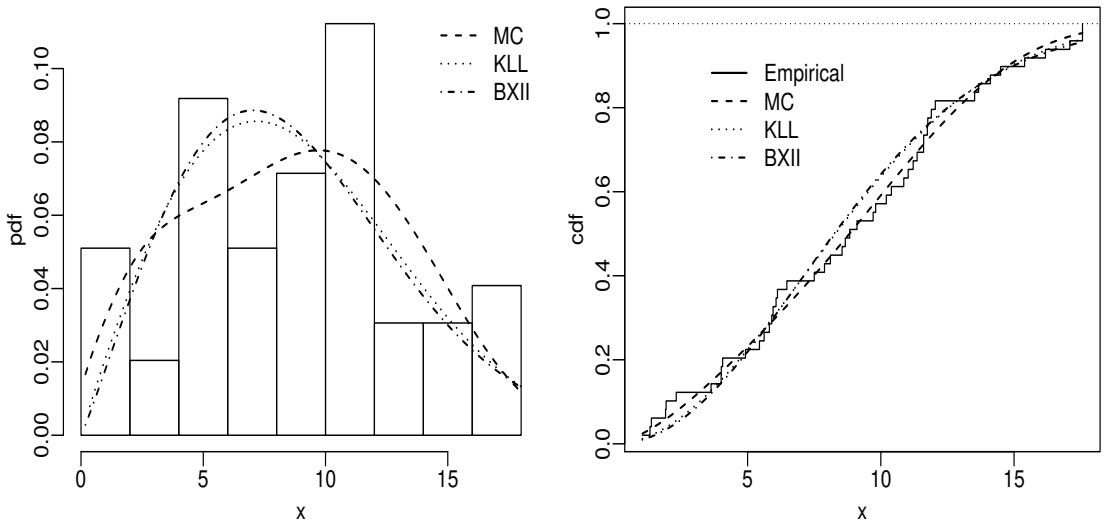
Table 8. ML estimates and their associated standard errors (SEs) in parentheses for the distributions fitted to Kevlar/epóxy data.

Distribution	Estimate				
MC$(a, b, c, \lambda, \beta)$	0.3290	0.1171	5.3114	0.1595	0.5822
	(0.0758)	(0.0248)	(0.0380)	(0.0137)	(0.0114)
BMW$(a, b, \alpha, \lambda, \gamma)$	0.7204	0.4003	0.0162	0.0814	1.7067
	(0.5976)	(1.1377)	(0.0308)	(0.1062)	(1.6739)
KLL(a, b, α, γ)	0.2718	13.0771	37.9120	7.1778	
	(0.0433)	(4.5558)	(0.4090)	(0.6379)	
BXII(s, d, c)	39.5646	18.5077	2.0830		
	(28.7962)	(25.6918)	(0.2439)		

The LR values for the tests MC vs BC ($c = 1$) and MC vs KC ($a = 1$) are 1.1182 and 0.8931, respectively, and therefore the two null hypotheses are not rejected. In Table 9, the more useful statistics W^* and A^* to compare nested and non-nested models indicate that the MC distribution is more appropriate for the current data. The KC distribution can also be chosen based on the AIC, CAIC and HQIC criteria. According to BIC criteria the BXII model is chosen. However, these criterions are more useful to compare nested models. The p-values of the KS statistic indicate that all models can be adopted to fit the current data, although it is higher for the BXII model.

Table 9. Statistics for the fitted models to Kevlar/epóxy data.

Distribution	W^*	A^*	AIC	CAIC	BIC	HQIC	KS	p-value (KS)
MC	0.0294	0.2228	291.3719	292.7673	300.8310	294.9607	0.0724	0.9593
BMW	0.0313	0.2304	291.6484	293.0438	301.1075	295.2372	0.0697	0.9711
KLL	0.0639	0.4196	291.8350	292.7441	299.4023	294.7060	0.0849	0.8716
BXII	0.0800	0.5221	291.4196	291.9530	297.0951	293.5729	0.0902	0.8198
BC	0.0350	0.2491	290.4901	291.3992	298.0574	293.3612	0.0750	0.9455
KC	0.0354	0.2494	290.2651	291.1742	297.8324	293.1361	0.0749	0.9463



(a) (b)

Figure 6. Histogram with estimated PDFs (a) and empirical CDF with estimated CDFs (b) of the MC, KLL and BXII models for Kevlar/epoxy data.

The histogram and estimated PDFs are reported in Figure 6(a), where the superiority of the MC distribution is noted, thus corroborating with the W^* and A^* statistics.

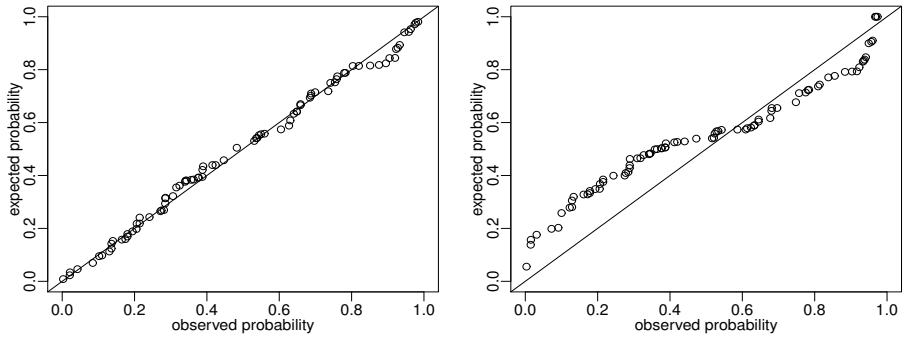
The estimated CDFs along and the empirical CDF are displayed in Figure 6(b). These plots reveal that the estimated CDF of the MC model is closer to the empirical one. Thus, the MC model has a better performance to explain the survival function of the data.

The probability-probability (PP) plots for windshields and Kevlar/epóxy data are given in Figures 7 and 8, respectively. For both data sets, the plot points are close to the diagonal line for the MC model, followed by the KLL distribution. This is a further indication that the MC distribution is the best model for these data sets. Plot of the profile log-likelihood function for windshields and Kevlar/epóxy data are shown in Figures 9 and 10, respectively. These plots were constructed by fixing the other parameters and varying the parameter of interest in a range covering the respective ML estimate. For example in Figure 9(a), the parameter a varies between 0.01 and 0.2, in Figure 9(b) $0.01 < b < 0.8$, in Figure 9(c) $10 < c < 40$, in Figure 9(d) $0.004 < \lambda < 0.05$ and Figure 9(e) $0.3 < \beta < 1.24$. The plots of the Figure 10 are constructed in an analogous way.

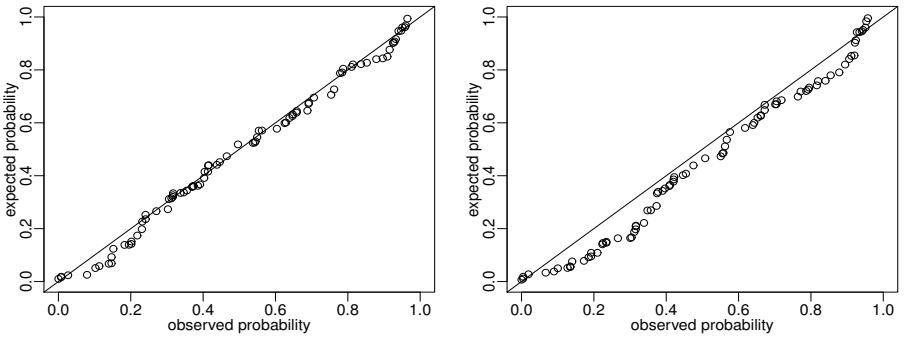
9. Conclusions, limitations, and future research

In this paper, the new McDonald-Chen distribution was proposed, which extended the Chen distribution and presented more flexibility. In the proposal, three shape parameters were added to the Chen distribution to obtain more flexibility and bimodality for the generated probability density function. Its failure or hazard rate function can be increasing, decreasing, upside-down bathtub, bathtub and increasing-decreasing-increasing shapes. Few distributions have this last form. As a result, the new distribution can accommodate several types of data sets, so providing a good alternative for fitting survival and fatigue data. Monte Carlo simulations evaluated the accuracy of the maximum likelihood estimators of the parameters. Finally, two real applications showed that the McDonald-Chen distribution provided better fits than three well-known models because it accommodates bimodality.

A limitation of the new distribution proposed here is its usefulness in fitting data with very small samples because this distribution has five parameters. This compromises the degrees of freedom for data with small samples.

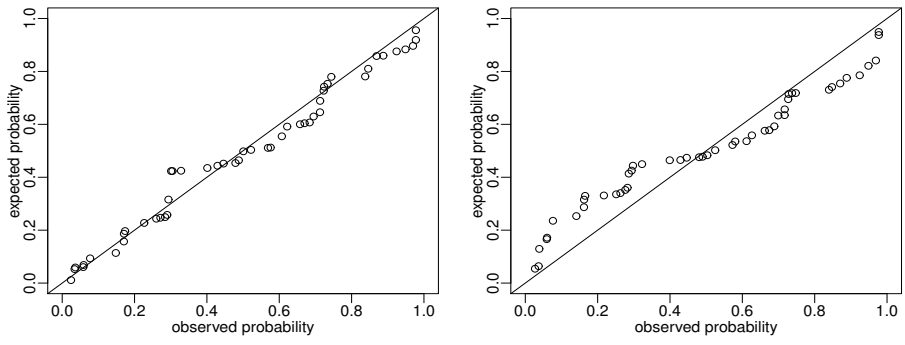


(a) MC (b) BMW

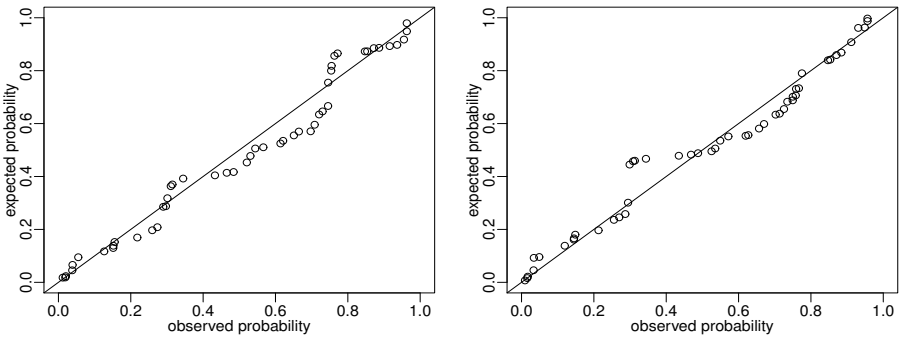


(c) KLL (d) BXII

Figure 7. PP-plots for windshields data.

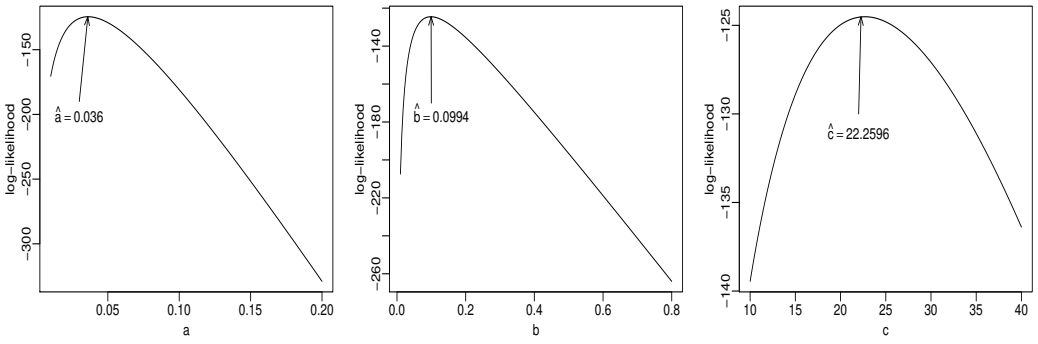


(a) MC (b) BMW

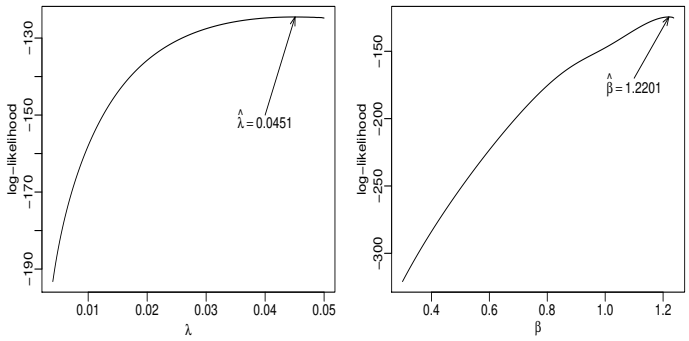


(c) KLL (d) BXII

Figure 8. PP-plots for Kevlar/epóxy data.

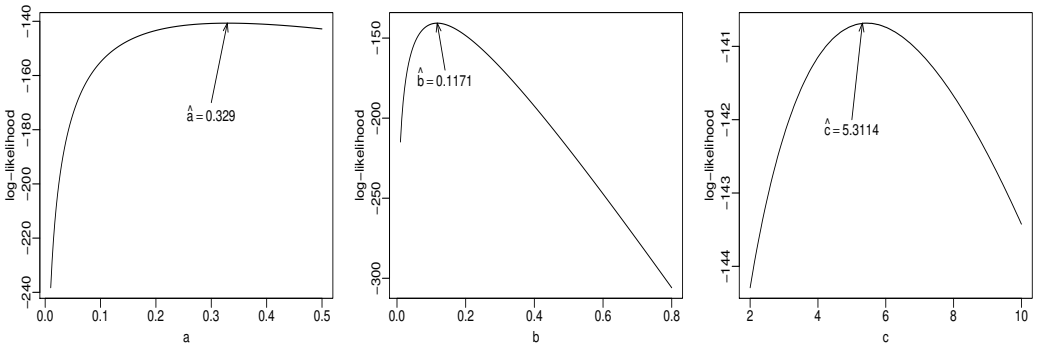


(a) ML estimate of a (b) ML estimate of b (c) ML estimate of c

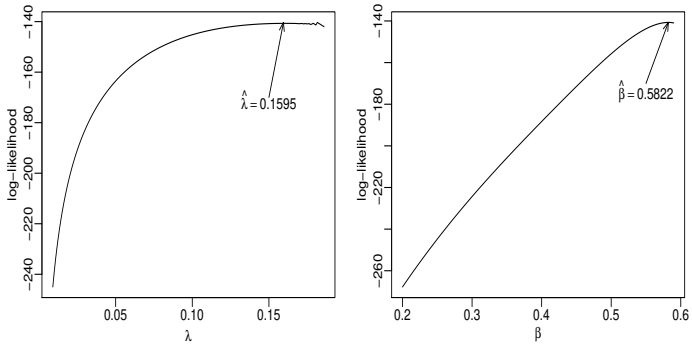


(d) ML estimate of λ (e) ML estimate of β

Figure 9. Profile log-likelihood functions for windshields data.



(a) ML estimate of a (b) ML estimate of b (c) ML estimate of c



(d) ML estimate of λ (e) ML estimate of β

Figure 10. Profile log-likelihood functions for Kevlar/epóxy data.

Future work can be directed to: (i) correct the maximum likelihood estimators analytically (if possible) or numerically (via bootstrap resampling); (ii) reparameterize the McDonald-Chen distribution in terms of the median and propose a regression model to model the median; and (iii) perform inference studies on the McDonald-Chen regression model and diagnostic analysis.

Author contributions Conceptualization, L.D.R.R., G.M.C., J.J.S.S.; methodology, L.D.R.R., G.M.C., J.J.S.S.; software, L.D.R.R., J.J.S.S.; validation, L.D.R.R., G.M.C., J.J.S.S.; formal analysis, L.D.R.R., J.J.S.S.; investigation, L.D.R.R., J.J.S.S.; data curation, L.D.R.R., J.J.S.S.; writing-original draft preparation, L.D.R.R., J.J.S.S.; writing-review and editing, G.M.C.; visualization, L.D.R.R., J.J.S.S., G.M.C.; supervision, G.M.C. All authors have read and agreed the published version of the paper.

Acknowledgements The authors would also like to thank the Editors-in-Chief and the anonymous reviewers for comments that improved the paper.

Funding The authors would like to thank the Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE), Brazil.

Conflicts of interest The authors declare no conflict of interest.

References

Alexander, C., Cordeiro, G.M., Ortega, E. M. M., and Sarabia, J.M., 2012. Generalized beta-generated distributions. Computational Statistics and Data Analysis, 56, 1880–1897.

Andrews, D.F. and Herzberg, A.M., 1985. Data: A Collection of Problems from Many Fields for the Student and Research Worker. Springer, New York, USA.

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

Chen, G. and Balakrishnan, N., 1995. A general purpose approximate goodness-of-fit test. Journal of Quality Technology, 27, 154–161.

Chen, Z., 2000. A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. Statistics and Probability Letters, 49, 155–161.

Cordeiro, G.M. and de Castro, M., 2011. A new family of generalized distributions. Journal of Statistical Computation and Simulation, 81, 883–898.

Cordeiro, G.M., Cintra, R.J., Rêgo, L.C., and Ortega, E.M.M., 2012a. The McDonald normal Distribution. Pakistan Journal of Statistics and Operation Research, 8, 301–329.

Cordeiro, G.M., Hashimoto, E.M., Ortega, E.M.M., and Pascoa, M.A.R., 2012b. The McDonald extended distribution: properties and applications. AStA Advances in Statistical Analysis, 96, 409–433.

Cordeiro, G.M., Aristizábal, W.D., Suárez, D.M., and Lozano, S., 2015. The gamma modified Weibull Distribution. Chilean Journal of Statistics, 6, 37–48.

Cordeiro, G.M., Lima, M.C.S., Ortega, E.M.M., and Suzuki, A.K., 2018. A new extended Birnbaum-Saunders model: Properties, regression and applications. Stats, 1, 32–47.

Cordeiro, G.M., Mansoor, and M., Provost, S.B., 2019. The Harris extended Lindley distribution for modeling hydrological data. Chilean Journal of Statistics, 10, 77–94.

de Santana, T.V.F., Ortega, E.M.M., Cordeiro, G.M., and Silva, G.O., 2012. The

Kumaraswamy-log-logistic distribution. Journal of Statistical Theory and Applications, 11, 265–291.

Dey, S., Kumar, D., Ramos, and P.L., Louzada, F., 2017. Exponentiated Chen distribution: properties and estimation. Communications in Statistics: Simulation and Computation, 46, 8118–8139.

Elbatal, I. and Aryal, G., 2015. Transmuted Dagum distribution with applications. Chilean Journal of Statistics, 6(2), 31–45.

Eugene, N., Lee, C., and Famoye, F., 2002. Beta-normal distribution and its applications. Communications in Statistics: Theory and Methods, 31, 497–512.

Kaplan, E.L. and Meier, P., 1958. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53, 457–481.

Kilbas, A.A., Srivastava, H.M., and Trujillo, J.J., 2006. Theory and Applications of Fractional Differential Equations. Elsevier, Amsterdam, Netherlands.

Lai, C.D., 2013. Constructions and applications of lifetime distributions. Applied Stochastic Models in Business and Industry, 29, 127–140.

Marinho, P.R.D., Silva, R.B., Bourguignon, M., Cordeiro, G.M., and Nadarajah, S., 2019. AdequacyModel: An R package for probability distributions and general purpose optimization. Plos One, 14, 1–30.

McDonald, J.B., 2008. Some generalized functions for the size distribution of income. In Modeling Income Distributions and Lorenz Curves. Springer, New York, USA, pp. 37–55.

Murthy, D.P., Xie, M., and Jiang, R. 2004. Weibull Models. Wiley, New York, USA.

Nedjar, S. and Zeghdoudi, H., 2016. On gamma Lindley distribution: Properties and simulations. Journal of Computational and Applied Mathematics, 298, 167–174.

Pogány, T.K., Cordeiro, G.M., Tahir, M.H., and Srivastava, H.M., 2017. Extension of generalized integro-exponential function and its application in study of Chen distribution. Applicable Analysis and Discrete Mathematics, 11, 434–450.

R Core Team., 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Silva, G.O., Ortega, E.M.M., and Cordeiro, G.M., 2010. The beta modified Weibull distribution. Lifetime Data Analysis, 16, 409–430.

Tahir, M.H. and Nadarajah, S., 2015. Parameter induction in continuous univariate distributions: Well-established G families. Anais da Academia Brasileira de Ciências, 87, 539–568.

Zimmer, W.J., Keats, J.B., and Wang, F., 1998. The Burr XII distribution in reliability analysis. Journal of Quality Technology, 30, 386–394.

Chilean Journal of Statistics
Volume 13, Number 1 (April 2022), Pages 113–132
DOI: 10.32372/ChJS.13-01-06

Distribution Theory
Research Paper

The arctan family of distributions: New results with applications

Emilio Gómez-Déniz[1,*], Enrique Calderín-Ojeda[2], and José María Sarabia[3]

[1]Department of Quantitative Methods in Economics and TiDES Institute, Universidad de Las Palmas de Gran Canaria, Las Palmas, Spain
[2]Centre for Actuarial Studies, Department of Economics, University of Melbourne, Melbourne, Australia
[3]Department of Quantitative Methods, Universidad CUNEF, Madrid, Spain

(Received: 09 March 2022 · Accepted in final form: 10 April 2022)

Abstract

In this paper, we explore the family of arctan transformation of a distribution function. We get some general properties such as those related to the right tail and scale transformation, among others. The results obtained are used to generalize the Pareto Type II (also known as Lomax) distribution, giving us a distribution with a long right-tail that admits the zero value in its support. We show some properties and provide closed-form expressions for the raw moments, the quantile function, the tail value at risk, and other analytical forms that can be helpful in financial and actuarial settings, such as the limited expected value, the mean excess function, and the integrated tail distribution. We also show three numerical illustrations including health expenditure for outpatients, automobile insurance claim size and to see how the new model works as compared to other distributions used in the applied statistical literature.

Keywords: Actuarial · arctan function · Claim size · Income · Pareto type II Distribution · Right tail

Mathematics Subject Classification: 62E10 · 62F10 · 62P05 · 62P25.

1. Introduction

Gómez-Déniz and Calderín-Ojeda (2015a) introduced a mechanism to add a shape parameter to a parent distribution by using the arctan trigonometric transformation of this parent model. They studied the case where the parent distribution was replaced by the classical Pareto cumulative distribution function (CDF). Due to this transformation, results for this new model were obtained including very nice properties. The case where the parent survival function (SF) is the exponential distribution was studied in Calderín-Ojeda et al. (2016). The discrete case was investigated in Gómez–Déniz et al. (2019), obtaining a generalization of the geometric distribution. Furthermore, this transformation was also used in income distribution by Gómez-Déniz (2016), getting the corresponding Lorenz and Leimkhuler curves.

*Corresponding author. Email: emilio.gomez-deniz@ulpgc.es

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

http://www.soche.cl/chjs

After showing the general properties of this family, one of its particular cases is investigated, the arctan Pareto type II distribution. We derive some essential properties, which are simple consequences of the properties of the general family. The flexibility of this distribution is illustrated by applying it to three empirical data sets and comparing the results to previously used distributions.

An apparent reason for generalizing a standard or parent distribution is that the generalized form provides greater flexibility as compared to the parent distribution. For example, consider the problem of determining a suitable model for a population for which it is desired to make a inference. A common way to carry out this task is to use a general model that includes a simpler one as a particular case or limit. After fitting both models, the one that yields the best inference is chosen. Experience indicates that the general model produces better results than the simplest model.

The rest of the paper is structured as follows. Properties of the family of the arctan transformation of a CDF are studied in Section 2. Section 3 is devoted to the specific subject of dealing with the Pareto type II distribution. Numerical applications are considered in Section 4, and finally, Section 5 concludes the work.

2. The arctan transformation and general properties

In this section we firstly illustrate the general procedure to derive the arctan family of distributions. Next, we present some relevant properties of this family. Finally we show that the arctan family of probability distributions can be ordered in terms of the usual stochastic order.

2.1 General methodology

Gómez-Déniz and Calderín-Ojeda (2015b) provided a method to add a scale parameter to a distribution (parent distribution), obtaining a more flexible distribution than the parent model. To make this paper self-contained, we reproduce here this methodology which is based on the $\tan^{-1}$ (arctan) transformation of the parent distribution.

The half-Cauchy distribution (Jacob and Jayakumar, 2012) truncated at $\alpha > 0$ has probability density function (PDF) given by

$$f(y) = \frac{1}{\tan^{-1}\alpha}\frac{1}{1+y^2}, \quad 0 < y < \alpha. \tag{2.1}$$

In the latter expression, $\tan^{-1}$ is the inverse of the circular tangent function. Let us consider now the transformation $y = \alpha\bar{F}_\Theta(x)$, where $\bar{F}_\Theta$ is the SF of a random variable X with support in $[a, b]$, whereas a and b can be finite or non-finite, and Θ is a parameter or vector of parameters. Then, the corresponding PDF of the random variable X obtained from Equation (2.1) results

$$f_{\Theta,\alpha}(x) = \frac{1}{\tan^{-1}\alpha}\frac{\alpha f_\Theta(x)}{1+[\alpha\bar{F}_\Theta(x)]^2}, \tag{2.2}$$

for $a \le x \le b$ and $\alpha > 0$. The SF of X, which is obtained from Equation (2.2) by integration, is stated as

$$\bar{F}_{\Theta,\alpha}(x) = \frac{\tan^{-1}(\alpha\bar{F}_\Theta(x))}{\tan^{-1}\alpha}. \tag{2.3}$$

Furthermore, it is simple to see that Equations (2.2) and (2.3) are proper PDF and SF, respectively, when the support of the parameter α is extended to $(-\infty, \infty)$ except for zero. In this case, we get that $\bar{F}_{\Theta,\alpha}(x) = \bar{F}_{\Theta,-\alpha}(x)$. Additionally, by taking in Equation (2.3) limit when the parameter α tends to zero and applying the L'Hospital rule, it is straightforward to derive that the parent SF, $\bar{F}_{\Theta}$, is obtained as a particular case, that is, $\bar{F}_{\Theta,\alpha}(x) \to \bar{F}_{\Theta}(x)$ when $\alpha \to 0$. Thus, this methodology can be considered a mechanism to add a scale parameter to a parent SF and, therefore, a mechanism to obtain a more flexible SF. In particular, the case where $\bar{F}$ is replaced by the CDF of the classical Pareto distribution was considered in Gómez-Déniz and Calderín-Ojeda (2015b) and Gómez-Déniz (2016) and the case where the parent SF is the classical exponential distribution was studied in Calderín-Ojeda et al. (2016). The discrete case was studied in Gómez–Déniz et al. (2019) obtaining a generalization of the classical geometric distribution. Also, in actuarial statistics, the arctan transformation was first used in Gómez-Déniz and Calderín-Ojeda (2015a).

2.2 Properties

The quantile function is easy to derive from Equation (2.3) and it is given by

$$x_\gamma = F^{-1}_{\Theta,\alpha}\left(1 - \alpha^{-1}\tan(\bar{\gamma}\tan^{-1}\alpha)\right), \tag{2.4}$$

where $\bar{\gamma} = 1-\gamma$, $0 < \gamma < 1$ and F^{-1} is the inverse of the CDF F. In particular, the median is expressed as

$$x_{0.5} = F^{-1}_{\Theta,\alpha}\left(1 - \alpha^{-1}\tan((0.5)\tan^{-1}\alpha)\right),$$

Proposition 2.1 Suppose that the parent SF depends on a vector of parameters $\Theta = (\theta_1, \ldots, \theta_s)$ satisfying $\bar{F}_{\Theta}(x/k) = \bar{F}_{\Theta_1}(x)$, being Θ_1 a vector of parameters for which the parameter j, for some $j \in \{1, \ldots, s\}$ is a scale or rate transformation of θ, with rate or scale value $k > 0$. Then, the arctan distribution preserves also the same transformation.

Proof By denoting $Y = k\,X$, and denoting the SF of Y as $\bar{F}^{Y}_{\Theta,\alpha}$, we have that

$$\bar{F}^{Y}_{\Theta,\alpha}(y) = \bar{F}^{Y}_{\Theta,\alpha}(kx) = \bar{F}_{\Theta,\alpha}(x/k) = \frac{\tan^{-1}(\alpha\bar{F}_{\Theta}(x/k))}{\tan^{-1}\alpha} = \frac{\tan^{-1}(\alpha\bar{F}_{\Theta_1}(x))}{\tan^{-1}\alpha} = \bar{F}_{\Theta_1,\alpha}(x),$$

where in the last equality we have used the assumption that $\bar{F}_{\Theta}(x/k) = \bar{F}_{\Theta_1}(x)$. □

To illustrate Proposition 2.1, consider the exponential distribution with mean $\Theta = 1/\lambda$ and $\lambda > 0$. Then, it is simple to verify that

$$P(kX > x) = P(X > x/k) = \exp(-x/(\lambda k)),$$

that is, the random variable kX follows an exponential distribution with parameter $\Theta_1 = 1/(\lambda k)$. Now, the arc transformation of the exponential distribution has SF given by

$$\bar{F}_{\Theta,\alpha}(x) = \frac{\tan^{-1}(\alpha\exp(-x/\lambda))}{\tan^{-1}\alpha},$$

which satisfies $F_{\Theta,\alpha}(x/k) = F_{\Theta_1,\alpha}(x)$ as it can be verified in a simple way.

It is already known that any probability distribution, that is specified through its CDF $F(x)$ on the real line, is heavy right-tailed (Rolski et al., 1999) if $\limsup_{x\to\infty}(-\log(\bar{F}(x)/x)) = 0$. Observe that $-\log(\bar{F}(x)$ is the hazard function of $F(x)$. Next, a result shows that, under mild condition, the family of SF provided in Equation (3.8) is a heavy-tailed distribution.

PROPOSITION 2.2 Suppose that the PDF of the parent distribution in the family stated in Equation (2.2) satisfies that

$$\limsup_{x\to\infty} f_\Theta(x) = 0. \tag{2.5}$$

Then, the CDF $F_{\Theta,\alpha}$ of the family defined in Equation (3.8) is a heavy-tailed distribution.

PROOF We have that

$$\begin{aligned}\limsup_{x\to\infty}\frac{1}{x}\log\bar{F}_{\Theta,\alpha}(x) &= -\frac{1}{\tan^{-1}\alpha}\limsup_{x\to\infty}\frac{\log(\tan^{-1}(\alpha\bar{F}_\Theta(x)))}{x}\\ &= \frac{\alpha}{\tan^{-1}\alpha}\limsup_{x\to\infty}\frac{f_\Theta(x)}{1+\alpha^2[\bar{F}_\Theta(x)]^2} = 0,\end{aligned}$$

after applying the L'Hospital rule. The fact that $\limsup_{x\to\infty}\bar{F}_\Theta(x) = 0$ and the assumption that $\limsup_{x\to\infty} f_\Theta(x) = 0$ conduct to the result.□

In this case, the distribution fails to possess any positive exponential moment, that is, $\int\exp(sx)\mathrm{d}F(x) = \infty$ for all $s > 0$ (Foss et al., 2011, Ch. 1, p. 2). Distributions of this type have moment generating function $M_F(s) = \infty$, for all $s > 0$, as occurs, for example, with the lognormal distribution. As a consequence of the last result, we have the following corollary.

COROLLARY 2.3 It is verified that $\limsup_{x\to\infty}\exp(sx)\bar{F}_{\Theta,\alpha}(x) = \infty$, for $s > 0$.

PROOF This is a direct consequence of Proposition 2.2.□

An important issue in extreme value theory is the regular variation (Bingham, 1987 and Konstantinides, 2018). This is, a fexible description of the variation of some function according to the polynomial form of the type $x^{-\delta} + o(x^{-\delta})$, $\delta > 0$. This concept is formalized in the following definition.

DEFINITION 2.4 A CDF (measurable function) is called regular varying at infinity with index $-\delta$ if it holds

$$\limsup_{x\to\infty}\frac{\bar{F}(\tau x)}{\bar{F}(x)} = \tau^{-\delta},$$

where $\tau > 0$ and the parameter $\delta \geq 0$ is called the tail index.

The next result establishes that if the SF of the parent distribution stated in Equation (3.8) is a regular variation Lebesgue measure, then the SF given in Equation (3.8) is also a regular variation Lebesgue measure.

PROPOSITION 2.5 Let $\bar{F}_\Theta(x)$ ibe a regular variation Lebesgue measure. Then, the SF given in Equation (2.3) is also a SF with regularly varying tails.

PROOF Consider the SF given in Equation (2.3). Then, we have

$$\limsup_{x\to\infty}\frac{\bar{F}_{\Theta,\alpha}(\tau x)}{\bar{F}_{\Theta,\alpha}(x)}=\limsup_{x\to\infty}\frac{f_{\Theta}(\tau x)}{f_{\Theta}(x)}\frac{1+\alpha^2[\bar{F}_{\Theta}(x)]^2}{1+\alpha^2[\bar{F}_{\Theta}(\tau x)]^2}=\tau^{-(\Theta+1)},$$

after applying the L'Hospital rule. The fact that $\limsup_{x\to\infty}\bar{F}_{\Theta,\alpha}(\tau x) = \limsup_{x\to\infty}\bar{F}_{\Theta,\alpha}(x) = 0$ and that $f_{\Theta}(\tau x)/f_{\Theta}(x) \to \tau^{\Theta}$ when $x \to \infty$, conduct to the result.□

In actuarial setting and also into the individual and collective risk models the practitioner is usually interested in the random variable $S_n = \sum_{i=1}^{n} X_i$ for $n \geq 1$. Although in practice, its PDF is difficult or impossible to calculate, we can approximate its probabilities by using the following Corollary, which is an immediate consequence of Proposition 2.5 (Jessen and Mikosch, 2006).

COROLLARY 2.6 Let $X_1, \ldots, X_n$ be independent identically distributed random variables with common SF given by Equation (2.3) and $S_n = \sum_{i=1}^{n} X_i,\ n \geq 1$. Then, we get

$$P(S_n > x) \sim P(X > x) \quad \text{as} \quad x \to \infty. \tag{2.6}$$

Therefore, if $P_n = \max_{i=1,\ldots,n} X_i$, for $n \geq 1$, we have that

$$P(S_n > x) \sim nP(X > x) \sim P(P_n > x).$$

This means that, for large x, the event $\{S_n > x\}$ is due to the event $\{P_n > x\}$. Therefore, exceedences of high thresholds by the sum S_n are due to the exceedence of this threshold by the largest value in the sample.

As Jessen and Mikosch (2006) pointed out, expression given in Equation (2.6) can be taken as the definition of a subexponential distribution. The class of those distributions is greater than the class of regularly varying distributions. The result given in Corollary 2.6 remains valid for subexponential distributions in the sense that subexponentiality of S_n implies subexponentiality of X_1. Usually, this property is referred to as convolution root closure of subexponential distributions. More details can be viewed in Embrechts and Goldie (1980) and Embrechts and Goldie (1982).

2.3 STOCHASTIC ORDERING

Next, a stochastic representation of the parameters of the given family in Equation (2.3) is studied. As it is well known, many parametric families of distributions can be stated by means of some stochastic orders according to the value of its parameters. For the general family of distributions given in Equation 2.3, it is difficult to establish an order in terms of the likelihood ratio order (Ross, 1996; Shaked and Shanthikumar, 2007). For a particular choice of the main distribution, this is possible (Gómez-Déniz and Calderín-Ojeda, 2015a). However, a weaker but also useful result may be obtained, as shown below.

DEFINITION 2.7 Let us consider two random variables X_{Θ_1} and X_{Θ_2}, with X_{Θ_1} preceding X_{Θ_2} in the stochastic dominance sense or X_{Θ_1} being smaller than X_{Θ_2}. In this case, the notation $X_{\Theta_1} \leq_{\text{ST}} X_{\Theta_2}$ is used, if and only if the CDF of X_{Θ_1} always exceeds X_{Θ_2}, that is,

$$F_{\Theta_1}(x) \geq F_{\Theta_2}(x), \quad -\infty < x < \infty,$$

where F_{Θ_1} and F_{Θ_2} are the CDFs of X_{Θ_1} and X_{Θ_2} respectively. Note that this expression is the same as

$$\bar{F}_{\Theta_1}(x) \leq \bar{F}_{\Theta_2}(x), \quad -\infty < x < \infty.$$

In the following, we provide two stochastic orderings. In the first one the order is given by fixing the shape parameter α and modifying the parameters of the parent distribution, whereas in the second one we have fixed the parameters vector of the parent distribution and changed the parameter α.

PROPOSITION 2.8 Let us consider two random variables X_1 and X_2 with CDFs $F_{\Theta_1}(x)$ and $F_{\Theta_2}(x)$, respectively such that X_1 is stochastically smaller than X_2 ($X_1 \leq_{\text{ST}} X_2$), that is, $F_{\Theta_1}(x) \geq F_{\Theta_2}(x)$ for $\Theta_1 \leq \Theta_2$. Then, the arctan transformation preserves this stochastic order, that is, $F_{\Theta_1,\alpha}(x) \geq F_{\Theta_2,\alpha}(x)$.

PROOF Since the arctan function is monotone, we have that

$$\begin{aligned} F_{\Theta_1}(x) \geq F_{\Theta_2}(x) &\Longrightarrow \alpha F_{\Theta_1}(x) \geq \alpha F_{\Theta_2}(x) \Longrightarrow \tan^{-1}(\alpha F_{\Theta_1}(x)) \geq \tan^{-1}(\alpha F_{\Theta_2}(x)) \\ &\Longrightarrow \frac{\tan^{-1}(\alpha F_{\Theta_1}(x))}{\tan^{-1}\alpha} \geq \frac{\tan^{-1}(\alpha F_{\Theta_2}(x))}{\tan^{-1}\alpha} \\ &\Longrightarrow F_{\Theta_1,\alpha}(x) \geq F_{\Theta_2,\alpha}(x). \end{aligned}$$

Hence, the result is obtained.□

THEOREM 2.9 Let X_1 and X_2 be two random variables with PDFs $f_{\Theta,\alpha_1}(x) > 0$ and $f_{\Theta,\alpha_2}(x) > 0$ obtained from Equation (2.2), respectively. If $\alpha_1 \leq \alpha_2$, then $X_1 \leq_{\text{LR}} X_2$.

PROOF Note that the ratio

$$\frac{f_{\Theta,\alpha_2}(x)}{f_{\Theta,\alpha_1}(x)} = \frac{\alpha_2 \tan^{-1}\alpha_1}{\alpha_1 \tan^{-1}\alpha_2} m_{\Theta,\alpha_1,\alpha_2}(x)$$

is non-decreasing if and only if $m'_{\Theta,\alpha_1,\alpha_2}(x) \geq 0$ for x in their support, where

$$m_{\Theta,\alpha_1,\alpha_2}(x) = \frac{1 + [\alpha_1 \bar{F}_\Theta(x)]^2}{1 + [\alpha_2 \bar{F}_\Theta(x)]^2}.$$

Some calculations show that

$$m'_{\Theta,\alpha_1,\alpha_2}(x) = \frac{2 f_\Theta(x) \bar{F}_\Theta(x)(\alpha_2^2 - \alpha_1^2) m_{\Theta,\alpha_1,\alpha_2}(x)}{(1 + [\alpha_1 \bar{F}_\Theta(x)]^2)(1 + [\alpha_2 \bar{F}_\Theta(x)]^2)}.$$

Now, taking into account that $\alpha_1 \leq \alpha_2$, then $m_{\Theta,\alpha_1,\alpha_2}(x) \geq 0$ and the result holds.□

We have now the following corollary.

COROLLARY 2.10 Let X_1 and X_2 be two random variables with PDFs $f_{\Theta,\alpha_1}(x) > 0$ and $f_{\Theta,\alpha_2}(x) > 0$ obtained from Equation (2.2), respectively and hazard rates $h_{\Theta,\alpha_1}(x)$ and $h_{\Theta,\alpha_2}(x)$, being $h_{\Theta,\alpha}(x) = f_{\Theta,\alpha}(x)/\bar{F}_{\Theta,\alpha}(x)$, respectively. If $\alpha_1 \leq \alpha_2$ then,

(i) $\mathrm{E}(X_1^k) \leq \mathrm{E}(X_2^k)$ for all $k > 0$,
(ii) $h_{\Theta,\alpha_1}(x) \leq h_{\Theta,\alpha_2}(x)$ for all x in their support.

PROOF It is well-known (Shaked and Shanthikumar, 2007) that

$$X_1 \leq_{\mathrm{LR}} X_2 \Longrightarrow X_1 \leq_{\mathrm{HR}} X_2 \Longrightarrow X_1 \leq_{\mathrm{ST}} X_2. \tag{2.7}$$

Therefore, (i) follows from Theorem 2.9 and Equation (2.7) by taking into account that $X_1 \leq_{\mathrm{ST}} X_2$ holds if and only if

$$\mathrm{E}\left[\Phi\left(X_1\right)\right] \leq \mathrm{E}\left[\Phi\left(X_2\right)\right] \text{ for all non-decreasing function } \Phi.$$

Similarly, (ii) follows by combining Theorem 2.9 and Equation (2.7). Thus, in consequence, if $\alpha_1 \leq \alpha_2$ we have that $\bar{F}_{\Theta,\alpha_1}(x) \leq \bar{F}_{\Theta,\alpha_2}(x)$.

3. The Lomax arctan distribution

In this section, we firstly introduce the Lomax arctan distribution (LAT hereafter) and derive some of its more relevant statistical and financial properties.

3.1 Specific model

A particular case of the Pareto Type II distribution is considered here. This distribution is essentially a classical Pareto distribution modified to get that the support begins at zero. As it is known, this distribution is widely employed as a model in business, economics, actuarial science, queueing theory, and internet traffic modeling, among others. Its SF is given by

$$\bar{F}_{\Theta}(x) = \left(\frac{\lambda}{\lambda + x}\right)^{\sigma}, \quad x \geq 0, \tag{3.8}$$

(Fisk, 1961; Suárez-Espinosa et al., 2018) which is a particular case of the Champernowne distribution (Champrenowne, 1952) and obviously is a scale transformation of the classical Pareto distribution (Arnold, 1983). A Lomax regression model with varying precision parameter was recently presented in Melo et al. (2021) In the rest of the paper we use $X \sim \mathrm{L}(\sigma, \alpha)$ to point out that X follows a Pareto Type II distribution with the PDF given in Equation (3.9).

An excellent property of this distribution, apart from having a very tractable SF, is a fascinating preservation property. That is, if $X \sim \mathrm{L}(\sigma, \lambda)$, then the random variable $kX \sim \mathrm{L}(\sigma, k\lambda)$, for $k > 0$. This property is very useful in economics and actuarial fields when dealing with inflation. The PDF, derived from Equation (3.8), results

$$f_{\Theta}(x) = \frac{\sigma\lambda^{\sigma}}{(x+\lambda)^{\sigma+1}}, \quad x \geq 0, \quad \sigma > 0, \quad \lambda > 0. \tag{3.9}$$

One of the advantages of working with the SF given in Equation (3.8) is the possibility of dealing with data that includes the zero value, the mode of the distribution. This is impossible for most classical continuous distributions, such as the gamma and the inverse Gaussian distribution. Nevertheless, the distribution has limited flexibility for adapting to empirical data whose modal value is not located at zero. To get a more flexible distribution, we consider here the $\tan^{-1}$ transformation of the Pareto Type II distribution. The resulting distribution, Pareto Type II arctan distribution, it is obtained by applying the Equation

(2.3) to Equation (3.8) to get the SF given by

$$\bar{F}_{\Theta,\alpha}(x) = \frac{\tan^{-1}(\alpha(1+x/\lambda)^{-\sigma})}{\tan^{-1}\alpha}. \tag{3.10}$$

Its PDF results

$$f_{\Theta,\alpha}(x) = \frac{\alpha\sigma}{\lambda \tan^{-1}\alpha} \frac{(1+x/\lambda)^{-\sigma-1}}{1+\alpha^2(1+x/\lambda)^{-2\sigma}}. \tag{3.11}$$

Figure 1 shows several graphs of the PDF given in Equation (3.11) for different values of its parameters. It is noted that when the scale parameter $\alpha < 1$ or the shape parameter $\sigma \leq 1$, the mode of the distribution is located at 0, and for values larger than one, the modal value moves to the right. Observe that the larger are the value of the parameters, the greater is the modal value.

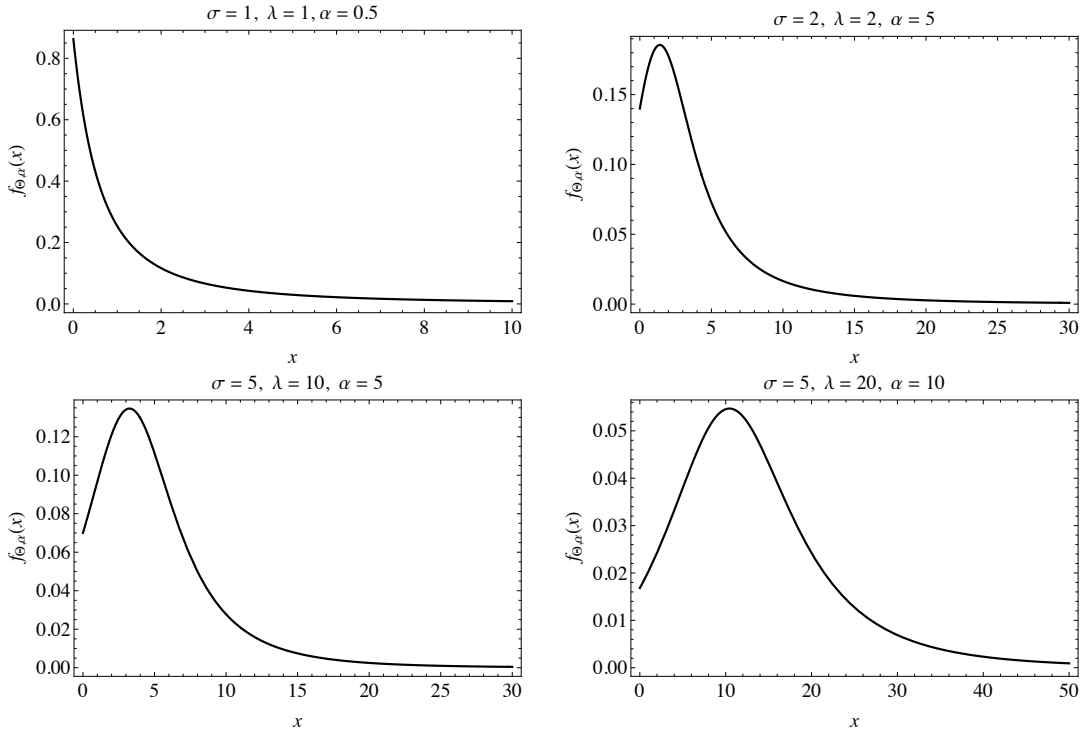


Figure 1. PDF the LAT distribution for selected values of parameters σ, λ and α

Since this distribution is a scale transformation of the Pareto arctan distribution studied in Gómez-Déniz and Calderín-Ojeda (2015a), we can easily obtain its row moments that are given by

$$\mathrm{E}(X^r) = \frac{\alpha\sigma\lambda^r}{\tan^{-1}\alpha} \sum_{j=0}^{r} \frac{(-1)^j}{\sigma - r + j} \binom{r}{j} {}_2F_1\left(1, \frac{\sigma - r + j}{2\sigma}; \frac{3\sigma - r + j}{2\sigma}; -\alpha^2\right),$$

where ${}_2F_1$ is the hypergeometric function defined as

$${}_2F_1(a,b;c;z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}\mathrm{d}t.$$

In particular, the mean takes the form

$$\mu = \mathrm{E}(X) = \frac{\alpha\lambda\sigma}{(\sigma-1)\tan^{-1}\alpha}\,{}_2F_1\left(1, \frac{\sigma-1}{2\sigma}; \frac{3\sigma-1}{2\sigma}; -\alpha^2\right) - \lambda, \quad \sigma > 1. \tag{3.12}$$

From Equation (2.4), the quantile function x_γ is simply derived as

$$x_\gamma = \lambda\left\{\left[\frac{1}{\alpha}\tan\left(\bar{\gamma}\tan^{-1}\alpha\right)\right]^{-1/\sigma} - 1\right\}, \tag{3.13}$$

and from Equation (3.13), the median can be easily obtained.

The mode, which can be obtained by differentiating Equation (3.11) with respect to the variable x, is expressed as

$$x_{Mo} = \lambda\left[\left(\frac{\alpha^2(\sigma-1)}{1+\sigma}\right)^{(2\sigma)^{-1}} - 1\right].$$

Then, the hazard rate function for the LAT distribution, $h_{\Theta,\alpha}(x) = f_{\Theta,\alpha}(x)/\bar{F}_{\Theta,\alpha}(x)$, which is obtained from Equations (3.10) and (3.11), has been plotted for the same values of parameters as considered in the previous Figure. This is shown in Figure 2. It can be observed that the hazard rate function has a variety of shapes. For example, for values of $\alpha < 1$, the hazard rate function is monotonically decreasing and for values of the scale parameters α and the shape parameter σ and scale parameter λ the function is firstly increasing and then decreasing.

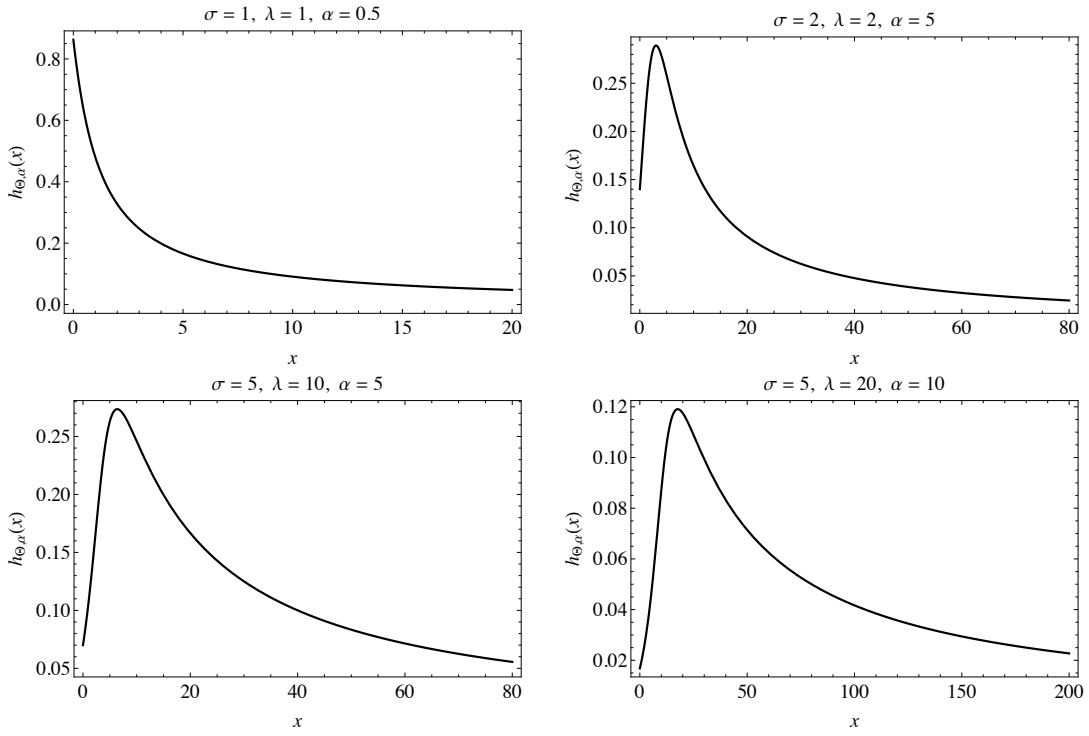


Figure 2. Failure rate function of LAT distribution for selected values of parameters σ, λ and α

We now provide some properties which are consequences of the results obtained in the previous section.

PROPOSITION 3.1 If $X \sim \text{LAT}(\sigma, \lambda, \alpha)$ then $kX \sim \text{LAT}(\sigma, k\lambda, \alpha)$.

PROOF It is a direct consequence of Proposition 2.1.□

PROPOSITION 3.2 The CDF $F_{\Theta,\alpha}(x)$ of the family stated in Equation (3.10), that is, the LAT distribution is a heavy-tailed distribution.

PROOF It is a direct consequence of applying Proposition 2.2 having into account that the PDF given in Equation (3.9) satisfies Equation (2.5).

PROPOSITION 3.3 The SF given in Equation (3.10) is a SF with regularly varying tails.

PROOF It is a consequence of the result provided in Proposition 2.5, having into account that the SF given in Equation (3.8) verifies

$$\limsup_{x\to\infty} \frac{\bar{F}_\Theta(\tau x)}{\bar{F}_\Theta(x)} = \tau^{-\sigma}.$$

Now, because $\sigma > 0$, we have the result.□

3.2 Further properties

We provide here some other properties which can be helpful in financial and actuarial fields. Let the random variable

$$Z = X \wedge \omega = \begin{cases} X, & X < \omega, \\ \omega, & X \geq \omega, \end{cases}$$

which is an amount used in excess of loss reinsurance context with excess level $\omega > 0$. Insurance companies widely use this tool to reduce the amount paid on larger claims. Its expected value, $\text{E}(X \wedge \omega)$, is referred to as the limited expected value in insurance context. Obviously, it is a right-censored variable for which it is easy to see (Hogg and Klugman, 1984; Boland, 2007) that can be computed as

$$\text{E}(X \wedge \omega) = \text{E}[\min(X, \omega)] = \int_0^\omega x\, f(x)\, \text{d}x + \omega\, \bar{F}(\omega). \tag{3.14}$$

Furthermore, it represents the expected amount per claim retained by the insured on a policy with a fixed amount deductible of ω. Thus, defining the expected dollar (or other monetary units) saving per incident when a deductible is imposed (Klugman et al., 2008, Ch. 3).

For the LAT distribution, the limited expected value given by Equation (3.14) is expressed as

$$\text{E}(X \wedge \omega) = (\omega + \lambda)\bar{F}_{\Theta,\alpha}(\omega) - \lambda + H^1_{\Theta,\alpha}(\omega) - H^2_{\Theta,\alpha}(\omega), \tag{3.15}$$

where

$$H^1_{\Theta,\alpha}(\omega) = \frac{\lambda+\omega}{\bar{F}(\omega)}\,{}_2F_1\left(1, \frac{1+\sigma}{2\sigma}; \frac{1}{2}\left(3+\frac{1}{\sigma}\right); (\alpha\bar{F}_{\Theta,\alpha}(\omega))^{-2}\right), \tag{3.16}$$

$$H^2_{\Theta,\alpha}(\omega) = \lambda\,{}_2F_1\left(1, \frac{1+\sigma}{2\sigma}; \frac{1}{2}\left(3+\frac{1}{\sigma}\right); (\alpha\bar{F}_{\Theta,\alpha}(\omega))^{-2}\right), \tag{3.17}$$

which can be obtained also by using a scale transformation of the classical Pareto distribution (Gómez-Déniz and Calderín-Ojeda, 2015a).

The value at risk (VaR) is defined as the amount of capital required to ensure that the insurer does not become insolvent with a high degree of certainty. The VaR of a random variable X which follows the LAT distribution is the $100q$th quantile and therefore coincides with Equation (3.13).

It is known that the use of the VaR is questionable due to the lack of subadditivity. For that reason, the expected loss given that the loss exceeds the $100q$th quantile of the distribution of X, that is, the tail value at risk (TVaR), is considered. Then, if X follows a LAT distribution, for any quantile q, the tail value at risk, can be obtained again by a scale transformation of the TVaR of the classical Pareto distribution and is given by

$$\begin{aligned}\mathrm{TVaR}(X;q) = \frac{1}{1-q}\int_q^1 \mathrm{VaR}(x;q)\,\mathrm{d}q = \frac{\alpha\lambda\sigma}{\bar{q}(\sigma-1)\tan^{-1}\alpha}\left[\frac{\tan(\bar{q}\tan^{-1}\alpha)}{\alpha}\right]^{1-1/\sigma}\\ \times\ {}_2F_1\left(1, \frac{\sigma-1}{2\sigma}; \frac{3}{2}-\frac{1}{2\sigma}; -\tan^2(\bar{q}\tan^{-1}\alpha)\right) - \lambda.\end{aligned}$$

The integrated tail distribution (also known as equilibrium distribution) is an important distribution that often appears in insurance and many other applied probability models.

Let $\bar{F}$ be the SF given in Equation (3.10). Then, the integrated tail distribution of F (for instance, Klüppelberg, 1988 and Yang, 2004) is defined as $F^I(x) = (1/\mathrm{E}(X))\int_0^x \bar{F}(y)\,\mathrm{d}y$. For the distribution proposed in this work, as proven in the following result, the integrated tail distribution can be written as a closed-form expression and given by

$$F^I_{\Theta,\alpha}(x) = \frac{1}{\mu}\left[(x+\lambda)\bar{F}_{\Theta,\alpha}(x) - \lambda\right] + \frac{\sigma}{\alpha\mu(\sigma+1)\tan^{-1}\alpha}\left[H^1_{\Theta,\alpha}(x) - H^2_{\Theta,\alpha}(x)\right], \tag{3.18}$$

where $H^j_{\Theta,\alpha}(x)$, for $j=1,2$, are given in Equations (3.16) and (3.17), respectively, whereas $\bar{F}_{\Theta,\alpha}$ and μ are defined in Equations (3.10) and (3.12), respectively. Under the classical model (Embrechts and Veraverbeke, 1982; Yang, 2004) and assuming a positive security loading, ρ, for the claim size distributions with regularly varying tails we have that, by using Equation (3.18), it is possible to obtain an approximation of the probability of ruin, $\Psi(u)$, when $u \to \infty$. In this case, the asymptotic approximation of the ruin function is stated as $\Psi(u) \sim (1/\rho)\bar{F}^I(u)$, for $u \to \infty$, where $\bar{F}^I(u) = 1 - F^I(u)$.

The failure rate of the integrated tail distribution, which is expressed as $\gamma_I(x) = \bar{F}(x)/\int_x^\infty \bar{F}(y)\,\mathrm{d}y$, is also obtained in closed-form. Furthermore, the reciprocal of γ_I is the mean residual life that can be easily derived. For a claim amount random variable X, the mean excess function (also known as the conditional mean exceedence) is the expected payment per claim for a policy with a fixed amount deductible of $x > 0$, where claims with amounts less than or equal to x are wholly ignored. Then, we have that

$$e(x) = \mathrm{E}(X - x | X > x) = \frac{1}{\bar{F}(x)}\int_x^\infty \bar{F}(u)\,\mathrm{d}u. \tag{3.19}$$

This function is also essential in an actuarial setting, when we deal with reinsurance (Albrecher et al., 2017). If X is a lifetime, as in demography or reliability, Equation (3.19) is recognized as the mean residual lifetime. The following result gives the mean excess function of the LAT distribution in a closed-form expression.

PROPOSITION 3.4 The mean excess function of the LAT distribution is given by

$$e_{\Theta,\alpha}(x) = \frac{1}{\bar{F}_{\Theta,\alpha}(x)}\left[\mu + \lambda - \frac{\sigma(H^1_{\Theta,\alpha}(x) - H^2_{\Theta,\alpha}(x))}{\alpha(1+\sigma)\tan^{-1}\alpha}\right] - (x+\lambda), \tag{3.20}$$

where $H^j_{\Theta,\alpha}(x)$, for $j = 1, 2$, are given in Equations (3.16) and (3.17), respectively, whereas $\bar{F}_{\Theta,\alpha}$ and μ are given in Equations (3.10) and (3.12), respectively.

PROOF Using the expression

$$e(x) = \frac{\mathrm{E}(X) - \mathrm{E}(X \wedge x)}{\bar{F}(x)},$$

which relates the mean excess function given in Equation (3.19) with the limited expected value function (Hogg and Klugman, 1984, p. 59), the result follows by using and Equations (3.12), (3.10), (3.15) and a some little algebra.□

Figure 3 shows the mean residual life function given in Equation (3.20) for special cases of parameters. It can be seen that this function can be increasing, decreasing, unimodal or anti-unimodal.

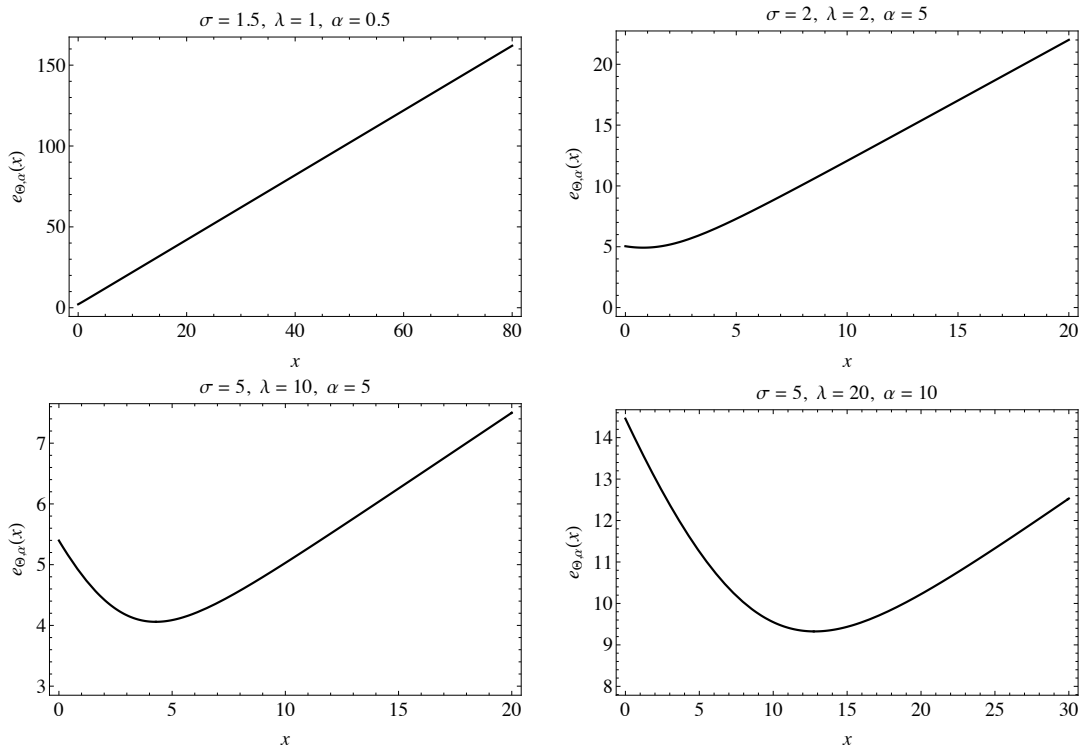


Figure 3. Mean residual life function of LAT distribution for selected values of parameters

4. Illustrative examples

In this section, we examine the practical performance of the LAT distribution in three examples that can be found in the personal web page of Professor E. Frees Frees (2010) (examples 1 and 3) and another one available in Klugman (1991) (example 2). All the data used in this work are displayed in Appendix.

The parameters are estimated using `WinRats` (Brooks, 2009) for examples 1 and 2, while `Mathematica` v.12.0 (Ruskeepaa, 2009) is used for example 3. The values of the supplied tests and the p-values were obtained using the `R` software. Graphical plots have been made employing `Mathematica` and `R`. All calculations were carried out on Windows-supported computers with an `i7-7700 CPU@3.60GHz` processor with response times for all examples standard.

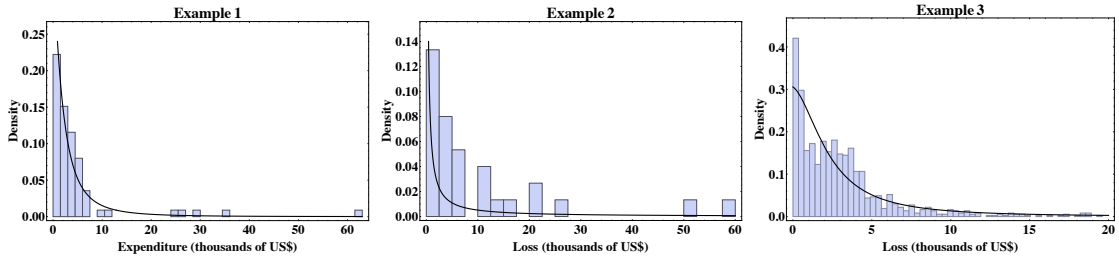
4.1 Example 1

The data were obtained from the Medical Expenditure Panel Survey (MEPS), conducted by the U.S. Agency of Health Research and Quality. MEPS is a probability survey that provides nationally representative estimates of health care use, expenditures, sources of payment, and insurance coverage for the U.S. civilian population. The variable of interest consist of amounts of expenditures for outpatient (EXPENDOP) visits. In the first row of Table 1, we report the descriptive statistics of the empirical data that seems to be unimodal and positively skewed. In Figure 4(al), it is displayed the histogram of the empirical data and the PDF plot corresponding to Example 1.

The log-likelihood function together with the normal equations, which provide the maximum likelihood estimates, are shown in Appendix of this article.

Table 1. Descriptive statistics of the data sets used in the indicated example.

Example	n	Mean	Standard deviation	Minimum	Maximum
1	75	4.95594	9.32897	0	62.8111
2	30	9.54	14.16	0	59
3	1091	5.3262	16.1746	0.005	273.604



(a) (b) (c)

Figure 4. Empirical histograms and PDF plots for examples 1 (a), 2 (b) and 3 (c).

We compare the LAT distribution introduced in this work with other competing models proposed in the literature that have the capacity to incorporate zero observations in the sample. As a benchmark, we consider the classical exponential distribution with mean $1/\lambda$, for $\lambda > 0$, the Lomax distribution and the generalized exponential distribution due to

Marshall and Olkin (1997), with SF given by

$$\bar{F}(x) = \frac{\lambda \exp(-\sigma x)}{1 - \bar{\lambda}\exp(-\sigma x)}, \quad x \geq 0, \sigma > 0, \lambda > 0,$$

and $\bar{\lambda} = 1 - \lambda$.

In Table 2 are exhibited the parameter estimates together with their standard errors (in brackets) for the four models considered. It can be seen that the LAT distribution provides the best fit to data in terms of the two measures of model selection examined, negative of the maximum of the log-likelihood function (NLL) and Akaike information criterion (AIC). Model selection was also assessed from a practical perspective using the Kolmogorov-Smirnov (KS) and the Crámer-von Mises (CM) goodness-of-fit tests to quantify the distance between the empirical CDF (ECDF) constructed from the data and the ones generated from the fitted models. Let $\widehat{F}$ denote the CDF of the fitted model, the original data by $x_1, \ldots, x_N$ and the ordered data in increasing magnitude by $x_{(1)}, \ldots, x_{(N)}$. Then the expressions of the KS and CM statistics are defined as:

(i) Kolmogorov-Smirnov test statistic: $D = \max(D^+, D^-)$, where

$$D^+ = \max_{1 \leq j \leq N} \left| \frac{j}{N} - \widehat{F}(x_{(j)}) \right|, \; D^- = \max_{1 \leq j \leq N} \left| \widehat{F}(x_{(j)}) - \frac{j-1}{N} \right|.$$

(ii) Crámer-von Mises test statistic:

$$W^2 = \sum_{j=1}^{N} \left[\widehat{F}(x_{(j)}) - \frac{2j-1}{2N} \right]^2 + \frac{1}{12N}.$$

Results on the goodness of fit of the four parametric models considered are also presented in last four rows of Table 2. Note that the LAT distribution yields lower values for both test statistics and it is not rejected for both tests as judged by the corresponding p-values.

Table 2. Example 1. Parameter estimates for the exponential (E), Lomax (L), generalized exponential (GE) and LAT distributions via maximum likelihood estimation. Standard errors are provided in parenthesis and p-values for the KS and CM tests between brackets.

Parameter	E	L	GE	LAT
$\widehat{\lambda}$	0.202	0.181	0.129	4.624
	(0.023)	(0.085)	(0.077)	(0.209)
$\widehat{\sigma}$		2.106	0.058	1.991
		(0.718)	(0.029)	(0.048)
$\widehat{\alpha}$				-0.581
				(0.238)
n	75	75	75	75
NLL	195.044	182.833	183.653	182.811
AIC	392.088	369.666	371.306	371.622
KS	0.157	0.747	0.107	0.077
	[0.10]	[< 0.001]	[0.652]	[0.97]
CM	0.752	0.227	0.191	0.127
	[0.007]	[0.237]	[0.293]	[0.460]

4.2 Example 2

In the second example, we use data that can be found in Appendix of Klugman (1991). In particular, we employ the data set 2 ,where the loss value for the first year in the 30 first classes have been taken. The second row of Table 1 shows the descriptive statistics of this second data set and in the middle panel of Figure 4 are illustrated the ECDF and the smooth CDF for the second example. In Table 3, we report the parameter estimates together with their standard errors (in brackets) for four of the models previously considered. Once again, it can be seen that the LAT distribution provides a marginal best fit data in terms of the negative of the NLL. However, when the AIC is considered, the GE distribution provides a slightly better fit to this dataset. Model selection was also assessed via KS and CM goodness-of-fit tests to quantify the distance between the ECDF constructed from the data and the ones generated from the fitted models. As judged by these tests, the LAT distribution is not rejected at usual significance levels.

Table 5 reports empirical and theoretical limited expected values for the LAT distribution with different values of the policy limit x, using the parameter estimates calculated for the dataset given in Example 2 and the expression defined in Equation (3.15). Note that for large values of x, that is, when x tends to infinity, the limited lev approaches to the mean of the distribution. Nevertheless, as in this case $\alpha < 1$, the mean does not exist.

Table 3. Example 2. Parameter estimates for the exponential (E), Lomax (L), generalized exponential (GE) and LAT distributions via maximum likelihood estimation. Standard errors are provided in parenthesis and p-values for the KS and CM tests between brackets.

Parameter	E	L	GE	LAT
$\widehat{\lambda}$	0.105	0.188	0.129	3.07×10^{-6}
	(0.019)	(0.174)	(0.123)	(0.003)
$\widehat{\sigma}$		1.340	0.034	0.176
		(0.751)	(0.027)	(0.016)
$\widehat{\alpha}$				-7.639
				(0.116)
n	30	30	30	30
NLL	97.644	93.726	93.034	63.969
AIC	197.288	191.452	190.068	133.938
KS	0.300	0.629	0.200	0.264
	[0.071]	[8.2×10^{-7}]	[0.586]	[0.134]

4.3 Example 3

The third dataset deals with automobile bodily injury claims data from the Insurance Research Council (IRC), a division of the American Institute for Chartered Property Casualty Underwriters and the Insurance Institute of America. The data, collected in 2002, contain information on demographic information about the claimant, attorney involvement and the economic loss (in thousands of US$). We consider a sample of 1091 losses from a single state. The third row of Table 1 reports descriptive statistics of this third data set, and in the bottom panel of Figure 4, the ECDF and the smooth CDF are displayed for the third example.

In Table 5, we report the parameter estimates together with their standard errors (in brackets) for the LAT and GE distributions and two models traditionally used to explain income data the lognormal (LO) distribution with parameters $\lambda \in (-\infty, \infty)$ and $\sigma > 0$ and

Table 4. Empirical and theoretical limited expected value for the LAT distribution and different values of the policy limit x for the second example dataset.

Policy limit (x)	Empirical	Fitted
0	0.00	0.00
2	1.50	0.99
4	2.70	1.82
6	3.57	2.58
8	4.27	3.31
10	4.94	4.02
12	5.47	4.70
14	5.90	5.37
16	6.27	6.02
18	6.60	6.66
20	6.94	7.29
22	7.20	7.91
24	7.40	8.53
26	7.60	9.13
28	7.77	9.73
30	7.90	10.32
32	8.04	10.90
34	8.17	11.48
36	8.30	12.05
38	8.44	12.62
40	8.57	13.18
42	8.70	13.74
44	8.84	14.30
46	8.97	14.85
48	9.10	15.40
50	9.24	15.94

the Singh-Maddala (SM) distribution with SF given by

$$\bar{F}(x) = \left[1 + \left(\frac{x}{\sigma}\right)^{\lambda}\right]^{-\alpha}, \quad x \geq 0, \sigma > 0, \lambda > 0, \alpha > 0. \tag{4.21}$$

Observe that the special case $\sigma = 1$ reduces Equation (4.21) to the Burr type XII distribution studied by Rezac et al. (2015). Once again, it can be seen that the LAT distribution provides the best fit to data in terms of the two measures of model selection examined, negative of the NLL and AIC. Model selection was also assessed via KS and CM goodness-of-fit tests to quantify the distance between the ECDF constructed from the data and the CDFs generated from the fitted models. As judged by the measures of model selection, the LAT distribution provides the best fit to the data. Moreover, although the LAT distribution is rejected in terms of the KS test and the CM test at the usual significance levels, the value of the test statistics are the lower among all the models considered.

Table 5. Example 3. Parameter estimates for the generalized exponential (GE), lognormal (LO), Singh-Maddala (SM) and LAT distributions via maximum likelihood estimation. Standard errors are provided in brackets and p-values between brackets.

Parameter	GE	LO	SM	LAT
$\widehat{\lambda}$	0.051	0.620	1.103	2.283
	(0.013)	(0.085)	(0.044)	(0.042)
$\widehat{\sigma}$	0.025	1.445	3.672	1.667
	(0.006)	(0.045)	(0.031)	(0.566)
$\widehat{\alpha}$			1.643	-1.709
			(0.188)	(0.514)
n	1091	1091	1091	1091
NLL	2637.87	2626.74	2601.69	2598.20
AIC	5279.74	5257.48	5209.38	5202.39
KS	0.089	0.093	0.062	0.052
	[< 0.001]	[< 0.001]	[< 0.001]	[0.005]
CM	2.715	2.446	1.200	1.002
	[< 0.001]	[< 0.001]	[< 0.001]	[0.004]

5. Conclusions, limitations, and future research

In this paper, we derive several properties related to the family of arctan transformation of a survival function, mainly those connected with the right tail of the distribution. After this, we introduced the arctan transformation of the Pareto Type II distribution, a scale transformation of the classical Pareto distribution. This is a model for non-negative continuous random variables, including the zero value in its support. We have provided in closed-form expression the raw moment, quantile function, the tail value at risk, and other functions which can be helpful in the financial and actuarial field, such as the integrated tail distribution, the limited expected value, and the mean excess function.

The performance of this new family of distributions has been illustrated by using three different data sets. The first one was associated with the expenditures for outpatients; the second one was related to the third party automobile insurance claims; and the final example considered automobile injury claims. Numerical results showed that the Lomax arctan distribution is helpful to explain heavy-tailed empirical data. However, although this distribution is able to capture the presence of zeros in the data, if the proportion of zeros is too high, the model has a worse performance in relation to the other models, as it is shown in the second example.

Further analysis of this probabilistic family remains as a topic for future studies. In this regard, investigation of the multivariate version of the arctan transformation is a topic that deserves to be examined in upcoming works in depth in upcoming works.

Author contributions Conceptualization, E.G-D., E.C-O., J.M.S.; methodology, E.G-D., E.C-O., J.M.S.; software, E.G-D., E.C-O., J.M.S.; validation, E.G-D., E.C-O., J.M.S.; formal analysis, E.G-D., E.C-O., J.M.S.; investigation, E.G-D., E.C-O., J.M.S.; data curation, E.G-D., E.C-O., J.M.S.; writing-original draft preparation, E.G-D., E.C-O., J.M.S.; writing-review and editing, E.G-D., E.C-O., J.M.S.; visualization, E.G-D., E.C-O., J.M.S.; supervision, E.G-D., E.C-O., J.M.S. All authors have read and agreed to the published version of the paper.

Acknowledgements We thank the two anonymous reviewers and the Editors for their valuable comments and suggestions, which have greatly helped us to improve the original paper.

Funding J.M.S. was partially funded by grant PID2019-105986GB-C22 (Ministerio de Ciencia en Innovación) from Spain.

Conflicts of interest The authors declare no conflict of interest.

References

Albrecher, H., Beirlant, J., and Teugels, J., 2017. Reinsurance: Actuarial and Statistical Aspects. Wiley, New York, USA.

Arnold, B., 1983. Pareto Distributions. International Cooperative Publishing House, Silver Spring, MA, USA.

Bingham, N., 1987. Regular Variation. Cambridge University Press, Cambridge, UK.

Boland, P., 2007. Statistical and Probabilistic Methods in Actuarial Science. Chapman and Hall, New York, USA.

Brooks, C., 2009. RATS Handbook to Accompany Introductory Econometrics for Finance. Cambridge University Press, Cambridge, UK.

Calderín-Ojeda, E., Azpitarte, F., and Gómez-Déniz, E., 2016. Modelling income data using two extensions of the exponential distribution. Physica A: Statistical Mechanics and its Applications, 461, 756–766.

Champernowne, D.G., 1952. The graduation of income distributions. Econometrica, 20, 591–615.

Embrechts, P. and Goldie, C.M., 1980. On closure and factorization properties of subexponential and related distributions. Journal of the Australian Mathematical Society, 29, 243–256.

Embrechts, P. and Goldie, C.M., 1982. On convolution tails. Stochastic Processes and their Applications, 13, 263–278.

Embrechts, P. and Veraverbeke, N., 1982. Estimates for the probability of ruin with special emphasis on the possibility of large claims. Insurance Mathematics and Economics, 1, 55–72.

Fisk, P.R., 1961. The graduation of income distributions. Econometrica, 29, 171–185.

Foss, S., Korshunov, D., and Zachary, S., 2011. An Introduction to Heavy-Tailed and Subexponential Distributions. Springer, New York, USA.

Frees, E.W., 2010. Regression Modeling with Actuarial and Financial Applications. Cambridge University Press. Cambridge, UK.

Gómez-Déniz, E. and Calderín-Ojeda, E., 2015a. Modelling insurance data with the Pareto arctan distribution. ASTIN Bulletin, 45, 639–660.

Gómez-Déniz, E. and Calderín-Ojeda, E., 2015b. On the use of the Pareto arctan distribution for describing city size in Australia and New Zealand. Physica A: Statistical Mechanics and its Applications, 436, 821–832.

Gómez-Déniz, E., 2016. A family of arctan Lorenz curves. Empirical Economics, 51, 1215–1233.

Gómez-Déniz, E.; Calderín–Ojeda, E. and Sarabia, J.M., 2019. The geometric ArcTan distribution with applications to model demand for health services. Communications in Statistics: Simulation and Computation, 48, 1101–1120.

Hogg, R. and Klugman, S., 1984. Loss Distributions. Wiley, New York, USA.

Jabob, E. and Jayakumar, K., 2012. On half-Cauchy distribution and process. International Journal of Statistika and Mathematika, 3, 77–81.

Jessen, A. and Mikosch, T., 2006. Regularly varying functions. Publications de l'Institut Mathématique, 80, 171–192.

Klugman, S.A., 1991. Bayesian Statistics in Actuarial Science: with Emphasis on Credibility. Kluwer Academic Publishers, Massachusetts, USA.

Klugman, S.A., Panjer, H.H., and Willmot, G.E., 2008. Loss Models. From Data to Decisions. Wiley, New Jersey, USA.

Klüppelberg, C., 1988. Subexponential distributions and integrated tails. Journal of Applied Probability, 25, 132–141.

Konstantinides, D., 2018. Risk Theory: A Heavy Tail Approach. World Scientific Publishing, New York, USA.

Marshall, A.W. and Olkin, I., 1997. A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. Biometrika, 84, 641–652.

Melo, M.S., Loose, L.H., and Carvalho, J. B., 2021. Lomax regression model with varying precision: Formulation, estimation, diagnostics, and application. Chilean Journal of Statistics, 12, 1891–204.

Rezac, J., Lio, Y.L., and Jiang, N., 2015. Burr type-XII percentile control charts. Chilean Journal of Statistics, 6, 67–87.

Rolski, T., Schmidli, H., Schmidt, V., and Teugel, J., 1999. Stochastic Processes for Insurance and Finance. Wiley, New York, USA.

Ross, S.M., 1996. Stochastic Processes. Wiley, New York, USA.

Ruskeepaa, H., 2009. Mathematica Navigator. Mathematics, Statistics, and Graphics. Academic Press, New York, USA.

Shaked, M. and Shanthikumar, J.G., 2007. Stochastic Orders. Springer, New York, USA.

Suárez-Espinosa, J., Villaseñor-Alva, J.A., Hurtado-Jamarillo, A., and Pérez-Rodríguez, P., 2018. A goodness of fit test for the Pareto distribution. Chilean Journal of Statistics, 9, 33–46.

Yang, H., (2004). Crámer-Lundberg asymptotics. In Encyclopedia of Actuarial Science, pp. 1–6. Wiley, New York, USA.

Appendix

Let us assume that $X_1, \ldots, X_n$ is a random sample selected from the distribution given in Equation (3.11), with their observations denoted by $x_1, \ldots, x_n$. The corresponding likelihood function is given by

$$\ell(\Theta, \alpha; \tilde{x}) = n(\log(\alpha) + \log(\sigma) - \log(\lambda) - \log(\tan^{-1}(\alpha)) - (\sigma+1)\sum_{i=1}^{n} \log(1 + x_i/\lambda) - \sum_{i=1}^{n} \log\left(1 + \alpha^2(1 + x_i/\lambda)^{-2\sigma}\right). \quad (5.22)$$

The normal equations obtained from Equation (5.22) are stated as

$$\frac{\partial \ell(\Theta, \alpha; \tilde{x})}{\partial \sigma} = \frac{n}{\sigma} - \sum_{i=1}^{n} \log(1 + x_i/\lambda) + 2\alpha^2 \sum_{i=1}^{n} \frac{(1 + x_i/\lambda)^{-2\sigma} \log(1 + x_i/\lambda)}{1 + \alpha^2(1 + x_i/\lambda)^{-2\sigma}} = 0, \quad (5.23)$$

$$\frac{\partial \ell(\Theta,\alpha;\tilde{x})}{\partial \lambda} = -\frac{n}{\lambda} + \frac{\sigma+1}{\lambda^2}\sum_{i=1}^{n}\frac{x_i}{1+x_i/\lambda} + \frac{2\sigma\alpha^2}{\lambda^2}\sum_{i=1}^{n}\frac{x_i(1+x_i/\lambda)^{-2\sigma-1}}{1+\alpha^2(1+x_i/\lambda)^{-2\sigma}} = 0,$$

$$\frac{\partial \ell(\Theta,\alpha;\tilde{x})}{\partial \alpha} = n\left[\frac{1}{\alpha} - \frac{1}{(1+\alpha^2)\tan^{-1}\alpha}\right] - 2\alpha\sum_{i=1}^{n}\frac{(1+x_i/\lambda)^{-2\sigma}}{1+\alpha^2(1+x_i/\lambda)^{-2\sigma}} = 0, \quad (5.24)$$

from which we can get the maximum likelihood estimates of the parameters by a numerical method such as Newton-Raphson. On taking the second partial derivatives of Equations (5.23)-(5.24), the Fisher information matrix $\mathcal{I}(\Theta,\alpha)$ can be obtained by taking the expectations of minus the second derivatives. The inverse of the matrix provides the variances for the maximum likelihood estimators.

Table 6. Data for example 1.

1.4683	35.9342	0	0	7.24614	4.62498	3.80673	1.95896
4.62158	3.72445	0.87823	0.28698	1.80114	6.73978	3.69576	0.06081
3.56721	24.2046	5.82075	6.46576	2.53495	0.69315	1.68874	0.82613
5.32987	3.46299	1.68822	0.03755	6.47876	2.58618	9.5353	0.54148
1.95018	1.18143	3.74168	0.77534	2.88031	2.40923	3.00777	0.36825
0.59158	0.05376	5.8413	0.17115	1.78891	0.47681	0.68236	
62.8111	0.12816	4.18265	5.37448	1.99109	3.76849	0.31383	
1.45670	3.44599	1.19869	2.56363	2.01848	2.56077	5.63908	
5.99927	3.08074	29.1859	5.20015	4.06117	2.13937	2.85663	
26.0717	0.11700	1.83426	0	10.8975	0.19800	4.37083	

Table 7. Data for example 2.

1	3	5	0	15	27	0	3	0	11
6	20	0	13	11	4	22	0	3	50
10	4	7	1	59	2	1	3	5	0

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://soche.cl/chjs/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