ISSN 1645-6726

REVSTAT

Statistical Journal

vol. 21 - n. 2 - April 2023

Instituto Nacional de Estatística
Statistics Portugal

REVSTAT-Statistical Journal, vol.21, n. 2 (April 2023)

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

Continues: Revista de Estatística = ISSN 0873-4275.

ISSN 1645-6726 ; e-ISSN 2183-0371

Editorial Board (2019-2023)

Editor-in-Chief – *Isabel FRAGA ALVES*

Co-Editor – *Giovani L. SILVA*

Associate Editors

Marília ANTUNES
Barry ARNOLD
Narayanaswamy BALAKRISHNAN
Jan BEIRLANT
Graciela BOENTE
Paula BRITO
Valérie CHAVEZ-DEMOULIN
David CONESA
Charmaine DEAN
Fernanda FIGUEIREDO
Jorge Milhazes FREITAS
Alan GELFAND
Stéphane GIRARD
Marie KRATZ
Victor LEIVA
Artur LEMONTE
Shuangzhe LIU
Maria Nazaré MENDES-LOPES
Fernando MOURA
John NOLAN
Paulo Eduardo OLIVEIRA
Pedro OLIVEIRA
Carlos Daniel PAULINO
Arthur PEWSEY
Gilbert SAPORTA
Alexandra M. SCHMIDT
Manuel SCOTTO
Lisete SOUSA
Milan STEHLÍK
María Dolores UGARTE

Executive Editor – *Olga BESSA MENDES*

Publisher – *Statistics Portugal*

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

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

Creative Commons Attribution 4.0 International (CC BY 4.0)
© Statistics Portugal, Lisbon. Portugal, 2023

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

INDEX

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

REVSTAT – Statistical Journal
Volume 21, Number 2, April 2023, 143–171
https://doi.org/10.57805/revstat.v21i2.407

On Goodness-of-Fit Tests for the Neyman Type A Distribution

Authors: Apostolos Batsidis ✉
– Department of Mathematics, University of Ioannina, 45110 Ioannina, Greece
abatsidis@uoi.gr

Artur J. Lemonte
– Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Brazil
arturlemonte@gmail.com

Received: April 2021 Revised: November 2021 Accepted: November 2021

Abstract:

- The two-parameter Neyman type A distribution is quite useful for modeling count data, since it corresponds to a simple, flexible and overdispersed discrete distribution, which is also zero-inflated. In this paper, we show that the probability generating function of the Neyman type A distribution is the only probability generating function which satisfies a certain differential equation. Based on an empirical counterpart of this specific differential equation, we propose and study a new goodness-of-fit test for this distribution. The test is consistent against fixed alternative hypotheses, while its null distribution can be consistently approximated by using parametric bootstrap. We investigate the finite sample performance of the proposed test numerically by means of Monte Carlo experiments, and comparisons with other existing goodness-of-fit tests are also considered. Empirical applications to real data are considered for illustrative purposes.

Keywords:

- *count data; empirical probability generating function; parametric bootstrap; probability generating function; Bell–Touchard distribution.*

AMS Subject Classification:

- 62F03, 62F40.

✉ Corresponding author.

1. INTRODUCTION

Modeling count data is an important issue in different disciplines and applied sciences such as medicine (see, for example, Joe and Zhu [25]), actuarial sciences (see, for example, Gossiaux and Lemaire [17], Lord *et al.* [32]), biology (see, for instance, Esnaola *et al.* [14]), health economics (see, for example, Zafakali and Ahmad [50]), among many others. With this aim, the one-parameter Poisson distribution and the two-parameter Negative Binomial distribution are commonly used. Nevertheless, observed count data often exhibit overdispersion (i.e., variance greater than the mean) and, therefore, the Poisson distribution is not adequate for fitting such data, since its variance is restricted to be equal to the mean. Additionally, a second usual feature of the observed count data is the presence of a high percentage of zero values (zero inflation or zero vertex). The zero-inflation index $zi = 1 + \log(p_0)/\mu$, where p_0 is the probability of zero, can be used to measure zero-inflation. Then $zi = 0$ for Poisson distribution, and $zi = 1 + \log(d)/(1-d) > 0$ for the Negative Binomial, where d denotes the Fisher dispersion index given by $d = \sigma^2/\mu$, where σ^2 and μ are the variance and mean, respectively [see 42]. Therefore, the Negative Binomial distribution is an improvement over the Poisson distribution, since it can model overdispersed and zero-inflated data.

Several other distributions have been presented in the statistical literature to handle both overdispersion and zero-inflation. In this frame, Neyman [39] developed the now well-known Neyman type A (NTA) distribution, which is overdispersed, because $d \geq 1$, and its zero-inflation index zi is always larger than the respective for the Negative Binomial for any fixed value of the dispersion index d (see Figure 1 in Puig and Valero [42]). For these reasons, the NTA distribution has been used in various disciplines such as bacteriology, ecology and entomology. The reader is referred to Johnson *et al.* [26, Chapter 9] and to Tripathi [49] for a list of applications of NTA distribution. Let $p_N(k;\tau,\delta)$ and $g_N(t;\tau,\delta)$ be the probability mass function (pmf) and probability generating function (pgf) of the NTA distribution, with parameters $\delta > 0$ and $\tau > 0$. We have that

$$\Pr(X = k) := p_N(k;\tau,\delta) = \frac{\tau^k \mathrm{e}^{\delta(\mathrm{e}^{-\tau}-1)}}{k!} m_k(\delta \mathrm{e}^{-\tau}), \quad k \in \mathbb{N}_0, \tag{1.1}$$

where $\mathbb{N}_0 = \mathbb{N} \cup \{0\} = \{0, 1, 2, ...\}$, $m_k(r) = \sum_{j=0}^{k} S(j,k) r^k$ is the k-th moment about zero for the Poisson distribution with parameter $r > 0$, and $S(k,j)$ are the Stirling numbers of second kind (see, for instance, Massé and Theodorescu [33] for further details). Also, $g_N(t;\tau,\delta) = \exp[\delta(\mathrm{e}^{\tau(t-1)} - 1)]$, $|t| \leq 1$. We shall use the notation $X \sim \mathrm{NTA}(\tau,\delta)$ to refer to this distribution.

Recently, Castellares *et al.* [8] on the basis of a series expansion presented in Touchard [48] and Bell [4, 5], obtained a two-parameter family of distributions (named as Bell–Touchard distribution) with pmf of the form

$$\Pr(X = k) := p(k;\theta) = \frac{\mathrm{e}^{b(1-\mathrm{e}^a)}\, a^k\, T_k(b)}{k!}, \quad k \in \mathbb{N}_0, \tag{1.2}$$

where $a > 0$ and $b > 0$, $\theta = (a,b) \in \Theta = (0,\infty) \times (0,\infty)$, and $T_k(\cdot)$ are the Touchard polynomials [48] defined by $T_k(b) = \mathrm{e}^{-b} \sum_{j=0}^{\infty} j^k\, b^j/j!$. We shall use the notation $X \sim \mathrm{BT}(a,b)$, or $X \sim \mathrm{BT}(\theta)$, to refer to the NTA distribution with this specific parameterization. If $X \sim \mathrm{BT}(a,b)$, then its pgf is given by

$$g(t;\theta) = \exp\{[b(\mathrm{e}^{ta} - \mathrm{e}^a)]\}, \quad |t| \leq 1. \tag{1.3}$$

The Touchard polynomials $T_k(b)$ corresponds to the k-th moment of the Poisson distribution with parameter equal to b and can be obtained for different values of k. For example, $T_0(b) = 1$, $T_1(b) = b$, $T_2(b) = b^2 + b$, $T_3(b) = b^3 + 3b^2 + b$, $T_4(b) = b^4 + 6b^3 + 7b^2 + b$, $T_5(b) = b^5 + 10b^4 + 25b^3 + 15b^2 + b$, $T_6(b) = b^6 + 15b^5 + 65b^4 + 90b^3 + 31b^2 + b$, and so on.

Remark 1.1. Note that when $b = 1$ in (1.2), the pmf of the Bell distribution introduced by Castellares *et al.* [7] is obtained as a special case, while the BT(a, b) distribution corresponds to the NTA$(\delta = b\mathrm{e}^a, \tau = a)$ distribution. So, the Bell–Touchard (BT) distribution is a reparameterization of the NTA distribution and, hence, in the whole paper the BT distribution stands for this reparameterization of the NTA distribution.

It is worth emphasizing that the two-parameter BT discrete distribution, or equivalently the NTA distribution, is very simple to deal with, since its pmf does not contain any complicated function. Tractability of the pmf may be a great advantage in computing the probabilities, as well as structural properties from that equation. The BT distribution has, among many other interesting properties the following properties:

- (**i**) it includes the one-parameter Bell distribution introduced by Castellares *et al.* [7] as a special case, which is also a reparameterization of the well-known NTA distribution;
- (**ii**) the Poisson distribution is not nested in the BT family, but it can be approximated for small values of a specific parameter of the BT distribution;
- (**iii**) it is a special case of a multiple Poisson process and can have a zero vertex;
- (**iv**) it is infinitely divisible;
- (**v**) it has variance larger than the mean;
- (**vi**) it is strongly unimodal for $b \geq 1$;
- (**vii**) it has an arbitrary number of modes when $b < 1$.

For a detailed description of the NTA distribution, the reader could consult Castellares *et al.* [8] and Johnson *et al.* [26, Chapter 9].

Based on the key features of the NTA distribution (or equivalently BT distribution), it can be easily justified why this distribution is a natural candidate and plays an important role in modeling count data with evidence of overdispersion and with high percentage of zero values. This implies that it is crucial to test the goodness-of-fit (gof) of this discrete distribution fitted to a given set of observations. A number of gof tests for count data are based on the pgf and the empirical pgf (epgf). To mention a few, but not limited to, we have the gof tests in Kocherlakota and Kocherlakota [29], Rueda *et al.* [46], Baringhaus and Henze [2], Epps [13], Rueda and O'Reilly [45], Meintanis and Bassiakos [36], Meintanis [35], Jiménez-Gamero and Alba-Fernandez [21], Batsidis *et al.* [3] and Milocevic *et al.* [37]. The motivation of using methods based on the pgf instead of the corresponding pmf when dealing with count data is, as argued by Nakamura and Perez-Abreu [38], that the pgf is usually simpler than the corresponding pmf. This is the case of the pgf of the BT distribution; compare expressions (1.2) and (1.3).

In this paper, we propose and study a consistent gof test for the two-parameter BT family of distributions; that is, based on Remark 1.1, it is equivalently to study a consistent gof test for the NTA distribution. Initially, it is shown that the pgf of the BT distribution is the only pgf satisfying a certain differential equation. Then, reasoning as Nakamura and Perez-Abreu [38] for testing Poisson distribution, Novoa-Muñoz and Jiménez-Gamero [41] for testing bivariate Poisson distribution, Jiménez-Gamero and Alba-Fernandez [21] for testing Poisson–Tweedie distribution, and Batsidis *et al.* [3] for testing Bell distribution, the proposed statistic is a function of the polynomial of an empirical version of the differential equation. In particular, the gof test proposed here can be considered as a generalization of the one in Batsidis *et al.* [3], since Bell distribution is a special case of the BT distribution. In addition, it can also be thought as a complement to the gof test for the Poisson–Tweedie distribution presented by Jiménez-Gamero and Alba-Fernandez [21], since NTA is a subset of the Poisson–Tweedie family of distributions. Additionally, for the first time, we apply some existing gof tests to the BT distribution and study their finite-sample properties from Monte Carlo simulation experiments. In particular, the numerical results reveal that two of the existing gof tests considered to the BT distribution present interesting results regarding size and power properties.

The paper is organized as follows. Section 2 contains some preliminaries related to existing gof tests. Section 3 introduces the test statistic and derives the asymptotic null distribution of the test statistic (i.e., the test statistic distribution under the null hypothesis), which depends on unknown quantities. To overcome this problem, it is shown that the parametric bootstrap consistently estimates the null distribution of the test statistic. Section 4 is devoted to study, with Monte Carlo simulation experiments, the finite sample performance of the proposed test and simultaneously to compare numerically the power of the new test with other two pgf-based tests introduced by Rueda and O'Reilly [45] and Meintanis [35]; that is, we also consider the pgf-based tests introduced by these authors to the BT distribution and study their finite sample properties in such a case. Apart from the previous gof tests, which are based on the pgf, the tests in Henze [19] and Klar [27], which are similar to that in Rueda and O'Reilly [45] but based on the distribution function and on the integrated distribution function, will also be considered in the comparison of the existing gof tests. Section 5 provides the application of the gof tests to real data sets. Section 6 closes up the paper with some concluding remarks. All technical proofs are deferred to Appendix.

Before ending this section we introduce some notation: all limits in this paper are taken when $n \to \infty$, where n denotes the sample size; $\xrightarrow{\mathcal{L}}$ denotes convergence in distribution; $\xrightarrow{\mathcal{P}}$ denotes convergence in probability; $\xrightarrow{a.s.}$ denotes the almost sure convergence; $I(A)$ denotes the indicator function of the set A; l^2 denotes the separable Hilbert space $l^2 = \{z = (z_0, z_1, z_2, ...),\ z_k \in \mathbb{R},\ \sum_{k\geq 0} z_k^2 < \infty\}$ with the usual inner product $\langle z, w\rangle_2 = \sum_{k\geq 0} z_k w_k$, and $\|\cdot\|_2$ stands for the associated norm; $\mathbb{E}_\theta$ and Cov_θ denote expectation and covariance by assuming that the data come from a BT distribution with parameter vector $\theta = (a, b)$; P_*, $\mathbb{E}_*$ and Cov_* denote the conditional probability law, the conditional expectation and the conditional covariance, respectively, given the data $X_1, ..., X_n$.

2. PRELIMINARIES AND EXISTING GOODNESS-OF-FIT TESTS

Let $X_1, ..., X_n$ be n independent and identically distributed random observations from a population X taking values in $\mathbb{N}_0$, with pgf $g(t) = \mathbb{E}(t^X)$, $|t| \leq 1$. Based on the sample $X_1, ..., X_n$, the objective is to test the composite, in the sense that the parameter vector $\theta = (a, b)$ is unknown, null hypothesis $H_0 : X \sim \mathrm{BT}(\theta)$, for some $\theta = (a, b) \in \Theta$ against the alternative hypothesis $H_1 : X \nsim \mathrm{BT}(\theta)$, $\forall\, \theta = (a, b) \in \Theta$. Obviously, based on Remark 1.1, the previous hypothesis is equivalent in testing the null hypothesis $H_0 : X \sim \mathrm{NTA}(\delta, \tau)$, for some $(\delta, \tau) \in (0, \infty) \times (0, \infty)$, against the alternative hypothesis $H_1 : X \nsim \mathrm{NTA}(\delta, \tau)$, $\forall\, (\delta, \tau) \in (0, \infty) \times (0, \infty)$.

It is well-known that the distribution of a random variable X taking values in $\mathbb{N}_0$ is fully and uniquely determined by its pgf. Also, the pgf can be consistently estimated by the epgf given by $g_n(t) = \frac{1}{n} \sum_{i=1}^{n} t^{X_i}$. It is worth stressing that Kocherlakota and Kocherlakota [29] were the first authors who proposed to base a gof test on the so-called epgf process with estimated parameter given by $K_n(\widehat{\theta}, t) = \sqrt{n}[g_n(t) - g(t; \widehat{\theta})]$, for $0 \leq t \leq 1$, where $g(t; \theta)$ is the pgf under the law in the null hypothesis; that is, in our special case, $g(t; \theta)$ is given in relation (1.3), and $\widehat{\theta} = (\widehat{a}, \widehat{b})$ is a consistent estimator of $\theta = (a, b)$.

Kocherlakota and Kocherlakota [29] exemplified their method with the Poisson-type distributions and NTA distribution. However, their method has the disadvantage that it depends on the choice of the value of t at which the pgf is evaluated. To overcome this problem, Rueda *et al.* [46] suggested the use of the following Cramér-von Mises type test statistic $R_{n,0}(\widehat{\theta}) = \int_0^1 K_n(\widehat{\theta}, t)^2 dt = n \int_0^1 [g_n(t) - g(t; \widehat{\theta})]^2 dt$. In addition, Rueda and O'Reilly [45] proposed a natural generalization of the Cramér-von Mises type test statistic by introducing a suitable weight function in order to make the test more sensitive to selected alternatives; see also Baringhaus *et al.* [1]. In this frame, they suggested the following test statistic $R_{n,w}(\widehat{\theta}) = n \int_0^1 [g_n(t) - g(t; \widehat{\theta})]^2 w(t) dt$, where $w(t)$ is a non-negative function on $(0, 1)$ such that $\int_0^1 w(t) dt < \infty$. By straightforward algebra, we have that $R_{n,w}(\widehat{\theta}) = \frac{1}{n} \sum_{j,k=1}^{n} \{\omega(1, X_{jk}) - \omega(g(t; \widehat{\theta}), X_j) - \omega(g(t; \widehat{\theta}), X_k) + \omega(g^2(t; \widehat{\theta}), 0)\}$, where $X_{jk} = X_j + X_k$, and $\omega(f, d) = \int_0^1 t^d f(t) w(t) dt$. Note that $R_{n,w}(\widehat{\theta})$ can be equivalently expressed in the form $R_{n,w}(\widehat{\theta}) = n \sum_{r,k=0}^{\infty} \{p(r; \theta) - \widehat{p}(r)\} \{p(k; \theta) - \widehat{p}(k)\} \int_0^1 t^{r+k} w(t) dt$, where $p(k; \theta)$ is given by (1.2), and

$$\widehat{p}(k) = \frac{1}{n} \sum_{j=1}^{n} I(X_j = k), \quad k = 0, 1, ... \tag{2.1}$$

Note that $\widehat{p}(k)$ corresponds to the empirical pmf for a given dataset. Hence, one rejects the null hypothesis H_0 for large values of the test statistic $R_{n,w}(\widehat{\theta})$.

After the pioneer work by Kocherlakota and Kocherlakota [29], a large number of gof tests for specific discrete distributions have been developed based on test statistics that utilize properties of the pgf of the law under the null hypothesis. In this context, Meintanis [35] presented a unified approach in testing the fit to any distribution belonging to the compound Poisson family of distributions. The compound Poisson family of distributions is defined as the distribution of $X = \sum_{j=1}^{N} Y_j$, where Y_j $(j = 1, ..., N)$ are independent and identically distributed with a common pgf $\psi(t; \xi)$, $\xi \in \mathbb{R}^p$ is a parameter vector, $N \sim \mathrm{Poisson}(\lambda)$ is

independent of Y_j $(j = 1, ..., N)$, and $\lambda > 0$. Meintanis [35] has noted that the pgf of any member of the compound Poisson family, say $\zeta(t)$, satisfies the following differential equation

$$\zeta'(t) - \lambda\psi'(t;\xi)\zeta(t) = 0, \tag{2.2}$$

where $\zeta'(t) = (d/dt)\zeta(t)$ and $\psi'(t;\xi) = (d/dt)\psi(t;\xi)$. Then, since the pgf and its derivatives can be consistently estimated by the epgf and the derivatives of the epgf (see, for example, Proposition 2 of Novoa-Muñoz and Jiménez-Gamero [40] for the uniform consistency of g_n and its derivatives), Meintanis [35] proposed the following test statistic

$$T_{n,w}(\widehat{\lambda}, \widehat{\xi}) = n \int_0^1 [\zeta_n'(t) - \widehat{\lambda}\psi'(t;\widehat{\xi})\zeta_n(t)]^2 w(t)dt, \tag{2.3}$$

where $\zeta_n'(t) = (d/dt)\zeta_n(t)$, and $\zeta_n(t)$ denotes the epgf. Note that the test statistic defined in (2.3) is an integral of the squared of an empirical counterpart of equation (2.2).

The general test statistic given in (2.3) can be exemplified in the special case of the BT distribution with parameter vector $\theta = (a, b)$, once the proposition below justifies that the BT distribution belongs to the compound Poisson family of distributions. This result can be found in Feller [15] and in Castellares *et al.* [8].

Proposition 2.1. *Let $X \sim \mathrm{BT}(a, b)$, where $a > 0$ and $b > 0$. Then, we have that $X = \sum_{j=1}^{N} Y_j$, where Y_j $(j = 1, ..., N)$ are independent and identically zero-truncated Poisson distributed random variables with parameter $a > 0$ and a common pgf $\psi(t;a) = \frac{\exp(at)-1}{\exp(a)-1}$, and $N \sim$ Poisson$(b(\mathrm{e}^a - 1))$ independent of Y_j $(j = 1, ..., N)$.*

In terms of the notation used by Meintanis [35], it is evident that the BT distribution belongs to the compound Poisson family with $\lambda = b(\mathrm{e}^a - 1)$, $\psi(t;\xi) = \frac{\exp(\xi t)-1}{\exp(\xi)-1}$, $\psi'(t;\xi) = \frac{\xi\exp(\xi t)}{\exp(\xi)-1}$, and $\xi = a$. Therefore, based on the work of Meintanis [35], the pgf $g(t;\theta)$ of the BT distribution defined in (1.3) satisfies the following differential equation

$$g'(t) - ba\mathrm{e}^{at}g(t) = 0, \quad \forall\, t \in [0, 1], \tag{2.4}$$

and so the null hypothesis H_0 is rejected for large values of the following test statistic $M_{n,w}(\widehat{\theta}) = n\int_0^1 G_n(t,\widehat{\theta})^2 w(t)dt$, where $G_n(t;\theta)$ is the empirical version of (2.4) given by

$$G_n(t;\widehat{\theta}) = g_n'(t) - \widehat{b}\widehat{a}\mathrm{e}^{\widehat{a}t}g_n(t), \tag{2.5}$$

with $g_n'(t) = (d/dt)g_n(t)$. By straightforward algebra (see also Meintanis [35, p. 753]), we have that $M_{n,w}(\widehat{\theta}) = \frac{1}{n}\sum_{j,k=1}^{n}\{X_jX_k\omega(1, X_{jk} - 2) + (\widehat{b}\widehat{a})^2\omega(\mathrm{e}^{2\widehat{a}t}, X_{jk}) - \widehat{b}\widehat{a}X_{jk}\omega(\mathrm{e}^{\widehat{a}t}, X_{jk} - 1)\}$. Note that $M_{n,w}(\widehat{\theta})$ can be equivalently expressed in the form $M_{n,w}(\widehat{\theta}) = n\sum_{r,k=0}^{\infty}\widehat{d}(r;\widehat{\theta})\widehat{d}(k;\widehat{\theta})\cdot\int_0^1 t^{r+k}w(t)dt$, where

$$\widehat{d}(k;\theta) = (k+1)\widehat{p}(k+1) - \sum_{u=0}^{k} coef(u;\theta)\widehat{p}(k-u), \quad k = 0, 1, ..., \tag{2.6}$$

and $coef(u;\theta) := coef(u;a,b) = \frac{ba^{u+1}}{u!}$ can be recursively calculated as follows: $coef(0;a,b) = ba$, and $coef(u;a,b) = coef(u-1;a,b)a/u$ for $u \geq 1$.

Remark 2.1. The asymptotic null distributions of the test statistics $R_{n,w}(\widehat{\theta})$ and $M_{n,w}(\widehat{\theta})$ are intractable (Rueda and O'Reilly [45] and Meintanis [35]) and, hence, the critical points required for the implementation of these test procedures can be determined via parametric bootstrap. It should be mentioned that the application of both tests requires the choice of a weight function. Specific choices of it, which are rather arbitrary, can lead to considerable computational simplification. In this frame, the choice of $w(t) = t^{\gamma}$, where $\gamma \geq 0$ denotes a constant, corresponds to an interesting choice. This weight function, apart from computational convenience, has the following interpretation: for large values of γ more weight is assigned to the values of $K_n(\widehat{\theta}, t)$ and $G_n(t; \widehat{\theta})$ near $t = 1$; hence, large values of γ should render the test sensitive to deviations from the moments of the hypothesized distribution; see, for instance, Gürtler and Henze [18].

Apart from the previous tests, which are based on the pgf, the tests in Henze [19] and Klar [27] denoted as H_n and W_n, which are similar to that in Rueda and O'Reilly [45] but they are based on the distribution function and on the integrated distribution function, respectively, will be also particularized for the BT distribution and will be also considered in the simulation studies of Section 4. Specifically, we consider the modified Cramér–von Mises statistic in expression (3.6) of Henze [19] given by

$$H_n = \sum_{k=0}^{X_{(n)}} [F_n(k) - F(k; \widehat{\theta})]^2 [F_n(k) - F_n(k-1)], \tag{2.7}$$

where $X_{(n)} = \max_{1 \leq j \leq n} X_j$, $F_n(x)$ stands for the empirical distribution function defined by $F_n(x) = n^{-1} \sum_{j=1}^{n} I(X_j \leq x)$, and $F(x; \theta)$ denotes the cumulative distribution function of the BT distribution with parameter θ. In contrast to the Cramér–von Mises statistic in expression (2.2) of Henze [19], whose practical calculation involves truncation, the calculation of H_n involves a finite sum and hence was preferred (see also Jiménez-Gamero and Alba-Fernandez [22]). Finally, following Henze [19], to perform the test based on H_n a parametric bootstrap is used and the null hypothesis is rejected for a large observed value of the test statistic H_n. We also consider the test statistic (see relation (1) in Klar [27]) $W_n = \sqrt{n} \sup_{t \geq 0} |Y_n(t) - \widehat{Y}(t)|$, where $Y(t) = \int_t^{+\infty} [1 - F(x)] dx$, $Y_n(t)$ denotes its empirical counterpart and $\widehat{Y}(t)$ equals $Y(t)$ with $F(x)$ replaced by $F(x; \widehat{\theta})$. In practice (see also Jiménez-Gamero and Alba-Fernandez [22]), we consider the expression (8) in Klar [27] given by

$$W_n = \sqrt{n} \sup_{1 \leq k \leq X_{(n)}} \left| \bar{X} - \mathbb{E}_{\widehat{\theta}}(X) + \sum_{j=0}^{k-1} [F_n(j) - F(j; \widehat{\theta})] \right|,$$

where $\bar{X}$ denotes the sample mean. For instance, if the moment estimator is used then the previous relation is simplified taking into account that $\mathbb{E}_{\widehat{\theta}}(X) = \bar{X}$. On the other hand, if the maximum likelihood (ML) estimator is used, then the relation is simplified taking into account that $\mathbb{E}_{\widehat{\theta}}(X) = \widehat{b}\widehat{a}\mathrm{e}^{\widehat{a}}$, where $\widehat{a}$ and $\widehat{b}$ are the ML estimates of a and b, respectively, since $\mathbb{E}(X) = ba\mathrm{e}^{a}$, when $X \sim \mathrm{BT}(a, b)$. Following Klar [27], to perform the test based on W_n a parametric bootstrap is used and hence the null hypothesis is rejected for a large value of the associated test statistic.

3. A NEW TEST STATISTIC

In this section, a new gof test statistic will be constructed based on the characterization of the BT distribution provided below and parallel with the tests discussed by Nakamura and Perez-Abreu [38] for testing Poisson distribution, Novoa-Muñoz and Jiménez-Gamero [41] for testing bivariate Poisson, Jiménez-Gamero and Alba-Fernandez [21] for testing Poisson–Tweedie, and Batsidis *et al.* [3] for testing Bell distribution. To be specific, the next proposition shows that the BT pgf is the unique solution of the differential equation given in (2.4).

Proposition 3.1. *Let $G = \{g : [0,1] \to \mathbb{R}$, such that g is a pgf and $g'(t) = (\partial/\partial t)g(t)$ exists $\forall\, t \in [0,1]\}$, which is equivalent to say that G is the set of probability generating functions associated with random variables taking values in $\mathbb{N}_0$ with finite mean. Let $g(t;\theta)$ be defined as in (1.3). Then, $g(t;\theta)$ is the only pgf in G satisfying the differential equation given in (2.4).*

Therefore, the BT pgf is the only pgf satisfying the differential equation (2.4). Also, the pgf $g(t)$ and its derivatives can be consistently estimated by the epgf and the derivatives of the epgf. Under the null hypothesis H_0, it then follows that the empirical version of (2.4) denoted by $G_n(t;\widehat{\theta})$ and given in (2.5) should be close to zero, $\forall\, t \in [0,1]$, where $\widehat{\theta} = (\widehat{a}, \widehat{b})$ is a consistent estimator of $\theta = (a,b)$. Additionally, $G_n(t;\widehat{\theta})$ can be expressed in the form $G_n(t;\widehat{\theta}) = \sum_{k\geq 0} \widehat{d}(k;\widehat{\theta})t^k$, where $\widehat{p}(k)$ and $\widehat{d}(k;\widehat{\theta})$ are defined in (2.1) and (2.6), respectively. It implies that (under the null hypothesis) $S_n(\widehat{\theta}) = \sum_{k\geq 0} \widehat{d}(k;\widehat{\theta})^2 \approx 0$. Note that $S_n(\widehat{\theta}) = \|\widehat{d}(\cdot;\widehat{\theta})\|_2^2$, where $\widehat{d}(\cdot;\widehat{\theta}) = (\widehat{d}(0;\widehat{\theta}), \widehat{d}(1;\widehat{\theta}), ...)$, and $\widehat{d}(k;\widehat{\theta})$ is given in (2.6). Also, $\widehat{d}(k;\theta) = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i;k,\theta)$, where

$$\phi(X;k,\theta) = (k+1)I(X = k+1) - b\sum_{u=0}^{k} \frac{a^{u+1}}{u!} I(X = k-u). \tag{3.1}$$

In this paper, we propose and study a new gof test for the BT family of distributions based on the statistic $S_n(\widehat{\theta})$. In order to give a solid justification of $S_n(\widehat{\theta})$ as a test statistic for testing H_0, we derive its limit distribution in the next theorem.

Theorem 3.1. *Let $X_1, ..., X_n$ be independent and identically distributed from X, a random variable taking values in $\mathbb{N}_0$ with probability mass function $p(k) = \Pr(X = k)$, $k \in \mathbb{N}_0$, so that $\mathbb{E}(X^2) < \infty$. Assume that $\widehat{\theta} \xrightarrow{a.s.(P)} \theta$, then $S_n(\widehat{\theta}) \xrightarrow{a.s.(P)} \eta = \|d(\cdot;\theta)\|_2^2$, where $d(\cdot;\theta) = (d(0;\theta), d(1;\theta), ...)$, and $d(k;\theta) = (k+1)p(k+1) - b\sum_{u=0}^{k} \frac{a^{u+1}}{u!} p(k-u), \quad k \in \mathbb{N}_0$.*

It should be noted that $\eta \geq 0$ and, from Proposition 3.1, $\eta = 0$ if and only if H_0 is true. Hence, the null hypothesis H_0 should be rejected for large values of the test statistic $S_n(\widehat{\theta})$. Now, to determine what is a large value we have to obtain the distribution of the test statistic $S_n(\widehat{\theta})$ under the null hypothesis H_0, or at least an approximation to it. With this aim, we next derive its asymptotic null distribution. We will assume that the estimator $\widehat{\theta} = (\widehat{a}, \widehat{b})$ satisfies the following regularity condition.

Assumption 1. Under H_0, if $\theta = (a,b) \in \Theta$ denotes the true parameter value, then $\sqrt{n}(\widehat{\theta} - \theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \ell(X_i;\theta) + o_P(1)$, with $\mathbb{E}_\theta\{\ell(X_i;\theta)\} = 0$ and $J(\theta) = \mathbb{E}_\theta\{\ell(X_i;\theta)^T \ell(X_i;\theta)\} < \infty$.

Assumption 1 implies that when the null hypothesis is true and θ denotes the true parameter value, then $\sqrt{n}(\widehat{\theta} - \theta)$ is asymptotically normally distributed. This assumption is not restrictive at all since it is fulfilled by commonly used estimators such as the the ML estimator and the moment estimator (see White [51] and Jiménez-Gamero and Kim [24], among others). In Appendix B, the form of the function ℓ is provided for the aforementioned estimators under the BT family of distributions, and we show that the conditions of Assumption 1 really holds for them.

The next theorem gives the asymptotic null distribution of $S_n(\widehat{\theta})$.

Theorem 3.2. *Let $X_1, ..., X_n$ be independent and identically distributed from $X \sim \mathrm{BT}(\theta)$, where $\theta = (a, b) \in \Theta$. Suppose that $\widehat{\theta}$ satisfies Assumption 1. Then, $nS_n(\widehat{\theta}) \xrightarrow{\mathcal{L}} \|S(\theta)\|_2^2$, where $\{S(\theta) = (S(0;\theta), S(1;\theta), ...)\}$ is a centered Gaussian process in l^2 with covariance kernel $\varrho(k, r) = \mathrm{Cov}_\theta\{Y(X;k,\theta), Y(X;r,\theta)\}$ for $k \in \mathbb{N}_0$ and $r \in \mathbb{N}_0$, $Y(X;k,\theta) = \phi(X;k,\theta) + (\mu_1(k;\theta), \mu_2(k;\theta))\ell(X;\theta)^T$, ϕ is defined in (3.1), $\mu_1(k;\theta) = \mathbb{E}_\theta\{(\partial/\partial a)\phi(X;k,\theta)\}$, and $\mu_2(k;\theta) = \mathbb{E}_\theta\{(\partial/\partial b)\phi(X;k,\theta)\}$.*

Remark 3.1. If someone specifies the function ℓ for a specific estimator, then the covariance kernel appeared in the statement of the previous theorem can be given explicitly since one has just to calculate an expectation. For the BT family of distributions, when the moment estimators are used, we have proved in Appendix B that the function ℓ can be obtained in a closed, but rather complicated, form. On the other hand, when the ML estimators are used, the function ℓ cannot be obtained in a closed form. For the previous reasons, we did not provide the form of the covariance kernel $\varrho(k, r)$ for the aforementioned estimators.

Note that the null distribution of $\|S(\theta)\|_2^2$ is that of $\sum_{j\geq 1} \lambda_j \chi_{1j}^2$, where χ_{11}^2, χ_{12}^2, ... are independent χ^2 variates with one degree of freedom, and the set $\{\lambda_j\}$ are the positive eigenvalues of the linear operator $f \mapsto \mathcal{C}f$ on l^2 associated with the kernel ϱ given in Theorem 3.2; that is, $(\mathcal{C}f)(k) = \sum_{r\geq 0} \varrho(r, k) f(r)$. Since these eigenvalues depend on the unknown θ, it is evident that the asymptotic null distribution of the test statistic $nS_n(\widehat{\theta})$ depends on the unknown true value of the parameter vector $\theta = (a, b)$. However, even if θ was known or replaced by an appropriate estimator $\widehat{\theta}$, to determine the eigenvalues of an operator is a quite hard problem and unfortunately we did not succeed in finding explicit expressions for such eigenvalues. For similar problems and arguments see Novoa-Muñoz and Jiménez-Gamero [41] and Jiménez-Gamero and Alba-Fernandez [22], among others. Based on the previous remarks, it is concluded that the asymptotic null distribution of $nS_n(\widehat{\theta})$ given in Theorem 3.2 does not provide a useful approximation to its null distribution. Therefore, one should find another way of approximating the null distribution of the test statistic $nS_n(\widehat{\theta})$.

A common approach is to consider a parametric bootstrap approach to estimate the null distribution of $\|S(\theta)\|_2^2$. In the sequel, the parametric bootstrap approach is defined. Given the data X_1, ..., X_n, let X_1^*, ..., X_n^* be independent and identically distributed from $X^* \sim \mathrm{BT}(\widehat{\theta})$. Let $S_n^*(\widehat{\theta}^*)$ be the bootstrap version of $S_n(\widehat{\theta})$ obtained by replacing $X_1, ..., X_n$ and $\widehat{\theta} = \widehat{\theta}(X_1, ..., X_n)$ with $X_1^*, ..., X_n^*$ and $\widehat{\theta}^* = \widehat{\theta}(X_1^*, ..., X_n^*)$, respectively, in the expression of $S_n(\widehat{\theta})$. Then, we approximate $P_\theta\{S_n(\widehat{\theta}) \leq x\}$ by means of its bootstrap version, i.e. $P_*\{S_n^*(\widehat{\theta}^*) \leq x\}$. In order to show that the parametric bootstrap consistently approximates the null distribution of $S_n(\widehat{\theta})$, we need the following assumption, which is a bit stronger than Assumption 1.

Assumption 2. Assumption 1 holds, and the functions $\ell(X;\theta)$ and $J(\theta)$ satisfy:

(**1**) $\sup_{\vartheta\in\Delta}\mathbb{E}_\vartheta\{\|\ell(X;\vartheta)\|^2 I(\|\ell(X;\vartheta)\|>\epsilon\sqrt{n})\}\longrightarrow 0, \forall\,\epsilon>0$, where $\Delta\subseteq\Theta$ is an open neighborhood of θ;

(**2**) $\ell(X;\vartheta)$ and $J(\vartheta)$ are continuous as functions of ϑ at $\vartheta=\theta$.

Theorem 3.3. *Let $X_1,...,X_n$ be independent and identically distributed from X, a random variable taking values in $\mathbb{N}_0$. Assume that $\widehat{\theta}\overset{a.s.(P)}{\longrightarrow}\theta$, for some $\theta\in\Theta$, and that Assumption 2 holds. Then,* $\sup_{x\in\mathbb{R}}\left|P_*\{S_n^*(\widehat{\theta}^*)\le x\}-P_\theta\{S_n(\widehat{\theta})\le x\}\right|\overset{a.s.(P)}{\longrightarrow}0$.

Theorem 3.3 holds whether H_0 is true or not. It states that the conditional distribution of $S_n^*(\widehat{\theta}^*)$ and the distribution of $S_n(\widehat{\theta})$ are close when the sample is drawn from a population with $\mathrm{BT}(\theta)$ distribution, $\theta=(a,b)$ being the limit of $\widehat{\theta}=(\widehat{a},\widehat{b})$. In particular, if the null hypothesis H_0 is true, then Theorem 3.3 states that the conditional distribution of $S_n^*(\widehat{\theta}^*)$ is close to the null distribution of $S_n(\widehat{\theta})$. Let $\alpha\in(0,1)$. Hence, the test function

$$\Psi^*=\begin{cases}1, & \text{if } S_n(\widehat{\theta})\ge s^*_{n,\alpha},\\ 0, & \text{otherwise,}\end{cases}$$

or, equivalently, the test that rejects H_0 when $p^*=P_*\{S_n^*(\widehat{\theta}^*)\ge S_{obs}\}\le\alpha$, is asymptotically correct in the sense that when H_0 is true, $\lim_{n\to\infty}P_\theta(\Psi^*=1)=\alpha$, where $s^*_{n,\alpha}=\inf\{x: P_*(S_n^*(\widehat{\theta}^*)\ge x)\le\alpha\}$ is the α upper percentile of the bootstrap distribution of $S_n(\widehat{\theta})$, and S_{obs} is the observed value of the test statistic obtained from a given dataset. An immediate consequence of Theorem 3.1 and Theorem 3.3 is that the test Ψ^* is consistent; that is, it is able to detect any fixed alternative, in the sense that $\Pr(\Psi^*=1)\to 1$ whenever $X\nsim\mathrm{BT}(\theta)$, for any $\theta\in\Theta$.

Remark 3.2. A parametric bootstrap estimator of the null distribution of $nS_n(\widehat{\theta})$ was previously discussed. As observed before, the most important difficulty with the distribution of $\|S(\theta)\|_2^2$ is the determination of the positive eigenvalues λ_j which, however, can be consistently (a.s.) approximated following Dehling and Mikosch [11]. In this context, another solution is to approximate the null distribution of $nS_n(\widehat{\theta})$ through the conditional distribution, given $X_1,...,X_n$, of $\sum_{j\ge1}\widehat{\lambda}_j\,\chi^2_{1j}$, where $\chi^2_{11},\chi^2_{12},...$ are independent χ^2 variates with one degree of freedom and $\widehat{\lambda}_j$ is a consistent estimator of the eigenvalue λ_j, by means of weighted bootstrap in the sense of Burke [6] (see also, for instance, Kojadinovic and Yan [30] and references therein). From a computational point of view, the weighted bootstrap is more efficient than the parametric bootstrap. On the other hand, it has the disadvantage that one needs to estimate the function ℓ (see, for instance, Jiménez-Gamero and Kim [24]). In this paper, we rely on parametric bootstrap similar to the existing gof tests described in Section 2.

Before closing this section, we have to note that the bootstrap p-value of any of the five tests, namely $S_n(\widehat{\theta})$, $R_{n,w}(\widehat{\theta})$, $M_{n,w}(\widehat{\theta})$, H_n and W_n cannot be exactly computed. In the sequel, let T denote any of the five test statistics and let T_{obs} stand for the observed value of such statistic. Then, the bootstrap p-value can be approximated as follows:

1. Calculate the observed values of the gof test statistics for the available dataset $X_1,...,X_n$, say $S_{obs}(\widehat{\theta})$, $M_{obs}(\widehat{\theta})$, $R_{obs}(\widehat{\theta})$, H_{obs} and W_{obs};

2. Generate B bootstrap samples $X_1^{*v}, ..., X_n^{*v}$ from $X^* \sim \mathrm{BT}(\widehat{\theta})$, for $v = 1, ..., B$;

3. Calculate the test statistics $S_n(\widehat{\theta})$, $M_{n,w}(\widehat{\theta})$, $R_{n,w}(\widehat{\theta})$, H_n and W_n for each bootstrap sample and denote them, respectively, by S_v^*, M_v^*, R_v^*, H_v^* and W_v^* for $v = 1, ..., B$;

4. Compute the p-values of the tests based on the statistics $S_n(\widehat{\theta})$, $M_{n,w}(\widehat{\theta})$, $R_{n,w}(\widehat{\theta})$, H_n and W_n by means, respectively, of the expressions

$$\widehat{p}_S = \frac{\#\{S_v^* \geq S_{obs}(\widehat{\theta})\}}{B},\ \widehat{p}_M = \frac{\#\{M_v^* \geq M_{obs}(\widehat{\theta})\}}{B},\ \widehat{p}_R = \frac{\#\{R_v^* \geq R_{obs}(\widehat{\theta})\}}{B},$$

$$\widehat{p}_H = \frac{\#\{H_v^* \geq H_{obs}\}}{B},\ \widehat{p}_W = \frac{\#\{W_v^* \geq W_{obs}\}}{B}.$$

For a good discussion of bootstrap p-values, see Efron and Tibshirani [12, Chapter 16].

4. FINITE-SAMPLE SIZE AND POWER PROPERTIES

The properties studied in the previous section related to the test statistic $S_n(\widehat{\theta})$ are asymptotic, which means that they describe the behavior of the proposed test when the sample size is large. In this section, we empirically investigate its performance in small and moderate sample sizes through Monte Carlo simulation experiments. We also include in the Monte Carlo studies the test statistics $R_{n,w}(\widehat{\theta})$, $M_{n,w}(\widehat{\theta})$, H_n and W_n for comparison. We have not considered the test statistic $K_n(t;\widehat{\theta})$ in the Monte Carlo experiments since the question on how to select t remains unsolved and its performance depends on different values of t. It is worth stressing that the numerical results regarding the existing gof tests applied in the BT distribution are new, and so it also represents an additional contribution of the current paper in studying the performance of these specific existing gof tests for this two-parameter discrete distribution. All computations were performed by using the R language [43]. In all cases, 10,000 Monte Carlo replications were considered. Without loss of generality, we consider $a = 0.8$ and 1.4, and $b = 0.6, 1.2$ and 1.8.

The computation of the test statistics $R_{n,w}(\widehat{\theta})$ and $M_{n,w}(\widehat{\theta})$ depend on the weight function $w(t)$. Here, we consider the weight function in the form $w(t) = t^\gamma$, where $t \in (0,1)$ and $\gamma = 0, 1, 2, 5$ and 10. It is interesting to note that $\gamma = 0$ corresponds to the probability density function of the uniform distribution on $(0,1)$ as a weight function. The resulting tests when $w(t) = t^\gamma$ is used as a weight function will be denoted by $R_{n,\gamma}(\widehat{\theta})$ and $M_{n,\gamma}(\widehat{\theta})$. In particular, we have that

$$R_{n,\gamma}(\widehat{\theta}) = \sum_{r,k=0}^{\infty} \frac{\{p(r;\theta) - \widehat{p}(r)\}\{p(k;\theta) - \widehat{p}(k)\}}{r+k+\gamma+1}$$

and

$$M_{n,\gamma}(\widehat{\theta}) = \sum_{r,k=0}^{\infty} \frac{\widehat{d}(r;\widehat{\theta})\widehat{d}(k;\widehat{\theta})}{r+k+\gamma+1}.$$

It should be emphasized that the test statistics $S_n(\widehat{\theta})$, $R_{n,\gamma}(\widehat{\theta})$ and $M_{n,\gamma}(\widehat{\theta})$ are defined by means of infinite sums and, hence, these sums have to be truncated at some finite value, say M. We have noted that $M = 20$ yields sufficiently precise values of these test statistics.

Random variates from BT(θ) distribution were generated by following Proposition 9 and Remark 13 in Castellares *et al.* [8]. To estimate $\theta = (a, b)$, we considered the ML method. Finally, we adopted the warp-speed method [16] for evaluating the proposed resampling scheme to reduce the computational burden. On the basis of the warp-speed method, instead of computing critical points for each Monte Carlo sample, one resample is generated for each Monte Carlo sample and each test statistic, say T, is computed for that sample, obtaining say T^*. Then, the resampling critical values for T are computed from the empirical distribution determined by the resampling repetitions of T^*. It is worth mentioning that the idea behind the warp-speed bootstrap method is that taking just *one* bootstrap draw for each simulated sample is sufficient to provide a useful approximation to the statistic of interest. Applying this insight to Monte Carlo evaluation of bootstrap-based tests yields evaluation methods that work with $B = 1$ [16]. Because of the resulting dramatic computational savings, Giacomini *et al.* [16] called their method as "Warp-Speed" Monte Carlo method.

4.1. Size properties

First, the type I error of the gof tests based on the statistics $R_{n,\gamma}(\widehat{\theta})$, $M_{n,\gamma}(\widehat{\theta})$, H_n, W_n and $S_n(\widehat{\theta})$ are investigated. We consider the sample sizes $n = 50, 70, 90$ and 150. The nominal levels of the tests are $\alpha = 0.10$ and 0.05. We report the null rejection rates of $H_0 : X \sim \text{BT}(\theta)$ for all the tests at the 10% and 5% nominal significance levels; i.e. the percentage of times that the corresponding statistics exceed the 10% and 5% upper points obtained from the reference distribution generated by parametric bootstrap. These rates estimate the type I error probability of the tests. The null rejection rates of the gof tests $R_{n,\gamma}(\widehat{\theta})$ and $M_{n,\gamma}(\widehat{\theta})$ are listed in Table 1, while Table 2 lists the null rejection rates of the gof tests $S_n(\widehat{\theta})$, H_n and W_n.

For $\gamma = 0$ (i.e., the weight function $w(t)$ corresponds to the probability density function of the uniform distribution on the unit interval), note that the gof tests based on the statistics $R_{n,0}(\widehat{\theta})$ and $M_{n,0}(\widehat{\theta})$ have not a good performance, mainly for small sample sizes and when the parameter a is less than 1 ($a < 1$). On the other hand, the performance of these gof tests improves considerably as γ increases for $a < 1$. It is also evident that values of γ greater than 5 have no effect on improving the performance of the gof tests based on the statistics $R_{n,\gamma}(\widehat{\theta})$ and $M_{n,\gamma}(\widehat{\theta})$ in such a case; compare the null rejection rates of the tests for $\gamma = 5$ and $\gamma = 10$ when $a < 1$. Hence, for $a < 1$, the weight function $w(t) = t^\gamma$ with $\gamma = 5$ seems to be a good choice for the test statistics $R_{n,\gamma}(\widehat{\theta})$ and $M_{n,\gamma}(\widehat{\theta})$ in the BT discrete distribution. It is interesting to note that the gof tests that use $R_{n,0}(\widehat{\theta})$ and $M_{n,0}(\widehat{\theta})$ as test statistics present better results when $a > 1$. However, the performance of these gof tests deteriorates as γ increases and when $a > 1$, and so the probability density function of the uniform distribution on the unit interval as weight function in such a case seems to be a good choice for these test statistics. In short, the numerical results in Table 1 reveals the difficulty of selecting the best value of γ for the gof tests based on the test statistics $R_{n,\gamma}(\widehat{\theta})$ and $M_{n,\gamma}(\widehat{\theta})$.

From Table 2, note that the null rejection rates of the gof tests that use H_n and W_n as test statistics are close to the significance levels considered. It is worth stressing that the proposed gof test that uses $S_n(\widehat{\theta})$ as test statistic also presents a good performance, mainly for small sample sizes, when compared with the existing gof tests and, hence, can be an interesting alternative to these gof tests.

Table 1: Null rejection rates of the gof tests $R_{n,\gamma} := R_{n,\gamma}(\widehat{\theta})$ and $M_{n,\gamma} := M_{n,\gamma}(\widehat{\theta})$ for some weight functions $w(t)$.

α	n	$a = 0.8$ and $b = 0.6$									
		$R_{n,0}$	$R_{n,1}$	$R_{n,2}$	$R_{n,5}$	$R_{n,10}$	$M_{n,0}$	$M_{n,1}$	$M_{n,2}$	$M_{n,5}$	$M_{n,10}$
0.10	50	.066	.077	.080	.082	.083	.066	.072	.078	.080	.082
	70	.077	.087	.090	.092	.092	.081	.085	.088	.091	.092
	90	.085	.092	.095	.094	.093	.078	.089	.093	.093	.093
	150	.091	.097	.098	.098	.098	.090	.098	.098	.098	.097
0.05	50	.025	.031	.034	.038	.038	.025	.031	.034	.036	.037
	70	.035	.040	.042	.045	.045	.035	.038	.042	.044	.044
	90	.036	.039	.041	.042	.042	.035	.037	.040	.041	.041
	150	.037	.040	.042	.042	.043	.041	.041	.043	.043	.043

α	n	$a = 0.8$ and $b = 1.2$									
		$R_{n,0}$	$R_{n,1}$	$R_{n,2}$	$R_{n,5}$	$R_{n,10}$	$M_{n,0}$	$M_{n,1}$	$M_{n,2}$	$M_{n,5}$	$M_{n,10}$
0.10	50	.060	.063	.069	.078	.086	.078	.073	.074	.080	.087
	70	.066	.075	.086	.089	.092	.083	.078	.084	.089	.093
	90	.066	.075	.086	.095	.099	.085	.083	.088	.097	.101
	150	.073	.078	.085	.090	.096	.086	.084	.086	.094	.096
0.05	50	.025	.027	.030	.036	.039	.033	.030	.033	.036	.038
	70	.028	.034	.036	.042	.043	.038	.039	.040	.042	.043
	90	.025	.031	.036	.041	.044	.038	.035	.039	.042	.045
	150	.028	.035	.038	.040	.041	.040	.040	.042	.042	.041

α	n	$a = 1.4$ and $b = 1.8$									
		$R_{n,0}$	$R_{n,1}$	$R_{n,2}$	$R_{n,5}$	$R_{n,10}$	$M_{n,0}$	$M_{n,1}$	$M_{n,2}$	$M_{n,5}$	$M_{n,10}$
0.10	50	.098	.086	.085	.071	.074	.088	.082	.080	.082	.085
	70	.094	.091	.084	.073	.078	.089	.082	.078	.078	.081
	90	.096	.093	.087	.078	.081	.091	.084	.079	.078	.079
	150	.106	.100	.091	.080	.080	.098	.092	.088	.086	.090
0.05	50	.046	.041	.038	.035	.036	.042	.039	.036	.035	.035
	70	.047	.046	.041	.040	.041	.043	.039	.038	.034	.035
	90	.047	.045	.039	.035	.039	.041	.038	.037	.036	.037
	150	.050	.049	.044	.036	.038	.049	.042	.041	.041	.043

Table 2: Null rejection rates of the gof tests H_n, W_n and $S_n := S_n(\widehat{\theta})$.

α	n	$a = 0.8$ and $b = 1.2$			$a = 0.8$ and $b = 1.2$			$a = 0.8$ and $b = 1.2$		
		H_n	W_n	S_n	H_n	W_n	S_n	H_n	W_n	S_n
0.10	50	.103	.098	.079	.101	.107	.083	.095	.099	.091
	70	.095	.099	.084	.097	.099	.090	.101	.103	.086
	90	.095	.099	.082	.110	.110	.087	.105	.111	.085
	150	.101	.099	.089	.100	.098	.094	.098	.105	.090
0.05	50	.049	.045	.036	.051	.052	.040	.050	.050	.041
	70	.047	.049	.037	.050	.050	.042	.050	.052	.041
	90	.048	.045	.037	.054	.054	.040	.054	.057	.039
	150	.047	.045	.043	.049	.048	.042	.048	.052	.045

4.2. Power properties

Next, the power of the tests based on the statistics $R_{n,\gamma}(\widehat{\theta})$, $M_{n,\gamma}(\widehat{\theta})$, $S_n(\widehat{\theta})$, H_n and W_n are investigated. To compute the powers of the tests, we carried out Monte Carlo simulation experiments similar to that described above, however, the data were generated from perturbed BT distributions, and from the geometric (Geo), binomial (Bin), discrete Weibull (dWei) and negative binomial (NB) distributions. We consider two kinds of perturbations for the BT distribution.

Table 3: Nonnull rejection rates of $R_{n,w}(\widehat{\theta})$ and $M_{n,w}(\widehat{\theta})$ for some weight functions $w(t)$: power.

Alternative	$n=60$				$n=80$			
	$R_{n,0}(\widehat{\theta})$		$M_{n,0}(\widehat{\theta})$		$R_{n,0}(\widehat{\theta})$		$M_{n,0}(\widehat{\theta})$	
	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05
Alt1	0.248	0.189	0.310	0.220	0.252	0.180	0.302	0.227
Alt2	0.842	0.787	0.861	0.812	0.885	0.847	0.900	0.867
Alt3	0.276	0.162	0.446	0.303	0.321	0.199	0.506	0.390
Alt4	0.947	0.931	0.970	0.952	0.971	0.959	0.988	0.976
Alt5	0.287	0.125	0.496	0.325	0.384	0.225	0.597	0.469
Alt6	0.870	0.798	0.936	0.880	0.924	0.867	0.972	0.943
Geo	0.680	0.551	0.658	0.521	0.779	0.713	0.764	0.689
Bin	0.785	0.780	0.800	0.784	0.802	0.789	0.825	0.806
dWei	0.952	0.922	0.961	0.935	0.969	0.955	0.974	0.962
NB	0.395	0.257	0.398	0.257	0.484	0.379	0.480	0.377

Alternative	$R_{n,2}(\widehat{\theta})$		$M_{n,2}(\widehat{\theta})$		$R_{n,2}(\widehat{\theta})$		$M_{n,2}(\widehat{\theta})$	
	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05
Alt1	0.291	0.218	0.340	0.240	0.294	0.218	0.339	0.257
Alt2	0.878	0.810	0.900	0.835	0.927	0.879	0.945	0.903
Alt3	0.344	0.178	0.493	0.308	0.406	0.254	0.567	0.428
Alt4	0.945	0.926	0.964	0.939	0.969	0.955	0.982	0.969
Alt5	0.424	0.181	0.611	0.395	0.551	0.357	0.712	0.590
Alt6	0.863	0.770	0.915	0.842	0.918	0.859	0.950	0.921
Geo	0.719	0.581	0.727	0.590	0.811	0.741	0.818	0.750
Bin	0.788	0.784	0.794	0.785	0.802	0.793	0.814	0.798
dWei	0.998	0.998	0.065	0.955	0.974	0.966	0.988	0.988
NB	0.375	0.217	0.418	0.217	0.464	0.359	0.500	0.387

Alternative	$R_{n,5}(\widehat{\theta})$		$M_{n,5}(\widehat{\theta})$		$R_{n,5}(\widehat{\theta})$		$M_{n,5}(\widehat{\theta})$	
	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05
Alt1	0.297	0.209	0.332	0.226	0.300	0.214	0.337	0.243
Alt2	0.881	0.803	0.897	0.821	0.933	0.876	0.946	0.894
Alt3	0.394	0.195	0.517	0.293	0.478	0.290	0.592	0.427
Alt4	0.929	0.907	0.943	0.914	0.954	0.937	0.963	0.947
Alt5	0.541	0.252	0.678	0.436	0.674	0.477	0.779	0.647
Alt6	0.892	0.783	0.932	0.839	0.944	0.883	0.966	0.931
Geo	0.725	0.585	0.739	0.600	0.816	0.741	0.822	0.752
Bin	0.779	0.775	0.781	0.775	0.792	0.782	0.797	0.784
dWei	0.999	0.998	0.998	0.998	0.999	0.981	0.999	0.994
NB	0.386	0.228	0.429	0.238	0.475	0.350	0.491	0.338

Let $X_1 \sim \mathrm{BT}(\theta)$ and X_2 be another random variable taking values on $\mathbb{N}_0$, not having a BT distribution and independent of X_1. Then, the random variables $X_1 + X_2$ and $\max\{X_1, X_2\}$

also take values on $\mathbb{N}_0$, but the corresponding distributions of these perturbed random variables do not belong to the BT family of distributions and, hence, they can be used as alternatives. In the Monte Carlo simulations, we consider X_2 as a discrete uniform random variable taking values on $\{0, 1, ..., k\}$, for $k = 2, 4$ and 5, being denoted as $dU2$, $dU4$ and $dU5$, respectively. Thus, we have the following alternative distributions: Alt1 $= X_1 + dU2$, Alt2 $= \max\{X_1, dU2\}$, Alt3 $= X_1 + dU4$, Alt4 $= \max\{X_1, dU4\}$, Alt5 $= X_1 + dU5$ and Alt6 $= \max\{X_1, dU5\}$.

Here, we consider $w(t) = t^\gamma$ with $\gamma = 0, 2, 5$, $n = 60, 80$, and $a = 0.8$ and $b = 0.6$. The Monte Carlo simulation results regarding the power of the gof tests $R_{n,w}(\widehat{\theta})$ and $M_{n,w}(\widehat{\theta})$ are listed in Table 3, and Table 4 lists the power results of the gof tests $S_n(\widehat{\theta})$, H_n and W_n. From Table 3, note that there is no great difference in powers when different weight functions are considered. It is interesting to note that the test based on the proposed statistic $S_n(\widehat{\theta})$ is the most powerful among the gof tests in the great majority of the cases; compare Tables 3 and 4. However, it is evident that no gof test provides the highest power against all alternatives; that is, for some alternative distributions, the new gof test exhibits the highest power, but for other ones, the existing gof tests yield greater power. In summary, there is no uniform superiority of one gof test with respect to the others, as expected from the theoretical results in [20]. As expected, as the sample size increases, the power of the tests increases. In short, the numerical results of this section reveal that the proposed gof test on the basis of the new statistic $S_n(\widehat{\theta})$ can be an interesting alternative to the existing gof tests based on the test statistics $R_{n,w}(\widehat{\theta})$, $M_{n,w}(\widehat{\theta})$, H_n and W_n. The main advantage of the test statistic $S_n(\widehat{\theta})$ in relation to the test statistics $R_{n,w}(\widehat{\theta})$, $M_{n,w}(\widehat{\theta})$ is that it is not necessary to consider a weight function for its computation. On the other hand, we have to truncate an infinite sum in a finite value to calculate the new test statistic.

Table 4: Nonnull rejection rates of $S_n(\widehat{\theta})$, H_n and W_n: power.

n	Alternative	$S_n(\widehat{\theta})$		H_n		W_n	
		$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$
60	Alt1	0.720	0.654	0.540	0.406	0.380	0.267
	Alt2	0.982	0.973	0.985	0.967	0.948	0.871
	Alt3	0.944	0.929	0.564	0.408	0.422	0.238
	Alt4	0.999	0.999	0.992	0.998	0.981	0.932
	Alt5	0.999	0.998	0.678	0.494	0.515	0.286
	Alt6	0.999	0.999	0.999	0.983	0.939	0.880
	Geo	0.919	0.826	0.557	0.426	0.648	0.512
	Bin	0.830	0.736	0.794	0.735	0.790	0.767
	dWei	0.928	0.907	0.997	0.952	0.999	0.999
	NB	0.682	0.546	0.378	0.252	0.375	0.233
80	Alt1	0.783	0.648	0.571	0.462	0.372	0.306
	Alt2	0.999	0.992	0.999	0.994	0.981	0.962
	Alt3	0.969	0.919	0.648	0.475	0.478	0.352
	Alt4	0.999	0.999	0.999	0.995	0.999	0.986
	Alt5	0.999	0.992	0.758	0.610	0.606	0.445
	Alt6	0.999	0.999	0.999	0.999	0.996	0.941
	Geo	0.939	0.890	0.646	0.497	0.774	0.674
	Bin	0.875	0.856	0.867	0.802	0.822	0.799
	dWei	0.983	0.909	0.999	0.996	0.999	0.999
	NB	0.734	0.596	0.440	0.303	0.458	0.345

Finally, we compute the powers of the gof tests by considering moment estimators. Castellares *et al.* [8] have provided the following moment estimators for a and b: $\widetilde{a} = \frac{s^2}{\bar{X}} - 1$, $\widetilde{b} = \frac{\bar{X}\exp(1-s^2/\bar{X})}{s^2/\bar{X}-1}$, where $\bar{X}$ and s^2 are the sample mean and standard deviation. Castellares *et al.* [8] proved that $\widetilde{a}$ and $\widetilde{b}$ are consistent estimators for a and b, respectively. The power results when using these estimators are presented in Tables 5 and 6. Note that the powers of the gof tests under the moment estimates are near the powers under the ML estimates. However, the powers under the ML estimates are in general greater than the ones under the moment estimates.

Table 5: Nonnull rejection rates of $R_{n,w}(\widetilde{\theta})$ and $M_{n,w}(\widetilde{\theta})$ for some weight functions $w(t)$: power under moment estimators.

Alternative	$n = 60$				$n = 80$			
	$R_{n,0}(\widetilde{\theta})$		$M_{n,0}(\widetilde{\theta})$		$R_{n,0}(\widetilde{\theta})$		$M_{n,0}(\widetilde{\theta})$	
	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05
Alt1	0.242	0.183	0.304	0.214	0.246	0.174	0.296	0.221
Alt2	0.811	0.756	0.830	0.781	0.854	0.816	0.869	0.836
Alt3	0.257	0.143	0.427	0.284	0.302	0.180	0.487	0.371
Alt4	0.909	0.893	0.932	0.914	0.933	0.921	0.950	0.938
Alt5	0.266	0.104	0.475	0.304	0.363	0.204	0.576	0.448
Alt6	0.822	0.750	0.888	0.832	0.876	0.819	0.924	0.895
Geo	0.672	0.543	0.650	0.513	0.771	0.705	0.756	0.681
Bin	0.745	0.740	0.760	0.744	0.762	0.749	0.785	0.766
dWei	0.948	0.918	0.957	0.931	0.965	0.951	0.970	0.958
NB	0.392	0.254	0.395	0.254	0.481	0.376	0.477	0.374

Alternative	$R_{n,2}(\widetilde{\theta})$		$M_{n,2}(\widetilde{\theta})$		$R_{n,2}(\widetilde{\theta})$		$M_{n,2}(\widetilde{\theta})$	
	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05
Alt1	0.265	0.192	0.314	0.214	0.268	0.192	0.313	0.231
Alt2	0.831	0.763	0.853	0.788	0.880	0.832	0.898	0.856
Alt3	0.322	0.156	0.471	0.286	0.384	0.232	0.545	0.406
Alt4	0.912	0.893	0.931	0.906	0.936	0.922	0.949	0.936
Alt5	0.408	0.165	0.595	0.379	0.535	0.341	0.696	0.574
Alt6	0.859	0.766	0.911	0.838	0.914	0.855	0.946	0.917
Geo	0.693	0.555	0.701	0.564	0.785	0.715	0.792	0.724
Bin	0.743	0.739	0.749	0.740	0.757	0.748	0.769	0.753
dWei	0.978	0.961	0.983	0.970	0.984	0.975	0.988	0.982
NB	0.372	0.214	0.415	0.214	0.461	0.356	0.497	0.384

Alternative	$R_{n,5}(\widetilde{\theta})$		$M_{n,5}(\widetilde{\theta})$		$R_{n,5}(\widetilde{\theta})$		$M_{n,5}(\widetilde{\theta})$	
	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05
Alt1	0.287	0.199	0.322	0.216	0.290	0.204	0.327	0.233
Alt2	0.843	0.765	0.859	0.783	0.895	0.838	0.908	0.856
Alt3	0.366	0.167	0.489	0.265	0.450	0.262	0.564	0.399
Alt4	0.914	0.892	0.928	0.899	0.939	0.922	0.948	0.932
Alt5	0.505	0.216	0.642	0.400	0.638	0.441	0.743	0.611
Alt6	0.879	0.770	0.919	0.826	0.931	0.870	0.953	0.918
Geo	0.708	0.568	0.722	0.583	0.799	0.724	0.805	0.735
Bin	0.743	0.739	0.745	0.739	0.756	0.746	0.761	0.748
dWei	0.989	0.979	0.992	0.984	0.991	0.986	0.994	0.990
NB	0.382	0.224	0.425	0.234	0.471	0.346	0.487	0.334

Table 6: Nonnull rejection rates of $S_n(\widetilde{\theta})$, H_n and W_n: power under moment estimators.

n	Alternative	$S_n(\widetilde{\theta})$		H_n		W_n	
		$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$
60	Alt1	0.706	0.605	0.490	0.392	0.331	0.217
	Alt2	0.961	0.941	0.967	0.946	0.916	0.853
	Alt3	0.942	0.901	0.552	0.406	0.394	0.226
	Alt4	0.997	0.994	0.981	0.969	0.954	0.921
	Alt5	0.980	0.960	0.634	0.475	0.477	0.242
	Alt6	0.998	0.996	0.977	0.953	0.922	0.836
	Geo	0.898	0.822	0.551	0.405	0.644	0.506
	Bin	0.794	0.700	0.769	0.699	0.754	0.742
	dWei	0.917	0.862	0.956	0.941	0.990	0.982
	NB	0.676	0.540	0.372	0.246	0.369	0.227
80	Alt1	0.733	0.634	0.522	0.412	0.358	0.257
	Alt2	0.982	0.971	0.986	0.976	0.960	0.930
	Alt3	0.957	0.917	0.620	0.463	0.476	0.324
	Alt4	0.999	0.998	0.991	0.984	0.975	0.959
	Alt5	0.987	0.973	0.720	0.566	0.587	0.407
	Alt6	0.999	0.999	0.991	0.978	0.966	0.924
	Geo	0.933	0.869	0.642	0.491	0.753	0.670
	Bin	0.850	0.820	0.831	0.777	0.786	0.763
	dWei	0.942	0.898	0.967	0.955	0.992	0.988
	NB	0.730	0.592	0.436	0.299	0.454	0.341

5. REAL DATA ILLUSTRATIONS

In this section, we apply the gof tests based on the test statistics $R_{n,w}(\widehat{\theta})$, $M_{n,w}(\widehat{\theta})$, $S_n(\widehat{\theta})$, H_n and W_n in some real datasets for the sake of illustration. We consider the weight function $w(t) = t^\gamma$ with $\gamma = 5$ to compute the test statistics $R_{n,w}(\widehat{\theta})$ and $M_{n,w}(\widehat{\theta})$. All computations were done using the R language [43]. The code used in the real data applications can be obtained from the authors upon request. The datasets we consider correspond to the number of chromatid aberrations in 24 hours [9, 10], the number of absences of workers in a particular division of a large steel corporation in an observational period of six months [47], the number of claims of automobile liability policies [28, pp. 244], and the number of hemocytometer yeast cell on European red mites on apple leaves [44]. Descriptive measures for these datasets are listed in Table 7.

Table 7: Descriptive measures.

	Chromatid	Absence	Claims	Cell
n	400	318	298	80
Mean ($\bar{x}$)	0.55	0.67	1.71	1.15
Variance (s^2)	1.13	1.53	3.67	2.10
Skewness	3.12	2.19	1.72	1.27
Kurtosis	15.68	7.72	6.90	3.96
CV	1.94	1.85	1.12	1.26
ID	2.05	2.29	2.15	1.83

CV: Coefficient of variation ($= s/\bar{x}$);
ID: Index of dispersion ($= s^2/\bar{x}$).

The ML estimates of the BT distribution parameters, asymptotic standard errors (SE), and the 90% confidence intervals (CI) for the model parameters for each dataset are presented in Table 8. Table 9 lists the bootstrap p-values (with $B = 5000$) of the gof tests on the basis of the test statistics $R_{n,w}(\widehat{\theta})$, $M_{n,w}(\widehat{\theta})$, $S_n(\widehat{\theta})$, H_n and W_n for testing gof to the BT distribution. It can be noted that the five gof tests agree that the two-parameter BT discrete distribution is not adequate for fitting the chromatid dataset, once the bootstrap p-value for all tests are < 0.01. In addition, the five gof tests agree that the BT distribution is adequate for fitting the absence data, claims data, and cell data; that is, the five tests agree that the null hypothesis cannot be rejected at any usual significance levels.

Table 8: ML estimates.

Parameter	Chromatid aberrations		
	ML estimate	SE	90% CI
a	0.6453	0.1112	(0.4630; 0.8277)
b	0.4450	0.1201	(0.2480; 0.6420)
Parameter	Absence proneness		
	ML estimate	SE	90% CI
a	1.2320	0.1589	(0.9714; 1.4926)
b	0.1586	0.0427	(0.0886; 0.2286)
Parameter	Claims of automobile		
	ML estimate	SE	90% CI
a	0.9795	0.1342	(0.7594; 1.1995)
b	0.6548	0.1728	(0.3714; 0.9382)
Parameter	Yeast cell		
	ML estimate	SE	90% CI
a	0.9340	0.2684	(0.4938; 1.3741)
b	0.4839	0.2596	(0.0582; 0.9096)

Table 9: Bootstrap p-values; $B = 5000$.

Dataset	$R_{n,w}(\widehat{\theta})$	$M_{n,w}(\widehat{\theta})$	$S_n(\widehat{\theta})$	H_n	W_n
Chromatid aberrations	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Absence proneness	0.5220	0.5290	0.1632	0.5540	0.3915
Claims of automobile	0.4614	0.3822	0.3050	0.5935	0.6100
Yeast cell	0.6804	0.6716	0.8694	0.7825	0.6355

A referee reminds us that the dataset regarding the absences of workers [47] was originally fitted with the Negative Binomial (NB) distribution. From Table 9, it is evident that the BT distribution (i.e., the NTA distribution) is not rejected by any of the gof tests, and so an interesting question is: which distribution fits better this dataset, BT or NB? The pmf of the two-parameter NB distribution, specified in terms of its mean, μ say, is given by

$$\Pr(Y = y) = \left(\frac{\varphi}{\varphi+\mu}\right)^{\varphi}\left(\frac{\mu}{\varphi+\mu}\right)^{y}\frac{\Gamma(y+\varphi)}{\Gamma(\varphi)\Gamma(y+1)}, \quad y = 0, 1, 2, ...,$$

where $\Gamma(\cdot)$ is the gamma function, and $\mu > 0$ and $\varphi > 0$. It can be shown that the variance can be written as $\mu + \mu^2/\varphi$ and hence the parameter φ is referred to as the "dispersion parameter". The ML estimates of μ and φ are (asymptotic SE between parentheses): $\widehat{\mu} = 0.6698(0.0754)$ and $\widehat{\varphi} = 0.3951(0.0752)$. The maximized log-likelihood function for the NB distribution is -347.95, and so the AIC is given by 699.89. The maximized log-likelihood function for the BT distribution is given by -345.60, which results in an AIC value of 695.20. On the basis of the AIC values, it seems that the two-parameter BT distribution fits better the absences of workers' data than the two-parameter NB distribution and, hence, should be preferred.

Finally, it is well-known that the NTA distribution is traditionally fitted to datasets from ecology, entomology, etc. For example, McGuire *et al.* [34] studied the distribution of larval populations of the European corn borer, *Pyrausta nubilalis* (Hbni.). A total of $n = 3205$ corn plants growing in an area located in Northwest Iowa were dissected and, hence, the data correspond to the number of borers per plant dissected; see Table 1 in McGuire *et al.* [34, p. 74]. The ML estimates of the BT distribution parameters are (asymptotic SE between parentheses): $\widehat{a} = 0.2756(0.0325)$ and $\widehat{b} = 7.1346(1.0695)$. The bootstrap p-values (with $B = 5000$) of the gof tests on the basis of the test statistics $R_{n,w}(\widehat{\theta})$, $M_{n,w}(\widehat{\theta})$, $S_n(\widehat{\theta})$, H_n and W_n for testing gof to the BT distribution are given, respectively, by 0.082, 0.098, 0.005, 0.034 and 0.056. Note that the gof tests deliver small p-values, which indicates that the two-parameter BT discrete distribution (i.e., the NTA distribution) seems not adequate for fitting these data. In short, this empirical application illustrates that the NTA distribution, which is quite common in ecology and entomology, should be used with some caution in these areas, since for some cases, as evidenced by the gof tests, it cannot be adequate to fit such datasets. This indeed reveals the importance of gof tests to the BT distribution (i.e., the NTA distribution).

6. CONCLUSIONS

In this paper, a new gof test for the Neyman type A distribution was introduced, which is based on the interesting property that its pgf is the unique pgf satisfying a certain differential equation. The new gof test statistic is a function of the coefficients of the polynomial of the resulting equation when one replaces the pgf with the empirical pgf in the aforementioned differential equation. Also, other four related gof test statistics already introduced in the statistical literature were particularized for the two-parameter Bell–Touchard distribution for the first time, and studied by means of Monte Carlo simulations. We have that these five tests (the four already proposed and the new one) are consistent against fixed alternative hypotheses. Also, the practical computation of p-values of these tests requires a parametric bootstrap approximation to the null distribution of the corresponding test statistics. We consider Monte Carlo simulation experiments to verify the performance of the gof tests in finite samples. The Monte Carlo simulation results indicate that the null rejection rates of the five tests are, in general, close to the nominal levels. In addition, the numerical results regarding the power of the tests reveals that no test provides the highest power against all alternatives considered: for some alternatives the new test exhibits the highest power, but for other ones the competing tests yield greater power. In short, there is no uniform superiority of one test with respect to the others. Finally, it is worth emphasizing that the new test statistic $S_n(\widehat{\theta})$ has no need of choosing a weight function for its computation, unlike the test statistics $R_{n,w}(\widehat{\theta})$ and $M_{n,w}(\widehat{\theta})$, which can be a great advantage in practice. On the other hand, we have to truncate an infinite sum in a finite value to calculate the new test statistic.

A. APPENDIX: Proofs

Here we prove the results provided in the previous sections.

Proof of Proposition 3.1: It can be checked that the pgf of $X \sim \mathrm{BT}(\theta)$ given in (1.3) satisfies the differential equation given in (2.4). Obviously, this part of the proof can also be obtained by the result given by Meintanis [35] since the $\mathrm{BT}(\theta)$ distribution belongs to the compound Poisson family of distributions. Next, we proof that it is the only pgf in G satisfying such differential equation. It is well-known that the solution of the linear differential equation of order one of the form $y' + p(t)y = 0$, where $y = y(t)$, $y' = (\partial/\partial t)y(t)$ and $p(t)$ is a continuous function in t, is given by $y = C\exp(-\int p(t)dt)$, where C is an arbitrary constant. Since the differential equation (2.4) is of this form, we have that $g(t) = C\exp(\int ab\mathrm{e}^{at}dt) = C\exp(b\mathrm{e}^{at})$. Taking into account that g is a pgf, it must satisfy $g(1) = 1$, implying that $C = \exp(-b\mathrm{e}^{a})$ and, hence, the desired result is obtained. □

Let $\phi(x;\theta) = (\phi(x;0,\theta), \phi(x;1,\theta), ...)$, and $f_r(a,b) = b^r \sum_{u\geq 0}(u+r)\frac{a^u}{u!} = b^r(a+r)\mathrm{e}^a$. We have the following lemmas.

Lemma A.1. *Let $X_1, ..., X_n$ be independent and identically distributed from X, a random variable taking values in $\mathbb{N}_0$ with probability mass function $p(k) = \Pr(X = k)$, $k \in \mathbb{N}_0$, so that $\mathbb{E}(X^2) < \infty$. Then, $\mathbb{E}(\|\phi(X;\theta)\|_2^2) \leq \mathbb{E}(X^2) + b^2 f_0^2(a,b) < \infty$, $\forall\, \theta = (a,b) \in \Theta$.*

Proof: By definition,

$$\|\phi(X;\theta)\|_2^2 = \sum_{k\geq 0}(k+1)^2 I(X = k+1) + \sum_{k\geq 0}\sum_{u=0}^{k}\frac{b^2 a^{2u+2}}{(u!)^2} I(X = k-u),$$

and, thus, $\mathbb{E}(\|\phi(X;\theta)\|_2^2) = \mathbb{E}(X^2) + \sum_{k\geq 0}\sum_{u=0}^{k}\frac{b^2 a^{2u+2}}{(u!)^2}p(k-u)$. To show the finiteness of $\mathbb{E}(\|\phi(X;\theta)\|_2^2)$, we must prove that $\sum_{k\geq 0}\sum_{u=0}^{k}\frac{b^2 a^{2u+2}}{(u!)^2}p(k-u) < \infty$. The rest of the proof is parallel with the one in Lemma 1 of Batsidis *et al.* [3] and for this reason is omitted. □

Let $\frac{\partial}{\partial\theta_i}\widehat{d}(\cdot;\theta) = \left(\frac{\partial}{\partial\theta_i}\widehat{d}(0;\theta), \frac{\partial}{\partial\theta_i}\widehat{d}(1;\theta), ...\right)$, where $i = 1, 2$, and so $\theta_1 := a$ and $\theta_2 := b$.

Lemma A.2. *Let $X_1, ..., X_n$ be independent and identically distributed from X, a random variable taking values in $\mathbb{N}_0$. Then, $\forall\, \theta = (a,b) \in \Theta$, we have that:*

(I)
$$\left\|\frac{\partial}{\partial\theta_1}\widehat{d}(\cdot;\theta)\right\|_2^2 \leqslant b^2(a+1)^2\mathrm{e}^{2a} = f_1^2(a,b) < \infty,$$
$$\left\|\frac{\partial}{\partial\theta_2}\widehat{d}(\cdot;\theta)\right\|_2^2 \leqslant a^2\mathrm{e}^{2a} = f_0^2(a,b) < \infty;$$

(II)
$$\left\|E\left\{\frac{\partial}{\partial\theta_i}\widehat{d}(\cdot;\theta)\right\}\right\|_2^2 < \infty, \quad i = 1, 2.$$

Proof: (I) We have that

$$\frac{\partial}{\partial a}\widehat{d}(k;\theta) = -b\sum_{u=0}^{k}\frac{(u+1)a^u}{u!}\widehat{p}(k-u). \tag{A.1}$$

Therefore,

$$\begin{aligned}\left\|\frac{\partial}{\partial a}\widehat{d}(\cdot;\theta)\right\|_2^2 &= b^2\sum_{u,v\geqslant 0}\frac{(u+1)a^u}{u!}\frac{(v+1)a^v}{v!}\sum_{k\geqslant\max\{u,v\}}\widehat{p}(k-u)\widehat{p}(k-v)\\ &\leqslant \left(b(a+1)\mathrm{e}^a\right)^2 = f_1^2(a,b) < \infty,\end{aligned}$$

once $\sum_{k\geqslant\max\{u,v\}}\widehat{p}(k-u)\widehat{p}(k-v) \leqslant \sum_{k\geqslant 0}\widehat{p}(k) = 1$ and $\sum_{l\geqslant 0}(l+1)\frac{a^l}{l!} = (a+1)\mathrm{e}^a$. Furthermore, we have that

$$\frac{\partial}{\partial b}\widehat{d}(k;\theta) = -\sum_{u=0}^{k}\frac{a^{u+1}}{u!}\widehat{p}(k-u). \tag{A.2}$$

Therefore,

$$\begin{aligned}\left\|\frac{\partial}{\partial b}\widehat{d}(\cdot;\theta)\right\|_2^2 &= \sum_{u,v\geqslant 0}\frac{a^{u+1}}{u!}\frac{a^{v+1}}{v!}\sum_{k\geqslant\max\{u,v\}}\widehat{p}(k-u)\widehat{p}(k-v)\\ &\leqslant \left(a\mathrm{e}^a\right)^2 = f_0^2(a,b) < \infty.\end{aligned}$$

(II) The result follows from part (I) by replacing $\widehat{p}(k-u)$ and $\widehat{p}(k-v)$ with $p(k-u)$ and $p(k-v)$, respectively. □

Lemma A.3. *Let $X_1, ..., X_n$ be independent and identically distributed from X, a random variable taking values in $\mathbb{N}_0$. For each $k \in \mathbb{N}_0$, let $\theta_l = (a_l, b_l)$ so that $\theta_l = \gamma_l\theta + (1-\gamma_l)\widehat{\theta}$, for some $\gamma_l \in [0,1]$. Then,*

$$\sum_{k\geq 0}\left\{\frac{\partial}{\partial\theta_i}\widehat{d}(k;\theta) - \frac{\partial}{\partial\theta_i}\widehat{d}(k;\theta_l)\right\}^2 \overset{a.s.(P)}{\longrightarrow} 0, \quad i = 1, 2.$$

Proof: From relation (A.1), and after some algebra, we have that

$$\begin{aligned}\Delta_1 &= \sum_{k\geq 0}\left\{\frac{\partial}{\partial a}\widehat{d}(k;\theta) - \frac{\partial}{\partial a}\widehat{d}(k;\theta_l)\right\}^2\\ &= \sum_{u,v\geqslant 0}\frac{u+1}{u!}(b_la_l^u - ba^u)\frac{v+1}{v!}(b_la_l^v - ba^v)M_1(u,v),\end{aligned}$$

with $0 \leq M_1(u,v) = \sum_{k\geq\max\{u,v\}}\widehat{p}(k-u)\widehat{p}(k-v) \leq 1$. By applying the mean value theorem, we have that $b_la_l^u = ba^u + u\widetilde{b}_u\widetilde{a}_u^{u-1}(\widetilde{a}_u - a) + \widetilde{a}_u^u(\widetilde{b}_u - b)$, $\forall u \geq 1$, where $\widetilde{\theta}_u = (\widetilde{a}_u, \widetilde{b}_u)$ with $\widetilde{\theta}_u = \gamma_u\theta_l + (1-\gamma_u)\theta$, for some $\gamma_u \in (0,1)$. Therefore, $\widetilde{a}_u - a = \gamma_u(a_l - a)$ and $\widetilde{b}_u - b = \gamma_u(b_l - a)$. Taking into further consideration that $a_u \leq \max\{a_l, a\} \leq \max\{\widehat{a}, a\} := \widetilde{a}$, $b_u \leq \max\{b_l, b\} \leq \max\{\widehat{b}, b\} := \widetilde{b}$, we have that $|\, b_la_l^u - ba^u \,| \leq u\widetilde{b}\widetilde{a}^{u-1}\,|\, a_l - a\,| + \widetilde{a}^u\,|\, b_l - b\,| \leq u\widetilde{b}\widetilde{a}^{u-1}\,|\,\widehat{a} - a\,| + \widetilde{a}^u\,|\,\widehat{b} - b\,|$, $\forall u \geq 1$. Similarly, we have that $|\, b_la_l^v - ba^v \,| \leq v\widetilde{b}\widetilde{a}^{v-1}\,|\,\widehat{a} - a\,| + \widetilde{a}^v\,|\,\widehat{b} - b\,|$, $\forall v \geq 1$.

From the above considerations we have that $|\Delta_1| \leq (\widehat{a}-a)^2(\widetilde{b}(\widetilde{a}+2)\mathrm{e}^{\widetilde{a}})^2 + 2\,|\widehat{a}-a|\,|\widehat{b}-b| \cdot \widetilde{b}(\widetilde{a}+1)\mathrm{e}^{\widetilde{a}}(\widetilde{a}+2)\mathrm{e}^{\widetilde{a}} + (\widehat{b}-b)^2((\widetilde{a}+1)\mathrm{e}^{\widetilde{a}})^2$. Taking into account that in the right-hand side of the above expression all the functions are continuous functions of θ, it follows that $(\widehat{a}-a)^2(\widetilde{b}(\widetilde{a}+2)\mathrm{e}^{\widetilde{a}})^2 \overset{a.s.(P)}{\longrightarrow} (a-a)^2(b(a+2)\mathrm{e}^{a})^2 = 0$, $|\widehat{a}-a|\,|\widehat{b}-b|\,\widetilde{b}(\widetilde{a}+1)\mathrm{e}^{\widetilde{a}}(\widetilde{a}+2)\mathrm{e}^{\widetilde{a}} \overset{a.s.(P)}{\longrightarrow} |a-a|\,|b-b|\,b(a+1)\mathrm{e}^{a}(a+2)\mathrm{e}^{a} = 0$, $(\widehat{b}-b)^2((\widetilde{a}+1)\mathrm{e}^{\widetilde{a}})^2 \overset{a.s.(P)}{\longrightarrow} (b-b)^2((a+1)\mathrm{e}^{a})^2 = 0$. Thus, $\Delta_1 \overset{a.s.(P)}{\longrightarrow} 0$.

From relation (A.2), and after some algebra, we have that

$$\begin{aligned}\Delta_2 &= \sum_{k\geq 0}\left\{\frac{\partial}{\partial b}\widehat{d}(k;\theta) - \frac{\partial}{\partial b}\widehat{d}(k;\theta_l)\right\}^2 \\ &= \sum_{u,v\geqslant 0}\frac{1}{u!}(a_l^{u+1}-a^{u+1})\frac{1}{v!}(a_l^{v+1}-a^{v+1})M_1(u,v).\end{aligned}$$

By applying the mean value theorem as done when studying Δ_1 and following similar steps, we get $|\Delta_2| \leq (\widehat{a}-a)^2((\widetilde{a}+1)\mathrm{e}^{\widetilde{a}})^2$. Then, it follows that $(\widehat{a}-a)^2((\widetilde{a}+1)\mathrm{e}^{\widetilde{a}})^2 \overset{a.s.(P)}{\longrightarrow} (a-a)^2 \cdot ((a+1)\mathrm{e}^{a})^2 = 0$, and, hence, $\Delta_2 \overset{a.s.(P)}{\longrightarrow} 0$. □

Lemma A.4. *Let $X_1, ..., X_n$ be independent and identically distributed from X, a random variable taking values in $\mathbb{N}_0$. Assume that $\widehat{\theta} \overset{a.s.(P)}{\longrightarrow} \theta$, for some $\theta \in \Theta$. Given the data, let $X_1^*, ..., X_n^*$ be independent and identically distributed from $X^* \sim BT(\widehat{\theta})$. Let $\widehat{d}^*(k;\theta)$ be defined as $\widehat{d}(k;\theta)$ with $\widehat{p}(k)$ replaced with $\widehat{p}^*(k) = \frac{1}{n}\sum_{j=1}^{n} I(X_j^* = k)$, $k \geq 0$. Then, for $i = 1, 2$,*

$$\textbf{(I)} \quad \sum_{k\geq 0}\left[\frac{\partial}{\partial\theta_i}\widehat{d}^*(k;\widehat{\theta}) - \mu_i(k;\widehat{\theta})\right]^2 \overset{P_*}{\longrightarrow} 0, \quad \text{a.s.}(P),$$

$$\textbf{(II)} \quad \sum_{k\geq 0}\left[\mu_i(k;\theta) - \mu_i(k;\widehat{\theta})\right]^2 \to 0, \quad \text{a.s.}(P).$$

Proof: (I) We have that

$$\sum_{k\geq 0}\left[\frac{\partial}{\partial a}\widehat{d}^*(k;\widehat{\theta}) - \mu_1(k;\widehat{\theta})\right]^2 = \sum_{k\geq 0}\left\{-\widehat{b}\sum_{v=0}^{k}(v+1)\frac{a^v}{v!}\left[\widehat{p}^*(k-v) - p(k-v;\widehat{\theta})\right]\right\}^2$$

$$= \widehat{b}^2\sum_{u,v\geq 0}(u+1)\frac{\widehat{a}^u}{u!}(v+1)\frac{\widehat{a}^v}{v!}\sum_{k\geq\max\{u,v\}}\left\{\widehat{p}^*(k-v) - p(k-v;\widehat{\theta})\right\}\left\{\widehat{p}^*(k-u) - p(k-u;\widehat{\theta})\right\}$$

$$\leq \left[\widehat{b}(\widehat{a}+1)\mathrm{e}^{\widehat{a}}\right]^2\sum_{k\geq 0}\left\{\widehat{p}^*(k) - p(k;\widehat{\theta})\right\}^2.$$

Since $[\widehat{b}(\widehat{a}+1)\mathrm{e}^{\widehat{a}}]^2$ is a continuous function of $\widehat{\theta} = (\widehat{a}, \widehat{b})$, we have that $\widehat{b}(\widehat{a}+1)\mathrm{e}^{\widehat{a}}]^2 \overset{a.s.(P)}{\longrightarrow} [b(a+1)\mathrm{e}^{a}]^2 < \infty, \quad \forall\,\theta \in \Theta$. We also have that (see proof of Lemma 4 in Batsidis *et al.* [3]) $\sum_{k\geq 0}\{\widehat{p}^*(k) - p(k;\widehat{\theta})\}^2 \overset{P_*}{\longrightarrow} 0$, and it follows that

$$\sum_{k\geq 0}\left[\frac{\partial}{\partial\theta_1}\widehat{d}^*(k;\widehat{\theta}) - \mu_1(k;\widehat{\theta})\right]^2 \overset{P_*}{\longrightarrow} 0, \quad a.s.(P).$$

Also, we have that

$$\sum_{k\geq 0}\left[\frac{\partial}{\partial b}\widehat{d}^*(k;\widehat{\theta})-\mu_2(k;\widehat{\theta})\right]^2=\sum_{k\geq 0}\left\{\sum_{v=0}^{k}\frac{a^{v+1}}{v!}[\widehat{p}^*(k-v)-p(k-v;\widehat{\theta})]\right\}^2$$

$$=\sum_{u,v\geq 0}\frac{\widehat{a}^{u+1}}{u!}\frac{\widehat{a}^{v+1}}{v!}\sum_{k\geq\max\{u,v\}}\{\widehat{p}^*(k-v)-p(k-v;\widehat{\theta})\}\{\widehat{p}^*(k-u)-p(k-u;\widehat{\theta})\}$$

$$\leq\left(\widehat{a}\mathrm{e}^{\widehat{a}}\right)^2\sum_{k\geq 0}\{\widehat{p}^*(k)-p(k;\widehat{\theta})\}^2.$$

Using similar arguments as above, we have $(\widehat{a}\mathrm{e}^{\widehat{a}})^2\overset{a.s.(P)}{\longrightarrow}(a\mathrm{e}^a)^2<\infty,\quad\forall\,\theta\in\Theta$. Then, taking into account that $\sum_{k\geq 0}\{\widehat{p}^*(k)-p(k;\widehat{\theta})\}^2\overset{P_*}{\longrightarrow}0$, we obtain

$$\sum_{k\geq 0}\left[\frac{\partial}{\partial\theta_2}\widehat{d}^*(k;\widehat{\theta})-\mu_2(k;\widehat{\theta})\right]^2\overset{P_*}{\longrightarrow}0,\quad a.s.(P).$$

(II) We have that $\sum_{k\geq 0}[\mu_1(k;\theta)-\mu_1(k;\widehat{\theta})]^2=\Delta_{11}+2\Delta_{12}+\Delta_{13}$, where

$$\Delta_{11}=\sum_{k\geq 0}\sum_{u,v=0}^{k}(u+1)\frac{\widehat{b}\widehat{a}^u}{u!}(v+1)\frac{\widehat{b}\widehat{a}^v}{v!}\{p(k-u;\widehat{\theta})-p(k-u;\theta)\}\{p(k-v;\widehat{\theta})-p(k-v;\theta)\},$$

$$\Delta_{12}=\sum_{k\geq 0}\sum_{u,v=0}^{k}(u+1)\frac{\widehat{b}\widehat{a}^u}{u!}\frac{v+1}{v!}\{p(k-u;\widehat{\theta})-p(k-u;\theta)\}p(k-v;\theta)\{\widehat{b}\widehat{a}^v-ba^v\},$$

$$\Delta_{13}=\sum_{k\geq 0}\sum_{u,v=0}^{k}\frac{u+1}{u!}\frac{v+1}{v!}p(k-u;\theta)p(k-v;\theta)\{\widehat{b}\widehat{a}^u-ba^u\}\{\widehat{b}\widehat{a}^v-ba^v\}.$$

It follows that

$$\Delta_{11}\leq(\widehat{b}(\widehat{a}+1)\mathrm{e}^{\widehat{a}})^2\sum_{k\geq 0}\{p(k;\widehat{\theta})-p(k;\theta)\}^2.$$

Since $(\widehat{b}(\widehat{a}+1)\mathrm{e}^{\widehat{a}})^2\overset{a.s.(P)}{\longrightarrow}(b(a+1)\mathrm{e}^a)^2$, it suffices to show that

$$\sum_{k\geq 0}\{p(k;\widehat{\theta})-p(k;\theta)\}^2\overset{a.s.(P)}{\longrightarrow}0,$$

then, $\Delta_{11}\overset{a.s.(P)}{\longrightarrow}0$. Taking into account that

$$\sum_{k\geq 0}\{p(k;\widehat{\theta})-p(k;\theta)\}^2\leq\sum_{k\geq 0}k^2\{p(k;\widehat{\theta})-p(k;\theta)\}^2,$$

and that $\mathbb{E}_\theta(X^2)=(ba\mathrm{e}^a)^2+ba\mathrm{e}^a(1+a)$, $\forall\theta\in\Theta$, the rest of the proof is parallel with the proof of Lemma 4 II given in Jiménez-Gamero and Alba-Fernandez [21] and, hence, it is omitted.

We now deal with Δ_{12}. After some algebra and by applying the mean value theorem as in the proof of Lemma A.2, we have that $|\Delta_{12}|\leq\widehat{b}(\widehat{a}+1)\mathrm{e}^{\widehat{a}}\widetilde{b}(\widehat{a}-a)(\widetilde{a}+2)\mathrm{e}^{\widetilde{a}}+\widehat{b}(\widehat{a}+1)\mathrm{e}^{\widehat{a}}\cdot(\widehat{b}-b)(\widetilde{a}+1)\mathrm{e}^{\widetilde{a}}$. Thus, $\Delta_{12}\overset{a.s.(P)}{\longrightarrow}0$.

Related to $|\Delta_{13}|$, note that after some algebra and following similar arguments as above, we have that

$$|\Delta_{13}| \leq \sum_{u\geq 1} \frac{u+1}{u!}\{u\widetilde{b}\widetilde{a}^{u-1}|\widehat{a}-a| + \widetilde{a}^{u}|\widehat{b}-b|\} \times \sum_{v\geq 1} \frac{v+1}{v!}\{v\widetilde{b}\widetilde{a}^{v-1}|\widehat{a}-a| + \widetilde{a}^{v}|\widehat{b}-b|\},$$

or

$$\begin{aligned} |\Delta_{13}| \leq\ & (\widehat{a}-a)^2\widetilde{b}\sum_{u,v\geq 1}\frac{u+1}{u!}uv\widetilde{a}^{u-1}\widetilde{a}^{v-1} \\ & + (\widehat{b}-b)^2\sum_{u,v\geq 1}\frac{u+1}{u!}\widetilde{a}^{u}\widetilde{a}^{v} \\ & + 2|\widehat{a}-a||\widehat{b}-b|\widetilde{b}\sum_{u,v\geq 1}\frac{u+1}{u!}u\widetilde{a}^{u-1}\widetilde{a}^{v}. \end{aligned}$$

Also, we have that $\sum_{k\geq 0}[\mu_2(k;\theta)-\mu_2(k;\widehat{\theta})]^2 = \Delta_{21}+2\Delta_{22}+\Delta_{23}$, where

$$\Delta_{21} = \sum_{k\geq 0}\sum_{u,v=0}^{k}\frac{\widehat{a}^{u+1}}{u!}\frac{\widehat{a}^{v+1}}{v!}\{p(k-u;\widehat{\theta})-p(k-u;\theta)\}\{p(k-v;\widehat{\theta})-p(k-v;\theta)\},$$

$$\Delta_{22} = \sum_{k\geq 0}\sum_{u,v=0}^{k}\frac{\widehat{a}^{u+1}}{u!}\frac{1}{v!}\{p(k-u;\widehat{\theta})-p(k-u;\theta)\}p(k-v;\theta)\{\widehat{a}^{v+1}-a^{v+1}\},$$

$$\Delta_{23} = \sum_{k\geq 0}\sum_{u,v=0}^{k}\frac{1}{u!}\frac{1}{v!}p(k-u;\theta)p(k-v;\theta)\{\widehat{a}^{u+1}-a^{u+1}\}\{\widehat{a}^{v+1}-a^{v+1}\}.$$

Similarly, $\Delta_{21}\leq(\widehat{a}\mathrm{e}^{\widehat{a}})^2\sum_{k\geq 0}\{p(k;\widehat{\theta})-p(k;\theta)\}^2$. Since $(\widehat{a}\mathrm{e}^{\widehat{a}})^2 \xrightarrow{a.s.(P)} (a\mathrm{e}^{a})^2$ and $\sum_{k\geq 0}\{p(k;\widehat{\theta})-p(k;\theta)\}^2 \xrightarrow{a.s.(P)} 0$, we have that $\Delta_{21} \xrightarrow{a.s.(P)} 0$. Also,

$$\begin{aligned} |\,\Delta_{22}\,| &\leq |\,\widehat{a}-a\,|\sum_{u\geq 0}\frac{\widehat{a}^{u+1}}{u!}\sum_{v\geq 0}\frac{v+1}{v!}\widetilde{a}^{v} \\ &= |\,\widehat{a}-a\,|\,\widehat{a}\mathrm{e}^{\widehat{a}}(\widetilde{a}+1)\mathrm{e}^{\widetilde{a}}. \end{aligned}$$

Since $|\,\widehat{a}-a\,|\,\widehat{a}\mathrm{e}^{\widehat{a}}(\widetilde{a}+1)\mathrm{e}^{\widetilde{a}} \xrightarrow{a.s.(P)} 0$, it follows that $\Delta_{22} \xrightarrow{a.s.(P)} 0$.

Finally, it holds that

$$\begin{aligned} |\,\Delta_{23}\,| &\leq \sum_{u,v\geq 0}\frac{\widehat{a}^{u+1}-a^{u+1}}{u!}\frac{\widehat{a}^{v+1}-a^{v+1}}{u!} \\ &\leq (\widehat{a}-a)^2\Big(\widetilde{a}+1)\mathrm{e}^{\widetilde{a}}\Big)^2, \end{aligned}$$

and since $(\widehat{a}-a)^2\big(\widetilde{a}+1)\mathrm{e}^{\widetilde{a}}\big)^2 \xrightarrow{a.s.(P)} 0$, it follows that $\Delta_{23} \xrightarrow{a.s.(P)} 0$. □

Proof of Theorem 3.1: By applying the mean value theorem, we get, for each $k \in \mathbb{N}_0$, that

$$\widehat{d}(k;\widehat{\theta}) = \widehat{d}(k;\theta) + \left\{\frac{\partial}{\partial\theta}\widehat{d}(k;\theta)\right\}(\widehat{\theta}-\theta)^T + \left\{\frac{\partial}{\partial\theta}\widehat{d}(k;\theta_l) - \frac{\partial}{\partial\theta}\widehat{d}(k;\theta)\right\}(\widehat{\theta}-\theta)^T, \tag{A.3}$$

with $\theta_l = \gamma_l\theta + (1-\gamma_l)\widehat{\theta}$, for some $\gamma_l \in (0,1)$. From Lemma A.1, $\mathbb{E}(\|\phi(X;\theta)\|_2^2) < \infty$ and thus by the strong law of large number (SLLN) in Hilbert spaces and the continuous mapping theorem, it follows that

$$\|\widehat{d}(k;\theta)\|_2^2 \xrightarrow{a.s.} \|E\{\phi(X;\theta)\}\|_2^2 = \eta < \infty. \tag{A.4}$$

Finally, the result follows from (A.3), (A.4) and Lemmas A.2 and A.3. □

Proof of Theorem 3.2: From expansion (A.3), Assumption 1 and Lemmas A.2 and A.3, it follows that

$$\sqrt{n}\widehat{d}(\cdot;\widehat{\theta}) = \sqrt{n}\widehat{d}(\cdot;\theta) + \left\{\frac{\partial}{\partial\theta}\widehat{d}(\cdot;\theta)\right\}\sqrt{n}(\widehat{\theta}-\theta)^T + r_1, \tag{A.5}$$

with $\|r_1\|_2 = o_P(1)$. Now, by applying the SLLN in Hilbert spaces and Assumption 1, we get

$$\sqrt{n}\widehat{d}(\cdot;\theta) + \left\{\frac{\partial}{\partial\theta}\widehat{d}(\cdot;\theta)\right\}\sqrt{n}(\widehat{\theta}-\theta)^T = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y(X_i;\cdot,\theta) + r_2, \tag{A.6}$$

with $\|r_2\|_2 = o_P(1)$. By the central limit theorem in Hilbert spaces,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y(X_i;\cdot,\theta) \xrightarrow{\mathcal{L}} S(\theta), \tag{A.7}$$

where $Y(X;\cdot,\theta) = (Y(X;0,\theta), Y(X;1,\theta), ...)$. The result follows from (A.5)–(A.7) and the continuous mapping theorem. □

Proof of Theorem 3.3: Proceeding as in the proof of Theorem 3.2, we have that

$$\sqrt{n}\widehat{d}^*(\cdot;\widehat{\theta}^*) = \sqrt{n}\widehat{d}^*(\cdot;\theta) + \left\{\frac{\partial}{\partial\theta}\widehat{d}^*(\cdot;\widehat{\theta})\right\}\sqrt{n}(\widehat{\theta}^*-\widehat{\theta})^T + r_1^*,$$

with $\|r_1^*\|_2 = o_{P_*}(1)$ a.s.(P). Let $Y_n^* = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y(X_i^*;\cdot,\widehat{\theta})$. By applying Lemma A.4 and Assumption 2, we get

$$\sqrt{n}\widehat{d}^*(\cdot;\theta) + \left\{\frac{\partial}{\partial\theta}\widehat{d}^*(\cdot;\widehat{\theta})\right\}\sqrt{n}(\widehat{\theta}^*-\widehat{\theta})^T = Y_n^* + r_2^*,$$

with $\|r_2^*\|_2 = o_{P_*}(1)$ *a.s.*(P). To prove the result we derive the asymptotic distribution of Y_n^*, showing that it coincides with the asymptotic distribution of $S_n(\widehat{\theta})$ when the data come from $X \sim \text{BT}(\theta)$. With this aim, we apply Theorem 1.1 in Kundu *et al.* [31]. So, we will show that conditions (i)–(iii) in that theorem hold. This can be done in a similar way with the proof of Theorem 3 in Jiménez-Gamero and Alba-Fernandez [21]. □

B. APPENDIX: Function ℓ

Here, the form of the function ℓ, appeared in Assumption 1, associated with the ML estimators, and the moment estimators are provided. Moreover, it is proved that the conditions given in Assumption 1 really hold for the aforementioned estimators. For details about the existence of the ML estimators, and ways of computing them in practice, we refer to Section 4.2 in Castellares *et al.* [8].

In this context, when the ML estimators of the BT distribution are used, particularized for this special distribution the general relation given in the the proof of Theorem 3.2 in White [51] (see also Jiménez-Gamero and Kim [24]), the ℓ function is given by $\ell(x;\theta) = -A(\theta)^{-1}\nabla \log f(x;\theta)$, with

$$A(\theta) = -\begin{pmatrix} ba^{-1}(1+a)\mathrm{e}^a & \mathrm{e}^a \\ \mathrm{e}^a & K_{bb} \end{pmatrix},$$

where K_{bb} cannot be obtained in closed-form and is provided in Castellares *et al.* [8, p. 4846], and $\nabla \log f(x;\theta) = \left(-b\mathrm{e}^{-a} + \frac{x}{a}, (1-\mathrm{e}^a) + \frac{\partial}{\partial b}\log T_x(b)\right)^T$. Note that $-A(\theta) = K(\theta)$ is the unit (per observation) expected Fisher information matrix. Despite the fact that $K(\theta)$ cannot be obtained in closed-form, we have from Castellares *et al.* [8, p. 4846] that $K_{bb} \leq \mathrm{e}^a b^{-1}$ and $\det(K(\theta)) < \infty$. This implies that the inverse of this matrix exists. Furthermore, we have from Castellares *et al.* [8] that $\mathbb{E}_\theta(\frac{\partial}{\partial\theta_1}\log f(x;\theta)) = 0$ and $\mathbb{E}_\theta(\frac{\partial}{\partial\theta_2}\log f(x;\theta)) = 0$. Therefore, the relation $\mathbb{E}_\theta\{\ell(X_i;\theta)\} = 0$ is fulfilled when the ML estimator is used. Finally, we have that $J(\theta) = \mathbb{E}_\theta\{\ell(X_i;\theta)^T\ell(X_i;\theta)\} = tr((K(\theta))^{-1}K(\theta)^{-1}\Sigma_1) = tr(K(\theta)^{-1}) < \infty$, where $tr(A)$ denotes the trace of the matrix A, and $\Sigma_1 = \mathrm{Cov}_\theta(\nabla \log f(X;\theta)) = K(\theta)$.

Now, we consider the moment estimators of the BT distribution parameters to find the expression ℓ and to confirm that the conditions given in Assumption 1 are satisfied. Initially, note that from Remark 12 in Castellares *et al.* [8], we have after some algebra that $(a,b)^T = (g_1(\mu_1,\mu_2), g_2(\mu_1,\mu_2))^T$, where

$$g_1(\mu_1,\mu_2) = \frac{\mu_2 - (\mu_1)^2}{\mu_1} - 1, \quad g_2(\mu_1,\mu_2) = \frac{\mu_1 \exp(1 - \frac{\mu_2-\mu_1^2}{\mu_1})}{\frac{\mu_2-(\mu_1)^2}{\mu_1} - 1},$$

with $\mu_k = \mathbb{E}(X^k)$, given in Remark 12 by Castellares *et al.* [8]. Therefore, since $g = (g_1, g_2)^T$ is continuously differential at $(\mu_1,\mu_2)^T$ and $\mathbb{E}(||X||^4) < \infty$, we have that (see for instance Jiménez-Gamero and Kim [24]) $\ell(x;\theta) = (\ell_1(x;\theta), \ell_2(x;\theta))^T$, and

$$\ell_1(x;\theta) = \left(\frac{\partial}{\partial\mu_1}g_1(\mu_1,\mu_2), \frac{\partial}{\partial\mu_2}g_1(\mu_1,\mu_2)\right)(x-\mu_1, x^2-\mu_2)^T,$$

$$\ell_2(x;\theta) = \left(\frac{\partial}{\partial\mu_1}g_2(\mu_1,\mu_2), \frac{\partial}{\partial\mu_2}g_2(\mu_1,\mu_2)\right)(x-\mu_1, x^2-\mu_2)^T.$$

Obviously, $\mathbb{E}_\theta\{\ell(X_i;\theta)\} = 0$ since $\mathbb{E}_\theta(X-\mu_1) = \mathbb{E}_\theta(X^2-\mu_2) = 0$. Therefore, the condition $\mathbb{E}_\theta\{\ell(X_i;\theta)\} = 0$ is fulfilled when the moment estimator is used. In the sequel, let us denote by $K_1(\theta)$ the 2×2 matrix with (i,j) element $(i,j=1,2)$ equal to $\frac{\partial}{\partial\mu_j}g_i(\mu_1,\mu_2)$. The elements of the matrix $K_1(\theta)$, which depend only on μ_1 and μ_2, are omitted here, however, they

are available upon request and can be given in closed-form. Finally, we have that $J(\theta) = E_\theta\{\ell(X_i;\theta)^T\ell(X_i;\theta)\} = tr(K_1(\theta)^T K_1(\theta)\Sigma_2)$, where

$$\Sigma_2 = \mathrm{Cov}_\theta\big(X-\mu_1, X^2-\mu_2\big)^T = \begin{pmatrix} \mu_2-\mu_1^2 & \mu_3-\mu_1\mu_2 \\ \mu_3-\mu_1\mu_2 & \mu_4-\mu_2^2 \end{pmatrix}.$$

Therefore, $J(\theta) < \infty$ since $tr((K_1(\theta))^T K_1(\theta)\Sigma_2) < \infty$.

ACKNOWLEDGMENTS

Artur Lemonte acknowledges the financial support of the Brazilian agency Conselho Nacional de Desenvolvimento Científico e Tecnológico (grant 304776/2019–0). Apostolos Batsidis would like to thank M.D. Jiménez-Gamero, Bojana Milosevic and Jochen Einbeck for valuable discussion during the preparation of this paper. We are very grateful to two anonymous referees for the valuable comments and suggestions which have improved the first version of the paper.

REFERENCES

[1] Baringhaus, L.; Gürtler, N. and Henze, N. (2000). Theory and methods: weighted integral test statistics and components of smooth tests of fit, *Australian and New Zealand Journal of Statistics*, **42**, 179–192.

[2] Baringhaus, L. and Henze, N. (1992). A goodness of fit test for the Poisson distribution based on the empirical generating function, *Statistics and Probability Letters*, **13**, 269–274.

[3] Batsidis, A.; Jiménez-Gamero, M.D. and Lemonte, A.J. (2020). On goodness-of-fit tests for the Bell distribution, *Metrika*, **83**, 297–319.

[4] Bell, E.T. (1934a). Exponential polynomials, *Annals of Mathematical*, **35**, 258–277.

[5] Bell, E.T. (1934b). Exponential numbers, *The American Mathematical Monthly*, **41**, 411–419.

[6] Burke, M. (2000). Multivariate tests-of-fit and uniform confidence bands using a weighted bootstrap, *Statistics and Probability Letters*, **46**, 13–20.

[7] Castellares, F.; Ferrari, S.L.P. and Lemonte, A.J. (2018). On the Bell distribution and its associated regression model for count data, *Applied Mathematical Modelling*, **56**, 172–185.

[8] Castellares, F.; Lemonte, A.J. and Moreno-Arenas, G. (2020). On the two-parameter Bell–Touchard discrete distribution, *Communications in Statistics – Theory and Methods*, **4**, 4834–4852.

[9] Catcheside, D.G.; Lea, D.E. and Thoday, J.M. (1946a). Types of chromosome structural change induced by the irradiation of Tradescantia microspores, *Journal of Genetics*, **47**, 113–136.

[10] Catcheside, D.G.; Lea, D.E. and Thoday, J.M. (1946b). The production of chromosome structural changes in Tradescantia microspores in relation to dosage, intensity and temperature, *Journal of Genetics*, **47**, 137–149.

[11] Dehling, H. and Mikosch, T. (1994). Random quadratic forms and the bootstrap for U-statistics, *Journal of Multivariate Analysis*, **51**, 392–413.

[12] Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York.

[13] Epps, T.W. (1995). A test of fit for lattice distributions, *Communications in Statistics – Theory and Methods*, **24**, 1455–1479.

[14] Esnaola, M.; Puig, P.; Gonzalez, D.; Castelo, R. and Gonzalez, J.R. (2013). A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated RNA-seq experiments, *BMC Bioinformatics*, **14**, 254.

[15] Feller, W. (1943). On a general class of contagious distributions, *The Annals of Mathematical Statistics*, **14**, 389–400.

[16] Giacomini, R.; Politis, D.N. and White, H. (2013). A warp-speed method for conducting Monte Carlo experiments involving bootstrap estimators, *Econometric Theory*, **29**, 567–589.

[17] Gossiaux, A. and Lemaire, J. (1981). Methodes d'ajustement de distributions de sinistres, *Bulletin of the Association of Swiss Actuaries*, **81**, 87–95.

[18] Gürtler, N. and Henze, N. (2000). Recent and classical goodness-of-fit tests for the Poisson distribution, *Journal of Statistical Planning and Inference*, **90**, 207–225.

[19] Henze, N. (1996). Empirical-distribution-function goodness-of-fit tests for discrete models, *The Canadian Journal of Statistics*, **24**, 81–93.

[20] Janssen, A. (2000). Global power functions of goodness of fit tests, *Annals Statistics*, **28**, 239–253.

[21] Jiménez-Gamero, M.D. and Alba-Fernández, M.V. (2019). Testing for the Poisson–Tweedie distribution, *Mathematics and Computers in Simulation*, **164**, 146–162.

[22] Jiménez-Gamero, M.D. and Alba-Fernández, M.V. (2021). A test for the geometric distribution based on linear regression of order statistics, *Mathematics and Computers in Simulation*, **186**, 103–123.

[23] Jiménez-Gamero, M.D. and Batsidis, A. (2017). Minimum distance estimators for count data based on the probability generating function with applications, *Metrika*, **80**, 503–545.

[24] Jiménez-Gamero, M.D. and Kim, H.-M. (2015). Fast goodness-of-fit tests based on the characteristic function, *Computational Statistics and Data Analysis*, **89**, 172–191.

[25] Joe, H. and Zhu, R. (2005). Generalized Poisson distribution: the property of mixture of Poisson and comparison with Negative Binomial distribution, *Biometrical Journal*, **47**, 219–229.

[26] Johnson, N.L.; Kotz, S. and Kemp, A. (1992). *Univariate Discrete Distributions*, 2nd edition, Wiley, New York.

[27] Klar, B. (1999). Goodness-of-fit tests for discrete models based on the integrated distribution function, *Metrika*, **49**, 53–69.

[28] Klugman, S.; Panjer, H. and Willmot, G. (1998). *Loss Models. From Data to Decisions*, John Wiley and Sons, New York.

[29] Kocherlakota, S. and Kocherlakota, K. (1986). Goodness of fit test for discrete distributions, *Communications in Statistics – Theory and Methods*, **15**, 815–829.

[30] Kojadinovic, I. and Yan, J. (2012). Goodness-of-fit testing based on a weighted bootstrap: a fast large sample alternative to the parametric bootstrap, *The Canadian Journal of Statistics*, **40**, 480–500.

[31] Kundu, S.; Majumdar, S. and Mukherjee, K. (2000). Central limit theorems revisited, *Statistics and Probability Letters*, **47**, 265–275.

[32] Lord, D.; Washington S.P. and Ivan, J.N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory, *Accident Analysis and Prevention*, **37**, 35–46.

[33] Massé, J. and Theodorescu, R. (2005). Neyman type A distribution revisited, *Statistica Neerlandica*, **59**, 206–213.

[34] McGuire, J.U.; Brindley, T.A. and Bancroft, T.A. (1957). The distribution of European corn borer larvae Pyrausta nubilalis (Hbn.), in field corn, *Biometrics*, **13**, 65–78.

[35] Meintanis, S. (2008). New inference procedures for generalized Poisson distributions, *Journal of Applied Statistics*, **35**, 751–762.

[36] Meintanis, S. and Bassiakos, Y. (2005). Goodness-of-fit test for additively closed count models with an application to the generalized Hermite distribution, *Sankhya*, **67**, 538–552.

[37] Milosevic, B.; Jiménez-Gamero, M.D. and Alba-Fernandez, M.V. (2021). Quantifying the ratio-plot for the geometric distribution, *Journal of Statistical Computation and Simulation*, **91**, 2153–2177.

[38] Nakamura, M. and Perez-Abreu, V. (1993). Empirical probability generating function. An overview, *Insurance: Mathematics and Economics*, **12**, 287–295.

[39] Neyman, J. (1939). On a new class of contagious distributions applicable in entomologyand bacteriology, *Annals of Mathematical Statistics*, **10**, 35–57.

[40] Novoa-Muñoz, F. and Jiménez-Gamero, M.D. (2014). Testing for the bivariate Poisson distribution, *Metrika*, **77**, 771–793.

[41] Novoa-Muñoz, F. and Jiménez-Gamero, M.D. (2016). A goodness-of-fit test for the multivariate Poisson distribution, *Sort*, **40**, 1–26.

[42] Puig, P. and Valero, J. (2006). Count data distributions: some characterizations with applications, *Journal of the American Statistical Association*, **101:473**, 332–340.

[43] R Core Team (2020). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

[44] Rashid, A.; Ahmad, Z. and Jan, T.R. (2016). A new count data model with application in genetics and ecology, *Electronic Journal of Applied Statistical Analysis*, **9**, 213–226.

[45] Rueda, R. and O'Reilly, F. (1999). Tests of fit for discrete distributions based on the probability generating function, *Communications in Statistics – Simulation and Computation*, **28**, 259–274.

[46] Rueda, R.; Perez-Abreu, V. and O'Reilly, F. (1991). Goodness of fit for the Poisson distribution based on the probability generating function, *Communications in Statistics – Theory and Methods*, **20**, 3093–3110.

[47] Sichel, H.S. (1951). The estimation of the parameters of a Negative Binomial distribution with special reference to psychological data, *Psychometrika*, **16**, 107–127.

[48] Touchard, J. (1933). Propriétés arihtmétiques de certains nombres récurrents, *Annales de la Société Scientifique de Bruxelles*, **53**, 21–31.

[49] Tripathi, R.C. (2004). *Neyman type A, B, and C Distributions*. In "Encyclopedia of Statistical Sciences" (S. Kotz, C.B. Read, N. Balakrishnan, B. Vidakovic and N.L. Johnson, Eds.).

[50] Zafakali, N.S. and Ahmad, W.M.A.W. (2013). Modeling and handling overdispersion health science data with zero-inflated Poisson model, *Journal of Modern Applied Statistical Methods*, **12**, Article 28.

[51] White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica*, **50**, 1–25.

REVSTAT – Statistical Journal
Volume 21, Number 2, April 2023, 173–196
https://doi.org/10.57805/revstat.v21i2.405

The Extended Chen–Poisson Lifetime Distribution

Authors: Ivo Sousa-Ferreira ✉
– Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Portugal
– CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal
ivo.ferreira@staff.uma.pt

Ana Maria Abreu
– Departamento de Matemática, Faculdade de Ciências Exatas e da Engenharia, Universidade da Madeira, Portugal
– CIMA – Centro de Investigação em Matemática e Aplicações, Portugal
abreu@staff.uma.pt

Cristina Rocha
– Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Portugal
– CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal
cmrocha@fc.ul.pt

Received: March 2021 Revised: November 2021 Accepted: November 2021

Abstract:

- A three-parameter lifetime distribution is proposed, named extended Chen–Poisson distribution, by compounding the Chen and zero-truncated Poisson distributions. The new distribution belongs to the unified Poisson family, where both distributions of the minimum and maximum are merged into one. Several properties of the distribution are studied. The proposed distribution is quite flexible since it accommodates different complex hazard shapes. Inference is based on the maximum likelihood method in the presence of a right-censoring mechanism. A simulation study is performed to evaluate the properties of the parameters estimators. Two real lifetime data sets are analysed for purposes of comparison with other generalizations of the Chen distribution, as well as with other members of the unified Poisson family. The obtained results allow to highlight the potential of the new distribution.

Keywords:

- *Chen distribution; compounding Poisson; maximum likelihood estimation; survival analysis; unified Poisson family.*

AMS Subject Classification:

- 62E15, 62N01, 62N02, 65C99.

✉ Corresponding author.

1. INTRODUCTION

In recent years, several researchers have proposed many generalizations of classical distributions by adding further parameters. Generally, the aim behind such generalized distributions is to improve goodness-of-fit. For instance, the choice for modelling a monotonic hazard function (hf) usually falls on the exponential, Weibull, gamma or others generalized exponential distributions. However, for complex phenomena in survival and reliability studies, the hazard behaviour is almost certainly not monotonic. Therefore, in a situation of non-monotonic hf, such as bathtub-shaped or unimodal, the aforementioned distributions are unreasonable or even unrealistic. These limitations have naturally increased the interest in developing new extensions or generalizations of the more traditional distributions.

In the current literature, the methods for generating new distributions can be divided into two main approaches. The first one consists in the introduction of shape parameter(s) in the baseline distribution to explore tail properties. Some well-known techniques are: Lehman alternatives (also known as exponentiated), Marshall–Olkin, Kumaraswamy, transmuted, among others. The second approach concerns compounding a baseline continuous lifetime distribution with a discrete distribution, namely Poisson, geometric, negative-binomial or logarithmic. One of the reasons for developing compounding distributions is that the lifetime of a system constituted by Z (discrete random variable) components can be characterized by the distribution of the minimum or maximum of the lifetimes of its components (non-negative continuous random variables), depending on whether they form a series or a parallel system, respectively. A detailed and comprehensive survey of the existing methods are presented in Tahir and Cordeiro [30], which also proposed some new distributions.

An interesting two-parameter lifetime distribution that exhibits an increasing or a bathtub-shaped hf was proposed by Chen [11]. Some merits of this distribution are related with the exact confidence intervals and exact joint confidence region for the parameters. Over the years, several generalizations of this distribution have been developed. One of the first extensions, named XTG distribution, was introduced by Xie *et al.* [34] by adding the lacking scale parameter. Although the resulting model provided a better fit to the analysed data, the variety of shapes of the hf was not enriched. Other researchers have proposed models with an increased number of alternative hazard shapes. The family of distributions given by Lehman alternatives was considered by Chaubey and Zang [10] and Sarhan and Apaloo [28], who obtained the exponentiated Chen and exponentiated XTG distributions, respectively. Nadarajah *et al.* [23] derived general properties of the Kumaraswamy family of distributions and illustrated the new results obtaining the Kumaraswamy versions of the Chen and XTG distributions. The Marshall–Olkin technique was applied by Alawadhi *et al.* [2] in order to develop the Marshall–Olkin Chen distribution. The Chen-geometric and Marshall–Olkin Chen distributions can be seen as similar models with the same number of parameters, but the parameter space of the former model takes a more limited range of values. Cordeiro *et al.* [13] proposed a new family of lifetime distributions compounding a given class of generalized Weibull distributions with the geometric distribution. Since the Chen and XTG distributions were shown to be members of such class of models, these authors described the Chen-geometric and XTG-geometric distributions as particular cases. Another compounding distribution was proposed by Pappas *et al.* [24], who studied the Chen-logarithmic distribution and also extended the parameter space of the logarithmic distribution to $\mathbb{R}^+\backslash\{0\}$.

The transmuted Chen distribution has already been developed and was reported in Tahir and Cordeiro [30]. For other recent generalized versions of the Chen distribution, the reader is referred to [3, 7, 31].

In the light of the above context, the aim of this paper is to propose a new flexible generalization of the Chen distribution [11] by compounding it with the zero-truncated Poisson (ZTP) distribution. The remainder of the paper is organized as follows. In Section 2, a brief review on the unified Poisson family of distributions discussed by Ramos *et al.* [26] is presented. Section 3 begins with the definition of the new lifetime distribution, followed by the study of its properties, including the shapes of the probability density function (pdf) and hf in Subsection 3.1, as well as the quantiles, moment generating function and mean residual life function in Subsection 3.2. In Subsection 3.3, the maximum likelihood (ML) method is applied in the presence of a right-censoring mechanism and the estimators performance is evaluated by a simulation study in Subsection 3.4. In Subsections 3.5 and 3.6, the usefulness of the new distribution is illustrated in two real data applications with uncensored and censored observations. Some final remarks are presented in Section 4.

2. THE UNIFIED POISSON FAMILY OF DISTRIBUTIONS: A BRIEF REVIEW

The new distribution arises on competitive and complementary risks (CCR) scenarios, wherein it is only possible to observe the minimum/maximum lifetime among all causes instead of observing the lifetime associated with a particular cause [5]. In these settings, a difficulty emerges if the causes are latent in the sense that there is no information about which cause was responsible for the occurrence of the event. On many situations, it is impossible to specify the true cause, even by an expert, because it is somehow masked. For instance, in the biomedical sciences the interest is often to study the time until death, which can occur due to several competing causes such as respiratory infection, cardiac arrest, stroke, cancer, diabetes, among others. This triggers a competitive risks problem (time-to-event of a series system) due to the fact that it is only possible to observe the minimum lifetime among all causes. In an opposite example, suppose that the death of a patient with a given infection is due to multiple organ failures such as in lungs, kidneys and liver. This is now a complementary risks problem (time-to-event of a parallel system) since only the maximum lifetime among all causes is observed. As mentioned by Basu and Klein [6], since a complementary risks problem is the dual of a competitive risks problem, in general it is sufficient to establish the results in terms of the distribution of the minimum or the maximum, although there are some situations where the distribution of the maximum is simpler to handle analytically.

Recently, Ramos *et al.* [26] showed that both distributions of the minimum and the maximum can be unified in a simple form using a latent variable with ZTP distribution. Let $X_1, ..., X_Z$ be the times to event associated with each cause and Z a random variable with ZTP distribution, with probability mass function $P(Z = z; \phi) = \phi^z \big(z!(\mathrm{e}^{\phi} - 1)\big)^{-1}$, $z \in \mathbb{N}$, $\phi \in \mathbb{R}^+$. Assume that the random variables X's and Z are independent and that $X_1, ..., X_Z$ are independent and identically distributed according to a continuous lifetime distribution with a generic baseline cumulative distribution function (cdf) $F_0(x; \boldsymbol{\theta})$, indexed by the parameters vector $\boldsymbol{\theta}$.

Defining $Y = \min\{X_1, ..., X_Z\}$ in a competitive risks problem, the conditional cdf of Y given that $Z = z$ is

$$F(y|z;\boldsymbol{\theta}) = 1 - P(Y > y|Z = z;\boldsymbol{\theta}) = 1 - \left[1 - F_0(y;\boldsymbol{\theta})\right]^z, \qquad y > 0.$$

Then, the marginal cdf of Y is

$$F(y;\boldsymbol{\theta},\phi) = \sum_{z=1}^{\infty} \frac{\phi^z}{z!(\mathrm{e}^{\phi}-1)}\left(1 - \left[1 - F_0(y;\boldsymbol{\theta})\right]^z\right) = \frac{1 - \mathrm{e}^{-\phi F_0(y;\boldsymbol{\theta})}}{1 - \mathrm{e}^{-\phi}}, \quad \phi > 0. \tag{2.1}$$

On the other hand, defining $T = \max\{X_1, ..., X_Z\}$ in a complementary risks problem, the conditional cdf of T given that $Z = z$ is

$$F(t|z;\boldsymbol{\theta}) = P(T \leq t|Z = z;\boldsymbol{\theta}) = \left[F_0(t;\boldsymbol{\theta})\right]^z, \qquad t > 0.$$

Consequently, the marginal cdf of T is

$$F(t;\boldsymbol{\theta},\phi) = \sum_{z=1}^{\infty} \frac{\phi^z}{z!(\mathrm{e}^{\phi}-1)}\left[F_0(t;\boldsymbol{\theta})\right]^z = \frac{1 - \mathrm{e}^{\phi F_0(t;\boldsymbol{\theta})}}{1 - \mathrm{e}^{\phi}}, \qquad \phi > 0. \tag{2.2}$$

Thus, the distribution obtained from (2.2) belongs to the same family of distributions presented in (2.1) if it is assumed that ϕ takes negative values. So, when the latent variable has a ZTP distribution, the distributions of the minimum and the maximum can be merged into one, giving rise to the unified Poisson family of distributions.

Thereafter, assume that T has a distribution from the unified Poisson family, wherein the parameter space is extended to $\mathbb{R}\backslash\{0\}$. Since the cdf of T is still defined by (2.2), the parameter ϕ of this family of models has a particular interpretation in CCR problems. When $\phi < 0$ ($\phi > 0$), T represents the minimum (maximum) lifetime among all causes.

A large number of compounded ZTP distributions has already been proposed considering separately the minimum or maximum, as reviewed by Tahir and Cordeiro [30]. Following the unified approach, some of these distributions can be merged or even extended. For instance, Ramos *et al.* [26] considered the extended Weibull–Poisson (EWP) distribution [16, 19] (that was initially derived only by taking the minimum) and showed that the exponential-Poisson [18] and Poisson-exponential [9] distributions (that were derived by taking the minimum and maximum, respectively) can be unified into a single distribution, named extended exponential-Poisson (EEP) distribution.

3. A NEW LIFETIME DISTRIBUTION

Let X be a random variable following a Chen distribution [11] with cdf and hf given by

$$F_0(x;\lambda,\gamma) = 1 - \mathrm{e}^{\lambda\left(1-\mathrm{e}^{x^{\gamma}}\right)}, \qquad x > 0, \tag{3.1}$$

and

$$h_0(x;\lambda,\gamma) = \lambda\gamma x^{\gamma-1}\mathrm{e}^{x^{\gamma}}, \qquad x > 0, \tag{3.2}$$

respectively, where $\lambda, \gamma > 0$. Since $h_0'(x;\lambda,\gamma) = [\gamma(x^\gamma + 1) - 1]h_0(x;\lambda,\gamma)x^{-1}$, only the parameter γ affects the shape of the hf, which is: i) bathtub-shaped for $\gamma < 1$ (decreasing for $0 < x \leq (1/\gamma - 1)^{1/\gamma}$ and increasing for $x > (1/\gamma - 1)^{1/\gamma}$); and ii) monotonically increasing for $\gamma \geq 1$.

By substituting (3.1) in the unified Poisson family of distributions (2.2), a new generalization of the Chen distribution arises with cdf given by

$$F(t;\lambda,\gamma,\phi) = \frac{1 - \mathrm{e}^{\phi\left[1-\mathrm{e}^{\lambda\left(1-\mathrm{e}^{t^\gamma}\right)}\right]}}{1 - \mathrm{e}^{\phi}}, \qquad t > 0, \tag{3.3}$$

where $\lambda, \gamma > 0$ and $\phi \in \mathbb{R}\backslash\{0\}$ are the parameters of the distribution. The corresponding pdf is

$$f(t;\lambda,\gamma,\phi) = \frac{\lambda\gamma\phi t^{\gamma-1}}{1 - \mathrm{e}^{-\phi}} \mathrm{e}^{t^\gamma + \lambda\left(1-\mathrm{e}^{t^\gamma}\right) - \phi\mathrm{e}^{\lambda\left(1-\mathrm{e}^{t^\gamma}\right)}}, \qquad t > 0. \tag{3.4}$$

Hereafter, the distribution of T will be referred to as extended Chen–Poisson (ECP) distribution, which is a customary name for distributions belonging to the unified Poisson family. In fact, this distribution unifies both the minimum ($\phi < 0$) and the maximum ($\phi > 0$) distributions, which correspond to the Chen–Poisson and Poisson-Chen distributions, respectively.

The survival function (sf) and hf of the ECP distribution are defined, respectively, as follows

$$S(t;\lambda,\gamma,\phi) = \frac{1 - \mathrm{e}^{-\phi\mathrm{e}^{\lambda\left(1-\mathrm{e}^{t^\gamma}\right)}}}{1 - \mathrm{e}^{-\phi}}, \qquad t > 0,$$

and

$$h(t;\lambda,\gamma,\phi) = \frac{\lambda\gamma\phi t^{\gamma-1}\mathrm{e}^{t^\gamma + \lambda\left(1-\mathrm{e}^{t^\gamma}\right)}}{\mathrm{e}^{\phi\mathrm{e}^{\lambda\left(1-\mathrm{e}^{t^\gamma}\right)}} - 1}, \qquad t > 0. \tag{3.5}$$

3.1. Shapes of the probability density function and hazard function

The pdf (3.4) and hf (3.5) for some combinations of parameters values are depicted in Figures 1 and 2, respectively. It is challenging to study analytically the theoretical behaviour of these functions due to their complex expressions. In addition, the monotonicity study is hampered by the fact that all three parameters, λ, γ and ϕ, affect both the density and hazard shapes.

Based on the analytical analysis of the pdf, and as illustrated on the graphical representation in Figure 1, the density shape can be: (a) monotonic decreasing; (b)–(c) unimodal; or (d) decreasing-increasing-decreasing (DID). In what concerns the hazard shape, Figure 2 suggests that it can be: (a) monotonic increasing; (b) monotonic decreasing; (c) unimodal; (d) bathtub; (e) increasing-decreasing-increasing (IDI); or (f) decreasing-increasing-decreasing-increasing (DIDI). Accordingly, the ECP distribution is shown to be quite flexible. Nonetheless, some care is needed as the monotonicity study of the hf should not be solely based on graphical analysis. Since $\lim_{t\to\infty} h(t;\lambda,\gamma,\phi) = \infty$, for all $\lambda, \gamma > 0$ and $\phi \in \mathbb{R}\backslash\{0\}$, the hf is ultimately increasing, so a pure monotonic decreasing or unimodal shape is impossible.

However, it was verified that when γ takes values close to zero the hf takes a long time to increase. In such cases it is usual to admit that, from the practical point of view, the hf has a generally decreasing right tail.

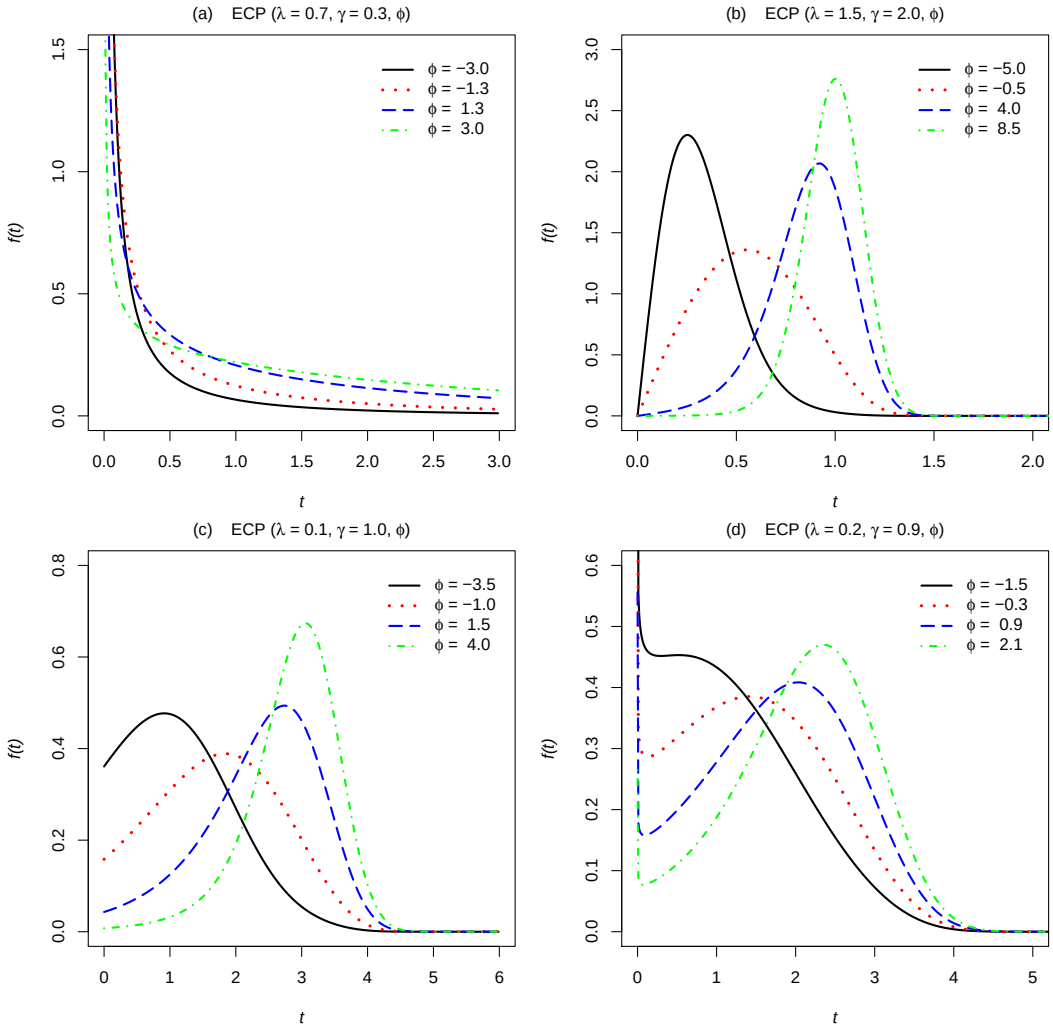

Figure 1: Probability density functions of the ECP distribution for different combinations of parameters values.

Proposition 3.1. *The Chen distribution is a limiting case of the ECP distribution, since when ϕ approaches 0 it follows that*

$$\lim_{\phi \to 0} h(t; \lambda, \gamma, \phi) = \lambda \gamma t^{\gamma - 1} \mathrm{e}^{t^{\gamma}},$$

which is the hf (3.2) of the Chen distribution.

Proposition 3.2. *The limiting behaviour of the pdf (3.4) and hf (3.5) of the ECP distribution is*

$$\textbf{(i)} \quad \lim_{t \to 0^+} f(t; \lambda, \gamma, \phi) = \lim_{t \to 0^+} h(t; \lambda, \gamma, \phi) = \begin{cases} \infty, & 0 < \gamma < 1, \\ \dfrac{\lambda \phi}{\mathrm{e}^{\phi} - 1}, & \gamma = 1, \\ 0, & \gamma > 1, \end{cases}$$

$\forall \lambda > 0$ *and* $\phi \in \mathbb{R} \backslash \{0\}$;

$$\textbf{(ii)} \quad \lim_{t \to \infty} f(t; \lambda, \gamma, \phi) = 0 \text{ and } \lim_{t \to \infty} h(t; \lambda, \gamma, \phi) = \infty, \ \forall \lambda, \gamma > 0 \text{ and } \phi \in \mathbb{R} \backslash \{0\}.$$

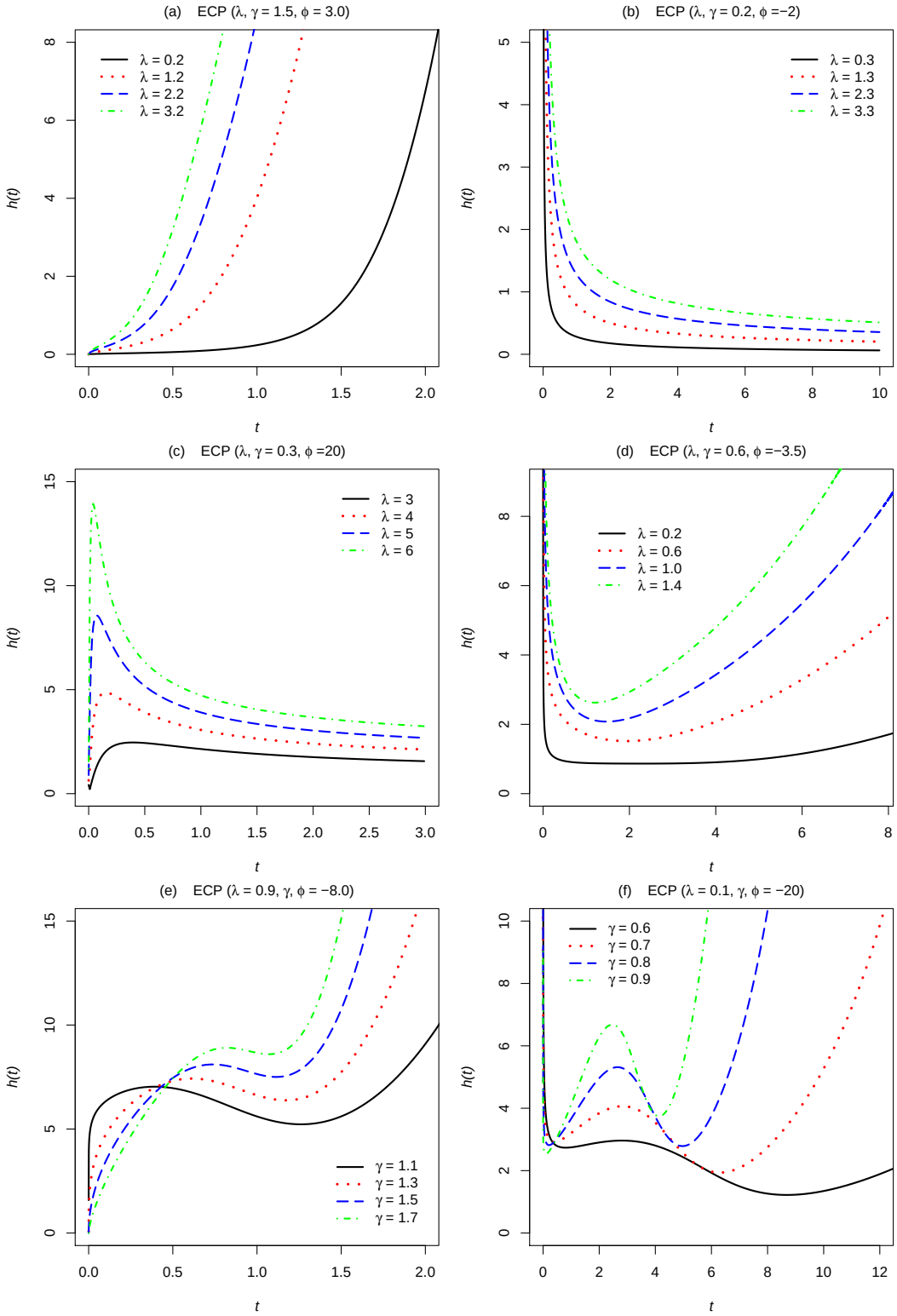

Figure 2: Hazard functions of the ECP distribution for different combinations of parameters values.

Proposition 3.3. *The theoretical behaviour of the pdf (3.4) of the ECP distribution may be characterized separately for the minimum ($\phi < 0$) and maximum ($\phi > 0$) distributions, as summarized in the following statements.*

(**i**) *Distribution of the minimum:*
- *For $\phi < 0$, $0 < \gamma \leq 1$ and $\lambda \geq (1-\phi)^{-1}$, the pdf is monotonically decreasing;*
- *For $\phi < 0$, $\gamma = 1$ and $0 < \lambda < (1-\phi)^{-1}$, the pdf is unimodal;*
- *For $\phi < 0$, $0 < \gamma < 1$ and $0 < \lambda < (1-\phi)^{-1}$, the pdf is monotonically decreasing or DID;*
- *For $\phi < 0$, $\gamma > 1$ and $\lambda > 0$, the pdf is unimodal;*

(**ii**) *Distribution of the maximum:*
- *For $0 < \phi \leq 1 - \lambda^{-1}$, $0 < \gamma \leq 1$ and $\lambda > 1$, the pdf is monotonically decreasing;*
- *For $\phi > 1 - \lambda^{-1}$, $\gamma = 1$ and $\lambda > 1$, the pdf is unimodal;*
- *For $\phi > 1 - \lambda^{-1}$, $0 < \gamma < 1$ and $\lambda > 1$, the pdf is monotonically decreasing or DID;*
- *For $\phi > 0$, $\gamma > 1$ and $\lambda > 1$, the pdf is unimodal;*
- *For $\phi > 0$, $\gamma \geq 1$ and $0 < \lambda \leq 1$, the pdf is unimodal;*
- *For $\phi > 0$, $0 < \gamma < 1$ and $0 < \lambda \leq 1$, the pdf is monotonically decreasing or DID.*

The proofs of Propositions 3.1 and 3.2 are straightforward and, therefore, are omitted. The proof of Proposition 3.3 is given in supplementary material file.

3.2. Quantiles, moments and mean residual life function

Some of the most important characteristics of a distribution, such as dispersion, skewness and kurtosis, can be studied through its quantiles and moments. By inverting the cdf (3.3), the quantile function of the ECP distribution is given by

$$Q(u;\lambda,\gamma,\phi) = \left\{ \log\left[1 - \lambda^{-1}\log\left(1 - \phi^{-1}\log\left((\mathrm{e}^{\phi}-1)u+1\right)\right)\right]\right\}^{1/\gamma}, \tag{3.6}$$

for $0 < u < 1$. This expression can be used for simulating pseudo-random values of $T \sim \mathrm{ECP}(\lambda,\gamma,\phi)$, considering that

$$T = \left\{ \log\left[1 - \lambda^{-1}\log\left(1 - \phi^{-1}\log\left((\mathrm{e}^{\phi}-1)U+1\right)\right)\right]\right\}^{1/\gamma}, \tag{3.7}$$

where U is a uniformly distributed random variable on $(0,1)$ interval.

The moment generating function of T can be defined as

$$M_T(w) = E(\mathrm{e}^{wT}) = \phi(1-\mathrm{e}^{-\phi})^{-1}\int_0^1 \exp\left\{w\left[\log\left(1-\lambda^{-1}\log(v)\right)\right]^{1/\gamma} - \phi v\right\}dv,$$

by making the change of variable $v = \mathrm{e}^{\lambda\left(1-\mathrm{e}^{t^\gamma}\right)}$. Then, the r-th raw moment of T is given by

$$E(T^r) = \phi(1-\mathrm{e}^{-\phi})^{-1}\int_0^1 \mathrm{e}^{-\phi v}\left[\log\left(1-\lambda^{-1}\log(v)\right)\right]^{r/\gamma} dv, \quad r = 1, 2, \ldots .$$

In particular, the mean and variance of ECP distribution are, respectively, given by

$$E(T) = \phi(1-\mathrm{e}^{-\phi})^{-1}\int_0^1 \mathrm{e}^{-\phi v}\left[\log\left(1-\lambda^{-1}\log(v)\right)\right]^{1/\gamma} dv,$$

and

$$\mathrm{Var}(T) = \phi(1-\mathrm{e}^{-\phi})^{-1}\int_0^1 \mathrm{e}^{-\phi v}\left[\log\left(1-\lambda^{-1}\log(v)\right)\right]^{2/\gamma} dv - [E(T)]^2.$$

The mean residual life function, as well as the hf, plays an important role in survival analysis for characterizing lifetime. While the latter represents the instantaneous event rate, the former summarizes the entire residual lifetime. The mean residual life function, $\mathrm{mrl}(t;\lambda,\gamma,\phi) = E(T-t|T \geq t)$, of the ECP distribution is given by

$$\mathrm{mrl}(t;\lambda,\gamma,\phi) = \phi(1-\mathrm{e}^{-\phi A})^{-1}\int_0^A \mathrm{e}^{-\phi v}\left[\log\left(1-\lambda^{-1}\log(v)\right)\right]^{1/\gamma} dv - t,$$

with $A = \mathrm{e}^{\lambda\left(1-\mathrm{e}^{t^\gamma}\right)}$.

The moments have no closed-form expressions and so they can only be obtained using numerical integration. Therefore, the classical measures of skewness and kurtosis based on moments are intractable. In this case, quantile-based measures are often considered, namely the Bowley skewness and Moors kurtosis that are given, respectively, by $B = [Q(3/4) - 2Q(1/2) + Q(1/4)]/[Q(3/4) - Q(1/4)]$ and $M = [Q(7/8) - Q(5/8) - Q(3/8) + Q(1/8)]/[Q(3/4) - Q(1/4)]$, where $Q(\cdot)$ comes from (3.6). These measures exist even for distributions without finite moments and are less sensitive to outliers.

3.3. Statistical inference

For statistical inference, the ML method is usually preferred due to the attractive properties of the resulting estimators, such as consistency, asymptotic efficiency, invariance property and asymptotic normality. Therefore, the ML method to estimate the three unknown parameters of the ECP distribution for the general case of right-censored time-to-event data is presented.

Let $\tilde{T}_i = \min\{T_i, C_i\}$, $i = 1, \ldots, n$, where T_i is the lifetime of i-th subject, following a ECP distribution, and C_i is the censoring time, assumed to have a distribution that does not depend on the parameters of T_i. Moreover, it is assumed that T_i and C_i are independent. So, the censoring mechanism is non-informative. The censoring indicator is defined as $\delta_i = I(T_i \leq C_i)$, taking the value 1 if T_i is a time-to-event and 0 if it is right-censored. Considering a random sample of n pairs, $(t_1, \delta_1), \ldots, (t_n, \delta_n)$, the log-likelihood function $\ell = \log L(\lambda, \gamma, \phi)$

is given by

$$\begin{aligned}\ell &= \sum_{i=1}^{n}\Big\{\delta_i \log f(t_i;\lambda,\gamma,\phi) + (1-\delta_i)\log S(t_i;\lambda,\gamma,\phi)\Big\}\\ &= n\log\left(\frac{\phi}{1-\mathrm{e}^{-\phi}}\right) + m\big(\lambda + \log(\lambda\gamma)\big) + (\gamma-1)\sum_{i=1}^{n}\delta_i\log(t_i) + \sum_{i=1}^{n}\delta_i t_i^{\gamma}\\ &\quad - \lambda\sum_{i=1}^{n}\delta_i \mathrm{e}^{t_i^{\gamma}} + \sum_{i=1}^{n}(1-\delta_i)\log\left(\frac{1-\mathrm{e}^{-\phi\mathrm{e}^{\lambda\left(1-\mathrm{e}^{t_i^{\gamma}}\right)}}}{\phi}\right) - \phi\sum_{i=1}^{n}\delta_i\mathrm{e}^{\lambda\left(1-\mathrm{e}^{t_i^{\gamma}}\right)},\end{aligned} \tag{3.8}$$

where $m=\sum_{i=1}^{n}\delta_i$ is the observed number of events. Some care must be taken when $\phi<0$, since the values of $\log(\phi)$ cannot be computed. This problem is easily overcome by considering the fact that $\log\big(\phi/(1-\mathrm{e}^{-\phi})\big)\in\mathbb{R}$, $\forall\phi\in\mathbb{R}\backslash\{0\}$, and $\log\big((1-\exp\{-\phi\mathrm{e}^{\lambda\left(1-\mathrm{e}^{t_i^{\gamma}}\right)}\})/\phi\big)\in\mathbb{R}$, $\forall\lambda,\gamma>0$ and $\phi\in\mathbb{R}\backslash\{0\}$.

The first-order partial derivatives of the log-likelihood function with respect to each of the three parameters are

$$\frac{\partial\ell}{\partial\lambda} = m\left(1+\frac{1}{\lambda}\right) - \sum_{i=1}^{n}\delta_i\mathrm{e}^{t_i^{\gamma}} - \sum_{i=1}^{n}(1-\delta_i)\frac{\phi\big(1-\mathrm{e}^{t_i^{\gamma}}\big)\mathrm{e}^{\lambda\left(1-\mathrm{e}^{t_i^{\gamma}}\right)}}{1-\mathrm{e}^{\phi\mathrm{e}^{\lambda\left(1-\mathrm{e}^{t_i^{\gamma}}\right)}}} - \phi\sum_{i=1}^{n}\delta_i\big(1-\mathrm{e}^{t_i^{\gamma}}\big)\mathrm{e}^{\lambda\left(1-\mathrm{e}^{t_i^{\gamma}}\right)},$$

$$\begin{aligned}\frac{\partial\ell}{\partial\gamma} &= \frac{m}{\gamma} + \sum_{i=1}^{n}\delta_i\log\big(t_i\big) + \sum_{i=1}^{n}\delta_i t_i^{\gamma}\log\big(t_i\big) - \lambda\sum_{i=1}^{n}\delta_i t_i^{\gamma}\log\big(t_i\big)\mathrm{e}^{t_i^{\gamma}}\\ &\quad + \lambda\phi\sum_{i=1}^{n}(1-\delta_i)\frac{t_i^{\gamma}\log\big(t_i\big)\mathrm{e}^{t_i^{\gamma}+\lambda\left(1-\mathrm{e}^{t_i^{\gamma}}\right)}}{1-\mathrm{e}^{\phi\mathrm{e}^{\lambda\left(1-\mathrm{e}^{t_i^{\gamma}}\right)}}} + \lambda\phi\sum_{i=1}^{n}\delta_i t_i^{\gamma}\log\big(t_i\big)\mathrm{e}^{t_i^{\gamma}+\lambda\left(1-\mathrm{e}^{t_i^{\gamma}}\right)},\end{aligned}$$

$$\frac{\partial\ell}{\partial\phi} = n\left(\frac{1}{\phi}+\frac{1}{1-\mathrm{e}^{\phi}}\right) - \frac{1}{\phi}\sum_{i=1}^{n}(1-\delta_i)\frac{1+\phi\mathrm{e}^{\lambda\left(1-\mathrm{e}^{t_i^{\gamma}}\right)} - \mathrm{e}^{\phi\mathrm{e}^{\lambda\left(1-\mathrm{e}^{t_i^{\gamma}}\right)}}}{1-\mathrm{e}^{\mathrm{e}^{\lambda\left(1-\mathrm{e}^{t_i^{\gamma}}\right)}}} - \sum_{i=1}^{n}\delta_i\mathrm{e}^{\lambda\left(1-\mathrm{e}^{t_i^{\gamma}}\right)}.$$

The ML estimates are determined by setting these partial derivatives equal to zero, obtaining a nonlinear system of equations that can only be solved using a numerical optimization method such as Newton–Raphson or Broyden–Fletcher–Goldfarb–Shanno (BFGS).

Under mild regularity conditions, the ML estimators of λ, γ and ϕ have an asymptotic multivariate normal distribution given by

$$(\hat{\lambda},\hat{\gamma},\hat{\phi}) \overset{a}{\sim} N\Big[(\lambda,\gamma,\phi),\ \boldsymbol{I}^{-1}(\lambda,\gamma,\phi)\Big], \quad \text{as} \quad n\longrightarrow\infty,$$

where the observed information matrix, $\boldsymbol{I}(\lambda,\gamma,\phi)$, is defined as

$$\boldsymbol{I}(\lambda,\gamma,\phi) = -\begin{bmatrix} \dfrac{\partial^2\ell}{\partial\lambda^2}, & \dfrac{\partial^2\ell}{\partial\lambda\partial\gamma}, & \dfrac{\partial^2\ell}{\partial\lambda\partial\phi} \\ \dfrac{\partial^2\ell}{\partial\gamma\partial\lambda}, & \dfrac{\partial^2\ell}{\partial\gamma^2}, & \dfrac{\partial^2\ell}{\partial\gamma\partial\phi} \\ \dfrac{\partial^2\ell}{\partial\phi\partial\lambda}, & \dfrac{\partial^2\ell}{\partial\phi\partial\gamma}, & \dfrac{\partial^2\ell}{\partial\phi^2} \end{bmatrix}.$$

The mathematical expressions of the elements of $\boldsymbol{I}(\lambda,\gamma,\phi)$ are given in supplementary material file.

For interval estimation and hypothesis testing, let $\widehat{\mathrm{Var}}(\hat{\lambda})$, $\widehat{\mathrm{Var}}(\hat{\gamma})$ and $\widehat{\mathrm{Var}}(\hat{\phi})$ denote the estimates of the main diagonal elements of the inverse of the observed information matrix, evaluated at the ML estimates of the parameters. The large-sample $(1-\alpha)100\%$ confidence intervals (CI) for λ, γ and ϕ are

$$\hat{\lambda} \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\lambda})}, \qquad \hat{\gamma} \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\gamma})} \qquad \text{and} \qquad \hat{\phi} \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\phi})},$$

respectively, where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

For computational implementation, the `optim` function available in `R` [25] statistical software (version 4.1.0) was used for direct maximization of the log-likelihood function (3.8).

3.4. Simulation study

In order to investigate the performance of ML estimators of the three parameters of the ECP distribution and to evaluate the accuracy of the resulting estimates, a simulation study was conducted through `R` [25] statistical software. In such simulation, the following steps were followed:

1. Specification of the parameters values $(\lambda, \gamma, \phi) =$ (0.2, 1.5, 3.0), (1.3, 0.2, -2.0), (3.0, 0.3, 20.0) and (0.6, 0.6, -3.5). These sets of parameters values were selected in order to yield increasing, decreasing, unimodal and bathtub shapes of the hazard function, respectively, as shown in Figure 2.
2. Specification of the sample size $n =$ 20, 50, 100, 500 and 1000.
3. Generation of a pseudo-random sample from (3.7), in the presence of random censoring (that has the types I and II of censoring mechanisms as special cases). Here, it is assumed that the event times follow an ECP distribution and the censoring times are uniformly distributed. The percentage of pseudo-random censoring is specified as 0%, 10% and 30%, following the procedures discussed in [27].
4. Computation of the ML estimates of the three parameters using the BFGS method and evaluation of the elements of the inverse of the observed information matrix at the ML estimates.
5. Repetition of the steps 1 to 4, $N = 1\,000$ times.
6. Calculation of the average of the N ML estimates and their standard errors.
7. Calculation of the bias, mean squared error (MSE) and coverage probability (CP) of the 95% CI for each parameter. The bias and MSE associated with the ML estimates of the parameter ϑ are, respectively, given by

$$\mathrm{Bias}_{\vartheta} = \frac{1}{N}\sum_{l=1}^{N}\left(\hat{\vartheta}_l - \vartheta\right) \qquad \text{and} \qquad \mathrm{MSE}_{\vartheta} = \frac{1}{N}\sum_{l=1}^{N}\left(\hat{\vartheta}_l - \vartheta\right)^2,$$

where $\hat{\vartheta}_l$ is the ML estimate obtained from the l-th sample, $l = 1, ..., N$, and $\boldsymbol{\vartheta} = (\lambda, \gamma, \phi)'$. The CP is the proportion of the N generated 95% CIs that include the real value of the parameter.

Table 1: The averages of the 1000 ML estimates for λ, γ and ϕ, their standard errors, bias, mean square errors (MSE) and coverage probabilities (CP) of the 95% CI. (Continues.)

Censoring	(λ, γ, ϕ)	n	Average			Standard error			Bias			MSE			CP		
			$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\phi}$
0%	(0.2, 1.5, 3.0)	20	0.268	1.580	6.593	0.240	0.394	7.865	0.068	0.080	3.593	0.093	0.156	588.379	80.7%	93.7%	99.1%
		50	0.215	1.551	3.260	0.140	0.278	2.727	0.015	0.051	0.260	0.021	0.071	6.965	88.5%	93.0%	98.9%
		100	0.204	1.538	3.021	0.102	0.209	1.831	0.004	0.038	0.021	0.010	0.040	2.814	90.1%	93.2%	98.0%
		500	0.196	1.515	2.940	0.045	0.087	0.721	−0.004	0.015	−0.060	0.002	0.007	0.487	96.2%	96.7%	97.9%
		1000	0.197	1.509	2.963	0.031	0.058	0.482	−0.003	0.009	−0.037	0.001	0.003	0.235	95.4%	96.7%	97.1%
	(1.3, 0.2, −2.0)	20	1.863	0.213	−1.591	1.311	0.057	3.013	0.563	0.013	0.409	1.461	0.002	1.685	99.6%	97.4%	100.0%
		50	1.507	0.203	−1.878	0.833	0.036	2.336	0.207	0.003	0.122	0.374	0.001	1.674	98.3%	97.8%	100.0%
		100	1.416	0.201	−1.988	0.666	0.026	1.953	0.116	0.001	0.012	0.229	0.000	1.502	97.2%	96.9%	99.8%
		500	1.337	0.199	−1.983	0.345	0.011	1.018	0.037	−0.001	0.017	0.079	0.000	0.701	95.6%	97.2%	97.9%
		1000	1.315	0.200	−2.000	0.252	0.008	0.725	0.015	0.000	0.000	0.036	0.000	0.287	97.0%	97.1%	97.7%
	(3.0, 0.3, 20.0)	20	3.213	0.432	58.535	0.823	0.215	29.647	0.213	0.132	38.535	0.872	0.130	12950.676	79.0%	79.6%	63.7%
		50	3.174	0.321	46.065	0.525	0.103	23.152	0.174	0.021	26.065	0.373	0.018	6438.042	82.9%	84.9%	74.1%
		100	3.150	0.304	35.482	0.388	0.068	18.428	0.150	0.004	15.482	0.216	0.006	2836.665	88.1%	89.9%	81.1%
		500	3.014	0.302	21.291	0.174	0.031	6.630	0.014	0.002	1.291	0.030	0.001	48.590	95.9%	95.8%	93.4%
		1000	3.005	0.302	20.479	0.121	0.022	4.310	0.005	0.002	0.479	0.014	0.000	18.554	95.4%	94.8%	92.9%
	(0.6, 0.6, −3.5)	20	0.987	0.634	−3.947	0.743	0.151	6.246	0.387	0.034	−0.447	0.526	0.020	34.523	93.1%	97.1%	99.7%
		50	0.882	0.604	−2.990	0.528	0.095	3.130	0.282	0.004	0.510	0.252	0.006	9.711	96.2%	97.0%	96.3%
		100	0.803	0.597	−3.101	0.431	0.066	2.539	0.203	−0.003	0.399	0.172	0.003	5.334	94.3%	96.4%	92.8%
		500	0.673	0.595	−3.400	0.233	0.028	1.450	0.073	−0.005	0.100	0.060	0.001	2.121	90.7%	96.5%	90.7%
		1000	0.644	0.597	−3.452	0.168	0.019	1.046	0.044	−0.003	0.048	0.034	0.000	1.255	91.1%	96.8%	91.4%

Table 1: (Continued.) The averages of the 1000 ML estimates for λ, γ and ϕ, their standard errors, bias, mean square errors (MSE) and coverage probabilities (CP) of the 95% CI. (Continues.)

Censoring	(λ, γ, ϕ)	n	Average			Standard error			Bias			MSE			CP		
			$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\phi}$
10%	(0.2, 1.5, 3.0)	20	0.282	1.575	7.146	0.274	0.421	10.866	0.082	0.075	4.146	0.110	0.168	502.776	82.6%	94.3%	99.4%
		50	0.219	1.556	3.340	0.153	0.302	2.990	0.019	0.056	0.340	0.025	0.080	8.137	88.3%	93.2%	99.3%
		100	0.200	1.549	2.965	0.103	0.210	1.826	0.000	0.049	−0.035	0.011	0.044	3.187	88.4%	92.3%	97.6%
		500	0.197	1.516	2.944	0.047	0.091	0.749	−0.003	0.016	−0.056	0.002	0.009	0.563	94.8%	95.9%	97.0%
		1000	0.197	1.511	2.947	0.033	0.062	0.508	−0.003	0.011	−0.053	0.001	0.004	0.255	95.7%	96.6%	97.3%
	(1.3, 0.2, −2.0)	20	1.810	0.214	−1.825	1.495	0.058	3.612	0.510	0.014	0.175	2.091	0.003	1.811	100.0%	95.7%	100.0%
		50	1.456	0.202	−2.055	1.042	0.036	2.929	0.156	0.002	−0.055	0.449	0.001	1.734	99.3%	98.0%	100.0%
		100	1.387	0.200	−2.058	0.845	0.027	2.428	0.087	0.000	−0.058	0.252	0.000	1.474	99.1%	97.1%	100.0%
		500	1.325	0.199	−2.051	0.472	0.012	1.357	0.025	−0.001	−0.051	0.111	0.000	0.889	96.6%	97.6%	99.1%
		1000	1.297	0.199	−2.061	0.387	0.008	1.080	−0.003	−0.001	−0.061	0.056	0.000	0.473	95.8%	97.8%	97.8%
	(3.0, 0.3, 20.0)	20	3.198	0.470	54.671	0.868	0.246	30.938	0.198	0.170	34.671	0.882	0.176	11453.983	79.0%	78.5%	62.5%
		50	3.224	0.324	51.708	0.534	0.110	24.354	0.224	0.024	31.708	0.431	0.026	8374.052	80.1%	81.5%	70.9%
		100	3.149	0.305	37.308	0.393	0.074	18.146	0.149	0.005	17.308	0.231	0.008	3257.225	86.7%	88.1%	78.5%
		500	3.029	0.301	21.980	0.182	0.034	7.398	0.029	0.001	1.980	0.036	0.001	71.255	96.1%	95.2%	92.8%
		1000	3.006	0.301	20.639	0.126	0.024	4.663	0.006	0.001	0.639	0.015	0.001	21.348	95.8%	95.5%	92.8%
	(0.6, 0.6, −3.5)	20	0.902	0.641	−4.820	0.853	0.157	8.715	0.302	0.041	−1.320	0.519	0.024	38.210	91.6%	96.0%	99.7%
		50	0.801	0.607	−3.962	0.596	0.096	4.970	0.201	0.007	−0.462	0.242	0.007	19.043	94.1%	96.4%	98.5%
		100	0.759	0.601	−3.582	0.470	0.068	3.349	0.159	0.001	−0.082	0.155	0.004	9.607	95.3%	96.7%	95.8%
		500	0.655	0.596	−3.586	0.282	0.031	1.849	0.055	−0.004	−0.086	0.067	0.001	2.374	90.9%	96.0%	91.0%
		1000	0.647	0.595	−3.517	0.230	0.022	1.431	0.047	−0.005	−0.017	0.049	0.000	1.808	89.4%	96.3%	89.7%

Table 1: (Continued.) The averages of the 1000 ML estimates for λ, γ and ϕ, their standard errors, bias, mean square errors (MSE) and coverage probabilities (CP) of the 95% CI.

Censoring	(λ, γ, ϕ)	n	Average			Standard error			Bias			MSE			CP		
			$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\phi}$
30%	(0.2, 1.5, 3.0)	20	0.309	1.594	11.608	0.307	0.465	13.178	0.109	0.094	8.608	0.173	0.223	2317.641	78.8%	92.2%	97.9%
		50	0.223	1.569	3.432	0.164	0.326	3.411	0.023	0.069	0.432	0.032	0.101	12.296	85.5%	93.6%	99.7%
		100	0.201	1.554	2.975	0.120	0.253	2.153	0.001	0.054	−0.025	0.013	0.053	3.814	89.4%	94.4%	98.8%
		500	0.196	1.520	2.922	0.055	0.111	0.873	−0.004	0.020	−0.078	0.003	0.012	0.700	94.7%	95.2%	96.2%
		1000	0.197	1.511	2.947	0.038	0.072	0.565	−0.003	0.011	−0.053	0.001	0.005	0.304	95.6%	96.1%	97.1%
	(1.3 0.2 −2.0)	20	1.771	0.215	−1.883	1.703	0.058	4.049	0.471	0.015	0.117	2.086	0.003	1.617	99.9%	95.7%	100.0%
		50	1.464	0.203	−2.015	1.249	0.037	3.321	0.164	0.003	−0.015	0.455	0.001	1.461	99.3%	97.5%	100.0%
		100	1.394	0.201	−2.045	1.041	0.028	2.871	0.094	0.001	−0.045	0.256	0.000	1.312	98.7%	97.2%	99.9%
		500	1.353	0.199	−1.979	0.592	0.012	1.595	0.053	−0.001	0.021	0.119	0.000	0.752	97.5%	97.4%	99.1%
		1000	1.314	0.199	−2.019	0.486	0.008	1.290	0.014	−0.001	−0.019	0.065	0.000	0.456	97.6%	97.6%	98.7%
	(3.0, 0.3, 20.0)	20	3.332	0.551	55.909	1.006	0.327	28.306	0.332	0.251	35.909	2.243	0.302	14102.353	80.7%	77.9%	57.8%
		50	3.266	0.342	65.858	0.562	0.141	24.938	0.266	0.042	45.858	0.504	0.049	15920.982	77.7%	77.2%	67.4%
		100	3.218	0.303	52.297	0.396	0.088	18.083	0.218	0.003	32.297	0.324	0.014	9011.748	79.6%	80.6%	71.9%
		500	3.046	0.299	23.555	0.200	0.042	9.380	0.046	−0.001	3.555	0.048	0.002	166.505	94.9%	94.9%	91.4%
		1000	3.016	0.300	21.256	0.139	0.030	5.855	0.016	0.000	1.256	0.018	0.001	37.586	96.6%	95.4%	94.1%
	(0.6 0.6 −3.5)	20	0.730	0.653	−6.067	1.076	0.170	13.444	0.130	0.053	−2.567	0.404	0.033	40.916	93.6%	95.8%	99.8%
		50	0.646	0.614	−5.323	0.857	0.105	9.228	0.046	0.014	−1.823	0.174	0.010	25.074	92.1%	96.6%	99.9%
		100	0.606	0.605	−5.033	0.703	0.076	7.275	0.006	0.005	−1.533	0.101	0.005	18.459	91.7%	96.4%	99.9%
		500	0.595	0.598	−4.091	0.496	0.039	3.805	−0.005	−0.002	−0.591	0.046	0.001	4.743	95.5%	96.6%	99.1%
		1000	0.588	0.595	−3.941	0.391	0.028	2.617	−0.012	−0.005	−0.441	0.037	0.001	2.665	92.6%	95.8%	99.3%

The results obtained from the simulation study are presented in Table 1. For samples generated with 0% of censoring, it is observed that the averages of the ML estimates of λ, γ and ϕ tend to the true value of the parameter as the sample size increases, as well as their standard errors tend to zero. Both the bias and MSE are smaller for larger sample sizes, reflecting that the ML estimators are asymptotically unbiased. Besides, the CP tends to be closer to the nominal level of 95%. However, it appears that ϕ has higher values for bias and MSE in comparison to the remaining parameters. This aspect is more visible for the set of parameters values corresponding to a unimodal hazard shape, but then it vanishes for large sample sizes and does not compromise the estimation of λ and γ.

In general, these results suggest that the estimation of parameters was performed consistently. Similar results were obtained for samples generated with 10% and 30%, despite the bias and MSE of all three parameters having slightly higher values. Although it is not shown here, the results were similar to the ones obtained for other choices of parameter values.

The programming codes of the simulation study, developed in `R`, are available in supplementary material file. Further research may be carried out to assess and explore other potential estimation procedures for the parameters of the ECP distribution, such as least-square estimators, minimum distance estimators, percentile based estimators, among others (see, for example, Dey *et al.* [14]).

3.5. Application to uncensored data: guinea pigs

In this section, the ECP distribution is applied to the (uncensored) guinea pigs data set reported by Bjerkedal [8]. The data represent the survival times, in days, of 72 guinea pigs infected with virulent tubercle bacilli. Dey *et al.* [15] analysed a transformed version of the original data (divided by 100), which is also considered in this work. Moreover, the adequacy of the ECP distribution is assessed in comparison with some other generalizations of the Chen distribution. Those models are listed in Table 2.

Table 2: List of distributions fitted to the guinea pigs data.

j-th Model, [ref.]	Probability density function, $f(t)$, $t > 0$
1 Chen, [11]	$\lambda_1\gamma_1 t^{\gamma_1-1}\mathrm{e}^{t^{\gamma_1}+\lambda_1(1-\mathrm{e}^{t^{\gamma_1}})}$, $\lambda_1, \gamma_1 > 0$
2 XTG, [34]	$\lambda_2\gamma_2(t/\phi_2)^{\gamma_2-1}\mathrm{e}^{(t/\phi_2)^{\gamma_2}+\lambda_2\phi_2(1-\mathrm{e}^{(t/\phi_2)^{\gamma_2}})}$, $\lambda_2, \gamma_2, \phi_2 > 0$
3 ECP	$\dfrac{\lambda_3\gamma_3\phi_3 t^{\gamma_3-1}}{1-\mathrm{e}^{-\phi_3}}\mathrm{e}^{t^{\gamma_3}+\lambda_3(1-\mathrm{e}^{t^{\gamma_3}})-\phi_3\mathrm{e}^{\lambda_3(1-\mathrm{e}^{t^{\gamma_3}})}}$, $\lambda_3, \gamma_3 > 0$, $\phi_3 \in \mathbb{R}\backslash\{0\}$
4 Chen-logarithmic, [24]	$\dfrac{\lambda_4\gamma_4(\phi_4-1)t^{\gamma_4-1}}{\left[1-(1-\phi_4)\mathrm{e}^{\lambda_4(1-\mathrm{e}^{t^{\gamma_4}})}\right]\log\phi_4}\mathrm{e}^{t^{\gamma_4}+\lambda_4(1-\mathrm{e}^{t^{\gamma_4}})}$, $\lambda_4, \gamma_4, \phi_4 > 0$
5 Exponentiated Chen, [10]	$\lambda_5\gamma_5\phi_5 t^{\gamma_5-1}\left[1-\mathrm{e}^{\lambda_5(1-\mathrm{e}^{t^{\gamma_5}})}\right]^{\phi_5-1}\mathrm{e}^{t^{\gamma_5}+\lambda_5(1-\mathrm{e}^{t^{\gamma_5}})}$, $\lambda_5, \gamma_5, \phi_5 > 0$
6 Marshall–Olkin Chen, [2]	$\dfrac{\lambda_6\gamma_6\phi_6 t^{\gamma_6-1}}{\left[1-(1-\phi_6)\mathrm{e}^{\lambda_6(1-\mathrm{e}^{t^{\gamma_6}})}\right]^2}\mathrm{e}^{t^{\gamma_6}+\lambda_6(1-\mathrm{e}^{t^{\gamma_6}})}$, $\lambda_6, \gamma_6, \phi_6 > 0$
7 Transmuted Chen, [30]	$\dfrac{\lambda_7\gamma_7 t^{\gamma_7-1}}{\left[1-\phi_7+2\phi_7\mathrm{e}^{\lambda_7(1-\mathrm{e}^{t^{\gamma_7}})}\right]^{-1}}\mathrm{e}^{t^{\gamma_7}+\lambda_7(1-\mathrm{e}^{t^{\gamma_7}})}$, $\lambda_7, \gamma_7 > 0$, $\phi_7 \in (-1,\ 1)$
8 Kumaraswamy Chen, [23]	$\dfrac{\lambda_8\gamma_8\phi_8\psi_8 t^{\gamma_8-1}(1-\mathrm{e}^{\lambda_8(1-\mathrm{e}^{t^{\gamma_8}})})^{\phi_8-1}}{\left[1-\left[1-\mathrm{e}^{\lambda_8(1-\mathrm{e}^{t^{\gamma_8}})}\right]^{\phi_8}\right]^{1-\psi_8}}\mathrm{e}^{t^{\gamma_8}+\lambda_8(1-\mathrm{e}^{t^{\gamma_8}})}$, $\lambda_8, \gamma_8, \phi_8, \psi_8 > 0$

The `AdequacyModel` [21] package was used for fitting models to the guinea pigs data. The ML estimates, their corresponding standard errors and −log-likelihood values of the fitted models are shown in Table 3. The `AdequacyModel` package also provides some useful statistics to assess the adequacy of the fitted models [22], such as the Cramér–von Mises (CM), Anderson–Darling (AD), Akaike information criterion (AIC), consistent Akaike information criterion (CAIC), Bayesian information criterion (BIC), Hannan–Quinn information criterion (HQIC) and in addition performs the Kolmogorov–Smirnov (KS) test. The obtained values are compiled in Table 4.

Table 3: ML estimates, standard errors and −log-likelihood values for the guinea pigs data.

Model	ML estimates				Standard error				$-\hat{\ell}$
	$\hat{\lambda}_j$	$\hat{\gamma}_j$	$\hat{\phi}_j$	$\hat{\psi}_j$	$\hat{\lambda}_j$	$\hat{\gamma}_j$	$\hat{\phi}_j$	$\hat{\psi}_j$	
Chen	0.208	0.759	—	—	0.034	0.043	—	—	104.241
XTG	0.391	0.322	0.010	—	0.165	0.023	0.005	—	100.839
ECP	1.225	0.407	12.094	—	0.256	0.061	5.158	—	93.537
Chen-logarithmic	0.208	0.758	1.008	—	0.131	0.094	1.395	—	104.241
Exponentiated Chen	0.995	0.444	7.209	—	0.306	0.080	4.095	—	94.186
Marshall–Olkin Chen	0.003	1.131	0.016	—	0.001	0.043	0.006	—	97.975
Transmuted Chen	0.117	0.809	0.753	—	0.025	0.045	0.203	—	102.617
Kumaraswamy Chen	0.896	0.339	9.229	2.364	0.391	0.324	11.159	6.413	94.108

Table 4: Goodness-of-fit statistics for the guinea pigs data.

Model	CM	AD	KS (p-value)	AIC	CAIC	BIC	HQIC
Chen	0.367	2.130	0.165 (0.040)	212.482	212.656	217.036	214.295
XTG	0.304	1.775	0.131 (0.172)	207.678	208.031	214.508	210.397
ECP	**0.085**	**0.514**	**0.082 (0.719)**	**193.075**	**193.428**	**199.905**	**195.794**
Chen-logarithmic	0.367	2.130	0.165 (0.040)	214.482	214.835	221.312	217.201
Exponentiated Chen	0.094	0.585	0.090 (0.601)	194.372	194.725	201.202	197.091
Marshall–Olkin Chen	0.199	1.153	0.137 (0.134)	201.652	202.005	208.482	204.371
Transmuted Chen	0.336	1.950	0.158 (0.055)	211.235	211.588	218.065	213.954
Kumaraswamy Chen	0.092	0.570	0.090 (0.610)	196.217	196.814	205.323	199.842

Bold values correspond to the best model.

The ECP distribution stands out as the best model among the fitted models, since its values of goodness-of-fit measures are the smaller ones and it has the highest p-value from the KS test. Interestingly, Dey *et al.* [15] showed that the alpha power transformed inverse Lindley (APTIL) distribution provides a better fit to the guinea pigs data, when compared to the fits of the inverse Lindley, generalized inverse Lindley, exponentiated generalized inverse Lindley, exponentiated inverse Lindley and inverse Weibull distributions. Nevertheless, the reported values of the AIC, BIC and KS statistic associated to the fit of the APTIL distribution are 234.817, 239.370 and 0.146, respectively, which are much higher than those obtained for the ECP distribution.

Additionally, the adequacy of the ECP distribution to model the guinea pigs data was informally evaluated through the two plots positioned on the upper panel of Figure 3, where plot (a) displays the empirical and model-based estimates of the sf; and plot (b) exhibits the histogram and model-based estimates of the pdf. In order to avoid a graphical overload, only the estimates of the Chen and ECP distributions are depicted. In both plots, the curves corresponding to the ECP distribution show close agreement, corroborating the fact that this distribution provides an adequate superior fitting to the survival times of guinea pigs with tuberculous infection.

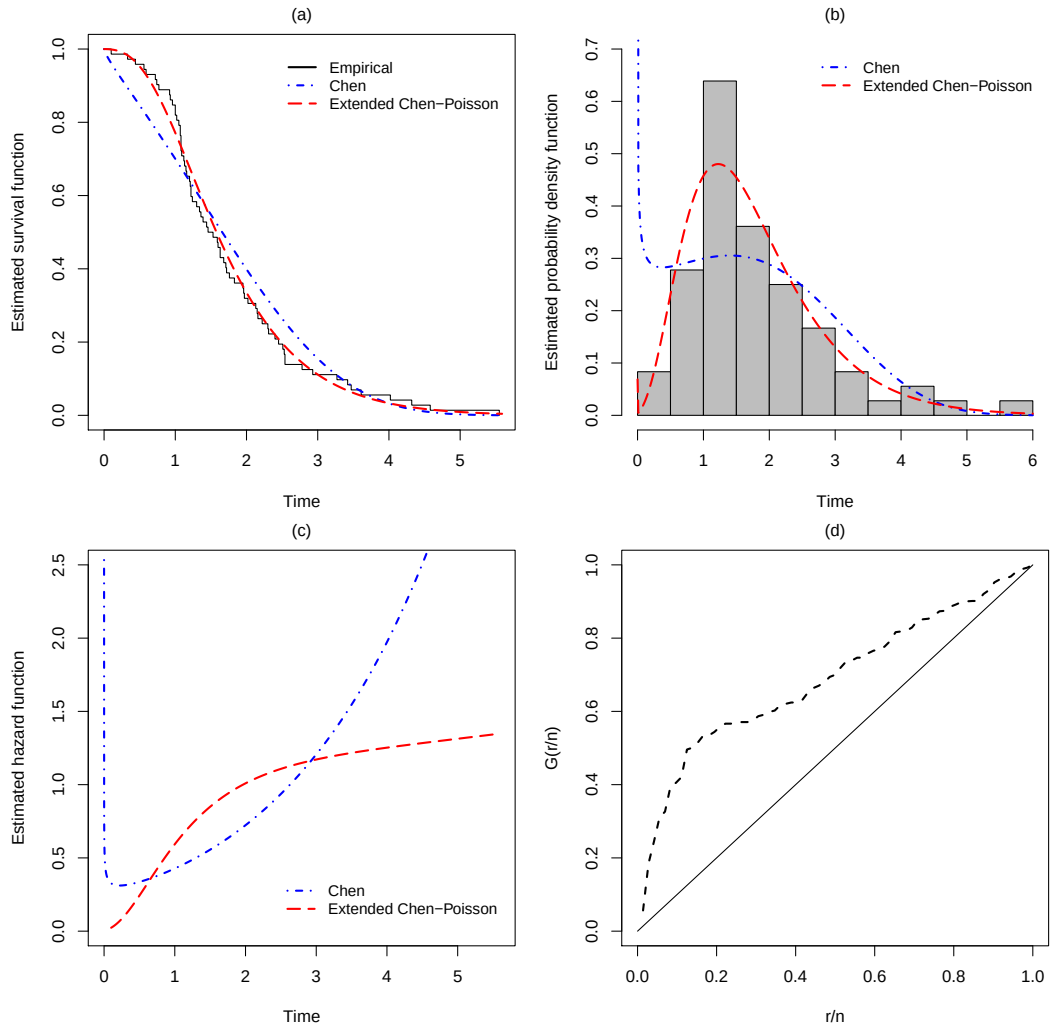

Figure 3: (a) Empirical and estimated survival functions of the Chen and ECP distributions; (b) Histogram and estimated probability density functions; (c) Estimated hazard functions; (d) Empirical scaled TTT-transform for the guinea pigs data.

The hf estimates of the referred distributions are shown in Figure 3 (c). With the purpose of identifying the hazard shape, a graphical method based on the total time on test (TTT) transform suggested by Aarset [1] was considered. The TTT plot is obtained by plotting the empirical scaled TTT-transform given by $G(r/n) = \left[\sum_{i=1}^{r} T_{i:n} + (n-r)T_{r:n}\right] / \left[\sum_{i=1}^{n} T_{i:n}\right]$ versus r/n, where $r = 1, ..., n$ and $T_{i:n}$ are the order statistics of the sample. It has been shown that the hf is increasing or decreasing if the TTT plot is concave or convex, respectively.

Although this is a sufficient but not a necessary condition, this graphical method is commonly used as a rough indicative of the hazard shape. Figure 3 (d) shows that the TTT plot is concave for the considered data, suggesting an increasing hf, which in theory would be properly accommodated by both distributions. However, the ML estimate of γ_1 of the Chen distribution is less than 1 (see Table 3), indicating that its hf is bathtub-shaped, as confirmed by Figure 3 (c). Hence, this distribution provides a poor fit. In fact, based on the p-value of the KS test (see Table 4), at significance level of 5%, there is evidence that the Chen distribution is not adequate for modelling this data. In contrast, the ECP distribution was able to capture an increasing hazard shape, reinforcing that it provides a good fit to the guinea pigs data.

Under the unified approach of Ramos *et al.* [26], it is possible to find whether the ECP distribution comes from the distribution of the minimum or maximum. Since the ML estimate of ϕ_3 is a positive value (see Table 3), the resulting distribution comes from the maximum of Chen distributions, that is, if T_i, $i = 1, ..., 72$, are the guinea pigs lifetimes, then $T_i = \max\{X_{i,1}, ..., X_{i,Z}\}$, where $X_{i,z}$, $z = 1, ..., Z$, follows a Chen distribution and Z is a non-observable random variable following a ZTP distribution.

3.6. Application to censored data: Rotterdam breast cancer

In this section, the ECP distribution is applied to the Rotterdam breast cancer data set reported by Sauerbrei *et al.* [29]. The data represent the relapse-free survival from 2982 patients with primary breast cancer whose records were included in the Rotterdam tumour bank. Here, the survival times (in years) since tumour removal until death from the disease is analysed. The maximum follow-up time is 19.283 years, the median (estimated by the reverse Kaplan–Meier method) is 9.273 years and the percentage of censoring is 57.3%. The Rotterdam data is also available in the `survival` [32] package.

The adequacy of the ECP distribution is assessed in comparison with some other members of the unified Poisson family [26], in particular with the EEP, EWP, generalized extended exponential-Poisson (GE2P) and extended exponenciated Weibull–Poisson (E2WP) distributions. Those models are listed in Table 5. Note that the E2WP distribution was proposed only by taking the maximum ($\phi_7 > 0$) [20], but we consider $\phi_7 \in \mathbb{R}\backslash\{0\}$ because it belongs to the unified Poisson family. Besides the Chen distribution being a limiting case of the ECP distribution (when $\phi_4 \to 0$), the Weibull distribution is a limiting case of the EWP (when $\phi_5 \to 0$) and E2WP (when $\psi_7 = 1$ and $\phi_7 \to 0$) distributions. For this reason, the Weibull distribution was also fitted to the Rotterdam data.

Given that in this application there are censored observations, the `maxLik` [33] package was conveniently used to maximize the log-likelihood function for censored data associated to each model, using the BFGS method. Table 6 compiles the ML estimates, their corresponding standard errors and $-$log-likelihood values. Here, it is verified that almost all fitted models come from the distribution of the maximum, except the GE2P distribution that comes from the distribution of the minimum ($\hat{\phi}_6 < 0$). Since the current application is not a CCR problem, the sign of $\hat{\phi}_j$ is not relevant. The observed values of the AIC, CAIC, BIC and HQIC statistics were also calculated in order to informally assess the adequacy of the fitted models,

as presented in Table 7. From these results it is seen that, although the ECP distribution has the smaller values of those criteria, the EWP distribution provides a similar fit. Thus, both ECP and EWP distributions are the best models among the fitted models to analyse the Rotterdam data.

Table 5: List of distributions fitted to the Rotterdam breast cancer data.

j-th Model, [ref.]	Probability density function, $f(t)$, $t > 0$
1 Chen, [11]	$\lambda_1\gamma_1 t^{\gamma_1-1}\mathrm{e}^{t^{\gamma_1}+\lambda_1(1-\mathrm{e}^{t^{\gamma_1}})}$, $\lambda_1, \gamma_1 > 0$
2 Weibull	$\lambda_2\gamma_2 t^{\gamma_2-1}\mathrm{e}^{-\lambda_2 t^{\gamma_2}}$, $\lambda_2, \gamma_2 > 0$
3 EEP, [18, 9]	$\dfrac{\lambda_3\phi_3}{1-\mathrm{e}^{-\phi_3}}\mathrm{e}^{-\lambda_3 t-\phi_3\mathrm{e}^{-\lambda_3 t}}$, $\lambda_3 > 0$, $\phi_3 \in \mathbb{R}\backslash\{0\}$
4 ECP	$\dfrac{\lambda_4\gamma_4\phi_4 t^{\gamma_4-1}}{1-\mathrm{e}^{-\phi_4}}\mathrm{e}^{t^{\gamma_4}+\lambda_4(1-\mathrm{e}^{t^{\gamma_4}})-\phi_4\mathrm{e}^{\lambda_4(1-\mathrm{e}^{t^{\gamma_4}})}}$, $\lambda_4, \gamma_4 > 0$, $\phi_4 \in \mathbb{R}\backslash\{0\}$
5 EWP, [16, 19, 26]	$\dfrac{\lambda_5\gamma_5\phi_5 t^{\gamma_5-1}}{1-\mathrm{e}^{-\phi_5}}\mathrm{e}^{-\lambda_5 t^{\gamma_5}-\phi_5\mathrm{e}^{-\lambda_5 t^{\gamma_5}}}$, $\lambda_5, \gamma_5 > 0$, $\phi_5 \in \mathbb{R}\backslash\{0\}$
6 GE2P, [4, 26]	$\dfrac{\lambda_6\gamma_6\phi_6}{1-\mathrm{e}^{-\phi_6}}\left(\dfrac{\mathrm{e}^{-\phi_6\mathrm{e}^{-\lambda_6 t}}-\mathrm{e}^{-\phi_6}}{1-\mathrm{e}^{-\phi_6}}\right)^{\gamma_6-1}\mathrm{e}^{-\lambda_6 t-\phi_6\mathrm{e}^{-\lambda_6 t}}$, $\lambda_6, \gamma_6 > 0, \phi_6 \in \mathbb{R}\backslash\{0\}$
7 E2WP, [20]	$\dfrac{\lambda_7^{\gamma_7}\gamma_7\phi_7\psi_7 t^{\gamma_7-1}\left[1-\mathrm{e}^{-(\lambda_7 t)^{\gamma_7}}\right]^{\psi_7-1}}{\left(\mathrm{e}^{\phi_7}-1\right)\mathrm{e}^{(\lambda_7 t)^{\gamma_7}-\phi_7\left[1-\mathrm{e}^{-(\lambda_7 t)^{\gamma_7}}\right]^{\psi_7}}}$, $\lambda_7, \gamma_7, \psi_7 > 0, \phi_7 \in \mathbb{R}\backslash\{0\}$

Table 6: ML estimates, standard errors and $-$log-likelihood values for the Rotterdam breast cancer data.

Model	ML estimates				Standard error				$-\hat{\ell}$
	$\hat{\lambda}_j$	$\hat{\gamma}_j$	$\hat{\phi}_j$	$\hat{\psi}_j$	$\hat{\lambda}_j$	$\hat{\gamma}_j$	$\hat{\phi}_j$	$\hat{\psi}_j$	
Chen	0.034	0.469	—	—	0.002	0.007	—	—	4913.724
Weibull	0.035	1.254	—	—	0.003	0.031	—	—	4817.114
EEP	0.101	—	1.479	—	0.006	—	0.194	—	4839.735
ECP	1.792	0.108	83.000	—	0.017	0.002	1.407	—	4780.796
EWP	0.227	2.330	39.353	—	0.005	0.035	1.047	—	4780.882
GE2P	1.609	0.046	−2.233	—	0.065	0.015	0.993	—	4797.261
E2WP	14.908	0.257	0.378	25.107	0.903	0.013	0.840	1.332	4780.297

Table 7: Goodness-of-fit statistics for the Rotterdam breast cancer data.

Model	AIC	CAIC	BIC	HQIC
Chen	9831.448	9831.452	9843.449	9835.766
Weibull	9638.228	9638.232	9650.229	9642.546
EEP	9683.471	9683.475	9695.472	9687.789
ECP	**9567.591**	**9567.599**	**9585.592**	**9574.068**
EWP	**9567.765**	**9567.773**	**9585.766**	**9574.242**
GE2P	9600.522	9600.530	9618.523	9606.999
E2WP	9569.527	9569.540	9593.528	9578.163

Bold values correspond to the best models.

In addition, the overall goodness-of-fit of the ECP distribution was informally evaluated through the two plots positioned on the upper panel of Figure 4, where plot (a) displays the estimates of the sf based on the Kaplan–Meier estimator and on the Chen and ECP distributions; and plot (b) exhibits the Cox–Snell residuals of the ECP distribution. The residuals are defined as $\hat{r}_i = \widehat{H}(t_i; \hat{\lambda}, \hat{\gamma}, \hat{\phi})$, $i = 1, ..., n$, where $\widehat{H}(t_i; \hat{\lambda}, \hat{\gamma}, \hat{\phi})$ is the estimated cumulative hazard function (chf) of the fitted model. When the model is adequate, the residuals behave approximately as a sample from a population with unit exponential distribution [12]. This assumption is informally checked through the graphical representation of $\big(\hat{r}_i, \widehat{H}_{NA}(\hat{r}_i)\big)$, where $\widehat{H}_{NA}(\hat{r}_i)$ is the Nelson-Aalen estimate of the chf of the residuals. There is a good fit when this representation yields a straight line through the origin with slope 1. In both (a) and (b) plots, the curves corresponding to the ECP distribution show general agreement, even though there are a few poorly fitted observations on the upper tail. This is acceptable since the 90-th quantile of the follow-up time (estimated by the reverse Kaplan–Meier method) is equal to 13.227 years, from which the model begins to provide a poor fit to the data.

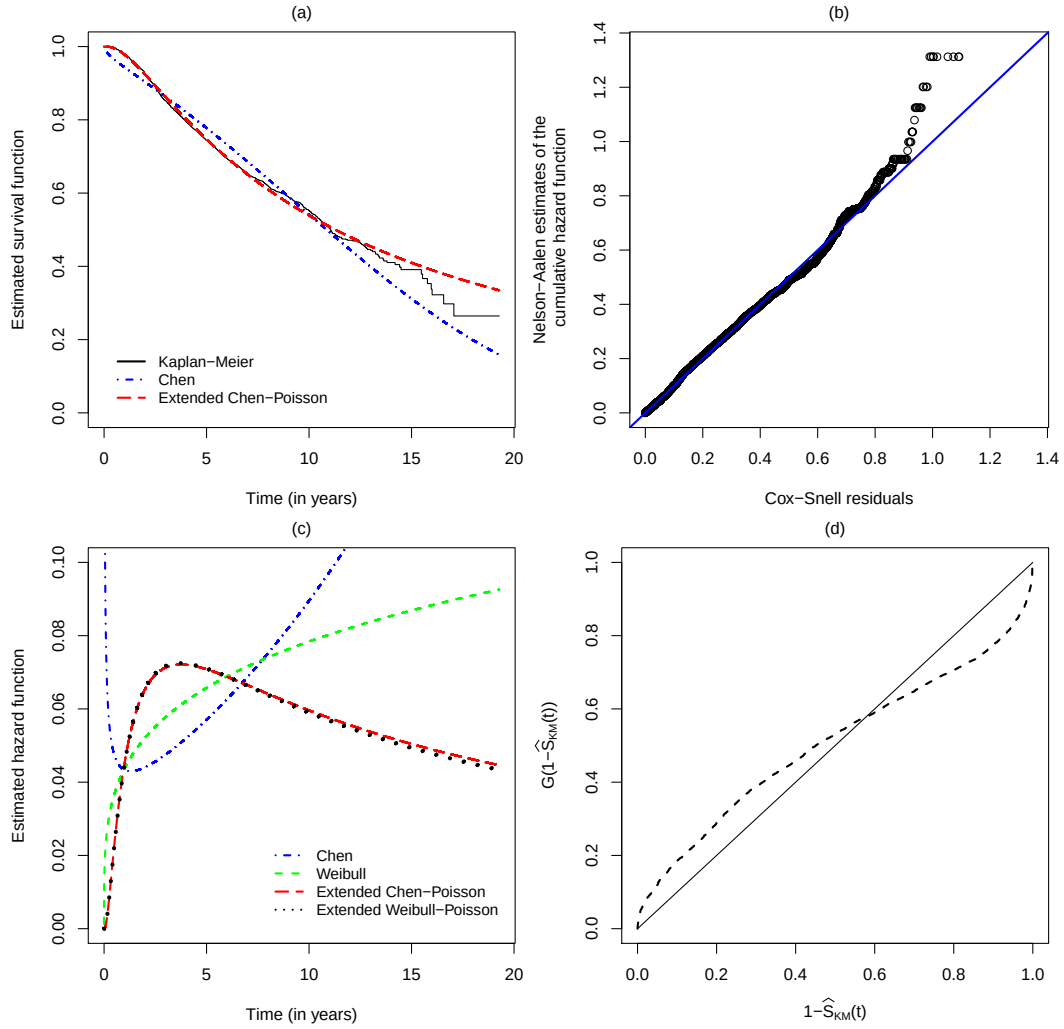

Figure 4: (a) Estimated survival functions based on the Kaplan–Meier estimator and on the Chen and ECP distributions; (b) Cox–Snell residuals of the ECP distribution; (c) Estimated hazard functions of the Chen, Weibull, ECP and EWP distributions; (d) Empirical scaled TTT-transform based on the Kaplan–Meier estimator for the Rotterdam breast cancer data.

The hf estimates of the Chen, Weibull, ECP and EWP distributions are depicted in Figure 4 (c). With the purpose of identifying the hazard shape, the TTT plot is once again considered. However, the existence of censored observations must be taken into account. As mentioned by Klefsjö [17], a natural generalization of the empirical scaled TTT-transform, $G(r/n)$, to accommodate right censored data consists in replacing the empirical cdf, r/n, by the estimator of the cdf based on the Kaplan–Meier estimator, $1-\widehat{S}_{\mathrm{KM}}(t)$. Figure 4 (d) shows that, in this case, the TTT plot is initially concave and then becomes convex, suggesting an unimodal hf. Both ECP and EWP distributions were able to capture an unimodal hazard shape, providing quite similar estimates. Therefore, in addition to both models being suitable for modelling the Rotterdam data, the proposed distribution is an adequate parametric alternative to the EWP distribution.

4. CONCLUDING REMARKS

In this paper, we introduce a new three-parameter lifetime distribution, named ECP distribution. The proposed distribution is a generalization of the Chen distribution [11] and arises from the unified Poisson compounding approach of Ramos *et al.* [26], where both distributions of the minimum and maximum are merged into one when it is assumed that the latent variable follows a ZTP distribution. Under this approach, the obtained distribution allows a practical interpretation in CCR settings. It was verified that if the parameter from the ZTP distribution takes a negative (or positive) value, then the random variable with ECP distribution represents the minimum (or maximum) lifetime among all unobservable causes. Several features of the new distribution are deduced, including the explicit expressions for the sf, pdf, hf, quantile function, moment generating function (particularly, for the mean and variance) and mean residual life function. The ECP distribution can take a richer variety of flexible hazard shapes regarding to the baseline distribution. In fact, the main advantage of the ECP distribution is that its hf can be monotonic increasing, monotonic decreasing, unimodal, bathtub, IDI or DIDI.

The estimation of the parameters is done by the ML method, considering a right-censoring mechanism. The results of the simulation study showed the effectiveness of the ML method, in which the bias and MSE of the parameters estimates are close to zero as the sample size increases. Additionally, two real data applications were presented with the following purposes:

- **i**) to assess the adequacy of the ECP distribution for modelling uncensored (guinea pigs) and censored (Rotterdam breast cancer) data;
- **ii**) to compare the proposed distribution with other generalizations of the Chen distribution, as well as with other members of the unified Poisson family.

In both applications, the ECP distribution clearly revealed to be a suitable parametric alternative for modelling the data, when compared with the competing models. It is noteworthy that some of the considered models have quite flexible hfs (such as the Marshal-Olkin Chen and E2WP distributions) but, for the analysed data sets, none was better than the ECP distribution. This fact emphasizes the potential and flexibility of the proposed model.

ACKNOWLEDGMENTS

This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia, under the projects UIDB/00006/2020 (CEAUL – Centro de Estatística e Aplicações) and UIDB/04674/2020 (Center for Research in Mathematics and Applications (CIMA) related with the Statistics, Stochastic Processes and Applications (SSPA) group). I. Sousa-Ferreira also acknowledges FCT for the PhD grant DFA/BD/6459/2020.

The authors would like to sincerely thank the Co-Editor and the three anonymous referees, for their valuable and very constructive comments. We would also like to thank P.L. Ramos for kindly providing us with the `R` programming codes of some alternative methodologies.

REFERENCES

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

[2] Alawadhi, F.A.; Sarhan, A.M. and Hamilton, D.C. (2016). Marshall–Olkin extended two-parameter bathtub-shaped lifetime distribution, *Journal of Statistical Computation and Simulation*, **86**(18), 3653–3666.

[3] Anzagra, L.; Sarpong, S. and Nasiru, S. (2020). Chen-G class of distributions, *Cogent Mathematics & Statistics*, **7**(1), 1–20.

[4] Barreto-Souza, W. and Cribari-Neto, F. (2009). A generalization of the exponential-Poisson distribution, *Statistics & Probability Letters*, **79**(24), 2493–2500.

[5] Basu, A.P. (1981). *Identifiability problems in the theory of competing and complementary risks – a survey.* In "Statistical Distributions in Scientific Work. NATO Advanced Study Institute Series (Series C – Mathematical and Physical Sciences)" (C. Taillie; G.P. Patil and B.A. Baldessari, Eds.), vol. 79, Springer, Dordrecht, 335–347.

[6] Basu, A.P. and Klein, J.P. (1982). Some recent results in competing risks theory, *Lecture Notes-Monograph Series*, **2**, 216–229.

[7] Bhatti, F.A.; Hamedani, G.G.; Najibi, S.M. and Ahmad, M. (2019). On the extended Chen distribution: development, properties, characterizations and applications, *Annals of Data Science*, 1–22.

[8] Bjerkedal, T. (1960). Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli, *American Journal of Hygiene*, **72**(1), 130–48.

[9] Cancho, V.G.; Louzada-Neto, F. and Barriga, G.D.C. (2011). The Poisson-exponential lifetime distribution, *Computational Statistics & Data Analysis*, **55**(1), 677–686.

[10] Chaubey, Y.P. and Zhang, R. (2015). An extension of Chen's family of survival distributions with bathtub shape or increasing hazard rate function, *Communications in Statistics – Theory and Methods*, **44**(19), 4049–4064.

[11] Chen, Z. (2000). A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function, *Statistics & Probability Letters*, **49**(2), 155–161.

[12] Collett, D. (2015). *Modelling Survival Data in Medical Research*, 3rd edition, Chapman and Hall/CRC.

[13] Cordeiro, G.M.; Silva, G.O. and Ortega, E.M.M. (2016). An extended-G geometric family, *Journal of Statistical Distributions and Applications*, **3**(3), 1–16.

[14] Dey, S.; Kumar, D.; Ramos, P.L. and Louzada, F. (2017). Exponentiated Chen distribution: properties and estimation, *Communications in Statistics – Simulation and Computation*, **46**(10), 8118–8139.

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

[16] Hemmati, F.; Khorram, E. and Rezakhah, S. (2011). A new three-parameter ageing distribution, *Journal of Statistical Planning and Inference*, **141**(7), 2266–2275.

[17] Klefsjö, B. (1991). TTT-plotting – a tool for both theoretical and practical problems, *Journal of Statistical Planning and Inference*, **29**(1–2), 99–110.

[18] Kuş, C. (2007). A new lifetime distribution, *Computational Statistics & Data Analysis*, **51**(9), 4497–4509.

[19] Lu, W. and Shi, D. (2012). A new compounding life distribution: the Weibull–Poisson distribution, *Journal of Applied Statistics*, **39**(1), 21–38.

[20] Mahmoudi, E. and Sepahdar, A. (2013). Exponentiated Weibull–Poisson distribution: model, properties and applications, *Mathematics and Computers in Simulation*, **92**, 76–97.

[21] Marinho, P.R.D.; Bourguignon, M. and Barros Dias, C.R.B. (2016). *AdequacyModel: adequacy of probabilistic models and general purpose optimization*, R package version 2.0.0. `https://CRAN.R-project.org/package=AdequacyModel`

[22] 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**(8), 1–30.

[23] Nadarajah, S.; Cordeiro, G.M. and Ortega, E.M.M. (2012). General results for the Kumaraswamy-G distribution, *Journal of Statistical Computation and Simulation*, **82**(7), 951–979.

[24] Pappas, V.; Adamidis, K. and Loukas, S. (2011). A three-parameter lifetime distribution, *Advances and Applications in Statistics*, **20**(2), 159–167.

[25] R Core Team (2021). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`

[26] Ramos, P.L.; Dey, D.K.; Louzada, F. and Lachos, V.H. (2020). An extended Poisson family of life distribution: a unified approach in competitive and complementary risks, *Journal of Applied Statistics*, **47**(2), 306–322.

[27] Ramos, P.L.; Guzman, D.C.F.; Mota, A.L.; Rodrigues, F.A. and Louzada, F. (2020). Sampling with censored data: a practical guide, *arXiv: 2011.08417* [stat.CO].

[28] Sarhan, A.M. and Apaloo, J. (2013). Exponentiated modified Weibull extension distribution, *Reliability Engineering & System Safety*, **112**, 137–144.

[29] Sauerbrei, W.; Royston, P. and Look, M. (2007). A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation, *Biometrical Journal*, **49**, 453–473.

[30] Tahir, M.H. and Cordeiro, G.M. (2016). Compounding of distributions: a survey and new generalized classes, *Journal of Statistical Distributions and Applications*, **3**(1), 1–35.

[31] Tarvirdizade, B. and Ahmadpour, M. (2019). A new extension of Chen distribution with applications to lifetime data, *Communications in Mathematics and Statistics*, 1–16.

[32] Therneau, T. (2021). *survival: a package for survival analysis in R*, R package version 3.2-11.
https://CRAN.R-project.org/package=survival

[33] Toomet, O. and Henningsen, A. (2021). *maxLik: Maximum Likelihood Estimation and Related Tools*, R package version 1.4-8.
https://CRAN.R-project.org/package=maxLik

[34] Xie, M.; Tang, Y. and Goh, T.N. (2002). A modified Weibull extension with bathtub-shaped failure rate function, *Reliability Engineering & System Safety*, **76**(3), 279–285.

REVSTAT – Statistical Journal
Volume 21, Number 2, April 2023, 197–217
https://doi.org/10.57805/revstat.v21i2.404

Orderings and Ageing of Reliability Systems with Dependent Components Under Archimedian Copulas

Authors: Ghobad Barmalzan ✉
– Department of Statistics, University of Zabol, Sistan and Baluchestan, Iran
ghbarmalzan@uoz.ac.ir

Ali Akbar Hosseinzadeh
– Department of Mathematics, University of Zabol, Sistan and Baluchestan, Iran
hosseinzadeh@uoz.ac.ir

Narayanaswamy Balakrishnan
– Department of Mathematics, McMaster University, Hamilton, Canada
bala@mcmaster.ca

Received: August 2020 Revised: November 2021 Accepted: November 2021

Abstract:

- In this paper, we have considered systems with dependent components having a joint distribution modeled by an Archimedean copula and with component lifetimes following accelerated failure time and modified proportional hazards distributions. We have then established characterization results specifically for series, fail-safe, 2-out-of-n and parallel systems through comparisons with average systems in terms of mean residual life, hazard rate and reversed hazard rate orders. We have also discussed various stochastic orderings and ageing results for the residual lives of parallel and series systems. The results established here are quite general, and several examples have been used to illustrate all the results and their reliability implications.

Keywords:

- *stochastic orders; ageing faster in hazard rate; ageing faster in reversed hazard rate; series systems; parallel systems; fail-safe systems.*

AMS Subject Classification:

- F60E15, 90B25.

✉ Corresponding author.

1. INTRODUCTION

Many coherent reliability systems, such as series, parallel, fail-safe and r-out-of-n systems, have all become useful and essential reliability structures in practice. For example, in the architecture of network circuits, series circuit configurations are often used to manage voltage drops to add to equal voltage, and for all the components in the circuit to share the same equal current and the resistance to sum to equal total resistance. Similarly, parallel circuit configurations are made use of so that all the components in the circuit can share the same equal voltage, and with branch current adding to equal total current and resistance diminishing to equal total resistance.

A fail-safe system is one that is designed so as to remain safe in the event of a failure; it is not designed to prevent failure, but it is intended to mitigate failure when it does occur. An elevator is a good example of a fail-safe system as it is designed with special brakes that are held back by the tension of the cable, so that if the cable does snap, the loss of tension would force the special brakes to be applied, thus averting an accident. Another recent practical application of fail-safe system (2-out-of-3 system, to be specific) is in the autonomous parking system in a car which consists of three computers and a sensor to determine an appropriate parking manoeuvre in a given situation. While the three computers take the specific information from the sensor into account and plan the steering and acceleration to successfully park, they would compare their results and only if at least two of them are in agreement, the car would park with that manoeuvre agreed by the majority of computers.

It is, therefore, quite important to understand the reliability and ageing characteristics of such coherent reliability systems commonly used in practice. Stochastic orders are useful tools for the purpose of comparative reliability evaluation and relative ageing of systems; one may refer to the book length accounts by Müller and Stoyan [26] and Shaked and Shanthikumar [33] for various stochastic orders, ageing notions and their applications to a wide range of problems arising from different fields. The earliest and pioneering work in this regard was carried out nearly five decades ago by Pledger and Proschan [28] and Proschan and Sethuraman [29]. There have been numerous subsequent developments in this direction, too many to list here, as a matter of fact. But, interested readers may refer to the following articles for some key results: Deshpande and Kochar [9], Saunders [32], Boland *et al.* [7], Kochar and Korwar [17], Dykestra *et al.* [11], Khaledi and Kochar [15], Kochar and Xu [18], Zhao and Balakrishnan [34], Zhao *et al.* [29], Balakrishnan *et al.* [2], and Barmalzan *et al.* [5]. Detailed reviews of all the developments in this regard have also been presented by Kochar [16] and Balakrishnan and Zhao [3].

Even though there is a huge body of literature on various types of comparisons of different reliability systems, as witnessed in the reviews of Kochar [16] and Balakrishnan and Zhao [4], most of the references cited therein and also all the papers mentioned above only deal with the case of independent and non-identical components. Very few papers have dealt with the case when the components in a system are dependent; see, for example, Rezapour and Alamatsaz [31], Li and Fang [21], Ding and Zhang [10], Cai *et al.* [8], Fang *et al.* [12], and Barmalzan *et al.* [6].

Many systems in practice will include a number of components that are homogeneous, like battery packs, circuits, airbags, etc.; but, the assumption that their lifetimes are independent may not be realistic and yet is one that is usually made in order to make the corresponding models and subsequent derivations simpler. As the components in a system will be functioning simultaneously, the functioning of one is likely to impact the functioning of others. Moreover, these components may all be manufactured by the same producer, and so may share the same manufacturing environment. It is, therefore, quite reasonable to expect some dependence between them!

In this work, we consider reliability systems with dependent components, with the joint distribution being modeled by a general Archimedean copula, and the lifetime of components following accelerated failure time and modified proportional hazards distributions. We then establish several characterization results for series, fail-safe, 2-out-of-n and parallel systems through comparisons with average systems in terms of hazard rate, reversed hazard rate and mean residual life orders.

There are several different ways to model dependence [see Kotz *et al.* [19]], and one convenient way is through the use of copulas [Nelsen [27]]. Here, in this work, we use an Archimedean copula to represent the joint distribution of the lifetimes of n components in the system, as it is a well-known family of copulas with many prominent copulas, such as independence, Ali-Mikhail-Haq, Gumbel-Hougaard, Clayton, and Frank copulas, all as special cases. It is for this reason that we assume the Archimedean copula to model the joint distribution of lifetimes of components.

The rest of this paper proceeds as follows. In Section 2, we briefly introduce some basic stochastic orders, ageing notions and copulas that are most pertinent for the discussions to follow in the subsequent sections; in addition, we provide a description of the accelerated failure time and modified proportional hazards families of distributions that are used to model the marginal distributions of lifetimes of components. In Section 3, we establish various stochastic orderings and ageing results for the residual lives of parallel systems. In Section 4, we similarly establish stochastic orderings and ageing results for the residual lives of series systems. In Section 5, we develop some characterization results for some coherent systems when the components follow an accelerated failure time model based on a comparison with an average system. Similarly, in Section 6, we present some characterization results for some coherent systems when the components follow a modified proportional hazards distribution based on a comparison with an average system. Finally, in Section 7, we present some concluding remarks and also some problems that will be of interest for further research.

2. DEFINITIONS AND KEY NOTIONS

We describe in this section some basic concepts about stochastic orders, copulas and two general families of lifetime distributions that are essential for subsequent developments. We assume through out that all random variables under consideration are lifetime variables and so are nonnegative, and we use "increasing" to mean "nondecreasing" and "decreasing" to mean "nonincreasing". We assume all the expectations involved to exist, and for ease of notation, we use $a \overset{sgn}{=} b$ to denote that both sides of an equality have the same sign.

2.1. Stochastic orders

Let X and Y be random variables with density functions f_X and f_Y, distribution functions F_X and F_Y, survival functions $\bar{F}_X = 1 - F_X$ and $\bar{F}_Y = 1 - F_Y$, hazard rate functions $h_X = f_X/\bar{F}_X$ and $h_Y = f_Y/\bar{F}_Y$, and reversed hazard rate functions $\tilde{h}_X = f_X/F_X$ and $\tilde{h}_Y = f_Y/F_Y$, respectively.

Definition 2.1. Then, X is said to be larger than Y in:

- **(i)** usual stochastic order (denoted by $X \geq_{\rm st} Y$) if $\bar{F}_X(t) \geq \bar{F}_Y(t)$, for all $t \in \mathbb{R}$, or equivalently, $\mathbb{E}[\phi(X)] \geq \mathbb{E}[\phi(Y)]$ for all increasing functions $\phi : \mathbb{R} \to \mathbb{R}$;
- **(ii)** hazard rate order (denoted by $X \geq_{\rm hr} Y$) if and only if $h_Y(t) \geq h_X(t)$, for all $t \in \mathbb{R}$, or equivalently, $\bar{F}_X(t)/\bar{F}_Y(t)$ is increasing in $t \in \mathbb{R}$;
- **(iii)** reversed hazard rate order (denoted by $X \geq_{\rm rh} Y$) if and only if $\tilde{h}_X(t) \geq \tilde{h}_Y(t)$, for all $t \in \mathbb{R}$, or equivalently, $F_X(t)/F_Y(t)$ is increasing in $t \in \mathbb{R}$;
- **(iv)** mean residual life order (denoted by $X \geq_{\rm mrl} Y$) if $E(X_t) \geq E(Y_t)$, for all $t \in \mathbb{R}$, where $E(X_t) = E(X - t|X > t)$ and $E(Y_t) = E(Y - t|Y > t)$ are the mean residual lives of X and Y, respectively.

Then, the following implications are well-known between these orders:

$$X \geq_{\rm hr[rh]} Y \Longrightarrow X \geq_{\rm st} Y;$$

see, for example, Müller and Stoyan [26] and Shaked and Shanthikumar [33] for extensive discussions on various stochastic orderings, their inter-relationships, and their properties and applications.

2.2. Ageing notions

Ageing, in reliability analysis, describes the variation in the performance of a unit over time. Several different measures and measure-based stochastic orders have been discussed in the literature pertaining to ageing characteristics of life distributions. Two most commonly used notions are through hazard and reversed hazard rates.

Definition 2.2. A random variable X is said to be ageing faster than Y in:

- **(i)** hazard rate (denoted by $X \geq_{\rm c} Y$) if $h_Y(t)/h_X(t)$ is increasing in $t \in \mathbb{R}$ (Kalashnikov and Rachev, [14]);
- **(ii)** reversed hazard rate (denoted by $X \geq_{\rm b} Y$) if $\tilde{h}_X(t)/\tilde{h}_Y(t)$ is increasing in $t \in \mathbb{R}$ (Rezaei *et al.*, [30]).

For more details on the relative ageing by increasing hazard ratio and reversed hazard ratio functions, one may refer to Lai and Xie [20], Misra and Francis [25] and Hazra and Misra [14].

2.3. Archimedean copulas

As mentioned earlier in Section 1, a plethora of stochastic orders and stochastic comparisons of random variables have been discussed in the literature; but, most of them involve only comparisons of marginal distributions of the underlying variables, without taking into account possible dependence between variables, with some exceptions, of course! Here, we consider characterizations of some reliability systems assuming the components to be dependent under an Archimedean copula.

Archimedean copulas are widely used for modeling dependence between variables due to their mathematical tractability as well as their ability to model a wide range of dependence structures. For a decreasing continuous function $\phi : [0, \infty) \longrightarrow [0, 1]$ with $\phi(0) = 1$, $\phi(+\infty) = 0$ and $\psi = \phi^{-1}$ being the pseudo-inverse,

$$C_\phi(u_1, ..., u_n) = \phi(\psi(u_1) + \cdots + \psi(u_n)), \qquad u_i \in [0, 1], \tag{2.1}$$

is said to be an Archimedean copula with generator ϕ if $(-1)^k \phi^{[k]}(x) \geq 0$ for $k = 0, ..., n-2$ and $(-1)^{n-2}\phi^{[n-2]}(x)$ is decreasing and convex, with $\phi^{[k]}(x)$ denoting the k-the derivative of the generator $\phi(x)$ with respect to x.

2.4. Accelerated failure time and modified proportional hazards distributions

Let $X_1, ..., X_n$ be random variables with X_i having $h_i(t)$, for $i = 1, ..., n$, as marginal hazard functions. Then, they are said to have an accelerated failure time family of distributions if, for all $t \geq 0$, $h_i(t) = h(\lambda_i t)$, for $i = 1, ..., n$, where $h(\cdot)$ is some baseline hazard function and $\lambda_i > 0$ are scale parameters (also called acceleration constants). Upon noting now that the cumulative hazard rate functions of X_i are given by $H_i(t) = \frac{1}{\lambda_i} H(\lambda_i t)$, and then using the relationship between cumulative hazard function and survival function of a distribution, we arrive at the form of cumulative distribution function for this family as

$$S_i(t) = e^{-H_i(t)} = e^{-\frac{1}{\lambda_i} H(\lambda_i t)} = \{e^{-H(\lambda_i t)}\}^{1/\lambda_i} = \{S(\lambda_i t)\}^{1/\lambda_i}, \tag{2.2}$$

for $t \geq 0$, and $i = 1, ..., n$; see, for example, Marshall and Olkin (2007) for details.

In the context of nonparametric rank tests, two families of distributions with

$$G_1(x) = (F(x))^\alpha, \ \alpha > 0, \qquad \bar{G}_2(x) = (S(x))^\beta, \ \beta > 0, \tag{2.3}$$

known as "Lehmann families", have been used extensively as nonparametric alternatives for tests for stochastic orderings. Upon combining the two families in (2.3), we can obtain an unified family of distributions with cumulative distribution function of the form

$$G(x) = 1 - \{1 - (F(x))^\alpha\}^\beta, \qquad \alpha, \beta > 0, \tag{2.4}$$

where $F(\cdot)$ is some baseline distribution function. Now, we may introduce acceleration constants λ_i $(i = 1, ..., n)$, as in (2.2), to arrive at a general form of accelerated failure time distribution with its cumulative distribution function as

$$F_i(t) = 1 - \{1 - (F(\lambda_i t))^\alpha\}^\beta, \quad t > 0, \alpha, \beta > 0, \tag{2.5}$$

for $i = 1, ..., n$. It is evident that the accelerated failure time model in (2.2) is a special case of (2.5) when $\alpha = 1$ and $\beta = 1/\lambda_i$.

Yet another flexible family of useful lifetime distributions, offered by Marshll and Olkin [23], has a survival function of the form

$$S^*(t) = \frac{\alpha S(t)}{1 - \bar{\alpha} S(t)}, \qquad t > 0,\ 0 < \alpha < 1, \bar{\alpha} = 1 - \alpha, \tag{2.6}$$

where S is some baseline survival function and α is referred to as a tilt parameter. Here again, by introducing acceleration constants λ_i $(i = 1, ..., n)$, as in (2.2), we arrive at a family of modified proportional hazards family of distributions with its survival function as

$$S_i(t) = \frac{\alpha S(\lambda_i t)}{1 - \bar{\alpha} S(\lambda_i t)}, \qquad t > 0, \lambda_i > 0, 0 < \alpha < 1, \bar{\alpha} = 1 - \alpha, \tag{2.7}$$

for $i = 1, ..., n$. The name "modified proportional hazards model" stems from the fact that the hazard functions of S and S^* in (2.6) satisfy the relationship

$$h_{S^*}(t) = h_S(t)\ \frac{1}{1 - \bar{\alpha} S(t)}, \tag{2.8}$$

which is indeed a modification of the proportional hazards assumption, with the multiplicative term varying over t, rather than being a constant.

3. RESULTS FOR RESIDUAL LIVES OF PARALLEL SYSTEMS

Let $X_{n:n}$ denote the lifetime of a parallel system consisting of n dependent components whose joint distribution is given by an Archimedean copula. Then, the survival function, density function, hazard rate function and reversed hazard rate function of the residual life variable $X_{n:n}(t)$ at x, given that the parallel system has survived till time t, are given by

$$F_{X_{n:n}(t)}(x) = \frac{\phi(n\psi[F(x+t)]) - \phi(n\psi[F(t)])}{1 - \phi(n\psi[F(t)])}, \qquad x, t \geq 0, \tag{3.1}$$

$$f_{X_{n:n}(t)}(x) = \frac{nf(x+t)\,\psi'[F(x+t)]\phi'(n\psi[F(x+t)])}{1 - \phi(n\psi[F(t)])}, \qquad x, t \geq 0, \tag{3.2}$$

$$h_{X_{n:n}(t)}(x) = \frac{nf(x+t)\,\psi'[F(x+t)]\phi'(n\psi[F(x+t)])}{1 - \phi(n\psi[F(x+t)])}, \qquad x, t \geq 0, \tag{3.3}$$

$$\tilde{h}_{X_{n:n}(t)}(x) = \frac{nf(x+t)\,\psi'[F(x+t)]\phi'(n\psi[F(x+t)])}{\phi(n\psi[F(x+t)]) - \phi(n\psi[F(t)])} \qquad x, t \geq 0, \tag{3.4}$$

where ϕ is the generator and $\psi = \phi^{-1}$. One question that we may ask here is, between two parallel systems with n and m components, which one is more reliable. Of course, this can be formulated using any particular stochastic order, as seen in the following theorems.

Theorem 3.1. *If $u \ln'[1 - \phi(u)]$ is decreasing in $u \in \mathbb{R}^+$, then for $m \geq n$, we have $X_{m:m}(t) \geq_{hr} X_{n:n}(t)$.*

Proof: With the hazard rate function of $X_{n:n}(t)$ as given in (3.3), for obtaining the desired result, it is sufficient to show that $h_{X_{n:n}(t)}(x) - h_{X_{m:m}(t)}(x) \leq 0$, for any $x \in \mathbb{R}^+$. We have

$$\begin{aligned} I(x) &= h_{X_{n:n}(t)}(x) - h_{X_{m:m}(t)}(x) \\ &= \frac{f(x+t)\psi'(F(x+t))}{\psi(F(x+t))}\left\{\frac{n\psi(F(x+t))\phi'(n\psi(F(x+t)))}{1-\phi(n\psi(F(x+t)))}\right. \\ &\quad \left. - \frac{m\psi(F(x+t))\phi'(m\psi(F(x+t)))}{1-\phi(m\psi(F(x+t)))}\right\} \\ &\overset{sgn}{=} u\ln'[1-\phi(u)]\big|_{u=n\psi(F(x+t))} - u\ln'[1-\phi(u)]\big|_{u=m\psi(F(x+t))}. \end{aligned} \tag{3.5}$$

Now, by using the decreasing property of $u\ln'[1-\phi(u)]$ with respect to $u \in \mathbb{R}^+$, for $m \geq n$, we readily observe from (3.5) that $h_{X_{n:n}(t)}(x) \geq h_{X_{m:m}(t)}(x)$, for $x \in \mathbb{R}^+$. Thus, the theorem gets established. □

Remark 3.1. Theorem 3.1 shows that, for some Archimedean copulas, parallel systems with more redundancy is more reliable in the sense of hazard rate order; that is, a parallel system with less (dependent) components will possess a higher hazard rate than a parallel system with less components.

Example 3.1. It should be mentioned that the condition "$u\ln'[1-\phi(u)]$ is decreasing" in Theorem 3.1 is quite general and holds for many Archimedean copulas. We now demonstrate this with the following examples:

1. If $\phi_1(u) = e^{-u^\theta}$, for $\theta \in \mathbb{R}^+$ (Gumbel copula, Nelsen [27]), we have

$$u\ln'[1-\phi_1(u)] = -\frac{t\phi_1'(u)}{1-\phi_1(u)} = \frac{\theta u^\theta e^{-u^\theta}}{1-e^{-u^\theta}},$$

   which is decreasing in $u \in \mathbb{R}^+$;

2. If $\phi_2(u) = 1-(1-e^{-u})^\theta$, for $\theta \in [0,1)$ (Li and Li [22]), we have

$$u\ln'[1-\phi_2(u)] = -\frac{u\phi_2'(u)}{1-\phi_2(u)} = \frac{\theta u e^{-u}}{1-e^{-u}},$$

   which is decreasing in $u \in \mathbb{R}^+$;

3. If $\phi_3(u) = \frac{1}{\sqrt{u}+1}$ (Li and Li [22]), we have

$$u\ln'[1-\phi_3(u)] = -\frac{u\phi_3'(u)}{1-\phi_3(u)} = \frac{1}{4(\sqrt{u}+1)},$$

   which is decreasing in $u \in \mathbb{R}^+$;

4. If $\phi_4(u) = \frac{1}{2}e^u\left(e^u - \frac{1}{2}\right)^{-1}$ (Ali-Mikhail-Haq copula, Nelsen [27]), we have

$$u\ln'[1-\phi_4(u)] = -\frac{u\phi_4'(u)}{1-\phi_4(u)} = \frac{ue^u}{2(e^u-\frac{1}{2})(e^u-1)},$$

   which is decreasing in $u \in \mathbb{R}^+$.

Example 3.2. Consider the standard exponential distribution as baseline distribution function. Assume that $\phi(u) = \frac{1}{\sqrt{u}+1}$, $t = 5$, $n = 5$ and $m = 10$.

Figure 1 presents plots of the hazard rate functions of $h_{X_{5:5}}(1/x-1)$ and $h_{X_{10:10}}(1/x-1)$, from which it can be observed that the value of $h_{X_{10:10(5)}}(1/x-1)$ is always smaller than that of $h_{X_{5:5(5)}}(1/x-1)$ on the interval $(0,1)$. Thus, the results of Theorem 3.1 is validated in this case.

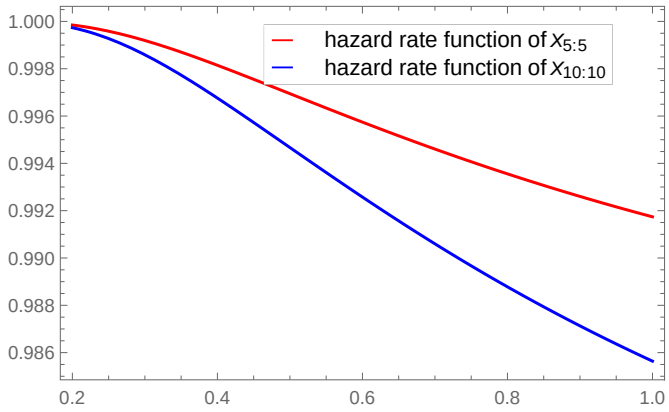

Figure 1: Plots of hazard rate functions of $h_{X_{5:5}}(1/x-1)$ and $h_{X_{10:10}}(1/x-1)$.

Theorem 3.2. *If $u\ln'[\phi(m\psi(F(t))) - \phi(u)]$ is increasing with respect to $u \in \mathbb{R}_+$, then for $m \geq n$, we have $X_{n:n}(t) \geq_{rh} X_{m:m}(t)$.*

Proof: With reversed hazard rate function of $X_{n:n}(t)$ as given in (3.4), for establishing the desired result, we need to show that $\tilde{h}_{X_{n:n}(t)}(x) \leq \tilde{h}_{X_{m:m}(t)}(x)$, for any $x \in \mathbb{R}^+$. Because $\phi'(x) \leq 0$, we have

$$
\begin{aligned}
I(x) &= \tilde{h}_{X_{n:n}(t)}(x) - \tilde{h}_{X_{m:m}(t)}(x) \\
&= \frac{f(x+t)\psi'(F(x+t))}{\psi(F(x+t))}\left\{\frac{n\psi(F(x+t))\phi'(n\psi(F(x+t)))}{\phi(n\psi[F(x+t)]) - \phi(n\psi[F(t)])}\right. \\
&\quad \left. - \frac{m\psi(F(x+t))\phi'(m\psi(F(x+t)))}{\phi(m\psi[F(x+t)]) - \phi(m\psi[F(t)])}\right\} \\
&\geq \frac{f(x+t)\psi'(F(x+t))}{\psi(F(x+t))}\left\{\frac{n\psi(F(x+t))\phi'(n\psi(F(x+t)))}{\phi(n\psi[F(x+t)]) - \phi(m\psi[F(t)])}\right. \\
&\quad \left. - \frac{m\psi(F(x+t))\phi'(m\psi(F(x+t)))}{\phi(m\psi[F(x+t)]) - \phi(m\psi[F(t)])}\right\} \\
&\overset{sgn}{=} u\ln'[\phi(u) - \phi(m\psi(F(t)))]\,|_{\,u=m\psi(F(x+t))} \\
&\quad - u\ln'[\phi(u) - \phi(m\psi(F(t)))]\,|_{\,u=n\psi(F(x+t))}. \qquad (3.6)
\end{aligned}
$$

Using the increasing property of $u\ln'[\phi(m\psi(F(t))) - \phi(u)]$ with respect to $u \in \mathbb{R}^+$, for $m \geq n$, we readily observe from (3.6) that $I(x) \geq 0$, for $x \in \mathbb{R}^+$. Thus, the theorem gets established. □

Remark 3.2. Theorem 3.2 shows that, for some Archimedean copulas, a parallel system with more (dependent) components will possess a higher reversed hazard rate than a parallel system with less components.

Theorem 3.3. *If* $u \ln'\left[-\frac{\phi'(u)}{1-\phi(u)}\right]$ *is decreasing in* $u \in \mathbb{R}^+$, *then for* $m \geq n$, *we have* $X_{n:n}(t) \geq_c X_{m:m}(t)$.

Proof: With the hazard rate functions of $X_{n:n}(t)$ and $X_{m:m}(t)$ as given in (3.3), we have

$$\begin{aligned} I(x) &= \frac{h_{X_{n:n}(t)}(x)}{h_{X_{m:m}(t)}(x)} \\ &= \frac{n}{m} \times \frac{\phi'(n\psi\{F(x+t)])}{1-\phi(n\psi[F(x+t)])} \times \left\{\frac{\phi'(m\psi[F(x+t)])}{1-\phi(m\psi[F(x+t)])}\right\}^{-1}. \end{aligned}$$

Because $\phi(x)$ is decreasing, we obtain, for $m \geq n$,

$$\begin{aligned} I'(x) &\overset{sgn}{=} \left\{\frac{\phi'(n\psi(F(x+t)))}{1-\phi(n\psi(F(x+t)))}\right\}' \times \frac{\phi'(m\psi(F(x+t)))}{1-\phi(m\psi(F(x+t)))} \\ &\quad - \frac{\phi'(n\psi(F(x+t)))}{1-\phi(n\psi(F(x+t)))} \times \left\{\frac{\phi'(m\psi(F(x+t)))}{1-\phi(m\psi(F(x+t)))}\right\}' \\ &\overset{sgn}{=} -n\psi(F(x+t))\left\{\frac{\phi''(n\psi(F(x+t)))}{\phi'(n\psi(F(x+t)))} + \frac{\phi'(n\psi(F(x+t)))}{1-\phi(n\psi(F(x+t)))}\right\} \\ &\quad + m\psi(F(x+t))\left\{\frac{\phi''(m\psi(F(x+t)))}{\phi'(m\psi(F(x+t)))} + \frac{\phi'(m\psi(F(x+t)))}{1-\phi(m\psi(F(x+t)))}\right\} \\ &= u \ln'\left[-\frac{\phi'(u)}{(1-\phi(u))}\right]\Bigg|_{u=m\psi(F(x+t))} - u \ln'\left[-\frac{\phi'(u)}{(1-\phi(u))}\right]\Bigg|_{u=n\psi(F(x+t))}. \end{aligned}$$

Due to the assumption that $u \ln'\left[-\frac{\phi'(u)}{1-\phi(u)}\right]$ is decreasing in $u \in \mathbb{R}^+$, we get the required result from the above equation. □

Remark 3.3. Theorem 3.3 shows that, for some Archimedean copulas, a parallel system with less redundancy (with dependence between components) ages faster in hazard rate than a parallel system with more redundancy. Some illustrations of the result in Theorem 3.3 can be seen in Part (i) of Example 3.4 of Ding and Zhang [10].

Theorem 3.4. *If* $u \ln'\left[-\frac{\phi'(u)}{\phi(u)-\phi(m\psi[F(t)])}\right]$ *is decreasing in* $u \in \mathbb{R}^+$, *then for* $m \geq n$, *we have* $X_{m:m}(t) \geq_b X_{n:n}(t)$.

Proof: With the reversed hazard rate functions of $X_{m:m}(t)$ and $X_{n:n}(t)$ as given in (3.4), we have

$$\begin{aligned} I(x) &= \frac{\tilde{h}_{X_{n:n}(t)}(x)}{\tilde{h}_{X_{m:m}(t)}(x)} \\ &= \frac{n}{m} \times \frac{\phi'(n\psi[F(x+t)])}{\phi(n\psi[F(x+t)]) - \phi(n\psi[F(t)])} \times \left\{\frac{\phi'(m\psi\{F(x+t)])}{\phi(m\psi[F(x+t)]) - \phi(m\psi[F(t)])}\right\}^{-1}. \end{aligned}$$

Because $\phi(x)$ is decreasing, we obtain, for $m \geq n$,

$$
\begin{aligned}
I'(x) &\overset{sgn}{=} \left\{\frac{\phi'(n\psi(F(x+t)))}{\phi(n\psi[F(x+t)]) - \phi(n\psi[F(t)])}\right\}' \times \frac{\phi'(m\psi(F(x+t)))}{\phi(m\psi[F(x+t)]) - \phi(m\psi[F(t)])} \\
&\quad - \frac{\phi'(n\psi(F(x+t)))}{\phi(n\psi[F(x+t)]) - \phi(n\psi[F(t)])} \times \left\{\frac{\phi'(m\psi(F(x+t)))}{\phi(m\psi[F(x+t)]) - \phi(m\psi[F(t)])}\right\}' \\
&\overset{sgn}{=} -n\psi(F(x+t))\left\{\frac{\phi''(n\psi(F(x+t)))}{\phi'(n\psi(F(x+t)))} - \frac{\phi'(n\psi(F(x+t)))}{\phi(n\psi[F(x+t)]) - \phi(n\psi[F(t)])}\right\} \\
&\quad + m\psi(F(x+t))\left\{\frac{\phi''(m\psi(F(x+t)))}{\phi'(m\psi(F(x+t)))} - \frac{\phi'(m\psi(F(x+t)))}{\phi(m\psi[F(x+t)]) - \phi(m\psi[F(t)])}\right\} \\
&\leq -n\psi(F(x+t))\left\{\frac{\phi''(n\psi(F(x+t)))}{\phi'(n\psi(F(x+t)))} - \frac{\phi'(n\psi(F(x+t)))}{\phi(n\psi[F(x+t)]) - \phi(m\psi[F(t)])}\right\} \\
&\quad + m\psi(F(x+t))\left\{\frac{\phi''(m\psi(F(x+t)))}{\phi'(m\psi(F(x+t)))} - \frac{\phi'(m\psi(F(x+t)))}{\phi(m\psi[F(x+t)]) - \phi(m\psi[F(t)])}\right\} \\
&= u\ln'\left[-\frac{\phi'(u)}{\phi(u) - \phi(m\psi[F(t)])}\right]\Bigg|_{u=m\psi(F(x+t))} - u\ln'\left[-\frac{\phi'(u)}{\phi(u) - \phi(m\psi[F(t)])}\right]\Bigg|_{u=n\psi(F(x+t))}.
\end{aligned}
$$

Due to assumption that $u\ln'\left[-\frac{\phi'(u)}{\phi(u)-\phi(m\psi[F(t)])}\right]$ is decreasing in $u \in \mathbb{R}^+$, from the above equation, we find $I(x)$ to be decreasing, as required. □

Remark 3.4. Theorem 3.4 shows that, for some Archimedean copulas, under the decreasing property of the function $u\ln'\left[-\frac{\phi'(u)}{\phi(u)-\phi(m\psi[F(t)])}\right]$ with respect to $u \in \mathbb{R}^+$, a parallel system with more redundancy ages faster in terms of the reversed hazard rate than a parallel system with less redundancy.

4. RESULTS FOR RESIDUAL LIVES OF SERIES SYSTEMS

Let $X_{1:n}$ denote the lifetime of a series system consisting of n dependent components whose joint distribution is given by an Archimedean copula. Then, the distribution function, density function, hazard rate function and reversed hazard rate function of residual life variable $X_{1:n}(t)$ at x, given that the series system has survived till time t, are given by

$$
\bar{F}_{X_{1:n}(t)}(x) = \frac{\phi\big(n\psi\big(\bar{F}(x+t)\big)\big)}{\phi\big(n\psi\big(\bar{F}(t)\big)\big)}, \qquad x, t > 0, \tag{4.1}
$$

$$
f_{X_{1:n}(t)}(x) = \frac{nf(x+t)\,\psi'\big(\bar{F}(x+t)\big)\phi'\big(n\psi\big(\bar{F}(x+t)\big)\big)}{\phi\big(n\psi\big(\bar{F}(t)\big)\big)}, \qquad x, t > 0, \tag{4.2}
$$

$$
h_{X_{1:n}(t)}(x) = \frac{nf(x+t)\,\psi'\big(\bar{F}(x+t)\big)\phi'\big(n\psi\big(\bar{F}(x+t)\big)\big)}{\phi\big(n\psi\big(\bar{F}(x+t)\big)\big)}, \qquad x, t > 0, \tag{4.3}
$$

$$
\tilde{h}_{X_{1:n}(t)}(x) = \frac{nf(x+t)\,\psi'\big(\bar{F}(x+t)\big)\phi'\big(n\psi\big(\bar{F}(x+t)\big)\big)}{\phi\big(n\psi\big(\bar{F}(t)\big)\big) - \phi\big(n\psi\big(\bar{F}(x+t)\big)\big)}, \qquad x, t > 0, \tag{4.4}
$$

respectively, where ϕ is the generator and $\psi = \phi^{-1}$. Now, we examine between two series systems with n and m components, which one is more reliable.

Theorem 4.1. *If $u \ln' \phi(u)$ is decreasing in $u \in \mathbb{R}^+$, then for $m \geq n$, we have $X_{1:n}(t) \geq_{hr} X_{1:m}(t)$.*

Proof: With the hazard rate functions of $X_{1:n}(t)$ and $X_{1:m}(t)$ as given in (4.3), we have

$$\begin{aligned} I(x) &= h_{X_{1:n}(t)}(x) - h_{X_{1:m}(t)}(x) \\ &= \frac{f(x+t)\psi'\big(\bar{F}(x+t)\big)}{\psi\big(\bar{F}(x+t)\big)} \\ &\quad \times \left\{ \frac{n\psi\big(\bar{F}(x+t)\big)\phi'\big(n\psi\big(\bar{F}(x+t)\big)\big)}{\phi\big(n\psi\big(\bar{F}(x+t)\big)\big)} - \frac{m\psi\big(\bar{F}(x+t)\big)\phi'\big(m\psi\big(\bar{F}(x+t)\big)\big)}{\phi\big(m\psi\big(\bar{F}(x+t)\big)\big)} \right\} \\ &\overset{sgn}{=} u \ln' \phi(u) \,|_{\, u = m\psi(\bar{F}(x+t))} - u \ln' \phi(u) \,|_{\, u = n\psi(\bar{F}(x+t))}. \end{aligned}$$

By using the decreasing property of $u \ln' \phi(u)$, for $m \geq n$, we readily observe that $I(x) \leq 0$. Thus, the theorem gets established. □

Remark 4.1. Theorem 4.1 shows that, for some Archimedean copulas, a series system with less (dependent) components is more reliable in the sense of hazard rate order; that is, a series system with less (dependent) components will possess a lower hazard function than a series system with more components.

Example 4.1. The condition "$u \ln' \phi(u)$ is decreasing" in Theorem 4.1 is quite general and can be verified for many well-known Archimedean copulas. For example, we consider the following:

1. If $\phi_1(u) = e^{-u^\theta}$, for $\theta \in \mathbb{R}^+$ (Gumbel copula, Nelsen [27]), we have

$$u \ln'[\phi_1(u)] = -\theta u^\theta,$$

   which is decreasing in $u \in \mathbb{R}^+$;

2. If $\phi_2(u) = (\theta u + 1)^{-\frac{1}{\theta}}$ (Clayton copula, Nelsen [27]), we have

$$u \ln'[\phi_2(u)] = -\frac{u}{\theta u + 1},$$

   which is decreasing in $u \in \mathbb{R}^+$.

Example 4.2. Consider the standard exponential distribution as baseline distribution function. Assume that $\phi(u) = (\theta u + 1)^{-\frac{1}{\theta}}$, $\theta = 2$, $t = 2$, $n = 4$ and $m = 10$.

Figure 2 presents plots of the hazard rate functions of $h_{X_{1:4}}(1/x - 1)$ and $h_{X_{1:10}}(1/x - 1)$, from which it can be observed that the value of $h_{X_{1:14}}(1/x - 1)$ is always smaller than that of $h_{X_{1:10}}(1/x - 1)$ on the interval $(0, 1)$. Thus, the result of Theorem 4.1 is validated in this case.

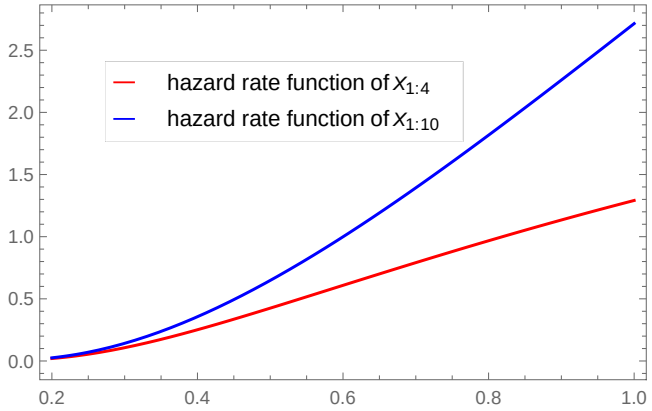

Figure 2: Plots of hazard rate functions of $h_{X_{1:4}}(1/x-1)$ and $h_{X_{1:10}}(1/x-1)$.

Theorem 4.2. *If* $u\ln'\big[\phi\big(n\psi\big(\bar{F}(t)\big)\big)-\phi(u)\big]$ *is decreasing in* $u\in\mathbb{R}^+$, *then for* $m\geq n$, *we have* $X_{1:n}(t)\geq_{rh} X_{1:m}(t)$.

Proof: With the reversed hazard rate functions of $X_{1:n}(t)$ and $X_{1:m}(t)$ as given in (4.4), for $m\geq n$, we have

$$\begin{aligned}
I(x) &= \tilde{h}_{X_{1:n}(t)}(x)-\tilde{h}_{X_{1:m}(t)}(x)\\
&= \frac{f(x+t)\psi'\big(\bar{F}(x+t)\big)}{\psi\big(\bar{F}(x+t)\big)}\Bigg\{\frac{n\psi\big(\bar{F}(x+t)\big)\phi'\big(n\psi\big(\bar{F}(x+t)\big)\big)}{\phi\big(n\psi\big(\bar{F}(t)\big)\big)-\phi\big(n\psi\big(\bar{F}(x+t)\big)\big)}\\
&\quad -\frac{n\psi\big(\bar{F}(x+t)\big)\phi'\big(m\psi\big(\bar{F}(x+t)\big)\big)}{\phi\big(m\psi\big(\bar{F}(t)\big)\big)-\phi\big(m\psi\big(\bar{F}(x+t)\big)\big)}\Bigg\}\\
&\geq \frac{f(x+t)\psi'\big(\bar{F}(x+t)\big)}{\psi\big(\bar{F}(x+t)\big)}\Bigg\{\frac{n\psi\big(\bar{F}(x+t)\big)\phi'\big(n\psi\big(\bar{F}(x+t)\big)\big)}{\phi\big(n\psi\big(\bar{F}(t)\big)\big)-\phi\big(n\psi\big(\bar{F}(x+t)\big)\big)}\\
&\quad -\frac{n\psi\big(\bar{F}(x+t)\big)\phi'\big(m\psi\big(\bar{F}(x+t)\big)\big)}{\phi\big(n\psi\big(\bar{F}(t)\big)\big)-\phi\big(m\psi\big(\bar{F}(x+t)\big)\big)}\Bigg\}\\
&\overset{sgn}{=} u\ln'\big[\phi\big(n\psi\big(\bar{F}(t)\big)\big)-\phi(u)\big]\,|_{u=n\psi(\bar{F}(x+t))}\\
&\quad -u\ln'\big[\phi\big(n\psi\big(\bar{F}(t)\big)\big)-\phi(u)\big]\,|_{u=m\psi(\bar{F}(x+t))}. \qquad (4.5)
\end{aligned}$$

Using the decreasing property of $u\ln'\big[\phi\big(n\psi\big(\bar{F}(t)\big)\big)-\phi(u)\big]$ in $u\in\mathbb{R}^+$, for $m\geq n$, we readily observe from (4.5) that $I(x)\geq 0$. Thus, the theorem gets established. □

Remark 4.2. Theorem 4.2 shows that, for some Archimedean copulas, a series system with less (dependent) components will possess lower reversed hazard rate than a series system with more components.

Theorem 4.3. *If* $u\ln'\Big[-\frac{\phi'(u)}{\phi(u)}\Big]$ *is decreasing (increasing) in* $u\in\mathbb{R}^+$, *then for* $m\geq n$, *we have* $X_{1:m}(t)\geq_c (\leq_c) X_{1:n}(t)$.

Proof: With the hazard rate functions of $X_{1:m}(t)$ and $X_{1:n}(t)$ as given in (4.3), we have

$$\begin{aligned} I(x) &= \frac{h_{X_{1:n}(t)}(x)}{h_{X_{1:m}(t)}(x)} \\ &= \frac{n}{m} \times \frac{\phi'\big(n\psi\big[\bar{F}(x+t)\big]\big)}{\phi\big(n\psi\big[\bar{F}(x+t)\big]\big)} \times \left\{ \frac{\phi'\big(m\psi\big[\bar{F}(x+t)\big]\big)}{\phi\big(m\psi\big[\bar{F}(x+t)\big]\big)} \right\}^{-1}. \end{aligned}$$

By differentiating this function, we find

$$\begin{aligned} I'(x) &\overset{sgn}{=} \left\{ \frac{\phi'\big(n\psi(\bar{F}(x+t))\big)}{\phi\big(n\psi(\bar{F}(x+t))\big)} \right\}' \times \frac{\phi'\big(m\psi(\bar{F}(x+t))\big)}{\phi\big(m\psi(\bar{F}(x+t))\big)} \\ &\quad - \frac{\phi'\big(n\psi(\bar{F}(x+t))\big)}{\phi\big(n\psi(\bar{F}(x+t))\big)} \times \left\{ \frac{\phi'\big(m\psi(\bar{F}(x+t))\big)}{\phi\big(m\psi(\bar{F}(x+t))\big)} \right\}' \\ &\overset{sgn}{=} n\psi(\bar{F}(x+t)) \left\{ \frac{\phi''(n\psi(\bar{F}(x+t)))}{\phi'(n\psi(\bar{F}(x+t)))} - \frac{\phi'(n\psi(\bar{F}(x+t)))}{\phi(n\psi(\bar{F}(x+t)))} \right\} \\ &\quad - m\psi(\bar{F}(x+t)) \left\{ \frac{\phi''(m\psi(\bar{F}(x+t)))}{\phi'(m\psi(\bar{F}(x+t)))} - \frac{\phi'(m\psi(\bar{F}(x+t)))}{\phi(m\psi(\bar{F}(x+t)))} \right\} \\ &= u \ln' \left[-\frac{\phi'(u)}{\phi(u)} \right] \Bigg|_{u=n\psi(\bar{F}(x+t))} - u \ln' \left[-\frac{\phi'(u)}{\phi(u)} \right] \Bigg|_{u=m\psi(\bar{F}(x+t))} \\ &\geq (\leq)\ 0, \end{aligned}$$

according to whether $u \ln' \left[-\frac{\phi'(u)}{\phi(u)} \right]$ is decreasing (or increasing) in $u \in \mathbb{R}^+$, for $m \geq n$. Thus, the theorem gets established. □

Remark 4.3. Theorem 4.3 shows that, for some Archimedean copulas, under the decreasing (increasing) property of the function $u \ln' \left[-\frac{\phi'(u)}{\phi(u)} \right]$, a series system with less (dependent) components ages faster (ages slower) in terms of hazard rate than a series system with more components. Some illustrations of the result in Theorem 4.3 can be seen in Part (ii) of Example 3.4 of Ding and Zhang [10].

Theorem 4.4. *If* $u \ln' \left[-\frac{\phi'(u)}{\phi\left(n\psi\left[\bar{F}(t)\right]\right) - \phi(u)} \right]$ *is decreasing in* $u \in \mathbb{R}^+$, *then for* $m \geq n$, *we have* $X_{1:n}(t) \geq_b X_{1:m}(t)$.

Proof: With the reversed hazard rate functions of $X_{1:m}(t)$ and $X_{1:n}(t)$ as given in (4.4), we have

$$\begin{aligned} I(x) &= \frac{\tilde{h}_{X_{1:n}(t)}(x)}{\tilde{h}_{X_{1:m}(t)}(x)} \\ &= \frac{n}{m} \times \frac{\phi'\big(n\psi\big[\bar{F}(x+t)\big]\big)}{\phi\big(n\psi\big[\bar{F}(t)\big]\big) - \phi\big(n\psi\big[\bar{F}(x+t)\big]\big)} \times \left\{ \frac{\phi'\big(m\psi\big[\bar{F}(x+t)\big]\big)}{\phi\big(m\psi\big[\bar{F}(t)\big]\big) - \phi\big(m\psi\big[\bar{F}(x+t)\big]\big)} \right\}^{-1}. \end{aligned}$$

As $\phi(x)$ is decreasing, for $m \geq n$, we obtain

$$
\begin{aligned}
I'(x) &\overset{sgn}{=} \left\{ \frac{\phi'\big(n\psi(\bar{F}(x+t))\big)}{\phi\big(n\psi\big[\bar{F}(t)\big]\big) - \phi\big(n\psi(\bar{F}(x+t))\big)} \right\}' \times \frac{\phi'\big(m\psi(\bar{F}(x+t))\big)}{\phi\big(m\psi\big[\bar{F}(t)\big]\big) - \phi\big(m\psi(\bar{F}(x+t))\big)} \\
&\quad - \frac{\phi'\big(n\psi(\bar{F}(x+t))\big)}{\phi\big(n\psi\big[\bar{F}(t)\big]\big) - \phi\big(n\psi(\bar{F}(x+t))\big)} \times \left\{ \frac{\phi'\big(m\psi(\bar{F}(x+t))\big)}{\phi\big(m\psi\big[\bar{F}(t)\big]\big) - \phi\big(m\psi(\bar{F}(x+t))\big)} \right\}' \\
&\overset{sgn}{=} n\psi(\bar{F}(x+t)) \left\{ \frac{\phi''(n\psi(\bar{F}(x+t)))}{\phi'(n\psi(\bar{F}(x+t)))} + \frac{\phi'(n\psi(\bar{F}(x+t)))}{\phi\big(n\psi\big[\bar{F}(t)\big]\big) - \phi(n\psi(\bar{F}(x+t)))} \right\} \\
&\quad - m\psi(\bar{F}(x+t)) \left\{ \frac{\phi''(m\psi(\bar{F}(x+t)))}{\phi'(m\psi(\bar{F}(x+t)))} + \frac{\phi'(m\psi(\bar{F}(x+t)))}{\phi\big(m\psi\big[\bar{F}(t)\big]\big) - \phi(m\psi(\bar{F}(x+t)))} \right\} \\
&\geq n\psi(\bar{F}(x+t)) \left\{ \frac{\phi''(n\psi(\bar{F}(x+t)))}{\phi'(n\psi(\bar{F}(x+t)))} + \frac{\phi'(n\psi(\bar{F}(x+t)))}{\phi\big(n\psi\big[\bar{F}(t)\big]\big) - \phi(n\psi(\bar{F}(x+t)))} \right\} \\
&\quad - m\psi(\bar{F}(x+t)) \left\{ \frac{\phi''(m\psi(\bar{F}(x+t)))}{\phi'(m\psi(\bar{F}(x+t)))} + \frac{\phi'(m\psi(\bar{F}(x+t)))}{\phi\big(n\psi\big[\bar{F}(t)\big]\big) - \phi(m\psi(\bar{F}(x+t)))} \right\} \\
&= u \ln' \left[-\frac{\phi'(u)}{\phi\big(n\psi\big[\bar{F}(t)\big]\big) - \phi(u)} \right] \Bigg|_{u=n\psi(\bar{F}(x+t))} - u \ln' \left[-\frac{\phi'(u)}{\phi\big(n\psi\big[\bar{F}(t)\big]\big) - \phi(u)} \right] \Bigg|_{u=m\psi(\bar{F}(x+t))}.
\end{aligned}
$$

Due to the assumption that $u \ln' \left[-\frac{\phi'(u)}{\phi\big(n\psi\big[\bar{F}(t)\big]\big) - \phi(u)} \right]$ is decreasing in $u \in \mathbb{R}^+$, we have $I'(x) > 0$. Thus, the theorem gets established. □

Remark 4.4. Theorem 4.4 shows that, for some Archimedean copulas, a series system with less (dependent) components ages faster in terms of reversed hazard rate than a series system with more components.

Example 4.3. We note that the condition "$u \ln' \left[-\frac{\phi'(u)}{\phi\big(m\psi\big[\bar{F}(t)\big]\big) - \phi(u)} \right]$ is decreasing" in Theorem 4.4 holds in many cases. For example, consider $\phi(u(x,t)) = e^{-u}$ and $0 < a(t) \leq 1$ and also $\phi(u(x,t)) < a(t)$ for all $t \in [0, \infty)$. We then have

$$
u \ln' \left[-\frac{\phi'(u)}{a - \phi(u)} \right] = u \left\{ \frac{\phi''(u)}{\phi'(u)} - \frac{\phi'(u)}{a - \phi(u)} \right\} = \frac{-au}{a - e^{-u}}
$$

to be decreasing in $u \in \mathbb{R}^+$.

5. SYSTEMS WITH DEPENDENT ACCELERATION FAILURE TIME COMPONENTS

One of the common reliability structures in practice is a r-out-of-n system. This system, consisting of n components, works iff at least r components work. It includes parallel, fail-safe and series systems all as special cases when $r = 1$, $r = n - 1$ and $r = n$, respectively. In this section, we develop some characterization results for these systems when the components are

dependent with an Archimedean copula and the component lifetimes follow an accelerated failure time distribution in (2.5) based on a comparison with the "average system". The results established here complete and extend some results of Cai *et al.* [8].

Using the copula representation for the joint distribution of $X_1, ..., X_n$ in (2.1), we have in this case

$$\bar{F}_{1:n}(x) = \phi\left(\sum_{k=1}^{n} \psi((1 - F^{\alpha}(\lambda_k x))^{\beta})\right), \quad x > 0, \tag{5.1}$$

$$\begin{aligned} \bar{F}_{2:n}(x) &= \sum_{l=1}^{n} \phi\left(\sum_{k=1,k\neq l}^{n} \psi((1 - F^{\alpha}(\lambda_k x))^{\beta})\right) \\ &\quad - (n-1)\phi\left(\sum_{k=1}^{n} \psi((1 - F^{\alpha}(\lambda_k x))^{\beta})\right), \quad x > 0, \end{aligned} \tag{5.2}$$

$$\bar{G}(x) = \frac{1}{n}\sum_{l=1}^{n} \phi\left(\sum_{k=1,k\neq l}^{n} \psi((1 - F^{\alpha}(\lambda_k x))^{\beta})\right), \quad x > 0. \tag{5.3}$$

The expressions in (5.1) and (5.2) correspond to the survival functions of the series system (i.e., $r = n$) and of the fail-safe system ($r = n - 1$), respectively. The expression in (5.3) corresponds to the survival function of an "average series system", whose lifetime is denoted by Y. This average series system can be explained by a randomization process as follows: From a series system comprising n components, one randomly selected component may be removed to obtain a series system with $(n - 1)$ remaining components; out of the n such $(n - 1)$-component series systems, we then randomly select one of them, and that is what the average series system is here. The expression of the survival function given in (5.3) then becomes clear.

Theorem 5.1. *We have:*

(i) $X_{1:n} \leq_{mrl} X_{2:n}$ *iff* $X_{1:n} \leq_{mrl} Y$;

(ii) $X_{1:n} \leq_{hr} X_{2:n}$ *iff* $X_{1:n} \leq_{hr} Y$;

(iii) $X_{1:n} \leq_{rh} X_{2:n}$ *iff* $X_{1:n} \leq_{rh} Y$.

Proof: (i) By definition, $X_{1:n} \leq_{mrl} X_{2:n}$ iff $\forall t > 0$, we have

$$\frac{\int_0^{\infty} \bar{F}_{2:n}(x+t)\,dx}{\bar{F}_{2:n}(t)} \geq \frac{\int_0^{\infty} \bar{F}_{1:n}(x+t)\,dx}{\bar{F}_{1:n}(t)}. \tag{5.4}$$

Upon using (5.1) and (5.2) in (5.4) and Theorem 2.A.6 of Shaked and Shanthikumar [33] and some simplifications, $\forall t > 0$,

$$\begin{aligned} &\phi\left(\sum_{i=1}^{n} \psi((1 - F^{\alpha}(\lambda_k t))^{\beta})\right) \times \int_0^{\infty} \left[\sum_{l=1}^{n} \phi\left(\sum_{k=1,k\neq l}^{n} \psi((1 - F^{\alpha}(\lambda_k x + \lambda_k t))^{\beta})\right)\right] dx \\ &\geq \sum_{l=1}^{n} \phi\left(\sum_{k=1,k\neq l}^{n} \psi((1 - F^{\alpha}(\lambda_k x + \lambda_k t))^{\beta})\right) \\ &\quad \times \int_0^{\infty} \left[\phi\left(\sum_{k=1}^{n} \psi((1 - F^{\alpha}(\lambda_k x + \lambda_k t))^{\beta})\right)\right] dx. \end{aligned} \tag{5.5}$$

Similarly, from (5.1) and (5.3), we see that $Y \geq_{mrl} X_{1:n}$ iff $\forall t > 0$,

$$\begin{aligned} &\frac{\int_0^\infty \frac{1}{n}\sum_{l=1}^n \phi\Big(\sum_{k=1,k\neq l}^n \psi((1-F^\alpha(\lambda_k x+\lambda_k t))^\beta)\Big)dx}{\frac{1}{n}\sum_{l=1}^n \phi\Big(\sum_{k=1,k\neq l}^n \psi((1-F^\alpha(\lambda_k t))^\beta)\Big)} \\ &\geq \frac{\int_0^\infty \phi\big(\sum_{i=1}^n \psi((1-F^\alpha(\lambda_k x+\lambda_k t))^\beta)\big)dx}{\phi(\sum_{k=1}^n \psi((1-F^\alpha(\lambda_k t))^\beta))}. \end{aligned} \tag{5.6}$$

The equivalence of the inequalities in (5.5) and (5.6) yields Part (i) immediately.

(ii) By definition, $X_{1:n} \leq_{hr} X_{2:n}$ iff $\forall x, t > 0$, we have

$$\frac{\bar{F}_{2:n}(x+t)}{\bar{F}_{2:n}(t)} \geq \frac{\bar{F}_{1:n}(x+t)}{\bar{F}_{1:n}(t)}. \tag{5.7}$$

Upon using (5.1) and (5.2) in (5.7) and simplification, $\forall x, t > 0$,

$$\begin{aligned} &\sum_{l=1}^n \phi\left(\sum_{k=1,k\neq l}^n \psi((1-F^\alpha(\lambda_k x+\lambda_k t))^\beta)\right) \times \left[\phi\left(\sum_{k=1}^n \psi((1-F^\alpha(\lambda_k t))^\beta)\right)\right] \\ &\geq \sum_{l=1}^n \phi\left(\sum_{k=1,k\neq l}^n \psi((1-F^\alpha(\lambda_k t))^\beta)\right) \times \left[\phi\left(\sum_{k=1}^n \psi((1-F^\alpha(\lambda_k x+\lambda_k t))^\beta)\right)\right]. \end{aligned} \tag{5.8}$$

Similarly, from (5.1) and (5.3), we see that $Y \geq_{hr} X_{1:n}$ iff $\forall x, t > 0$,

$$\begin{aligned} &\frac{\frac{1}{n}\sum_{l=1}^n \phi\Big(\sum_{k=1,k\neq l}^n \psi((1-F^\alpha(\lambda_k x+\lambda_k t))^\beta)\Big)}{\frac{1}{n}\sum_{l=1}^n \phi\Big(\sum_{k=1,k\neq l}^n \psi((1-F^\alpha(\lambda_k t))^\beta)\Big)} \\ &\geq \frac{\phi\big(\sum_{k=1}^n \psi((1-F^\alpha(\lambda_k x+\lambda_k t))^\beta)\big)}{\phi(\sum_{k=1}^n \psi((1-F^\alpha(\lambda_k t))^\beta))}. \end{aligned} \tag{5.9}$$

The equivalence of the inequalities in (5.8) and (5.9) yields Part (ii) immediately.

(iii) This can be proved in a manner similar to Part (ii). □

Next, from the copula representation for the joint distribution of $X_1, ..., X_n$ in (2.1), we have, in this case, for $x > 0$,

$$F_{n:n}(x) = \phi\left(\sum_{k=1}^n \psi(1-(1-F^\alpha(\lambda_k x))^\beta)\right), \tag{5.10}$$

$$\begin{aligned} F_{n-1:n}(x) &= \sum_{l=1}^n \phi\left(\sum_{k=1,k\neq l}^n \psi(1-(1-F^\alpha(\lambda_k x))^\beta)\right) \\ &\quad - (n-1)\phi\left(\sum_{k=1}^n \psi(1-(1-F^\alpha(\lambda_k x))^\beta)\right), \end{aligned} \tag{5.11}$$

and let Z have its distribution function as

$$H(x) = \frac{1}{n}\sum_{l=1}^n \phi\left(\sum_{k=1,k\neq l}^n \psi(1-(1-F^\alpha(\lambda_k x))^\beta)\right), \qquad x > 0. \tag{5.12}$$

The expression in (5.10) corresponds to the survival function of a parallel system (i.e., $r = 1$), while the expression in (5.11) corresponds to the survival function of a 2-out-of-n system. The expression in (5.12) corresponds to the survival function of an "average parallel system", whose lifetime is denoted here by Z. This average parallel system can once again be explained by a randomization process as follows: From a parallel system consisting of n components, one randomly selected component may be removed to obtain a parallel system with $(n-1)$ remaining components; out of the n such $(n-1)$-component parallel systems, we randomly select one of them, and that is what the average parallel system is here. The expression of the survival function given in (5.12) then becomes clear.

Theorem 5.2. *In the special case when $n = 2$, we have:*

(i) $X_{n-1:n} \leq_{mrl} X_{n:n}$ *iff* $Z \leq_{mrl} X_{n:n}$;

(ii) $X_{n-1:n} \leq_{hr} X_{n:n}$ *iff* $Z \leq_{hr} X_{n:n}$;

(iii) $X_{n-1:n} \leq_{rh} X_{n:n}$ *iff* $Z \leq_{rh} X_{n:n}$.

Proof: This can be established in a manner analogous to Theorem 5.1, and we therefore do not present it here for the sake of brevity. □

We now present a complete characterization result for the special case when $n = 2$.

Theorem 5.3. *We have:*

(i) $X_{1:2} \leq_{mrl} Y \iff X_{1:2} \leq_{mrl} X_{2:2} \iff Z \leq_{mrl} X_{2:2}$;

(ii) $X_{1:2} \leq_{hr} Y \iff X_{1:2} \leq_{hr} X_{2:2} \iff Z \leq_{hr} X_{2:2}$;

(iii) $X_{1:2} \leq_{rh} Y \iff X_{1:2} \leq_{rh} X_{2:2} \iff Z \leq_{rh} X_{2:2}$.

Proof: In Theorem 3.1, we have characterization between $X_{1:n}$ and $X_{2:n}$ based on characterization between $X_{1:n}$ and Y. For the case when $n = 2$, it is simply a characterization between $X_{1:2}$ and $X_{2:2}$ based on characterization between $X_{1:2}$ and Y. Similarly, in Theorem 3.2, we have characterization between $X_{n-1:n}$ and $X_{n:n}$ based on characterization between Z and $X_{n:n}$, which in the case when $n = 2$, is simply a characterization between $X_{1:2}$ and $X_{2:2}$ based on characterization between Z and $X_{2:2}$. As the left hand sides of both results are the same variables, the characterization results on the right hand sides must be equivalent. Thus, the characterization of $X_{1:2}$ and Y must be equivalent to the characterization of Z and $X_{2:2}$. □

6. SYSTEMS WITH DEPENDENT MODIFIED PROPORTIONAL HAZARDS COMPONENTS

In this section, we assume that the n components in a reliability system are dependent with their component lifetimes following a modified proportional hazards model in (2.7) and their joint distribution being represented by an Archimedean copula in (2.1). We then establish some characterization results for series, fail-safe, 2-out-of-n and parallel systems in this general setup using mean residual life, hazard rate and reversed hazard orders based on a comparison with the "average system". The results established here complete and extend some results of Cai *et al.* [8].

In this case, from (2.1), we have

$$\bar{F}_{1:n}(x) = \phi\left(\sum_{k=1}^{n}\psi\left(\frac{\alpha\bar{F}(\lambda_k x)}{1-\bar{\alpha}\bar{F}(\lambda_k x)}\right)\right), \qquad x>0, \tag{6.1}$$

$$\begin{aligned}\bar{F}_{2:n}(x) &= \sum_{l=1}^{n}\phi\left(\sum_{l=1,k\neq l}^{n}\psi\left(\frac{\alpha\bar{F}(\lambda_k x)}{1-\bar{\alpha}\bar{F}(\lambda_k x)}\right)\right)\\ &\quad -(n-1)\phi\left(\sum_{k=1}^{n}\psi\left(\frac{\alpha\bar{F}(\lambda_k x)}{1-\bar{\alpha}\bar{F}(\lambda_k x)}\right)\right), \qquad x>0,\end{aligned} \tag{6.2}$$

$$\bar{G}(x) = \frac{1}{n}\sum_{l=1}^{n}\phi\left(\sum_{k=1,k\neq l}^{n}\psi\left(\frac{\alpha\bar{F}(\lambda_k x)}{1-\bar{\alpha}\bar{F}(\lambda_k x)}\right)\right), \qquad x>0, \tag{6.3}$$

where ϕ is the generator and $\psi=\phi^{-1}$. The expressions in (6.1)–(6.3) correspond to the survival functions of series, fail-safe and average series systems in this case, respectively. We use Y to denote the lifetime of the average series system whose survival function is given in (6.3)

Theorem 6.1. *We have:*

(i) $X_{1:n} \leq_{mrl} X_{2:n}$ *iff* $X_{1:n} \leq_{mrl} Y$;

(ii) $X_{1:n} \leq_{hr} X_{2:n}$ *iff* $X_{1:n} \leq_{hr} Y$;

(iii) $X_{1:n} \leq_{rh} X_{2:n}$ *iff* $X_{1:n} \leq_{rh} Y$.

Proof: This can be established in a manner analogous to Theorem 5.1, and we therefore do not present it here for the sake of brevity. □

Next, from the copula representation for the joint distribution of $X_1,...,X_n$ in (2.1), we find in this case

$$F_{n:n}(x) = \phi\left(\sum_{k=1}^{n}\psi\left(\frac{1-\bar{F}(\lambda_k x)}{1-\bar{\alpha}\bar{F}(\lambda_k x)}\right)\right), \qquad x>0, \tag{6.4}$$

$$\begin{aligned}F_{n-1:n}(x) &= \sum_{l=1}^{n}\phi\left(\sum_{k=1,k\neq l}^{n}\psi\left(\frac{1-\bar{F}(\lambda_k x)}{1-\bar{\alpha}\bar{F}(\lambda_k x)}\right)\right)\\ &\quad -(n-1)\phi\left(\sum_{k=1}^{n}\psi\left(\frac{1-\bar{F}(\lambda_k x)}{1-\bar{\alpha}\bar{F}(\lambda_k x)}\right)\right), \qquad x>0,\end{aligned} \tag{6.5}$$

and let Z be a random variable with its distribution function as

$$H(x) = \frac{1}{n}\sum_{l=1}^{n}\phi\left(\sum_{k=1,k\neq l}^{n}\psi\left(\frac{1-\bar{F}(\lambda_k x)}{1-\bar{\alpha}\bar{F}(\lambda_k x)}\right)\right), \qquad x>0. \tag{6.6}$$

The expressions in (6.4)–(6.6) correspond to the distribution functions of parallel, 2-out-of-n and average parallel systems in this case.

Theorem 6.2. *We have:*

(i) $X_{n-1:n} \leq_{mrl} X_{n:n}$ *iff* $Z \leq_{mrl} X_{n:n}$;

(ii) $X_{n-1:n} \leq_{hr} X_{n:n}$ *iff* $Z \leq_{hr} X_{n:n}$;

(iii) $X_{n-1:n} \leq_{rh} X_{n:n}$ *iff* $Z \leq_{rh} X_{n:n}$.

Proof: This can be proved in a manner analogous to Theorem 6.1, and we therefore do not present the proof here for the sake of brevity. □

Theorem 6.3. *In the special case when* $n = 2$, *we have:*

(i) $X_{1:2} \leq_{mrl} Y \iff X_{1:2} \leq_{mrl} X_{2:2} \iff Z \leq_{mrl} X_{2:2}$;

(ii) $X_{1:2} \leq_{hr} Y \iff X_{1:2} \leq_{hr} X_{2:2} \iff Z \leq_{hr} X_{2:2}$;

(iii) $X_{1:2} \leq_{rh} Y \iff X_{1:2} \leq_{rh} X_{2:2} \iff Z \leq_{rh} X_{2:2}$.

Proof: This can be proved in a way similar to Theorem 5.3, and we therefore do not describe it here. □

7. CONCLUDING REMARKS

In this work, we have considered reliability systems with dependent components having accelerated failure time and modified proportional hazards distributions and having a joint distribution represented by a general Archimedean copula. We have focused especially on series, fail-safe, 2-out-of-n and parallel systems, and have then established some characterization results for these systems through comparisons with average systems in terms of mean residual life, hazard rate and reversed hazard rate orders. It will naturally be of interest to extend these results to the case of general $(n - r + 1)$-out-of-n systems and sequential $(n - r + 1)$-out-of-n systems as discussed by Barmalzan *et al.* [6] under the general setting considered here; one may see Misra and Francis [25] for some results in this regard under a restricted setting. We are currently working on these problems and hope to report the findings in a future paper.

ACKNOWLEDGMENTS

We express our sincere thanks to the Editor and anonymous reviewers for their useful comments and suggestions on an earlier version of this manuscript which led to this improved version. This work has been supported by University of Zabol, grant number: UOZ-GR-3389.

REFERENCES

[1] Balakrishnan, N.; Barmalzan, G. and Haidari, A. (2014). Stochastic orderings and ageing properties of residual life lengths of live components in $(n-k+1)$-out-of-n systems, *Journal of Applied Probability*, **51**, 58–68.

[2] Balakrishnan, N.; Barmalzan, G. and Haidari, A. (2018). On stochastic comparisons of k-out-of-n systems with Weibull components, *Journal of Applied Probability*, **55**, 834–844.

[3] Balakrishnan, N. and Zhao, P. (2013a). Hazard rate comparison of parallel systems with heterogeneous gamma components, *Journal of Multivariate Analysis*, **113**, 153–160.

[4] Balakrishnan, N. and Zhao, P. (2013b). Ordering properties of order statistics from heterogeneous populations: a review with an emphasis on some recent developments, *Probability in the Engineering and Informational Sciences*, **27**, 403–443.

[5] Barmalzan, G.; Ayat, S.M.; Balakrishnan, N. and Roozegar, R. (2020). Stochastic comparisons of series and parallel systems with dependent heterogeneous extended exponential components under Archimedean copula, *Journal of Computational and Applied Mathematics*, **380**, Article 112965.

[6] Barmalzan, G.; Haidari, A. and Balakrishnan, N. (2018). Univariate and multivariate stochastic orderings of residual lifetimes of live components in sequential $(n-r+1)$-out-of-n systems, *Journal of Applied Probability*, **55**, 834–844.

[7] Boland, P.J.; El-Neweihi, E. and Proschan, F. (1994). Applications of the hazard rate ordering in reliability and order statistics, *Journal of Applied Probability*, **31**, 180–192.

[8] Cai, N.; Ni, W. and Li, C. (2019). Some ordering properties of series and parallel systems with dependent component lifetimes, *Communications in Statistics – Theory and Methods*, **48**, 4764–4779.

[9] Deshpande, J.V. and Kochar, S.C. (1983). Dispersive ordering is the same as tail ordering, *Advances in Applied Probability*, **15**, 686–687.

[10] Ding, W. and Zhang, Y. (2018). Relative ageing of series and parallel systems: effects of dependence and heterogeneity among components, *Operations Research Letters*, **46**, 219–224.

[11] Dykstra, R.A.; Kochar, S.C. and Rojo, J. (1997). Stochastic comparisons of parallel systems of heterogeneous exponential components, *Journal of Statistical Planning and Inference*, **65**, 203–211.

[12] Fang, L.; Balakrishnan, N. and Jin, Q. (2020). Optimal grouping of heterogeneous components in series-parallel and parallel-series systems under Archimedean copula dependence, *Journal of Computational and Applied Mathematics*, **377**, Article 112916.

[13] Hazra, N.K. and Misra, N. (2020). On relative ageing of coherent systems with dependent identically distributed components, *Advances in Applied Probability*, **52**, 348–376.

[14] Kalashnikov, V.V. and Rachev, S.T. (1986). *Characterization of queueing models and their stability*. In "Probability Theory and Mathematical Statistics" (S. Watanabe and V.V. Prokhorov, Eds.), Vol. 2, pp. 37–53, Amsterdam, VNU Science Press.

[15] Khaledi, B.E. and Kochar, S. (2000). Some new results on stochastic comparisons of parallel systems, *Journal of Applied Probability*, **37**, 1123–1128.

[16] Kochar, S.C. (2012). Stochastic comparisons of order statistics and spacings: a review, *ISRN Probability and Statistics*, **2012**, Article ID 839473.

[17] Kochar, S.C. and Korwar, R. (1996). Stochastic orders for spacings of heterogeneous exponential random variables, *Journal of Multivariate Analysis*, **57**, 69–83.

[18] Kochar, S. and Xu, M. (2009). Comparisons of parallel systems according to the convex transform order, *Journal of Applied Probability*, **46**, 342–352.

[19] Kotz, S.; Balakrishnan, N. and Johnson, N.L. (2000). *Continuous Multivariate Distributions*, Vol. 1 (2nd ed.), New York, John Wiley & Sons.

[20] Lai, C.D. and Xie, M. (2003). Relative ageing for two parallel systems and related problems, *Mathematical and Computer Modelling*, **38**, 1339–1345.

[21] Li, X. and Fang, R. (2015). Ordering properties of order statistics from random variables of Archimedean copulas with applications, *Journal of Multivariate Analysis*, **133**, 304–320.

[22] Li, C. and Li, X. (2015). Likelihood ratio order of sample minimum from heterogeneous Weibull random variables, *Statistics & Probability Letters*, **97**, 46–53.

[23] 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.

[24] Marshall, A.W. and Olkin, I. (2007). *Life Distributions: Structure of Nonparametric, Semiparametric, and Parametric Families*, New York, Springer.

[25] Misra, N. and Francis, J. (2015). Relative ageing of $(n-k+1)$-out-of-n systems, *Statistics & Probability Letters*, **106**, 272–280.

[26] Müller, A. and Stoyan, D. (2002). *Comparison Methods for Stochastic Models and Risks*, Hoboken, Chichester, England, John Wiley & Sons.

[27] Nelsen, R.B. (2006). *An Introduction to Copulas*, New York, Springer.

[28] Pledger, P. and Proschan, F. (1971). *Comparisons of order statistics and of spacings from heterogeneous distributions*. In "Optimizing Methods in Statistics" (J.S. Rustagi, Ed.), pp. 89–113, New York, Academic Press.

[29] Proschan, F. and Sethuraman, J. (1976). Stochastic comparisons of order statistics from heterogeneous populations, with applications in reliability, *Journal of Multivariate Analysis*, **6**, 608–616.

[30] Rezaei, M.; Gholizadeh, B. and Izadkhah, S. (2015). On relative reversed hazard rate order, *Communications in Statistics – Theory and Methods*, **44**, 300–308.

[31] Rezapour, M. and Alamatsaz, M.H. (2014). Stochastic comparison of lifetimes of two $(n-k+1)$-out-of-n systems with heterogeneous dependent components, *Journal of Multivariate Analysis*, **130**, 240–251.

[32] Saunders, D.J. (1984). Dispersive ordering of distributions, *Advances in Applied Probability*, **16**, 693–694.

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

[34] Zhao, P. and Balakrishnan, N. (2009). Characterization of MRL order of fail-safe systems with heterogeneous exponential components, *Journal of Statistical Planning and Inference*, **139**, 3027–3037.

[35] Zhao, P.; Li, X. and Balakrishnan, N. (2009). Likelihood ratio order of the second order statistic from independent heterogeneous exponential random variables, *Journal of Multivariate Analysis*, **100**, 952–962.

REVSTAT – Statistical Journal
Volume 21, Number 2, April 2023, 219–233
https://doi.org/10.57805/revstat.v21i2.403

Performance Comparison of Independence Tests in Two-Way Contingency Tables

Authors: Ebru Ozturk
– Department of Biostatistics, Faculty of Medicine, Hacettepe University, Ankara, Turkey
ebru.ozturk3@hacettepe.edu.tr

Merve Basol ✉
– Department of Biostatistics, Faculty of Medicine, Erciyes University, Kayseri, Turkey
merve.basol@erciyes.edu.tr

Dincer Goksuluk
– Department of Biostatistics, Faculty of Medicine, Erciyes University, Kayseri, Turkey
dincergoksuluk@erciyes.edu.tr

Sevilay Karahan
– Department of Biostatistics, Faculty of Medicine, Hacettepe University, Ankara, Turkey
sevilaykarahan@gmail.com

Received: March 2021 Revised: November 2021 Accepted: November 2021

Abstract:

- Several test statistics are available for testing the independence of categorical variables from two-way contingency tables. A vast majority of published articles used the Pearson's chi-squared test for such purposes; however, this test statistic may lead to biased conclusions under certain conditions. Therefore, we aimed to compare the performance of test statistics via a comprehensive simulation study considering several factors in contingency tables. We also evaluated the performance of each test statistic on a real-life dataset. This study contributes to the literature guiding researchers to select an appropriate test statistic under different conditions.

Keywords:

- *contingency table; power; power divergence; simulation study; type one error.*

AMS Subject Classification:

- 62H17, 62E10.

✉ Corresponding author.

1. INTRODUCTION

The data type of measured variables is important to determine the statistical methods for summarizing and testing the relationship or independence between variables [9]. Analyzing categorical data is generally less tractable and may require much effort for selecting appropriate statistical methods, such as log-linear models, logistic regression, and chi-square tests. The contingency table approach is one of the frequently used methods to summarize the joint distribution of two categorical variables. An example of r-by-c contingency table showing the joint distribution of categorical variables X and Y is given in Table 1. Here, n_{ij} $(i = 1, 2, ..., r$ and $j = 1, 2, ..., c)$ represents the frequencies of joint occurrences, $n_{i+} = \sum_{j=1}^{c} n_{ij}$ and $n_{+j} = \sum_{i=1}^{r} n_{ij}$ are row and column totals (i.e., row/column marginals), and $n = \sum_{i=1}^{r} n_{i+} = \sum_{j=1}^{c} n_{+j} = \sum_{j=1}^{c} \sum_{i=1}^{r} n_{ij}$ is the grand total of contingency table that also refers to sample size.

Table 1: An example of r-by-c contingency table.

	Y_1	Y_2	$\cdots$	Y_c	Total
X_1	n_{11}	n_{12}	$\cdots$	n_{1c}	n_{1+}
X_2	n_{21}	n_{21}	$\cdots$	n_{2c}	n_{2+}
$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	n_{r+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+c}	n

Specification of the joint probability distribution of Table 1 is crucial since it plays a key role in the type of statistical analysis used. The distribution of a contingency table may be one of multinomial, product multinomial, hypergeometric, and Poisson based on the cell counts that are fixed such that row/column marginals or totals. The inference about the independence between categorical variables can be evaluated using the appropriate sampling distribution and statistical hypotheses. The hypotheses for testing the independence of categorical variables in Table 1 is defined as

$$\begin{aligned} &H_0 : \ \pi_{1j} = \pi_{2j} = ... = \pi_{rj}, \\ &H_1 : \ \pi_{ij} \neq \pi_{kj} \quad \text{at least one } i, j, k, \quad i \neq k, \end{aligned} \tag{1.1}$$

where π_{ij} is the hypothesized cell probability of the i-th row and the j-th column, and $\hat{\pi}_{ij}$ is the estimated cell probability from sampling distribution. There are several methods for estimating cell probabilities, i.e., π_{ij}, and testing a hypothesis (1.1) depending on the joint distributions [13, 8].

Pearson's chi-square test statistic is widely used for testing the hypothesis (1.1). However, it is not a gold standard and may not be appropriate for small samples [1]. There exist various test statistics proposed to test the independence, where each performs better under certain conditions, such as sample size, number of rows and columns, sampling methods, etc. In this study, we used the most common of these methods, which are:

(**i**) Pearson's chi-square test;

(**ii**) likelihood ratio test;

(**iii**) Freeman–Tukey test;

(**iv**) Cressie–Read test;

(**v**) Fisher–Freeman–Halton's exact test.

A hypothesis established from a contingency table, considering the purpose of the study, could be tested using different statistical test procedures. The results of the hypothesis tests might be in the opposite direction for the variety of hypothesis tests. It is a crucial issue since it may mislead the researcher in their studies. Therefore, it is essential to choose appropriate statistical tests or methods to achieve correct and unbiased conclusions. In this study, we aimed to compare different test procedures and related test statistics under various scenarios for the power $(1-\beta)$ and the type-I error rate (α) of the test statistic. We conducted a comprehensive simulation study using the combinations of sample size, effect size, and sampling design. Furthermore, we applied each method to a real-life dataset for making a fair comparison between simulation and real-life data results. This study contributed to the literature by considering each test procedure under several conditions and comparing the performances of each test statistic via a comprehensive simulation study. Furthermore, the current study compared the simulation results with a real-life dataset and showed the concordance (or discordance) between the simulation study and the real-life example. All the analyses were performed on the R programming language (`https://cran.r-project.org/`) through self-written codes available upon request to the correspondent author.

The plan of this study is as follows. The methods, statistical background, simulation scenarios, and real datasets are explained in detail in the Material and Methods section. The results of simulated and real datasets are presented in the Results section, and finally, we discussed the results in the Discussion section with conclusions and future work.

2. MATERIAL AND METHODS

The statistical methods proposed to test the hypothesis (1.1) are detailed in subsection 2.1. These methods use the observed (n_{ij}) and expected (E_{ij}) frequencies to compute test statistics. All test statistics are asymptotically chi-square distributed with degrees of freedom $(r-1)(c-1)$.

2.1. Test Statistics

The most common test statistic proposed to test independence between categorical variables is the Pearson's chi-square statistic [1],which takes the difference between observed and expected frequencies into account. The test statistic (χ^2) is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-E_{ij})^2}{E_{ij}}. \tag{2.1}$$

The likelihood ratio test statistic is another approach to test independence [1]. Unlike Pearson's chi-square statistic, it is based on the ratio of the observed and expected frequencies. The test statistic is

$$G^2 = 2 \times \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \times \log\left(\frac{n_{ij}}{E_{ij}}\right). \tag{2.2}$$

The Freeman and Tukey test statistic aims to approximate Binomial or Poisson distribution to normal distribution by stabilizing the variance [7, 2]. It is based on the differences between the square root of observed and expected frequencies. The test statistic is

$$\mathrm{FT}^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left(\sqrt{n_{ij}} + \sqrt{n_{ij}+1} - \sqrt{4 \times E_{ij} + 1}\right)^2. \tag{2.3}$$

Cressie and Read [4] proposed the power divergence family as a generalization of goodness-of-fit test. It is flexible and converges to other well-known test statistics based on the choice of tuning parameter λ. The family of power divergence test statistic is

$$PD = \frac{2}{\lambda \times (\lambda + 1)} \times \sum_{i=1}^{r} \sum_{j=1}^{c} \pi_{ij} \times \left[\left(\frac{n_{ij}}{E_{ij}}\right)^{\lambda} - 1\right]. \tag{2.4}$$

The power divergence test statistic converges to Pearson's chi-square, likelihood ratio, and Freeman–Tukey statistics when λ equals 1, 0 and 0.5, respectively. They [4] suggested taking λ as 2/3, called the Cressie–Read test statistic, as being an excellent compromise between Pearson's chi-square and likelihood ratio test statistics [4]. The test statistic is

$$\mathrm{CR} = \frac{9}{5} \times \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \times \left[\left(\frac{n_{ij}}{E_{ij}}\right)^{2/3} - 1\right]. \tag{2.5}$$

In addition to the above-mentioned test statistics, we evaluated the Fisher–Freeman Halton (FFH) exact test statistic [6], which is the extension of Fisher's exact test to r-by-c tables. The Fisher–Freeman–Halton test statistic gives the exact p-value, which is calculated from sequentially generated contingency tables until one of the cells in the given margin is equal 0. This method becomes computationally intensive as the sample size increases. To overcome this problem, the Monte Carlo approach that selects samples randomly from the contingency tables is recommended [1]. In this study, we used large sample sizes. However, we benefited from the Monte-Carlo approach to decrease the computation time of the FFH test statistic.

2.2. Simulation Scenarios

We conducted a comprehensive simulation study using the R language environment [12]. We considered several factors such as sample size (n), effect size (w), and sampling design in the simulation. We used two different contingency tables, with dimensions of 5-by-5 and

5-by-2, in all simulation scenarios. Simulation scenarios consist of all possible combinations of:

- Sample size (n): $\{100, 200, 500\}$ for the 5-by-5 table and $\{40, 80, 200\}$ for the 5-by-2 table as *small*, *medium* and *large*, respectively,
- Effect size: (w): $\{0.10, 0.30, 0.50\}$ as *small*, *medium*, and *large* [3],
- Sampling design: balanced $(0.20, 0.20, 0.20, 0.20, 0.20)$,
  almost balanced $(0.15, 0.15, 0.20, 0.25, 0.25)$ and
  imbalanced row margins $(0.05, 0.05, 0.30, 0.30, 0.30)$,

where different sample sizes were used for 5-by-5 and 5-by-2 contingency tables while effect sizes and sampling designs were similar. The sample sizes were chosen so that the contingency tables were not sparse. Furthermore, the effect sizes were specified as in the literature [3]. Data were generated under product multinomial distribution via an `R` package `rTableICC` [5] by setting row marginal and total sample size fixed. Cell probabilities were specified according to changing effect size and sampling design. We compared each method using type I error rate and power. Each simulation scenario was repeated $10,000$ times. Each generated contingency table was tested with the Pearson's chi-square test, likelihood ratio test, Freeman–Tukey test, Cressie–Read test, and Fisher–Freeman–Halton's exact test. The type-I error rate of each test statistic was calculated as the proportion of false rejection obtained from $10,000$ replications when the null hypothesis was true, i.e., the effect size is $w = 0$. The power of each test, on the other hand, was calculated as the proportion of rejection obtained from $10,000$ replications assuming that the null hypothesis was false, i.e., the effect size is $w \neq 0$. The power and type-I error rate of the Pearson's chi-square test, likelihood ratio test, Freeman–Tukey test, and Cressie–Read tests statistics were obtained using the underlying Chi-square distribution. The comparison for the result of the Fisher–Freeman–Halton's exact test was evaluated using a p-value against the level of statistical significance. The statistical significance was taken as $p < 0.05$ in all simulation scenarios.

2.3. Real-life datasets

In addition to the simulation study, we evaluated the selected methods on real datasets. The first of the datasets is related to suicides. Suicides adversely affect not only the person who committed suicide, but also the people around the person, communities, and countries. According to the World Health Organization [17], suicide leads to a serious public health issue. Therefore, we decided to examine the specific causes of suicide within education level in Turkey in the year 2018. The datasets were provided by the Turkish Statistical Institute [15] and are represented in Table 2.

Nowadays, one of the major issues in the world, which is the infection of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV2), also known as COVID-19, has led to the global pandemic. Therefore, another dataset, which is taken from Ozsurekci *et al.* [10], was chosen to be used in this study. The children who were infected with or exposed to COVID-19 might have developed multisystem inflammatory syndrome (MIS-C) due to the triggering of the immune system. They compared children with MIS-C ($n = 30$) and severe/critical cases with COVID-19 ($n = 22$) in terms of respiratory support systems. This information is given in Table 3.

Table 2: Contingency table between causes of suicide and education level.

Education Level	Causes				
	Marital Conflict	Financial Difficulty	Disease	Emotional	Other
Never received formal education	9 (7.14%)	4 (1.63%)	53 (7.91%)	4 (4.71%)	53 (6.31%)
Primary School	27 (21.43%)	53 (21.54%)	155 (23.13%)	10 (11.76%)	174 (20.71%)
Secondary School	60 (47.62%)	74 (30.08%)	197 (29.40%)	37 (43.57%)	269 (32.02%)
High School	22 (17.93%)	81 (32.93%)	170 (25.37%)	23 (27.06%)	205 (24.40%)
Graduate	8 (6.35%)	34 (13.82%)	95(14.18%)	11 (12.92%)	139 (16.55%)

Table 3: Contingency table between disease group and respiratory support system.

Respiratory Support	Group	
	Cases with MIS-C	Severe/Critical cases with COVID-19
None	14 (46.67%)	6 (27.27%)
Oxygen Only	7 (23.33%)	8 (36.36%)
High Flow Support	0 (0.00%)	2 (9.09%)
Non-invasive ventilation	6 (20.00%)	0 (0.00%)
Invasive mechanical ventilation	3 (10.00%)	6 (27.27%)

3. RESULTS

The performance of the test statistics was compared according to type-I error rate and power. The power of test statistics were presented in Figures 1 and 2 while the type-I error rates were presented in Figures 3 and 4[1]. In each figure, effect sizes and sampling designs were given in the rows and columns, respectively. The test statistics were given on the x-axis and the sample size was indicated using different line type within each figure. Although we graphically presented the power and type-I error rate results in Figures 1–4, it was not easy to read exact values from corresponding figures when the points and lines were overlapped or test statistics slightly differed. Therefore, we provided the findings of Figures 1–4 with supplementary tables in the Appendix section.

When the power results are examined in Figures 1 and 2 (Tables 5 and 6 in the Appendix) for 5-by-5 and 5-by-2 contingency tables, we observe that both the effect size and the sample size have a positive effect on power of test statistics. The statistical power of methods increases with the increasing sample size and effect size. However, the sampling design has no or a considerably small effect on power for each method. Among the methods considered, the likelihood ratio test has the highest power in almost all scenarios. The Pearson's chi-square and the Cressie–Read test statistics had less power in almost all designs when the sample size was small. The power of Freeman–Tukey test decreased as the sampling design became imbalanced. We also observed that the power of the Fisher–Freeman–Halton test was higher in the imbalanced design, except for the likelihood ratio test.

[1] Figures were generated using the `ggplot2` [16] package in the `R` programming language.

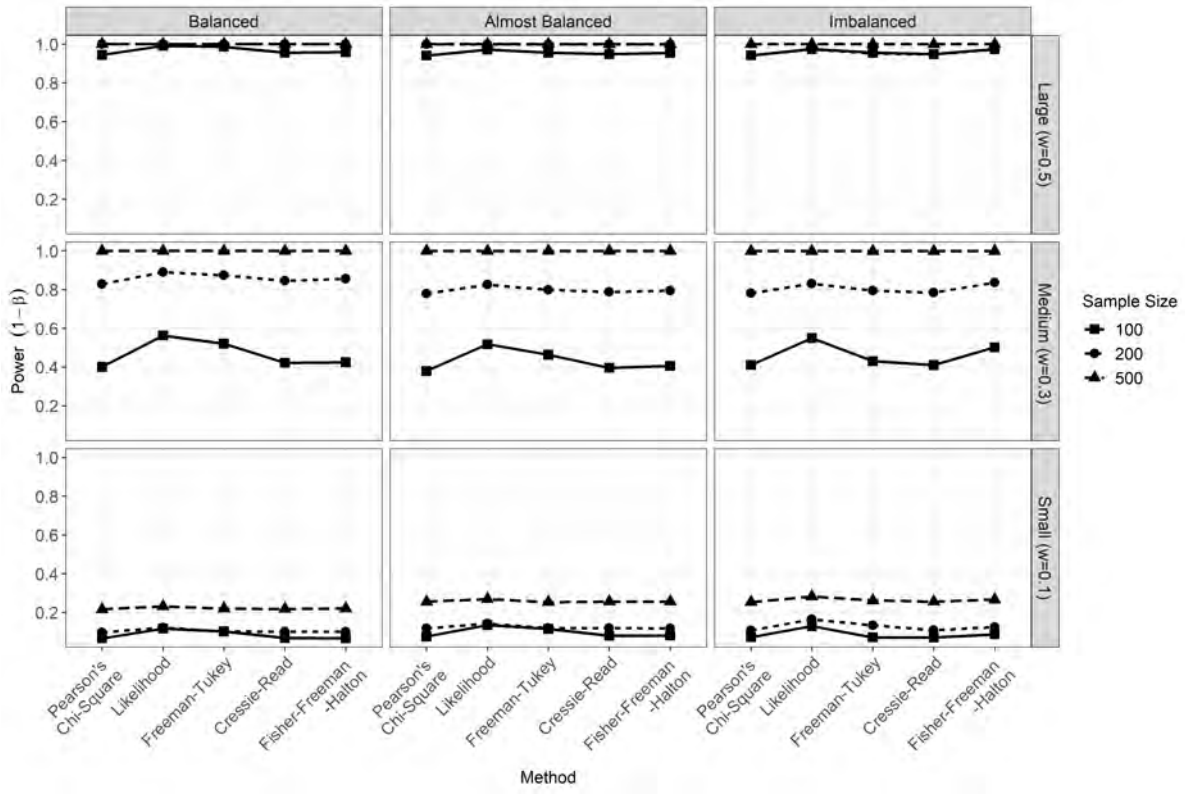

Figure 1: Simulation results – Power of tests in 5-by-5 contingency table.

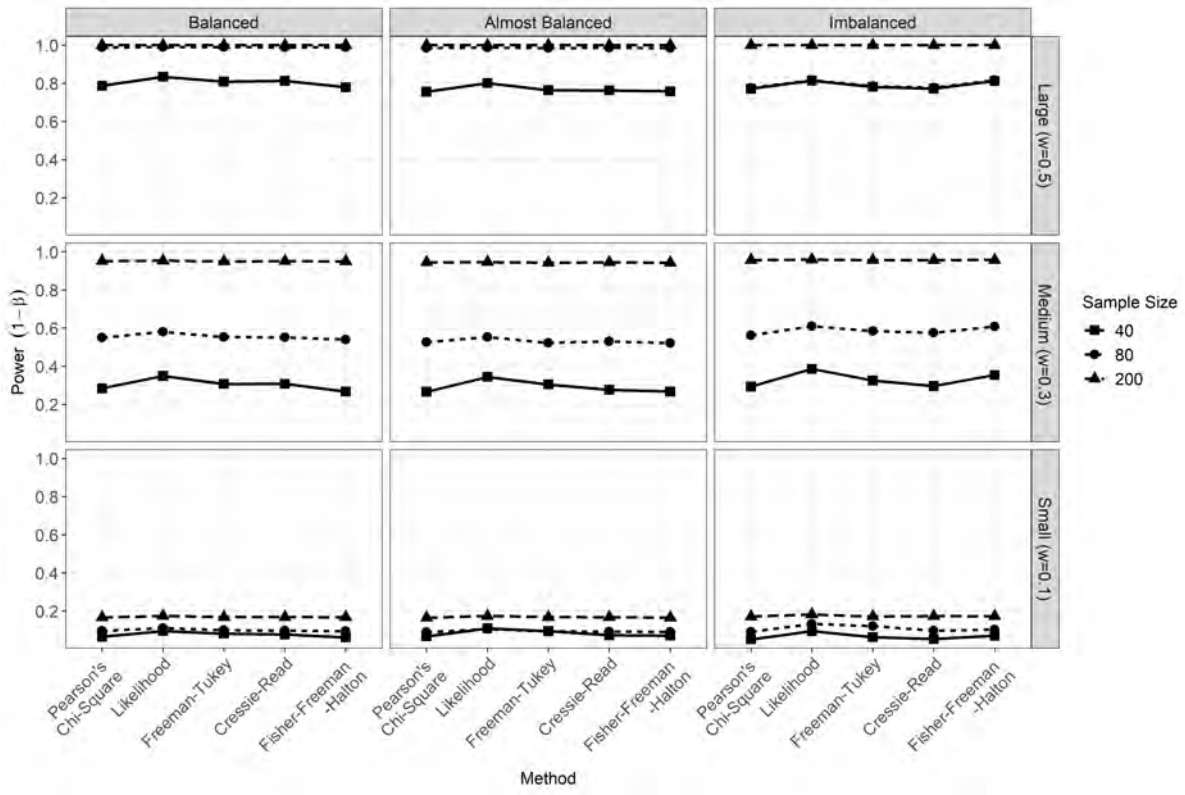

Figure 2: Simulation results – Power of tests in 5-by-2 contingency table.

The type-I error rate results of the 5-by-5 and 5-by-2 tables are given in Figures 3 and 4 (Tables 7 and 8 in the Appendix). According to the results, the likelihood ratio test was generally liberal generating type I error rates above the nominal level. Nevertheless, we observed that the type-I error rate of the likelihood ratio test was close to the nominal level as the sample size increased. In the balanced sampling design with the larger sample sizes, the type-I error rate of all test statistics, except for the likelihood ratio test statistic, was close to the nominal level. The Freeman–Tukey test statistic had a remarkably higher type-I error rate than the nominal level in small samples for balanced and almost balanced designs. However, it had the lowest type-I error rate below the nominal level in the imbalanced sampling design with a small sample size. In balanced and almost balanced designs, the Pearson's chi-square test, Cressie–Read test, and Fisher–Freeman–Halton test were better at controlling the type-I error rate at the nominal level in almost all sample sizes. However, in the imbalanced sampling design, Cressie–Reed and Pearson's chi-square test statistics were generally conservative for the small sample size and had type-I error rates closer to the nominal level as the sample size increased. Finally, the Fisher–Freeman–Halton test statistic had type-I error rates very close to the nominal level for the imbalanced sampling design.

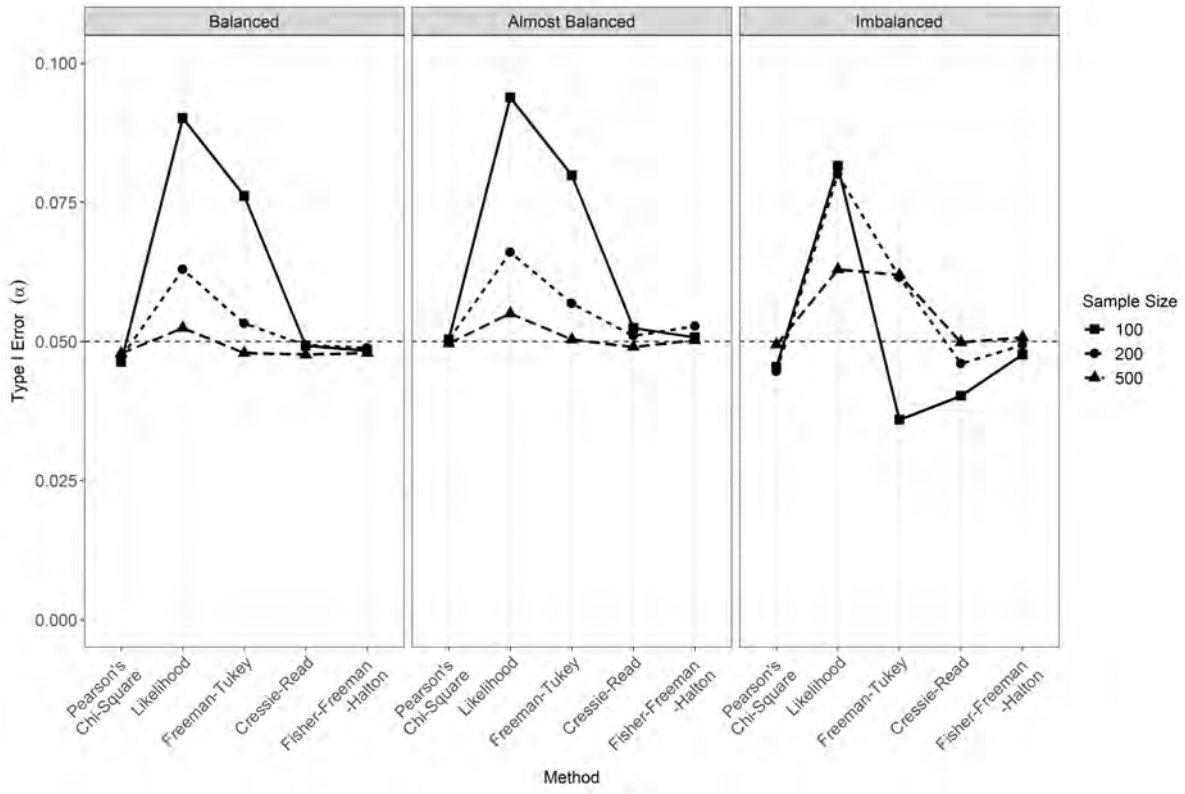

Figure 3: Simulation results – Type I error rates in 5-by-5 contingency table.

The results of real datasets are represented in Table 4. The suicide dataset (Table 2) had small effect size (i.e., $w = 0.16$), large sample size (i.e., $n = 1967$), and imbalanced design according to the row probabilities (i.e., $0.063, 40.2131, 0.324, 0.255$, and 0.146). Therefore, the suicide dataset corresponds to the simulation combination that was small effect size, large sample size, and imbalanced sampling design with the 5-by-5 table (bottom-right panel of Figure 1).

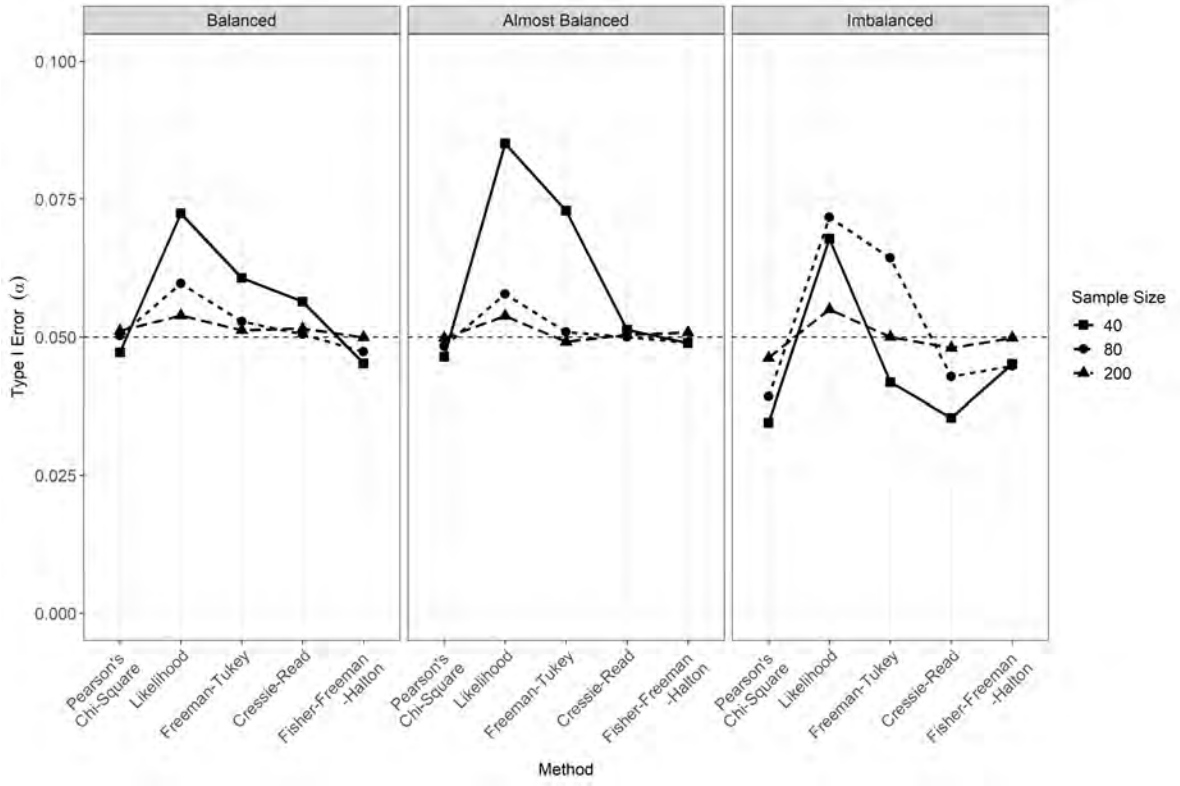

Figure 4: Simulation results – Type I error rates in 5-by-2 contingency table.

Although we found a statistically significant association between education level and suicide ($p < 0.001$ for all test statistics), the degree of association was not high ($w = 0.16$). Under this simulation scenario, the power of the Pearson and Cressie–Read test statistics was lower than the likelihood ratio test, which was similar to the real dataset results. On the other hand, the COVID-19 dataset had a large effect size (i.e., $w = 0.46$), small sample size (i.e., $n = 52$), and imbalanced sampling design according to the row probabilities (i.e., $0.385, 0.289, 0.039, 0.115$, and 0.173). This dataset corresponds to the simulation combination of large effect size, small sample size, and imbalanced sampling design with the 5-by-2 table (upper-right panel of Figure 2). In the COVID-19 dataset, all test statistics found a significant association between disease group and respiratory support system. According to the simulation results, there were slight differences between methods under a similar scenario in the COVID-19 dataset. Nonetheless, the power of likelihood ratio and Fisher–Freeman–Halton test statistics were higher than other methods. We observed results similar to simulation results in the COVID-19 dataset. The power of the likelihood ratio test statistic was the highest as compared to other methods. In addition, we saw that the Freeman–Tukey and Fisher–Freeman–Halton tests were almost similar to the likelihood ratio test.

Table 4: Results of real datasets.

Datasets	Methods								
	χ^2	p-value	G^2	p-value	FT^2	p-value	CR	p-value	FFH
Causes / Education level	48.66	<0.001	52.75	<0.001	54.01	<0.001	49.67	<0.001	0.001
Res. Support / Group	11.30	0.023	14.23	0.007	13.94	0.007	11.74	0.019	0.016

4. DISCUSSION

Previous studies in the literature evaluated the performance of various test statistics for r-by-c contingency tables. Rudas [14] compared the Pearson's chi-square, Cressie–Read, and likelihood ratio statistics for 2-by-2 and 3-by-3 tables. They reported that the Pearson's chi-square test statistic outperformed the likelihood ratio test when the sample size was small. Furthermore, they showed that the Cressie–Read and Pearson's chi-square test statistics had similar results. Parshall *et al.* [11] conducted a Monte Carlo simulation study to compare the type-I error rate and power of Pearson's chi-square, likelihood ratio, and Cressie–Read test statistics. They generated datasets from uniform distribution and found that the likelihood ratio test statistic failed to control the type I error rate at the nominal level. In addition to the previously published studies, this study considered the effects of sample size, effect size, and sampling design on the performance of various test statistics of contingency tables. A comprehensive simulation study were conducted and the findings showed that (Figures 1–4):

- The effect size and sample size were positively associated with the power of tests. The statistical power of each method increased as the number of samples or effect size increased.
- Sampling design did not affect the power of tests or slightly changed it.
- The likelihood ratio test had higher type-I error rates than the nominal level in almost all simulation scenarios. However, its statistical power was higher than other methods. We concluded that the likelihood ratio test was generally liberal, and the rejected null hypothesis should be validated using alternative methods.
- The Pearson's chi-square and Cressie–Read statistics had similar results in almost all scenarios. We mainly suggest these methods for balanced or almost balanced sampling designs when the sample size is large.
- The Fisher–Freeman–Halton (FFH) test had similar results with Pearson's chi-square and Cressie–Read tests in balanced sampling designs. However, results were promising and better than other methods in the imbalanced sampling designs. Hence, we suggest using the FFH test when the sampling design is imbalanced.
- The Freeman–Tukey (FT) test had decreased power as the sampling design became imbalanced. Even the type-I error rate was higher than the nominal level, except for the imbalanced sampling design with a small sample size, the FT test was better at controlling the type-I error rate than the likelihood ratio test.

To test the independence between variables in two-way contingency tables, one should be aware of the sampling design, the sample size, and the effect size. The power and type-I error rate are affected by those factors. The Pearson's chi-square test is a frequently used method for testing the independence in two-way contingency tables. However, we showed in our study that the Cressie–Read and Fisher–Freeman–Halton tests are efficient alternatives to the Pearson's chi-square test since they are good at controlling type-I error rates at the nominal level under certain conditions. Moreover, the power of these test statistics is as good as or better than the Pearson's chi-square test statistic. Therefore, researchers should consider the effect of the above-mentioned factors before selecting the appropriate test statistic for testing the independence in a contingency table.

Another significant issue in the analysis of the contingency tables is whether there are cells with zero observed frequencies and expected frequencies below 5. These cell frequencies affect the choice of the appropriate test statistic. In this study, we counted both the number of cells with zeros and the cells with an expected value of less than 5 for 10,000 replication data in each simulation scenario. The average number of cells with zeros and the average number of cells with expected counts below 5 were calculated specifically in the small sample size and imbalanced design for both 5-by-5 and 5-by-2 tables. The average number of cells with zeros was 4 (16%) and the average number of cells with the expected value less than 5 was 14 (56%) in the 5-by-5 tables. For the 5-by-2 tables, these values were 1 (10%) and 5 (50%), respectively. The amount of cells with lower expected counts were in the majority as expected. However, the amount of zero inflation were slight to moderate in some of simulation scenarios. This study did not account for the effect of zero inflation since it was not severe in the generated datasets. However, the effect of zero inflation should carefully considered before selecting an appropriate test statistic in contingency tables. Lydersen [8] indicated that when no more than 20 percent of the cells have an expected value below 5, the Fisher's exact test was recommended. In this study, for the small sample size and imbalanced design, we also observed that the performance of the Fisher–Freeman–Halton test statistic was better than other test statistics according to the both type-I error level and power. Therefore, we observed that simulation results are concordant with the literature [8]. As a result, for a small sample size with an imbalanced sampling design, we could say that the Fisher–Freeman–Halton test statistic is more convenient for these conditions when considering the results.

This study considered two-way contingency tables with dimensions 5-by-5 and 5-by-2. In practice, researchers wish to work with contingency tables with lower dimensions due to simplicity and less sample size. However, one may be required to work with a contingency table having rows or columns above three. For example, in medical sciences, a binary response variable such as death versus alive or healthy versus diseased might be compared between five groups which can be summarized in a contingency table with dimensions 5-by-2. Furthermore, a response variable with five categories like a 5-point Likert scale or reasons of suicides as in Table 2 might be associated with another categorical variable with five categories such as the education level. Although high-dimensional contingency tables are not frequently used or preferred in researches, they may have to be used in some studies. Therefore, the performance of test statistics in high-dimensional contingency tables should be carefully considered for selecting an appropriate test statistic. Our study provided detailed results of test statistics in high-dimensional contingency tables. Furthermore, this study can be extended to a more general case by considering the dimension of contingency tables as a new factor in the simulation scenarios.

The problem of selecting the appropriate method for testing the independence in a contingency table is not a recent topic; however, it is an ongoing issue since the performance of each method is unclear for most of the scenarios. In this study, we conducted a comprehensive simulation study considering several factors, and compared the simulation results with real data examples. We aimed to provide comparative results and bring attention to other statistical methods than Pearson's chi-square test, which is the most common in practice. We highlighted that researchers should consider various factors such as sampling design, sample size, and effect size before selecting the statistical procedures to test the independence in contingency tables. Although we covered many scenarios in the simulation study, there still exist scenarios that are not covered and the performances are unclear. Our study was not able to reflect the performance of selected methods in sparse contingency tables. We leave this topic for further research.

APPENDIX

Table 5: Simulation results – Power of tests in 5-by-5 contingency table.

Effect Size	Sampling Design	Sample Size	Pearson's Chi-Square	Likelihood	Freeman –Tukey	Cressie –Read	Fisher–Freeman –Halton
Low ($w = 0.1$)	Balanced	100	0.0658	0.1165	0.1018	0.0677	0.0667
		200	0.0968	0.1214	0.1024	0.1011	0.1002
		500	0.2161	0.2315	0.2193	0.2179	0.2200
	Almost Balanced	100	0.0757	0.1346	0.1151	0.0791	0.0801
		200	0.1172	0.1416	0.1189	0.1187	0.1169
		500	0.2550	0.2688	0.2531	0.2557	0.2561
	Imbalanced	100	0.0723	0.1289	0.0709	0.0703	0.0868
		200	0.1069	0.1638	0.1332	0.1084	0.1245
		500	0.2531	0.2820	0.2616	0.2559	0.2658
Medium ($w = 0.3$)	Balanced	100	0.4006	0.5628	0.5205	0.4221	0.4244
		200	0.8280	0.8898	0.8742	0.8449	0.8556
		500	1.0000	1.0000	1.0000	1.0000	1.0000
	Almost Balanced	100	0.3793	0.5174	0.4635	0.3958	0.4053
		200	0.7792	0.8260	0.7988	0.7874	0.7940
		500	0.9990	0.9991	0.9992	0.9990	0.9990
	Imbalanced	100	0.4104	0.5494	0.4300	0.4096	0.5020
		200	0.7810	0.8312	0.7939	0.7839	0.8366
		500	0.9985	0.9986	0.9985	0.9986	0.9988
Large ($w = 0.5$)	Balanced	100	0.9448	0.9910	0.9867	0.9566	0.9586
		200	1.0000	1.0000	1.0000	1.0000	1.0000
		500	1.0000	1.0000	1.0000	1.0000	1.0000
	Almost Balanced	100	0.9412	0.9713	0.9565	0.9487	0.9547
		200	0.9999	0.9999	0.9999	0.9999	0.9999
		500	1.0000	1.0000	1.000	1.0000	1.0000
	Imbalanced	100	0.9423	0.9745	0.9560	0.9482	0.9745
		200	0.9421	0.9734	0.9522	0.9477	0.9738
		500	1.0000	1.0000	1.0000	1.0000	1.0000

Table 6: Simulation results – Power of tests in 5-by-2 contingency table.

Effect Size	Sampling Design	Sample Size	Pearson's Chi-Square	Likelihood	Freeman –Tukey	Cressie –Read	Fisher–Freeman –Halton
Low ($w = 0.1$)	Balanced	40	0.0656	0.0950	0.0819	0.0769	0.0619
		80	0.0964	0.1103	0.1002	0.0971	0.0946
		200	0.1661	0.1743	0.1683	0.1681	0.1680
	Almost Balanced	40	0.0673	0.1089	0.0948	0.0735	0.0713
		80	0.0899	0.1055	0.0936	0.0921	0.0906
		200	0.1653	0.1735	0.1684	0.1679	0.1656
	Imbalanced	40	0.0514	0.0949	0.0634	0.0531	0.0703
		80	0.0912	0.1333	0.1211	0.0966	0.1027
		200	0.1709	0.1826	0.1720	0.1737	0.1729
Medium ($w = 0.3$)	Balanced	40	0.2837	0.3494	0.3071	0.3074	0.2674
		80	0.5500	0.5809	0.5527	0.5521	0.5403
		200	0.9513	0.9534	0.9504	0.9515	0.9502
	Almost Balanced	40	0.2663	0.3442	0.3044	0.2768	0.2672
		80	0.5260	0.5529	0.5231	0.5306	0.5216
		200	0.9449	0.9462	0.9430	0.9452	0.9429
	Imbalanced	40	0.2932	0.3855	0.3250	0.2962	0.3539
		80	0.5625	0.6094	0.5853	0.5755	0.609
		200	0.9567	0.9595	0.9571	0.9571	0.9575
Large ($w = 0.5$)	Balanced	40	0.7870	0.8343	0.8088	0.8124	0.7790
		80	0.9890	0.9907	0.9894	0.9891	0.9888
		200	1.0000	1.0000	1.0000	1.0000	1.0000
	Almost Balanced	40	0.7558	0.8012	0.7636	0.7630	0.7579
		80	0.9852	0.9868	0.9849	0.9856	0.9848
		200	1.0000	1.0000	1.0000	1.0000	1.0000
	Imbalanced	40	0.7710	0.8137	0.7815	0.7723	0.8123
		80	0.7758	0.8171	0.7850	0.7766	0.8169
		200	1.0000	1.0000	1.0000	1.0000	1.0000

Table 7: Simulation results – Type I error rates in 5-by-5 contingency table.

Sampling Design	Sample Size	Pearson's Chi-Square	Likelihood	Freeman –Tukey	Cressie –Read	Fisher–Freeman –Halton
Balanced	100	0.0463	0.0901	0.0761	0.0492	0.0483
	200	0.0471	0.0629	0.0532	0.0493	0.0488
	500	0.0478	0.0524	0.0479	0.0476	0.0479
Almost Balanced	100	0.0503	0.0938	0.0798	0.0524	0.0507
	200	0.0503	0.0660	0.0568	0.0510	0.0527
	500	0.0496	0.0550	0.0503	0.0490	0.0502
Imbalanced	100	0.0454	0.0815	0.0359	0.0402	0.0476
	200	0.0446	0.0800	0.0615	0.0460	0.0493
	500	0.0494	0.0629	0.0619	0.0498	0.0507

Table 8: Simulation results – Type I error rates in 5-by-2 contingency table.

Sampling Design	Sample Size	Pearson's Chi-Square	Likelihood	Freeman –Tukey	Cressie –Read	Fisher–Freeman –Halton
Balanced	40	0.0472	0.0724	0.0607	0.0564	0.0452
	80	0.0502	0.0597	0.0528	0.0505	0.0473
	200	0.0511	0.0539	0.0512	0.0515	0.0499
Almost Balanced	40	0.0464	0.0851	0.0729	0.0513	0.0489
	80	0.0483	0.0578	0.0509	0.0500	0.0490
	200	0.0498	0.0538	0.0491	0.0504	0.0508
Imbalanced	40	0.0345	0.0678	0.0418	0.0354	0.0451
	80	0.0392	0.0717	0.0643	0.0428	0.0447
	200	0.0462	0.0550	0.0500	0.0480	0.0498

ACKNOWLEDGMENTS

The authors are grateful and would like to thank the reviewers for their valuable comments that have increased the quality of our manuscript.

REFERENCES

[1] Agresti, B. (2002). *Categorical Data Analysis*, John Wiley & Sons, New Jersey.

[2] Bishop, Y.M.M. and Mosteller, F. (1969). *Smoothed contingency-table analysis*. In "National Halothane Study: a Study of the Possible Association Between Halothane Anesthesia and Postoperative Hepatic Necrosis" (J.P. Bunker; W.H. Forrest; F. Mosteller and L.D. Vandam, Eds.), National Research Council, Washington, DC, The National Academies Press, 237–286.

[3] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, Routledge, United Kingdom.

[4] Cressie, N.A.C. and Read, T.R.C. (1984). Multinomial goodness-of-fit tests, *Journal of the Royal Statistical Society Series B (Methodological)*, **46**(3), 141–149.

[5] Demirhan, H. (2016). rTableICC: an R package for random generation of 2×2×K and R×C contingency tables, *The R Journal*, **8**(1), 48–63.

[6] Freeman, G.H. and Halton, J.H. (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance, *Biometrika*, **38**(1/2), 141–149.

[7] Freeman, M.F. and Tukey, J.W. (1950). Transformations related to the angular and the square root, *The Annals of Mathematical Statistics*, **21**(4), 607–611.

[8] Lydersen, P.S.; Senchaudhuri, P.V. and Laake, P. (2007). Choice of test for association in small sample unordered r × c tables, *Statistics in Medicine*, **26**(23), 4328–4343.

[9] Oyeyemi, G.M.; Adewara, A.A.; Adebola, F.B. and Salau, S.I. (2010). On the estimation of power and sample size in test of independence, *Asian Journal of Mathematics & Statistics*, **3**(3), 139–146.

[10] Ozsurekci, Y.; Gürlevik, S.; Kesici, S.; Akca, U.K.; Oygar, P.D.; Aykac, K.; Karacanoglu, D.; Saritas, N.O.; Ilbay, S.; Katlan, B.; Ertugrul, İ.; Cengiz, A.B.; Basaran, O.; Cura, Y.B.C.; Karakaya, J.; Bilginer, Y.; Bayrakci, B.; Ceyhan, M. and Ozen, S. (2021). Multisystem inflammatory syndrome in children during the COVID-19 pandemic in Turkey: first report from the Eastern Mediterranean, *Clin Rheumatol*, **12**, 1–11.

[11] Parshall, C.D.; Kromrey, J.D. and Dailey, R. (1999). Comparative performance of three statistical tests of homogeneity for sparse i × j contingency tables, *Communications in Statistics – Simulation and Computation*, **28**(1), 275–289.

[12] R Core Team. (2018). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Austria.
`http://www.R-project.org/`

[13] Read, T.R.C. and Cressie, N.A.C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer-Verlag, New York.

[14] Rudas, T. (1986). A Monte Carlo comparison of the small sample behaviour of the Pearson, the likelihood ratio and the Cressie–Read statistics, *Journal of Statistical Computation and Simulation*, **24**(2), 107–120.

[15] TURKSTAT. (2019). *TURKSTAT: Distribution of selected causes of suicides with respect to gender and education level in 2018*, Turkish Statistical Institute.

[16] Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag, New York.

[17] World Health Organization. (2019). *World Health Organization: Suicide*, World Health Organization.

REVSTAT – Statistical Journal
Volume 21, Number 2, April 2023, 235–266
https://doi.org/10.57805/revstat.v21i2.402

Conditional Evaluations of Sums of Sample Maxima and Records

Authors: Tomasz Rychlik
– Institute of Mathematics, Polish Academy of Sciences,
Śniadeckich 8, 00656 Warsaw, Poland
trychlik@impan.pl

Magdalena Szymkowiak ✉
– Institute of Automatic Control and Robotics, Poznan University of Technology,
Plac Marii Skłodowskiej-Curie 5, 60965 Poznań, Poland
magdalena.szymkowiak@put.poznan.pl

Received: October 2020 Revised: November 2021 Accepted: November 2021

Abstract:

- We consider sequences of independent and identically absolutely continuously distributed random variables assuming that they have finite expectation and variance. We determine sharp lower and upper bounds on the expectation of the sum of n first sample maxima and n first upper record values under the condition that the value of the j-th $(1 \leq j \leq n)$ sample maximum and record value, respectively, are known and equal to a given quantile of the parent distribution. The bounds are expressed in terms of the expectation and standard deviation of the parent distribution. Analogous evaluations are presented for the sum of record values in n observations, when the j-th sample maximum is known. The theoretical results are numerically compared.

Keywords:

- *sample maximum; upper record; conditional expectation; bound.*

AMS Subject Classification:

- 60E15, 62G32.

✉ Corresponding author.

1. INTRODUCTION

Let $X_1, ..., X_n, ...$ denote i.i.d. random variables with a common absolutely continuous distribution function F and density function f, say. We assume that they have a finite second moment. Let $M_n = \max\{X_1, ..., X_n\}$, $n = 1, 2, ...$, stand for the maximum of the first n observations. For fixed $1 \le j \le n$ and $0 < q < 1$, we determine tight lower and upper bounds for the standardized versions of the conditional expectations

$$\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n} M_k - n\mu \middle| M_j = F^{-1}(q)\right) \tag{1.1}$$

over all parent distribution functions F where μ and σ denote the respective mean and standard deviation. It is clear that manipulating with location and scale of the parent distribution function F, we may obtain arbitrarily large and small values of the conditional expectation in (1.1), and a proper standardization allows to get rid of the trivial extremes. We chose the mean and standard deviation of the parent distribution as the most classic location and scale parameters, respectively. Normalization (1.1) allows us to get rid of dependence on location and scale, and its variability depends only on the shape of the parent distribution. It is also intuitively clear that the conditional expectation depends on the location of M_j in the support of X_1, and its distribution over the support. This is well expressed by the order of respective quantile.

A similar problem is solved for the upper records. We define the first record time and value as $T_1 = 1$ and $R_1 = X_1$, respectively. The further record times and values are determined recursively $T_n = \min\{k > T_{n-1} : X_k > M_{k-1}\}$, and $R_n = X_{T_n} = M_{T_n}$. By definition, the sequence of upper records is the maximal increasing subsequence of the non-decreasing sequence of sample maxima, arisen by crossing out all the repetitions. The second problem we cope with here is evaluating

$$\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n} R_k - n\mu \middle| R_j = F^{-1}(q)\right), \qquad 1 \le j \le n,\ 0 < q < 1. \tag{1.2}$$

For describing our last problem, we introduce the record indicators $\eta_k = 1$ if $X_k > M_{k-1}$ and $\eta_k = 0$ otherwise. Our purpose is to evaluate

$$\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n} X_k\eta_k - n\mu \middle| M_j = F^{-1}(q)\right), \qquad 1 \le j \le n,\ 0 < q < 1. \tag{1.3}$$

Expression $\sum_{k=1}^{n} X_k\eta_k$ represents the sum of all the record values observed among the first n observations. Condition $M_j = F^{-1}(q)$ means that the actual record value after j observations amounts to $F^{-1}(q)$.

Exemplary applications of our problems are connected with sponsoring and rewarding sportsmen.

Example. Some sports disciplines consists in gaining the greatest possible results. The examples are here the track and field competitions in jumping and throwing. The sportsmen receive scholarships and rewards proportional to (or linearly dependent on) their achievements. Suppose that due to an agreement with a sponsor a person receives a scholarship in the period of n months based on sports level which is measured by his/her personal best result.

(In the meantime he/she can achieve worse results, but it is known that he/she is able to attain the results on the level of his/her personal best.) Therefore his/her joint earnings in n months are proportional to $\sum_{i=1}^{n} M_i$.

In the second case, the sponsoring company signs the agreement with the organizer of a competition series that it pays honoraria for n consecutive records during the competitions of the amounts linearly dependent on the values of records. The sum of payoffs is a linear function of $\sum_{i=1}^{n} R_i$ then.

Another variant of the agreement is that the company sponsors n track and field meetings so that it pays a random number of honoraria to the people who gain new records during these events. The total amount of the rewards is proportional to $\sum_{i=1}^{n} X_i\eta_i$, where X_i is the result of the winner in the i-th competition, and η_i is the respective record indicator.

We try to evaluate the total sums of payments in these three models on the basis of knowledge of j-th value of the payment, $1 \le j \le n$, which are M_j, R_j and M_j again, respectively, but we do not know a substantially random mechanism generating the results. However, such generally stated problems do not have nontrivial solutions. We should know at least approximate values of the location and scale parameters. Therefore we included the mean and standard deviation in the models, which are the most popular parameters of location and scale. Also, one other factor specific to a given sport discipline should be taken into account. For instance, it is obvious that one can expect more progress in the triple jump or hammer throw in the ladies competitions rather than among the men, because the women version of these sport competitions were introduced quite recently. Mathematically, the tendency of the given discipline for gaining new records is expressed the small value of the quantile order q of the parent distribution.

We solve our three problems using a similar approach. We represent expressions (1.1)–(1.3) in integral forms, depending on indices j and n, quantile order q, and parent distribution function F. Then for fixed j, n, and q, we determine the lower and upper bounds on the integrals representing (1.1)–(1.3), and distribution functions F which attain the bounds. These distributions have atoms, and formally do not satisfy the continuity assumptions. However, if we skilfully spread out (uniformly, for simplicity) the atom masses over their small neighborhoods preserving the parent mean and variance, we may attain values of conditional expectations arbitrarily close to the respective bounds by an absolutely continuous distributions. This means that our bounds are optimal: there are sequences of continuous distributions tending weakly to a discontinuous ones which approach respective bounds arbitrarily close. For brevity of presentation, we merely present these limiting discontinuous distributions, and imprecisely write that the bounds are attained by them.

The integral bounds are calculated with use of the method proposed in Moriguti [21] who used it for evaluating the expectations of order statistics from i.i.d. samples and their differences.

Lemma 1.1. *Let H be a non-decreasing right-continuous function on an interval $[a, b]$, and continuous at a and b. Let $\overline{H}$ and $\underline{H}$ be the smallest concave majorant, and the greatest convex minorant of H, respectively. Let $\overline{h}$ and $\underline{h}$ denote the the right-hand side derivatives of $\overline{H}$ and $\underline{H}$, respectively. Then for every non-decreasing function f on $[a, b]$ we have*

$$\int_a^b f(x)\overline{h}(x)dx \le \int_a^b f(x)H(dx) \le \int_a^b f(x)\underline{h}(x)dx \tag{1.4}$$

under the assumption that the integrals exist and are finite. The lower (upper) bound is attained iff f is constant in every interval of the open set $\{x \in [a,b] : \overline{H}(x) > H(x)\}$ ($\{x \in [a,b] : \underline{H}(x) < H(x)\}$, respectively), and $f(x)$ is left-continuous (right-continuous, resp.) at every discontinuity point (if any) of H.

Moriguti ([21], Theorem 1) determined the upper bound in (1.4) under a more general assumption that H has a bounded variation on $[a,b]$, and is continuous at the interval ends. The lower one is easily concluded from Theorem 1 of Moriguti [21]:

$$-\int_a^b f(x)H(dx) = \int_a^b f(x)(-H)(dx) \le -\int_a^b f(x)\overline{h}(x)dx,$$

because $-\overline{h}$ is the derivative of the greatest convex minorant of $-H$.

Order statistics, especially sample extremes, and records were the objects of extensive studies. Arnold *et al.* [2], and David and Nagaraja [9] are the most popular textbooks devoted to order statistics. Comprehensive studies of records were presented in Arnold *et al.* [3] and Nevzorov [23]. Gumbel [13] and Hartley and David [14] independently derived sharp upper mean-variance bounds on the maxima of i.i.d. random variables. Analogous estimates for the record values were presented in Nagaraja [22]. These bounds were determined with use of the Schwarz inequality. Applying the same tool one can establish analogous bounds on sums of maxima and records, but the respective analytic formulae are complicated.

Predictions of order statistics and record values were analyzed by Raqab and Balakrishnan [28], Ahmadi and Balakrishnan [1], MirMostafaee and Ahmadi [20], and Volterman *et al.* [31], among others. In particular, Rychlik [29] and Klimczak [17] determined bounds on conditional expectations of future order statistics and records. Balakrishnan *et al.* [5], Asgharzadeh *et al.* [4], Khatib and Ahmadi [15], and Khatib *et al.* [16] studied reconstructions of previous failure times and records in various models. Klimczak and Rychlik [18] presented evaluations of conditional expectations of previous order statistics and records.

Conditional expectations of (1.1), (1.2), and (1.3) are studied in survival analysis, the gambling, finance, and reliability theories. A problem of prediction of the sum of minima (dual to (1.1)) was treated by Nevzorov *et al.* [24]. Problem (1.3) is a modification of a classical secretary choice problem which consist in maximizing the probability of finding the maximal record value in a finite sequence of i.i.d. observation in an on-line decision procedure (see, e.g., Gilbert and Mosteller [11] or Chow *et al.* [8]). Various generalizations of the secretary problem can be found in Freeman [10] and Samuels [30]. Recent developments in the subject are presented in Ramsey [27], Kuchta [19], Woryna [32], and Grau Ribas [12]. Sums of records in fixed numbers of trials were treated in Bel'kov and Nevzorov [6]. For a fixed parent distribution, they maximized $\mathbb{E}\left(\sum_{k=j}^{n} X_k\eta_k | X_1, ...X_j\right)$ with respect to $j = 1, ..., n-1$. Nevzorov and Tovmasyan [26] analyzed a similar problem if the number of upper records was maximized instead of the sum of their values. Bel'kov and Nevzorov [7] maximized the joint sum of upper and lower records in the analogous model. Nevzorov and Stepanov [25] maximized the expected sum of maxima by choosing an optimal starting time.

Evaluations of (1.1), (1.2) and (1.3) are presented in Sections 2, 3, and 4, respectively. Section 5 is devoted to numerical comparisons.

2. SUMS OF SAMPLE MAXIMA

Lemma 2.1. *Let $X_1, ..., X_n$ be i.i.d. with an absolutely continuous distribution function F and density f. Then $\mathbb{E}\big(\frac{1}{n}\sum_{k=1}^{n} M_k \big| M_j = x\big)$ is identical with the expectation of the distribution function*

$$
\begin{aligned}
F_{j,n,F}(y|x) &= \begin{cases} \dfrac{1}{jn}\displaystyle\sum_{k=1}^{j-1}(j-k)\frac{F^k(y)}{F^k(x)}, & y < x, \\ \dfrac{j}{n} + \dfrac{1}{n}\displaystyle\sum_{k=1}^{n-j} F^k(y), & y \geq x. \end{cases} \\
&= \begin{cases} \dfrac{1}{jn}\,\dfrac{F(x)F(y)}{[F(y)-F(x)]^2}\left[j-1-j\dfrac{F(y)}{F(x)}+\dfrac{F^j(y)}{F^j(x)}\right], & y < x, \\ \dfrac{j}{n} + \dfrac{1}{n}\,\dfrac{F(y)}{1-F(y)}[1-F^{n-j}(y)], & y \geq x. \end{cases}
\end{aligned} \tag{2.1}
$$

We adhere to the convention that $\sum_{k=i}^{j} a_i = 0$ for $j < i$.

Proof: For $j < k$ we have

$$\mathbb{P}(M_k = x | M_j = x) = \mathbb{P}(X_i \leq x,\ i = j+1, ..., k) = F^{k-j}(x),$$

and for $x < y$ yields

$$\mathbb{P}(M_j \leq x, M_k \leq y) = \mathbb{P}(X_i \leq x,\ i = 1, ..., j,\ X_i \leq y,\ i = j+1, ..., k) = F^j(x)F^{k-j}(y).$$

Therefore the joint density function of M_j and M_k is

$$f_{M_j,M_k}(x,y) = j(k-j)F^{j-1}(x)F^{k-j-1}(y)f(x)f(y), \qquad x < y. \tag{2.2}$$

Since

$$f_{M_j}(x) = jF^{j-1}(x)f(x), \tag{2.3}$$

the conditional density of M_k under condition $M_j = x$ has the form

$$f_{M_k|M_j}(y|x) = (k-j)F^{k-j-1}(y)f(y), \qquad y > x,$$

and the respective conditional distribution function is

$$F_{M_k|M_j}(y|x) = \begin{cases} 0, & y < x, \\ F^{k-j}(y), & y \geq x. \end{cases} \tag{2.4}$$

Take now $k < j$. Using the exchangeability argument we conclude that $\mathbb{P}(M_k = x \,|\, M_j = x) = \frac{k}{j}$ for any x. Applying (2.2) and (2.3) we also obtain

$$f_{M_k|M_j}(y|x) = \frac{k(j-k)}{j}\,\frac{F^{k-1}(y)}{F^k(x)}, \qquad y < x.$$

It follows that the conditional distribution function of M_k with respect to $M_j = x$ is

$$F_{M_k|M_j}(y|x) = \begin{cases} \dfrac{j-k}{j}\,\dfrac{F^k(y)}{F^k(x)}, & y < x, \\ 1, & y \geq x. \end{cases} \tag{2.5}$$

Obviously, the distribution of M_j given $M_j = x$ is the degenerate measure concentrated at x. Combing this fact with (2.4) and (2.5), we get

$$\sum_{k=1}^{n} F_{M_k|M_j}(y|x) = \begin{cases} \dfrac{1}{j}\displaystyle\sum_{k=1}^{j-1}(j-k)\frac{F^k(y)}{F^k(x)}, & y < x, \\ j + \displaystyle\sum_{k=1}^{n-j} F^k(y), & y \geq x. \end{cases}$$

Finally,

$$\mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n} M_k \middle| M_j = x\right) = \int_{\mathbb{R}} y \sum_{k=1}^{n} F_{M_k|M_j}(dy|x) = \int_{\mathbb{R}} y\, F_{j,n,F}(dy|x). \qquad \square$$

Distribution function (2.1) in the standard uniform case has the form

$$\begin{aligned} F_{j,n}(u|q) &= \begin{cases} \dfrac{1}{jn}\displaystyle\sum_{k=1}^{j-1}(j-k)\frac{u^k}{q^k}, & 0 < u < q, \\ \dfrac{j}{n} + \dfrac{1}{n}\displaystyle\sum_{k=1}^{n-j} u^k, & q \leq u < 1, \end{cases} \\ &= \begin{cases} \dfrac{1}{jn}\,\dfrac{uq}{(u-q)^2}\left[j - 1 - j\dfrac{u}{q} + \dfrac{u^j}{q^j}\right], & u < q, \\ \dfrac{j}{n} + \dfrac{1}{n}\,\dfrac{u}{1-u}[1-u^{n-j}], & u \geq q, \end{cases} \qquad 0 < q < 1. \end{aligned} \tag{2.6}$$

It has the density function

$$\begin{aligned} f_{j,n}(u|q) &= \begin{cases} \dfrac{1}{jnq}\displaystyle\sum_{k=0}^{j-2}(j-k-1)(k+1)\frac{u^k}{q^k}, & 0 < u < q, \\ \dfrac{1}{n}\displaystyle\sum_{k=0}^{n-j-1}(k+1)u^k, & q \leq u < 1, \end{cases} \\ &= \begin{cases} \dfrac{(j+1)q}{jn(q-u)^2}\left[1 - j\dfrac{u^{j-1}}{q^{j-1}} + (j-1)\dfrac{u^j}{q^j}\right] - \dfrac{q^2}{jn(q-u)^3} \\ \times\left[2 - j(j+1)\dfrac{u^{j-1}}{q^{j-1}} + 2(j-1)(j+1)\dfrac{u^j}{q^j} - j(j-1)\dfrac{u^{j+1}}{q^{j+1}}\right], & 0 < u < q, \\ \dfrac{1}{(1-u)^2}[1 - (n+1-j)u^{n-j} + (n-j)u^{n+1-j}], & q \leq u < 1, \end{cases} \end{aligned} \tag{2.7}$$

when $0 < q < 1$, and the jump of height $\frac{j+1}{2n} + \frac{1}{n}\sum_{k=1}^{n-j} q^k = \frac{j+1}{2n} + \frac{1}{n}\,\frac{q}{1-q}(1-q^{n-j})$ at q.

We also note that

$$F_{j,n}(q-|q)=\frac{j-1}{2n},\quad F_{j,n}(q|q)=\frac{j}{n}+\frac{1}{n}\sum_{k=1}^{n-j}q^k=\frac{j}{n}+\frac{1}{n}\,\frac{q}{1-q}(1-q^{n-j}), \tag{2.8}$$

$$f_{j,n}(0|q)=\frac{j-1}{jnq},\quad f_{j,n}(1|q)=\frac{(n-j)(n-j+1)}{2n}, \tag{2.9}$$

$$f_{j,n}(q-|q)=\frac{(j-1)(j+1)}{6nq},\quad f_{j,n}(q+|q)=\frac{1}{n}\sum_{k=0}^{n-j-1}(k+1)q^k$$

$$=\frac{1-(n+1-j)q^{n-j}+(n-j)q^{n+1-j}}{n(1-q)^2}.$$

Before we formulate the main results of this section, we define some auxiliary notions. Put

$$j_*=j_*(n)=\frac{2n+1-\sqrt{8n+1}}{2}, \tag{2.10}$$

$$If_{j,n}(u|q)=\int_0^u f_{j,n}^2(v|q)dv$$

$$=\frac{1}{(jnq)^2}\int_0^u\left[\sum_{k=0}^{j-2}(j-k-1)(k+1)\frac{v^k}{q^k}\right]^2dv$$

$$=\frac{1}{(jnq)^2}\sum_{r=0}^{2j-4}\frac{1}{(r+1)q^r}\left[\sum_{k=\max\{1,r-j+3\}}^{\min\{r+1,j-1\}}k(j-k)(r-k+2)(j-r+k-2)\right]u^{r+1} \tag{2.11}$$

for $0<u\le q$, and

$$Jf_{j,n}(u|q)=\int_u^1 f_{j,n}^2(v|q)dv$$

$$=\frac{1}{n^2}\int_u^1\left[\sum_{k=0}^{n-j-1}(k+1)u^k\right]^2dv$$

$$=\frac{1}{n^2}\sum_{r=0}^{2(n-j-1)}\frac{1}{(r+1)}\left[\sum_{k=\max\{1,r-n+j+2\}}^{\min\{r+1,n-j\}}k(r-k+2)\right](1-u^{r+1}) \tag{2.12}$$

for $q\le u<1$. We first describe the upper bounds for $2\le j\le n-1$. The extreme cases $j=1$ and $j=n$ are treated separately.

Theorem 2.1. *Let $X_1,\dots,X_n$ be i.i.d. with some distribution and density functions F and f, mean μ and variance σ^2. Fix $2\le j\le n-1$, and $0<q<1$.*

(i) *If $j_*\le j\le n-1$ (see (2.10)), and $q\le\frac{j-1}{jn}$, then*

$$\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n}M_k-n\mu\,\middle|\,M_j=F^{-1}(q)\right)\le 0. \tag{2.13}$$

The bound is attained in limit by sequences of continuous distributions tending to degenerate ones.

(ii) *Assume that $q>\frac{j-1}{jn}$ and either of two conditions holds. One is $j_*\le j\le n-1$, and the other is $2\le j<j_*$ with the assumption that the equation*

$$f_{j,n}(1|q)(u-1)+1=F_{j,n}(u|q) \tag{2.14}$$

has a solution in $(0,q)$.

(**a**) *If moreover* $\frac{j-1}{jn} < q < \frac{(j-1)(j+1)}{6n+(j-1)(j-2)}$ *then the equation*

$$1 - F_{j,n}(u|q) = (1-u) f_{j,n}(u|q) \tag{2.15}$$

has a unique solution $0 < u_* < q$, *and then*

$$\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n} M_k - n\mu \middle| M_j = F^{-1}(q)\right) \le nA_{j,n}(q), \tag{2.16}$$

where

$$A^2_{j,n}(q) = If_{j,n}(u_*|q) + f^2_{j,n}(u_*|q)(1-u_*) - 1. \tag{2.17}$$

The equality in (2.16) is attained by the parent distribution with the quantile function

$$F^{-1}(u) = \mu + \frac{\sigma}{A_{j,n}(q)}[f_{j,n}(\min\{u, u_*\}|q) - 1]. \tag{2.18}$$

(**b**) *However, if* $q \ge \frac{(j-1)(j+1)}{6n+(j-1)(j-2)}$, *then (2.16) holds with*

$$A^2_{j,n}(q) = If_{j,n}(q|q) + \frac{[1 - F_{j,n}(q - |q)]^2}{1-q} - 1, \tag{2.19}$$

and attainability condition

$$F^{-1}(u) = \mu + \frac{\sigma}{A_{j,n}(q)} \times \begin{cases} f_{j,n}(u|q) - 1, & u < q, \\ \dfrac{1 - F_{j,n}(q-|q)}{1-q}, & u \ge q. \end{cases} \tag{2.20}$$

(**iii**) *Suppose that* $2 \le j < j_*$, *and either of two assumptions holds. The first is* $q \le \frac{j-1}{jn}$. *The other admits* $q > \frac{j-1}{jn}$, *but demands that the equation*

$$f_{j,n}(0|q)u = F_{j,n}(u|q) \tag{2.21}$$

has a solution in $(q, 1)$ *then. In consequence, the equation*

$$F_{j,n}(u|q) = uf_{j,n}(u|q) \tag{2.22}$$

has a unique solution $q < u_{**} < 1$, *and (2.16) holds with*

$$A^2_{j,n}(q) = f^2_{j,n}(u_{**}|q)u_{**} + Jf_{j,n}(u_{**}|q) - 1. \tag{2.23}$$

In this case the bound in (2.16) is attained by the distribution with the quantile function

$$F^{-1}(u) = \mu + \frac{\sigma}{A_{j,n}(q)}[f_{j,n}(\max\{u, u_{**}\}|q) - 1]. \tag{2.24}$$

(**iv**) *Finally, let* $2 \le j < j_*$, *and* $q > \frac{j-1}{jn}$, *and equations (2.15) and (2.22) do not have solutions in* $(0, q)$ *and* $(q, 1)$, *respectively.*

(**a**) *If moreover the equation*

$$f_{j,n}(q - |q)(u - q) + F_{j,n}(q - |q) = F_{j,n}(u|q) \tag{2.25}$$

has a solution in $(q, 1)$, *though, then there exist unique* $0 < u_* < q < u_{**} < 1$ *satisfying the equations*

$$f_{j,n}(u_*|q) = f_{j,n}(u_{**}|q) = \frac{F_{j,n}(u_{**}|q) - F_{j,n}(u_*|q)}{u_{**} - u_*}, \tag{2.26}$$

and (2.16) holds with

$$A^2_{j,n}(q) = If_{j,n}(u_*|q) + f^2_{j,n}(u_*|q)(u_{**} - u_*) + Jf_{j,n}(u_{**}|q) - 1. \tag{2.27}$$

The equality in (2.16) is attained then if

$$F^{-1}(u) = \mu + \frac{\sigma}{A_{j,n}(q)} \times \begin{cases} f_{j,n}(u_*|q) - 1, & u_* \le u \le u_{**}, \\ f_{j,n}(u|q) - 1, & \text{otherwise.} \end{cases} \tag{2.28}$$

(**b**) *If (2.25) does not have a solution in* $(q, 1)$, *then there exists a unique* $q < u_{**} < 1$ *such that*

$$f_{j,n}(q - |q) < \frac{F_{j,n}(u_{**}|q) - F_{j,n}(q - |q)}{u_{**} - q} = f_{j,n}(u_{**}|q), \tag{2.29}$$

and (2.16) holds with

$$A^2_{j,n}(q) = If_{j,n}(q|q) + \frac{[F_{j,n}(u_{**}|q) - F_{j,n}(q - |q)]^2}{u_{**} - q} + Jf_{j,n}(u_{**}|q) - 1, \tag{2.30}$$

and the equality in (2.10) holds for

(2.31)

$$F^{-1}(u) = \mu + \frac{\sigma}{A_{j,n}(q)} \times \begin{cases} \dfrac{F_{j,n}(u_{**}|q) - F_{j,n}(q - |q)}{u_{**} - q} - 1, & q \le u < u_{**}, \\ f_{j,n}(u|q) - 1, & \text{otherwise.} \end{cases}$$

Proof: By Lemma 2.1,

$$\begin{aligned} n\mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n} M_k - \mu \middle| M_j = F^{-1}(q)\right) &= n\int_{\mathbb{R}} (y - \mu) F_{j,n,F}(dy|F^{-1}(q)) \\ &= n\int_0^1 [F^{-1}(u) - \mu] F_{j,n}(du|q). \end{aligned} \tag{2.32}$$

For using Lemma 1.1, we need to determine the greatest convex minorant of (2.6). This distribution function is convex on the intervals $[0, q)$ and $[q, 1]$, and has a jump up at q. We easily notice that the greatest convex minorant may have four possible shapes. It is certainly linear near q and possibly identical with $F_{j,n}(u|q)$ at the ends of $[0, 1]$. However, it may happen that the linear part reaches either of the end-points of the interval, or even the line may cover the whole interval.

(i) Then problem is most simple when $f_{j,n}(1|q) \le 1 \le f_{j,n}(0|q)$, i.e. when $j \ge j_*$ and $q \le \frac{j-1}{jn}$ (cf. (2.9)). Then the straight line $\ell(u) = u$, $0 \le u \le 1$, connects the points $(0, F_{j,n}(0|q)) = (0,0)$ and $(1, F_{j,n}(1|q)) = (1,1)$, and runs beneath $F_{j,n}(u|q)$ in between. It follows that the line is the greatest convex minorant of (2.6), and its derivative amounts to constant 1. Therefore

$$\mathbb{E}\left(\sum_{k=1}^{n} M_k - n\mu \middle| M_j = F^{-1}(q)\right) \le n\int_0^1 [F^{-1}(u) - \mu]du = 0.$$

This proves inequality (2.13). In order to prove its optimality, for simplicity we consider the family of two-point distributions with the quantile functions

$$F_\varepsilon^{-1}(u) = \mu + \sigma \times \begin{cases} -\sqrt{\dfrac{1-\varepsilon}{\varepsilon}}, & u < \varepsilon, \\ \sqrt{\dfrac{\varepsilon}{1-\varepsilon}}, & u \ge \varepsilon, \end{cases} \qquad 0 < \varepsilon < 1. \tag{2.33}$$

Applying the de l'Hospital rule and boundedness of $f_{j,n}(u|q)$ near 0, we obtain

$$\begin{aligned} &\lim_{\varepsilon\to 0}\int_0^1 [F_\varepsilon^{-1}(u) - \mu]F_{j,n}(du|q) \\ &= \sigma \lim_{\varepsilon\to 0}\left[-\sqrt{\varepsilon(1-\varepsilon)}\,\frac{F_{j,n}(\varepsilon|q)}{\varepsilon} + \sqrt{\frac{\varepsilon}{1-\varepsilon}}\,[1 - F_{j,n}(\varepsilon|q)]\right] = 0. \end{aligned}$$

This argument shows that the zero bound is optimal if the greatest convex minorant is linear. We shall not repeat it in the future proofs. Note that here the same conclusion could be derived if we locate the vanishing atom on the right.

Now we observe that each of equations (2.14) and (2.21) has at most two solutions in $(0,q)$ and $(q,1)$, respectively, because their left-hand sides are linear, and the right-hand sides are strictly convex. We also note that existence of solutions to (2.14) excludes that for (2.21) and vice-versa. Assume for instance that u_0 is the solution (the single one or the smaller of two) to (2.14). It follows that

$$F_{j,n}(u|q) > f_{j,n}(1|q)(u-1) + 1, \qquad q < u < 1. \tag{2.34}$$

The straight line $f_{j,n}(0|q)u$ runs below the point $(u_0, F_{j,n}(u_0|q))$, and line $f_{j,n}(1|q)(u-1)+1$ right to u_0. By (2.34), it cannot meet $F_{j,n}(u|q)$ in $(q,1)$. When (2.21) has a solution, we argue in a similar way to exclude that of (2.14).

(ii) Let $f_{j,n}(0|q) < 1$, i.e., $q > \frac{j-1}{jn}$. Assume moreover that either $f_{j,n}(1|q) \le 1$ $(j \ge j_*)$ or $f_{j,n}(1|q) > 1$ $(j < j_*)$ holds together with existence of solution to (2.14). It follows that then the greatest convex minorant of $F_{j,n}(u|q)$ coincides first with the function itself, and then with the straight line $f_{j,n}(1|q)(u-1)+1$ (at least on $[q,1)$). The change point u_* amounts to q if

$$f_{j,n}(q-|q) \le \frac{1 - F_{j,n}(q-|q)}{1-q}. \tag{2.35}$$

(a) If $\frac{j-1}{2n} < q < \frac{(j-1)(j+1)}{6n+(j-1)(j-2)}$, we have the reversed inequality in (2.35). Function

$$\underline{F}_{j,n}(u|q) = \begin{cases} F_{j,n}(u|q), & u \le q, \\ \dfrac{1 - F_{j,n}(q-|q)}{1-q}\,(u-1) + 1, & u \ge q, \end{cases} \tag{2.36}$$

is not convex then. However, there exist $u_* < q$ such that the line $\frac{1-F_{j,n}(u_*|q)}{1-u_*} \times (u-1)+1$ connecting the points of the graph of $F_{j,n}(u|q)$ at u_* and 1 runs below the graph, and is tangent to it at u_*. This provides the change point of the minorant, and is certainly determined by (2.15). When $\frac{j-1}{jn} < q \le \frac{j-1}{2n}$, we have $F_{j,n}(q-|q) \ge q$ which again implies that the change point is u_* defined in (2.15). It follows that for $\frac{j-1}{jn} < q < \frac{(j-1)(j+1)}{6n+(j-1)(j-2)}$ the derivative of the greatest convex minorant of (2.6) has the form $\underline{f}_{j,n}(u|q) = f_{j,n}(\min\{u, u_*\}|q)$. Coming back to (2.32) we obtain

$$\begin{aligned}\mathbb{E}\left(\sum_{k=1}^{n} M_k - n\mu \middle| M_j = F^{-1}(q)\right) &\le n\int_0^1 [F^{-1}(u)-\mu][\underline{f}_{j,n}(u|q)-1]du \\ &\le n\left(\int_0^1 [F^{-1}(u)-\mu]^2 du \int_0^1 [\underline{f}_{j,n}(u|q)-1]^2 du\right)^{\frac{1}{2}} \\ &= n\sigma\left(\int_0^1 \underline{f}^2_{j,n}(u|q)du - 1\right)^{\frac{1}{2}}. \qquad (2.37)\end{aligned}$$

The last equality follows from the fact that $\underline{f}_{j,n}(u|q)$ integrates to $\underline{F}_{j,n}(1|q) = 1$ on the interval $[0,1]$. Using (2.11) we easily check that $\int_0^1 \underline{f}^2_{j,n}(u|q)du - 1 = A^2_{j,n}(q)$ defined in (2.17). The equality in the latter inequality of (2.37) holds when

$$F^{-1}(u) - \mu = \alpha[\underline{f}^2_{j,n}(u|q) - 1], \qquad 0 < u < 1, \qquad (2.38)$$

for some $\alpha > 0$. Note that the right-hand side is right-continuous and integrates to 0, which allows to preserve the expectation condition for the left-hand side. The variance assumption implies $\alpha = \frac{\sigma}{A_{j,n}(q)}$. Observe that condition (2.38) for the equality in the latter Schwarz inequality of (2.37) preserves constancy intervals of the derivative of the greatest convex minorant which is necessary for satisfying the first equality condition of Lemma 1.1 in the first upper Moriguti inequality of (2.37). The other is satisfied as well since we defined the right continuous version of the quantile function in (2.38). It follows that (2.18) actually defines the parent distribution for which the bound (2.16) with the right-hand side defined in (2.17) is attained.

This approach is used in our further investigations. Determination of the greatest convex minorant is the crucial step in the evaluation method. The upper bound coincides with the Hilbert norm of its derivative decreased by one, and a proper linear modification of this function defines the quantile function of the distribution which satisfies the moment conditions and attains the bound. For brevity, in our further studies we stop calculations once we define the greatest convex minorant of a suitable integrand, and tacitly refer to the procedure described in the previous paragraph.

(b) Using (2.8) and (2.9) we check that (2.35) is satisfied when $q \ge \frac{(j-1)(j+1)}{6n+(j-1)(j-2)}$. Note that (2.35) implies $F_{j,n}(q-|q) < q$. Indeed, relation $F_{j,n}(q-|q) \ge q$ forces $\frac{1-F_{j,n}(q-|q)}{1-q} \le 1$, and $f_{j,n}(q-|q) > 1$. The latter is a consequence of the fact that $F_{j,n}(u|q)$ crosses then the line $\ell(u) = u$ from bottom to top in $(0,q)$. Its derivative is necessarily greater than 1 at the crossing point, and increases later on. Also, relation $q \ge \frac{(j-1)(j+1)}{6n+(j-1)(j-2)}$ implies that (2.36)

is actually a convex function, and it forms the greatest convex minorant of $F_{j,n}(u|q)$. Its right-hand side derivative is

$$\underline{f}_{j,n}(u|q) = \begin{cases} f_{j,n}(u|q), & u < q, \\ \dfrac{1 - F_{j,n}(q-|q)}{1-q} = \dfrac{2n+1-j}{2n(1-q)}, & u \geq q. \end{cases}$$

Following the arguments presented above we conclude that in this case we obtain the bound in (2.16) defined by (2.19) and its attainability condition are described in (2.20).

(iii) Under the assumptions of this point, $F_{j,n}(u|q)$ runs below the line $\ell(u) = u$ in some left neighborhood of 1. If $q \leq \frac{j-1}{jn}$, the function is located above the line for all $0 < u < q$. Therefore the greatest convex minorant has to be first linear, and then identical with $F_{j,n}(u|q)$. When $q > \frac{j-1}{jn}$ and $f_{j,n}(0|q) < 1$ in consequence, but (2.21) holds for some $q < u < 1$, then $F_{j,n}(u|q)$, $0 < u < q$, lies above the line $f_{j,n}(0|q)u$, but this is not true for some $u \in (q, 1)$. Again we deduce that the minorant is first linear and eventually strictly convex. The change point belongs to $(q, 1)$, and q is impossible. This point is determined by (2.22) which means that the linear part of the greatest convex minorant is tangent to $F_{j,n}(u|q)$ at the change point u_{**}. The derivative of the convex minorant is then $\underline{f}_{j,n}(u|q) = f_{j,n}(\max\{u, u_{**}\}|q)$. Proceeding as in the previous part on the proof we determine the mean-variance bound for the conditional expectation and the condition of its attainability.

(iv) The assumptions mean that $F_{j,n}(u|q)$ goes below $\ell(u) = u$ in some neighborhoods of 0 and 1. Moreover, the lines tangent to $F_{j,n}(u|q)$ at 0 and 1 run below the graph of the function. This implies that the greatest convex minorant of $F_{j,n}(u|q)$ cannot be linear at vicinities of the end-points. So the linear part may appear only in the central part, and it contains q.

(a) If (2.25) has a solution then the derivative of the greatest convex minorant of $F_{j,n}(u|q)$ can be written as

$$\underline{f}_{j,n}(u|q) = \begin{cases} f_{j,n}(u_*|q), & u_* \leq u \leq u_{**}, \\ f_{j,n}(u|q), & \text{elsewhere}, \end{cases}$$

where $0 < u_* < q < u_{**} < 1$ are determined from the tangency conditions (2.26). In the standard way we establish the bound in (2.16) with the right-hand side described in (2.27), and the attainability condition (2.28).

(b) The lack of solution to (2.25) implies that all the lines $u \mapsto f_{j,n}(v|q)(u-v) + F_{j,n}(v|q)$, tangent to $F_{j,n}(u|q)$ at $v < q$ run below $F_{j,n}(u|q)$ for $u < v$. The only candidate for the change point of the minorant from $F_{j,n}(u|q)$ into a line is q. Consider the functions $\ell_\alpha(u) = \alpha(u-q) + F_{j,n}(q-|q)$, and increase the slopes α starting from $f_{j,n}(q-|q)$ until we touch any point of $F_{j,n}(u|q)$ for $u \geq q$. Obviously q cannot be the first meeting point, because the line connecting $F_{j,n}(q-|q)$ and $F_{j,n}(q|q)$ is vertical. It cannot be 1, either, because then $\alpha = \frac{1-F_{j,n}(q-|q)}{1-q} \geq f_{j,n}(1|q)$ which contradicts the assumption that (2.14) does not

have a solution in $(0, q)$. Consequently, our assumptions imply

$$\underline{f}_{j,n}(u|q) = \begin{cases} \dfrac{F_{j,n}(u_{**}|q) - F_{j,n}(q-|q)}{u_{**} - q}, & q \le u \le u_{**}, \\ f_{j,n}(u|q), & \text{elsewhere}, \end{cases}$$

where u_{**} is determined by solving (2.29). This allows us to conclude (2.16) with (2.30) and (2.31) assuring the equality in (2.16). □

We separately consider the extreme cases $j = 1$ and $j = n$, for which the distribution function (2.6) does not have any mass on the left and right, respectively, of q. This allows us to simplify the arguments of the above proof in order to get desired conclusions. The details of the reasoning are left to the reader.

Theorem 2.2. *Let the assumptions of Theorem 2.1 hold.*

(**i**) *Let* $M_1 = F^{-1}(q)$ *for some* $0 < q < 1$.

(**a**) *If* $q < \frac{n-3}{n-1}$, *then there exists* $q < u_{**} < 1$ *solving the equation*

$$\begin{aligned} \frac{u}{1-u}(1-u^{n-1}) &= nF_{1,n}(u|q) = n(u-q)f_{1,n}(u|q) \\ &= \frac{u-q}{(1-u)^2}[1 - nu^{n-1} + (n-1)u^n], \end{aligned}$$

and then we have (2.16) with $j = 1$ *and*

$$A^2_{1,n}(q) = f^2_{1,n}(u_{**}|q)(u_{**} - q) + Jf_{1,n}(u_{**}|q) - 1$$

(see (2.12)). The equality is attained if

$$F^{-1}(u) = \mu + \frac{\sigma}{A_{1,n}(q)} \times \begin{cases} -1, & u < q, \\ f_{1,n}(u_{**}|q) - 1, & q \le u \le u_{**}, \\ f_{1,n}(u|q) - 1, & u \ge u_{**}. \end{cases}$$

(**b**) *If* $q \ge \frac{n-3}{n-1}$, *then* $A_{1,n}(q) = \sqrt{\frac{q}{1-q}}$, *and the bound is attained by the two-point parent distribution on* $\mu - \sigma\sqrt{\frac{1-q}{q}}$ *and* $\mu + \sigma\sqrt{\frac{q}{1-q}}$ *with respective probabilities* q *and* $1 - q$.

(**ii**) *Under the condition* $M_n = F^{-1}(q)$, *there are three possible cases.*

(**a**) *When* $q \le \frac{n-1}{n^2}$ *(cf. Theorem 2.1*(i)*), the optimal upper bound on the standardized expectation of the first* n *maxima is equal to* 0.

(**b**) *If* $\frac{n-1}{n^2} < q < \frac{n-1}{n+2}$ *then the statements of Theorem 2.1*(iia) *hold with* j *replaced by* n.

(**c**) *If* $q \ge \frac{n-1}{n+2}$ *then the statements of Theorem 2.1*(iib) *hold with* j *replaced by* n.

In the following theorem we describe the lower bounds on the conditional expectations of sample maxima for all $1 \le j \le n$.

Theorem 2.3. *Assume the conditions of Theorem 2.1.*

(i) *Let $0 < q_* \le 1$ be the unique solution to*

$$j + \frac{q}{1-q}(1-q^{n-j}) = nq. \tag{2.39}$$

If either $j_* \le j \le n$ (comp. (2.10)) or $1 \le j < j_*$ and $q < q_*$ defined above, then

$$\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n} M_k - n\mu \middle| M_j = F^{-1}(q)\right) \ge -\frac{j + \frac{q}{1-q}(1-q^{n-j}) - nq}{\sqrt{q(1-q)}}. \tag{2.40}$$

The lower bound in (2.40) is attained by the two-point distribution supported on $\mu - \sigma\sqrt{\frac{1-q}{q}}$ and $\mu + \sigma\sqrt{\frac{q}{1-q}}$ with respective probabilities q and $1-q$.

(ii) *If $1 \le j < j_*$ and $q \ge q_*$ then the optimal bound is*

$$\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n} M_k - n\mu \middle| M_j = F^{-1}(q)\right) \ge 0.$$

Note that $q_* = 1$ for $j = n$, and so (2.40) holds for all $0 < q < 1$ with the right hand-side simplified to $-n\sqrt{\frac{1-q}{q}}$.

Proof: We rewrite representation (2.32), and apply the lower estimate of Lemma 1.1. In consequence of the shape of the distribution function (2.6), the only possible shapes of its smallest concave majorant is the linear function $\ell(u) = u$ when $F_{j,n}(q|q) \le q$ or a broken line with the break point q otherwise. Inequality $F_{j,n}(q|q) \le q$ is equivalent to

$$j + \sum_{k=1}^{n-j} q^k - nq \le 0. \tag{2.41}$$

The left-hand side function is strictly convex, positive at 0, and vanishing at 1. Its derivative at 1 amounts to $\frac{(n-j)(n-j+1)}{2} - n = \frac{1}{2}\left[j^2 - (2n+1)j + n(n-1)\right]$ which is non-positive for $j \ge j_*$ (see (2.10)), and positive otherwise. Accordingly, inequality (2.41) is false for all $0 < q < 1$ when $j \ge j_*$. If $j < j_*$, (2.41) holds only for sufficiently large q. Precisely, this is true for $q \ge q_*$ defined in (2.39).

(i) Assume so that either $j_* \le j \le n$ or $1 \le j < j_*$ and $q < q_*$. Then the smallest concave majorant has the form

$$\overline{F}_{j,n}(u|q) = \begin{cases} \dfrac{F_{j,n}(q|q)}{q}\,u = \dfrac{j + \sum_{k=1}^{n-j} q^k}{nq}\,u, & u \le q, \\ \dfrac{1 - F_{j,n}(q|q)}{1-q}(u-1) + 1 = \dfrac{n - j - \sum_{k=1}^{n-j} q^k}{n(1-q)}(u-1) + 1, & u \ge q. \end{cases}$$

We use

$$\overline{f}_{j,n}(u|q) - 1 = \begin{cases} \dfrac{j + \sum_{k=2}^{n-j} q^k - (n-1)q}{nq}, & u \le q, \\ \dfrac{(n-1)q - j - \sum_{k=2}^{n-j} q^k}{n(1-q)} & u > q, \end{cases} \tag{2.42}$$

for establishing the following lower mean-variance bound

$$n\mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n} M_k \middle| M_j = F^{-1}(q)\right) \geq n\int_0^1 [F^{-1}(u) - \mu][\overline{f}_{j,n}(u|q) - 1]du$$

$$\geq -n\left[\int_0^1 [F^{-1}(u) - \mu]^2 du\right]^{1/2}\left[\int_0^1 [\overline{f}_{j,n}(u|q) - 1]^2 du\right]^{1/2} = -n\sigma a_{j,n}(q), \tag{2.43}$$

where

$$a_{j,n}^2(q) = \int_0^1 [\overline{f}_{j,n}(u|q) - 1]^2 du = \frac{1}{n^2 q(1-q)}\left(j + \sum_{k=1}^{n-j} q^k - nq\right)^2.$$

Note that under the assumption the expression in the parentheses is positive. Now set

$$F^{-1}(u-) - \mu = -\frac{\sigma}{a_{j,n}(q)}[\overline{f}_{j,n}(u|q) - 1] \tag{2.44}$$

which asserts the equalities in both the inequalities of (2.43). Note that the right-hand side of (2.44) is non-decreasing and left-continuous. Moreover, its integral over $[0, 1]$ is equal to 0, and the integral of its square amounts to 1. This implies that the left-hand side determines the standardized lower quantile function of a distribution with mean μ and variance σ^2. Plugging (2.42) into (2.44) we obtain

$$F^{-1}(u-) = \mu + \sigma \times \begin{cases} -\sqrt{\dfrac{1-q}{q}}, & u \leq q, \\ \sqrt{\dfrac{q}{1-q}}, & u > q, \end{cases}$$

which describes the two-point distribution defined in the first part of Theorem 2.3.

(ii) Otherwise, if $1 \leq j < j_*$ and $q \geq q_*$, the derivative of the smallest concave majorant $\overline{F}_{j,n}(u|q) = u$, $0 \leq u \leq 1$, of $F_{j,n}(u|q)$ is equal to 1. Consequently,

$$n\mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n} M_k \middle| M_j = F^{-1}(q)\right) \geq n\int_0^1 [F^{-1}(u) - \mu]du = 0. \qquad \square$$

Seemingly, the conditions for attaining the upper bounds in Theorem 2.2(ib) and Theorem 2.3(i) pretend to be identical. There are subtle differences between them, though. In the first case, the strictly increasing quantile functions $F_n^{-1}(u)$ should tend to the right-continuous version of the two-valued extreme quantile function. In the other one, they should tend to the left-continuous lower quantile function. We omit presenting elementary constructions of such sequences.

3. SUMS OF UPPER RECORDS

Lemma 3.1. *Let $X_1, ..., X_n, ...$ be i.i.d. with an absolutely continuous distribution function F and density f, and let $R_1, ..., R_n$ denote the values of the first upper records in the sequence. Then $\mathbb{E}\big(\frac{1}{n}\sum_{k=1}^{n} R_k \big| R_j = x\big)$ for some $1 \le j \le n$ is identical with the expectation of the distribution function*

$$G_{j,n,F}(y|x) = \begin{cases} \dfrac{j-1}{n}\,\dfrac{-\ln[1-F(y)]}{-\ln[1-F(x)]}, & y < x, \\ 1 - \dfrac{1}{n}\,\dfrac{1-F(y)}{1-F(x)} \displaystyle\sum_{k=0}^{n-j-1} \frac{n-j-k}{k!}\left[-\ln\frac{1-F(y)}{1-F(x)}\right]^k, & y \ge x. \end{cases} \tag{3.1}$$

Proof: The density function of the single record value R_j, and the joint density of a pair (R_j, R_k), $j < k$, have the forms

$$f_{R_j}(x) = \frac{\{-\ln[1-F(x)]\}^{j-1}}{(j-1)!}\, f(x), \tag{3.2}$$

$$f_{R_j,R_k}(x,y) = \frac{\{-\ln[1-F(x)]\}^{j-1}}{(j-1)!}\,\frac{\left[-\ln\frac{1-F(y)}{1-F(x)}\right]^{k-j-1}}{(k-j-1)!}\,\frac{f(x)f(y)}{1-F(x)}, \quad x < y, \tag{3.3}$$

respectively (see, e.g., Arnold *et al.* [3], p. 11). It follows that for $j < k$

$$f_{R_k|R_j}(y|x) = \frac{\left[-\ln\frac{1-F(y)}{1-F(x)}\right]^{k-j-1}}{(k-j-1)!}\,\frac{f(y)}{1-F(x)}, \qquad y > x,$$

is the conditional density function of R_k under the condition that $R_j = x$. We see that the conditional distribution is identical with the unconditional distribution of the $(k-j)$-th record value from a sequence with the left-truncated parent distribution function $\frac{1-F(y)}{1-F(x)}$, $y > x$. The respective distribution function is

$$F_{R_k|R_j}(y|x) = 1 - \frac{1-F(y)}{1-F(x)} \sum_{i=0}^{k-j-1} \frac{[-\ln[\frac{1-F(y)}{1-F(x)}]^i}{i!}, \qquad x < y.$$

We also note that

$$\sum_{k=j+1}^{n} F_{R_k|R_j}(y|x) = n - j - \frac{1-F(y)}{1-F(x)} \sum_{k=0}^{n-j-1} \frac{n-j-k}{k!}\left[-\ln\frac{1-F(y)}{1-F(x)}\right]^k, \quad x < y. \tag{3.4}$$

Referring again to (3.2) and (3.3), we obtain

$$\begin{aligned} f_{R_k|R_j}(y|x) = {} & \frac{(j-1)!}{(k-1)!(j-k-1)}\left[\frac{-\ln[1-F(y)]}{-\ln[1-F(x)]}\right]^{k-1} \\ & \times \left[1 - \frac{-\ln[1-F(y)]}{-\ln[1-F(x)]}\right]^{j-k-1} \frac{-f(y)}{[1-F(y)]\ln[1-F(x)]} \end{aligned}$$

for $y < x$ and $k < j$. This coincides with the density function of the k-th order statistic from an i.i.d. sample of size $j-1$ from the right-truncated distribution function $\frac{-\ln[1-F(y)]}{-\ln[1-F(x)]}$, $y < x$.

Obviously, the sum of ordered variables is identical with that of the original unordered ones. Therefore

$$\mathbb{E}\left(\sum_{k=1}^{j-1} R_k \middle| R_j = x\right) = (j-1)\int_{-\infty}^{x} y\,\frac{-\ln[1-F(dy)]}{-\ln[1-F(x)]} \tag{3.5}$$

Combining (3.4), (3.5) with the trivial fact $\mathbb{E}(R_j|R_j = x) = x = \int_{\mathbb{R}} y\mathbf{1}_{[x,+\infty)}(dy)$, we conclude

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n} R_k \middle| R_j = x\right)\\
&= \frac{j-1}{n}\int_{-\infty}^{x} y\,\frac{-\ln[1-F(dy)]}{-\ln[1-F(x)]} + \frac{1}{n}\int_{\mathbb{R}} y\mathbf{1}_{[x,+\infty)}(dy) + \frac{1}{n}\int_{x}^{+\infty} y\sum_{k=j+1}^{n} F_{R_k|R_j}(dy|x)\\
&= \int_{\mathbb{R}} yF_{j,n,F}(dy|x),
\end{aligned}$$

which proves our statement. □

In the standard uniform case (3.1) takes on the form

$$G_{j,n}(u|q) = \begin{cases} \dfrac{j-1}{n}\,\dfrac{-\ln(1-u)}{-\ln(1-q)}, & 0 < u < q < 1,\\[2ex] 1 - \dfrac{1}{n}\,\dfrac{1-u}{1-q}\displaystyle\sum_{k=0}^{n-j-1}\frac{n-j-k}{k!}\left(-\ln\frac{1-u}{1-q}\right)^k, & 0 < q \le u < 1. \end{cases} \tag{3.6}$$

It has the density function

$$g_{j,n}(u|q) = \begin{cases} \dfrac{j-1}{n}\,\dfrac{1}{-\ln(1-q)}\,\dfrac{1}{1-u}, & 0 < u < q < 1,\\[2ex] \dfrac{1}{n(1-q)}\displaystyle\sum_{k=0}^{n-j-1}\frac{1}{k!}\left(-\ln\frac{1-u}{1-q}\right)^k, & 0 < q \le u < 1, \end{cases}$$

and an atom with the weight $\frac{1}{n}$ at q. In particular we have

$$G_{j,n}(q-|q) = \frac{j-1}{n}, \quad G_{j,n}(q|q) = \frac{j}{n},$$

$$g_{j,n}(0|q) = \frac{j-1}{-n\ln(1-q)}, \quad g_{j,n}(1-|q) = \begin{cases} +\infty, & j < n-1,\\ \dfrac{1}{n(1-q)}, & j = n-1,\\ 0, & j = n, \end{cases} \tag{3.7}$$

$$g_{j,n}(q-|q) = \frac{j-1}{-n(1-q)\ln(1-q)}, \quad g_{j,n}(q+|q) = \frac{1}{n(1-q)}.$$

We also define

$$Ig_{j,n}(u|q) = \int_0^u g_{j,n}^2(v|q)dv = \left[\frac{j-1}{-n\ln(1-q)}\right]^2\int_0^u\frac{1}{(1-v)^2}dv = \left[\frac{j-1}{-n\ln(1-q)}\right]^2\frac{u}{1-u}, \tag{3.8}$$

for $0 < u \le q$, and

$$
\begin{aligned}
Jg_{j,n}(u|q) &= \int_u^1 g_{j,n}^2(v|q)dv = \frac{1}{n^2(1-q)^2}\int_u^1 \left[\sum_{k=0}^{n-j-1}\frac{1}{k!}\left(-\ln\frac{1-v}{1-q}\right)^k\right]^2 dv \\
&= \frac{1}{n^2(1-q)^2}\sum_{r=0}^{2(n-j-1)}\left[\sum_{k=\max\{0,r-n+j+1\}}^{\min\{r,n-j-1\}}\binom{r}{k}\right]\frac{1}{r!}\int_u^1\left(-\ln\frac{1-u}{1-q}\right)^r dv \\
&= \frac{1-u}{n^2(1-q)^3}\sum_{r=0}^{2(n-j-1)}\left[\sum_{k=\max\{0,r-n+j+1\}}^{\min\{r,n-j-1\}}\binom{r}{k}\right]\left[\sum_{k=0}^{r}\frac{1}{k!}\left(-\ln\frac{1-u}{1-q}\right)^k\right] \qquad (3.9)
\end{aligned}
$$

for $q \le u < 1$. Note that for $r = 0, ..., n-j-1$, the sum of binomial coefficients in the first square brackets of (3.9) amounts to 2^r.

Theorem 3.1. *Let $X_1, ..., X_n, ...$ be i.i.d. with some distribution and density functions F and f, mean μ and variance σ^2. Fix $2 \le j \le n-2$, and $0 < q < 1$.*

(i) *Suppose that either of two assumptions holds. One is $q \le 1 - \exp(-\frac{j-1}{n})$. The other is $q > 1 - \exp(-\frac{j-1}{n})$ and the equation*

$$g_{j,n}(0|q)u = G_{j,n}(u|q) \qquad (3.10)$$

has a solution in $(q, 1)$. Then the equation

$$G_{j,n}(u|q) = ug_{j,n}(u|q)$$

has a unique solution $q < u_{**} < 1$, and we have

$$\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n} R_k - n\mu \middle| R_j = F^{-1}(q)\right) \le nB_{j,n}(q) \qquad (3.11)$$

with

$$B_{j,n}^2(q) = g_{j,n}^2(u_{**}|q)u_{**} + Jg_{j,n}(u_{**}|q) - 1. \qquad (3.12)$$

In this case the bound in (3.11) is attained by the distribution with the quantile function

$$F^{-1}(u) = \mu + \frac{\sigma}{B_{j,n}(q)}[g_{j,n}(\max\{u, u_{**}\}|q) - 1]. \qquad (3.13)$$

(ii) *Assume that $q > 1 - \exp(-\frac{j-1}{n})$ and the equation (3.10) does not have a solution in $(q, 1)$.*

(a) *If moreover there exists in $(q, 1)$ a solution to the equation*

$$g_{j,n}(q-|q)(u-q) + G_{j,n}(q-|q) = G_{j,n}(u|q) \qquad (3.14)$$

then there is a unique pair $0 < u_* < q < u_{**} < 1$ satisfying the equations

$$g_{j,n}(u_*|q) = g_{j,n}(u_{**}|q) = \frac{G_{j,n}(u_{**}|q) - G_{j,n}(u_*|q)}{u_{**} - u_*}, \qquad (3.15)$$

and (3.11) holds with

$$B^2_{j,n}(q) = Ig_{j,n}(u_*|q) + g^2_{j,n}(u_*|q)(u_{**} - u_*) + Jg_{j,n}(u_{**}|q) - 1.$$

The equality in (3.11) is attained then if

$$F^{-1}(u) = \mu + \frac{\sigma}{B_{j,n}(q)} \times \begin{cases} g_{j,n}(u_*|q) - 1, & u_* \le u \le u_{**}, \\ g_{j,n}(u|q) - 1, & \text{otherwise.} \end{cases}$$

(**b**) *If (3.14) does not have a solution in $(q, 1)$, then there is a unique $q < u_{**} < 1$ such that*

$$g_{j,n}(q - |q) < \frac{G_{j,n}(u_{**}|q) - G_{j,n}(q - |q)}{u_{**} - q} = g_{j,n}(u_{**}|q), \tag{3.16}$$

and (3.11) holds with

$$B^2_{j,n}(q) = Ig_{j,n}(q|q) + \frac{[G_{j,n}(u_{**}|q) - G_{j,n}(q - |q)]^2}{u_{**} - q} + Jg_{j,n}(u_{**}|q) - 1, \tag{3.17}$$

whereas the equality in (3.11) is attained for

$$F^{-1}(u) = \mu + \frac{\sigma}{B_{j,n}(q)} \times \begin{cases} \dfrac{G_{j,n}(u_{**}|q) - G_{j,n}(q - |q)}{u_{**} - q} - 1, & q \le u < u_{**}, \\ g_{j,n}(u|q) - 1, & \text{otherwise.} \end{cases} \tag{3.18}$$

The idea of proof of Theorem 3.1 as well as the following results is similar to that of Theorem 2.1. Therefore we sketch only the main points focusing merely on the differences.

Proof: Since $g_{j,n}(1 - |q) = +\infty$ for $j \le n - 2$, we can exclude the possibilities that the greatest convex minorant of (3.6) is linear at the neighborhood of 1.

(i) Suppose that either $g_{j,n}(0|q) \ge 1$ (i.e., $q \le 1 - \exp\left(-\frac{j-1}{n}\right)$, comp. (3.7)) or $g_{j,n}(0|q) < 1$ but the line tangent to $G_{j,n}(u|q)$ at 0 meets $G_{j,n}(u|q)$ somewhere in $(q, 1)$. This implies that there is a line located below it in the positive half-axis which runs through $(0, 0)$ and is tangent to $G_{j,n}(u|q)$ at some u_{**} in $(q, 1)$. Its segment joining $(0, 0)$ with $(u_{**}, G_{j,n}(u_{**}|q))$ extended by $G_{j,n}(u|q)$ itself on the right composes the greatest convex minorant of $G_{j,n}(u|q)$. This observation allows us to determine the bound (3.12) in (3.11), and the condition of its attainability (3.13) (comp. (2.16), (2.23) and (2.24)).

(ii) If $q > 1 - \exp\left(-\frac{j-1}{n}\right)$ and $g_{j,n}(0|q)u$ runs below $G_{j,n}(u|q)$ on $(q, 1)$, then the convex minorant should coincide with the original function on a right neighborhood of 0 as well as that of 1, and be linear in between. There are two possible subcases.

(a) If the line tangent to $G_{j,n}(u|q)$ at $q-$ runs above $G_{j,n}(u|q)$ on the whole $(q, 1)$ (i.e., (3.14) does hold), the point where the minorant transforms into a line has to be less than q. The linear part should be tangent to the graph of $G_{j,n}(u|q)$ at the both its ends. Therefore the end points $u_* < q < u_{**}$ are determined by equations (3.15). Once we fix the convex minorant we are in a position to calculate the sharp upper bound on the conditional expectation, and the parent distribution which attains it.

(b) In the opposite case, the linear part starts at q, and its right end u_{**} is determined by the tangency condition (3.16). This provides the bound defined in (3.17) and its attainability condition described in (3.18). □

Below we present without a proof the upper bounds for conditional expectations of the sum of first upper records under condition $R_j = F^{-1}(q)$ for remaining $j = 1$, $n-1$, and n.

Theorem 3.2. *Suppose that $X_1, ..., X_n, ...$ satisfy the assumptions of Theorem 3.1.*

(i) *There exists $q < u_{**} < 1$ solving the equation*

$$g_{1,n}(u|q)(u-q) = G_{1,n}(u|q)$$

such that

$$\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n} R_k - n\mu \middle| R_1 = F^{-1}(q)\right) \le nB_{1,n}(q) \tag{3.19}$$

where

$$B_{1,n}^2(q) = g_{1,n}^2(u_{**}|q)(u_{**}-q) + Jg_{1,n}(u_{**}|q) - 1.$$

The equality in (3.19) holds for the distribution function F satisfying

$$F^{-1}(u) = \mu + \frac{\sigma}{B_{1,n}(q)} \times \begin{cases} -1, & u < q, \\ g_{1,n}(u_{**}|q) - 1, & q \le u < u_{**}, \\ g_{j,n}(u|q) - 1, & u \ge u_{**}. \end{cases}$$

(ii) *For $j = n-1$ and $j = n$ we have the following.*

(a) *If $q \le 1 - \exp\left(-\frac{j-1}{n}\right)$, then the optimal bound is*

$$\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n} R_k - n\mu \middle| R_j = F^{-1}(q)\right) \le 0.$$

(b) *If $1 - \exp\left(-\frac{j-1}{n}\right) < q < 1 - \exp\left(-\frac{j-1}{n+1-j}\right)$, then (3.11) holds with*

$$B_{j,n}^2(q) = Ig_{j,n}(u_*|q) + g_{j,n}^2(u_*|q)(1-u_*) - 1$$

for $0 < u_* < q$ satisfying the equation

$$g_{j,n}(u|q)(1-u) = 1 - G_{j,n}(u|q).$$

The condition for getting the equality in (3.11) is

$$F^{-1}(u) - \mu + \frac{\sigma}{B_{j,n}(q)}[g_{j,n}(\min\{u, u_*\}|q)].$$

(c) *Finally, if $q \ge 1 - \exp\left(-\frac{j-1}{n+1-j}\right)$, then (3.11) holds with*

$$B_{j,n}^2(q) = Ig_{j,n}(q|q) + \frac{[1 - G_{j,n}(q-|q)]^2}{1-q} - 1.$$

The equality in (3.11) holds then if

$$F^{-1}(u) = \mu + \frac{\sigma}{B_{j,n}(q)} \times \begin{cases} g_{j,n}(u|q) - 1, & u < q, \\ \dfrac{1 - G_{j,n}(q-|q)}{1-q} - 1, & u \ge q. \end{cases}$$

The lower bounds on the conditional expectations of the sums of consecutive record values are presented below. The proof mimics the proof of Theorem 2.3, and it is omitted.

Theorem 3.3. *Assume the conditions of Theorem 3.1. For any $1 \le j \le n$, we have two cases.*

(i) *If $q < \frac{j}{n}$, then*

$$\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n} R_k - n\mu \middle| R_j = F^{-1}(q)\right) \ge -\frac{j-nq}{\sqrt{q(1-q)}}. \tag{3.20}$$

The equality in (3.20) is attained by the two-point distribution supported on the points $\mu - \sigma\sqrt{\frac{1-q}{q}}$ and $\mu + \sigma\sqrt{\frac{q}{1-q}}$ with probabilities q and $1-q$, respectively.

(ii) *If $q \ge \frac{j}{n}$ then*

$$\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n} R_k - n\mu \middle| R_j = F^{-1}(q)\right) \ge 0.$$

is optimal.

A more precise description of the attainability conditions in case (i) is presented in the comment below Theorem 2.3. Note that for $j = n$ only this case occurs.

4. SUMS OF RECORDS IN FINITE SEQUENCES

The problem of maximizing the conditional expectation of $\sum_{k=1}^{n} X_k\eta_k$ makes sense if X_k are positive.

Lemma 4.1. *Let $X_1, ..., X_n$ be positive i.i.d. with an absolutely continuous distribution function F, and finite expectation. Then $\mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\eta_k \middle| M_j = x\right)$ for some $1 \le j \le n$ is identical with the expectation of the distribution function*

$$H_{j,n,F}(y|x) = \begin{cases} 0, & y < 0, \\ \dfrac{n-1}{n} - \displaystyle\sum_{k=1}^{n-j} \frac{1-F^k(x)}{nk} - \left(\sum_{k=2}^{j} \frac{1}{nk}\right)\left[1 - \frac{-\ln[1-F(y)]}{-\ln[1-F(x)]}\right], & 0 \le y < x, \\ 1 - \displaystyle\sum_{k=1}^{n-j} \frac{1-F^k(y)}{nk}, & y \ge x. \end{cases} \tag{4.1}$$

Proof: Since we may observe at most j records among $X_1, ..., X_j$, we have

$$\begin{aligned} \mathbb{E}\left(\sum_{k=1}^{j} X_k\eta_k \middle| M_j = x\right) &= \sum_{k=1}^{j} \mathbb{E}\left(\sum_{i=1}^{k} R_i \middle| M_j = R_k = x, \sum_{i=1}^{j}\eta_i = k\right) \mathbb{P}\left(\sum_{i=1}^{j}\eta_i = k\right) \\ &= \sum_{k=1}^{j} \mathbb{E}\left(\sum_{i=1}^{k-1} R_i + x \middle| R_k = x, \sum_{i=1}^{j}\eta_i = k\right) \mathbb{P}\left(\sum_{i=1}^{j}\eta_i = k\right) \\ &= x + \sum_{k=2}^{j} \mathbb{E}\left(\sum_{i=1}^{k-1} R_i \middle| R_k = x, \sum_{i=1}^{j}\eta_i = k\right) \mathbb{P}\left(\sum_{i=1}^{j}\eta_i = k\right). \end{aligned}$$

Let $Y_1(x), ..., Y_{j-1}(x)$ denote i.i.d. random variables with a common distribution function $F_x(y) = \frac{-\ln[1-F(y)]}{-\ln[1-F(x)]}$, $y < x$. By arguments of the proof of Lemma 3.1 we notice that

$$\begin{aligned}
\mathbb{E}\left(\sum_{k=1}^{j} X_k\eta_k \middle| M_j = x\right) &= x + \sum_{k=2}^{j} \mathbb{E}\left(\sum_{i=1}^{k-1} Y_i(x)\right) \mathbb{P}\left(\sum_{i=1}^{j} \eta_i = k\right) \\
&= x + \mathbb{E}Y_1(x) \sum_{k=2}^{j} (k-1)\ \mathbb{P}\left(\sum_{i=2}^{j} \eta_i = k-1\right) \\
&= x + \mathbb{E}Y_1(x)\ \mathbb{E}\left(\sum_{i=2}^{j} \eta_i\right) \\
&= x + \mathbb{E}Y_1(x) \sum_{k=2}^{j} \frac{1}{k},
\end{aligned}$$

because $\mathbb{P}(\eta_k = 1) = \frac{1}{k} = 1 - P(\eta_k = 0)$. Note that under the condition $M_j = x$, just one among $X_k\eta_k$, $k = 1, ..., j$, has value x for sure. The other ones take on either some values in $(0, x)$ as the order statistics from the sample with the distribution function F_x, or they amount to 0. The first ones appear with probabilities $\frac{1}{k}$, and the others with probabilities $1 - \frac{1}{k}$, $k = 2, ..., j$. Therefore we can write

$$\begin{aligned}
\mathbb{E}\left(\sum_{k=1}^{j} X_k\eta_k \middle| M_j = x\right) &= \int_{\mathbb{R}} y\, \mathbf{1}_{[x,\infty)}(dy) + \sum_{k=2}^{j} \frac{1}{k} \int_0^x y\, \frac{-\ln[1-F(dy)]}{-\ln[1-F(x)]} \\
&\quad + \left(j - \sum_{k=1}^{j} \frac{1}{k}\right) \int_{\mathbb{R}} y\, \mathbf{1}_{[0,\infty)}(dy). \qquad (4.2)
\end{aligned}$$

For $k > j$, the conditional distribution of $X_k\eta_k$ has an atom at 0 with probability

$$\begin{aligned}
\mathbb{P}(X_k\eta_k = 0 | M_j = x) &= \mathbb{P}(X_k \le x) + \mathbb{P}(x < X_k \le \max\{X_{j+1}, ..., X_{k-1}\}) \\
&= F(x) + \int_x^\infty \mathbb{P}(\max\{X_{j+1}, ..., X_{k-1}\} \ge y) f(y) dy \\
&= F(x) + \int_x^\infty [1 - F^{k-j-1}(y)] f(y) dy \ = \ 1 - \frac{1 - F^{k-j}(x)}{k-j}.
\end{aligned}$$

Moreover, for $y > x$ we have

$$\begin{aligned}
\mathbb{P}(X_k\eta_k > y | M_j = x) &= \mathbb{P}(X_k > \max\{y, X_{j+1}, ..., X_{k-1}\}) \\
&= \int_y^\infty \mathbb{P}(\max\{X_{j+1}, ..., X_{k-1}\} < t) f(t) dt \\
&= \int_y^\infty F^{k-j-1}(t) f(t) dt \ = \ \frac{1 - F^{k-j}(y)}{k-j}.
\end{aligned}$$

Summing up, we obtain

$$\mathbb{P}(X_k\eta_k \le y | M_j = x) = \begin{cases} 0, & y < 0, \\ 1 - \dfrac{1 - F^{k-j}(x)}{k-j}, & 0 \le y \le x, \\ 1 - \dfrac{1 - F^{k-j}(y)}{k-j}, & y \ge x. \end{cases} \qquad (4.3)$$

Combining (4.2) and (4.3) yields

$$\begin{aligned}\mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n}X_k\eta_k\middle|M_j=x\right) &= \frac{1}{n}\left[j-\sum_{k=1}^{j}\frac{1}{k}+n-j-\sum_{k=j+1}^{n}\frac{1-F^{k-j}(x)}{k-j}\right]\int_{\mathbb{R}}y\,\mathbf{1}_{[0,\infty)}(dy)\\ &\quad+\sum_{k=2}^{j}\frac{1}{nk}\int_0^x y\,\frac{-\ln[1-F(dy)]}{-\ln[1-F(x)]}+\frac{1}{n}\int_{\mathbb{R}}y\,\mathbf{1}_{[x,\infty)}(dy)\\ &\quad+\frac{1}{n}\int_x^\infty y\left[n-j+\sum_{k=j+1}^{n}\frac{1-F^{k-j}(dy)}{k-j}\right] = \int_{\mathbb{R}}yH_{j,n,F}(dy|x).\end{aligned}\tag{4.4}$$

This completes the proof. □

If $X_1,...,X_n$ are standard uniform, (4.1) simplifies to

$$H_{j,n}(u|q)=\begin{cases}\dfrac{n-1}{n}-\displaystyle\sum_{k=1}^{n-j}\frac{1-q^k}{nk}-\left(\sum_{k=2}^{j}\frac{1}{nk}\right)\left[1-\frac{-\ln(1-u)}{-\ln(1-q)}\right], & 0\le u<q,\\ 1-\displaystyle\sum_{k=1}^{n-j}\frac{1-u^k}{nk}, & q\le u\le 1.\end{cases}\tag{4.5}$$

It has two atoms at 0 and q with respective probabilities $1-\sum_{k=1}^{j}\frac{1}{nk}-\sum_{k=1}^{n-j}\frac{1-q^k}{nk}$ and $\frac{1}{n}$, and the density function

$$h_{j,n}(u|q)=\begin{cases}\left(\displaystyle\sum_{k=2}^{j}\frac{1}{nk}\right)\frac{1}{-(1-u)\ln(1-q)}, & 0<u<q,\\ \dfrac{1}{n}\displaystyle\sum_{k=0}^{n-j-1}u^k=\frac{1-u^{n-j}}{n(1-u)}, & q<u<1.\end{cases}\tag{4.6}$$

Below we use the following values

$$H_{j,n}(q-|q)=\frac{n-1}{n}-\sum_{k=1}^{n-j}\frac{1-q^k}{nk},\qquad h_{j,n}(q-|q)=\frac{\sum_{k=2}^{j}\frac{1}{k}}{-n\ln(1-q)(1-q)},\tag{4.7}$$

and the the following function (comp. (3.8))

$$Ih_{j,n}(u|q)=\int_0^u h_{j,n}^2(v|q)dv=\left[\frac{\sum_{k=2}^{j}\frac{1}{k}}{-n\ln(1-q)}\right]^2\int_0^u\frac{1}{(1-v)^2}dv=\left[\frac{\sum_{k=2}^{j}\frac{1}{k}}{-n\ln(1-q)}\right]^2\frac{u}{1-u}.\tag{4.8}$$

Theorem 4.1. *Let $X_1,...,X_n$ be positive i.i.d. with an absolutely continuous distribution function F, and finite variance σ^2. Let μ stand for the respective expectation.*

(i) *Let $0<q_*<1$ be the unique solution to the equation*

$$1+\sum_{k=1}^{n-1}\frac{1-q^k}{k}-n(1-q)=0.\tag{4.9}$$

(**a**) *If $q \le q_*$, then the bound*

$$\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n} X_k\eta_k - n\mu \middle| M_1 = F^{-1}(q)\right) \le 0 \tag{4.10}$$

is sharp and attained in limit by the degenerate distribution.

(**b**) *If $q > q_*$, then*

(4.11)

$$\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n} X_k\eta_k - n\mu \middle| M_1 = F^{-1}(q)\right) \le nC_{j,n}(q) = \frac{1+\sum_{k=1}^{n-1}\frac{1-q^k}{k} - n(1-q)}{\sqrt{q(1-q)}},$$

and the equality is attained by the parent distribution assigning the masses q and $1-q$ to the points 0 and $\frac{\sigma}{\sqrt{q(1-q)}}$, respectively.

(**ii**) *Assume $2 \le j \le n$.*

(**a**) *If*

$$H_{j,n}(u|q) \ge u, \tag{4.12}$$

(comp. (4.5)) for all $0 < u < q$, then the optimal inequality is

$$\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n} X_k\eta_k - n\mu \middle| M_j = F^{-1}(q)\right) \le 0.$$

Otherwise we have three possibilities.

(**b**) *If*

$$h_{j,n}(q-|q) \le \frac{H_{j,n}(q-|q)}{q} < 1, \tag{4.13}$$

then (4.11) holds with M_1 and $\sum_{k=1}^{n-1}\frac{1-q^k}{k}$ replaced by M_j and $\sum_{k=1}^{n-j}\frac{1-q^k}{k}$, respectively, and identical conditions of attainability.

(**c**) *If*

$$1 > \frac{H_{j,n}(q-|q)}{q} < h_{j,n}(q-|q) \le \frac{1-H_{j,n}(q-|q)}{1-q}, \tag{4.14}$$

then there exists a unique $0 < u_* < q$ solving the equation

$$H_{j,n}(u|q) = uh_{j,n}(u|q)$$

(see (4.5) and (4.6)), and then

$$\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n} X_k\eta_k - n\mu \middle| M_j = F^{-1}(q)\right) \le nC_{j,n}(q), \tag{4.15}$$

where

$$C_{j,n}^2(q) = h_{j,n}^2(u_*)u_* + Ih_{j,n}(q|q) - Ih_{j,n}(u_*|q) + \frac{[1-H_{j,n}(q-|q)]^2}{1-q} - 1.$$

(see also (4.8)). The equality in (4.15) holds for F with the quantile function

$$F^{-1}(u) = \frac{\sigma}{C_{j,n}(q)} \times \begin{cases} 0, & 0 < u < u_*, \\ h_{j,n}(u|q) - h_{j,n}(u_*|q), & u_* \le u < q, \\ \dfrac{1-H_{j,n}(q-|q)}{1-q} - h_{j,n}(u_*|q), & q \le u < 1. \end{cases}$$

(**d**) *Finally, if*

$$h_{j,n}(q-|q) > \min\left\{\frac{H_{j,n}(q-|q)}{q}, \frac{1-H_{j,n}(q-|q)}{1-q}\right\}$$

then there also exists $u_* < u_{**} < q$ *satisfying the equation*

$$(1-q)h_{j,n}(u|q) = 1 - H_{j,n}(u|q),$$

and then (4.15) holds with

$$C^2_{j,n}(q) = h^2_{j,n}(u_*)u_* + Ih_{j,n}(u_{**}|q) - Ih_{j,n}(u_*|q) + h^2_{j,n}(u_{**})(1-u_{**}) - 1,$$

and the equality condition

$$F^{-1}(u) = \frac{\sigma}{C_{j,n}(q)} \times \begin{cases} 0, & 0 < u < u_*, \\ h_{j,n}(u|q) - h_{j,n}(u_*|q), & u_* \le u < u_{**}, \\ h_{j,n}(u_{**}|q) - h_{j,n}(u_*|q), & u_{**} \le u < 1. \end{cases}$$

Proof: We first notice that in contrast to the maximization problems studied in Sections 2 and 3, one treated here is not location-scale invariant. Indeed, if we translate the parent distribution by $c > 0$ to the right, we obtain

$$\begin{aligned} &\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n}(X_k + c)\eta_k - n(\mu + c)\middle| M_j = x + c\right) \\ = \ &\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n} X_k\eta_k - n\mu\middle| M_j = x\right) - c\left(n - \sum_{k=1}^{j}\frac{1}{k} - \sum_{k=1}^{n-j}\frac{1-F^k(x)}{k}\right) \\ < \ &\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n} X_k\eta_k - n\mu\middle| M_j = x\right) \end{aligned}$$

(see (4.1)). Accordingly, it suffices to restrict our investigations to the distributions whose supports start from 0. Alternatively, we can consider the problem modification where the lack of record gives the gain equal to the minimal value of the distribution support, and then remove the solutions for which $F^{-1}(0) \neq 0$.

Distribution functions (4.5) contain atoms at their left-end points of the supports, and hence they do not satisfy the assumptions of Lemma 1.1. We show that we get the sharp right-hand inequality in (1.4) if we replace the greatest convex minorant of $H_{j,n}(u|q)$ by the greatest convex minorant $\underline{H}_{j,n,0}(u|q)$ of $H_{j,n}(u|q)$ and the point $(0,0)$. Take $\varepsilon > 0$ sufficiently small so that $\underline{H}_{j,n,0}(u|q)$ is also the greatest convex minorant of $H_{j,n,\varepsilon}(u|q) = \min\{H_{j,n}(u|q), \frac{u}{\varepsilon}\} \le H_{j,n}(u|q)$. For every non-deceasing function f yields

$$\int_0^1 f(u)H_{j,n}(du|q) \le \int_0^1 f(u)H_{j,n,\varepsilon}(du|q) \le \int_0^1 f(u)\underline{h}_{j,n,0}(u|q)\,du,$$

where $\underline{h}_{j,n,0}(u|q)$ denotes the right derivative of $\underline{H}_{j,n,0}(u|q)$. Let f_0 satisfy the equality conditions in the latter inequality: f_0 is constant on each interval of $\{H_{j,n,\varepsilon}(u|q) < \underline{H}_{j,n,0}(u|q)\}$ and right-continuous. In particular, it is constant on $\{H_{j,n,\varepsilon}(u|q) < H_{j,n}(u|q)\}$, and 0 can be

attached to this interval by right-continuity of f_0. For brevity, denote the extended interval by $[0,\delta)$. Therefore

$$\int_0^\delta f_0(u)H_{j,n}(du|q) = \int_0^\delta f_0(u)H_{j,n,\varepsilon}(du|q) = f_0(0)H_{j,n,\varepsilon}(\delta-) = f_0(0)H_{j,n}(\delta-).$$

The respective integrals over $[\delta,1)$ are identical, because $H_{j,n}(u|q)$ and $H_{j,n,\varepsilon}(u|q)$ are identical there. Consequently,

$$\int_0^1 f_0(u)H_{j,n}(du|q) = \int_0^1 f_0(u)H_{j,n,\varepsilon}(du|q) = \int_0^1 f_0(u)\underline{h}_{j,n,0}(u|q)\,du,$$

which proves sharpness of the upper bound.

It follows that for proving our bounds, we need to determine the greatest convex minorants of the functions $H_{j,n,0}(u|q)$, which amount to 0 at $u=0$, and coincide with $H_{j,n}(u|q)$ otherwise. Note that each $H_{j,n}(u|q)$ is convex non-decreasing in $(q,1)$, and its derivative satisfies $h_{j,n}(1-|q) = 1-\frac{j}{n} < 1$. So this part of the function runs above the line $\ell(u)=u$, and does not affects the convex minorant.

(i) Function $H_{1,n}(u|q)$ is constant on the interval $(0,q)$. Therefore the greatest convex minorant of $H_{1,n,0}(u|q)$ is either the straight line $\underline{H}_{1,n,0}(u|q) = u$, $0<u<1$, when $H_{1,n}(q-|q) \ge q$, or the broken line

$$\underline{H}_{1,n,0}(u|q) = \begin{cases} \dfrac{H_{1,n}(q-|q)}{q}\,u, & u<q, \\ \dfrac{1-H_{1,n}(q-|q)}{1-q}(u-1)+1, & u\ge q, \end{cases} \tag{4.16}$$

otherwise. Function

$$H_{1,n}(q-|q) - q = \frac{n-1}{n} - \sum_{k=1}^{n-1}\frac{1-q^k}{nk} - q, \qquad 0<q<1,$$

amounts to $\frac{n-1}{n} - \sum_{k=1}^{n-1}\frac{1}{nk} > 0$ at 0, and to $-\frac{1}{n}<0$ at 1. Moreover, its derivative $\frac{1}{n}\sum_{k=0}^{n-2} q^k - 1$ is negative for all $0<q<1$. Therefore $\underline{H}_{1,n.0}(u|q) = u$ for $q \le q_*$ defined in (4.9), and has the form (4.16) for $q>q_*$.
Repeating the reasoning of the previous proofs we determine the sharp bounds (4.10) and (4.11). In the modified location-scale invariant problem, the former is attained by (2.33) with $\varepsilon\to 0$. In order to obey the restriction $F_\varepsilon^{-1}(0)=0$ we put $\mu = \sigma\sqrt{\frac{1-\varepsilon}{\varepsilon}}$. In the latter, the modified problem has solution (2.33) with ε replaced by q. Again, the support requirement narrows the attainability condition to the last statement of Theorem 4.1(i).

(ii) Relation (4.12) implies that $H_{j,n,0}(u|q) \ge u = \underline{H}_{j,n,0}(u|q)$, $0<u<1$. It follows that zero provides the optimal bound for (1.3). Otherwise we obtain non-trivial evaluations. Under condition (4.13) the line $\frac{H_{j,n}(q-|q)}{q}\,u$ runs beneath function $H_{j,n,0}(u|q)$ on $(0,q)$, and connects its end-points. Another linear function $\frac{1-H_{j,n}(q-|q)}{1-q}(u-1)+1$ minorizes $H_{j,n,0}(u|q)$ in $[q,1]$. Gluing together the lines we obtain the greatest convex minorant of $H_{j,n,0}(u|q)$ (note that the inequalities $\frac{H_{j,n}(q-|q)}{q} < 1 < \frac{1-H_{j,n}(q-|q)}{1-q}$ guarantee convexity and compare with (4.16)).

Mimicking the arguments of the previous proofs we calculate the upper bounds and determine the location-scale family of two-point distributions attaining the bounds in the location-scale invariant problem. Under the support restriction, we distinguish the scale family of distributions with the left support end-point equal to 0. If $\frac{H_{j,n}(q-|q)}{q} < h_{j,n}(q-|q)$ (see (4.7) and (4.14)), then the right part of the line $\frac{H_{j,n}(q-|q)}{q}\,u$, $0 < u < q$, lies above $H_{j,n}(u|q)$, and cannot constitute a part of the minorant. It should be replaced by a line with a smaller slope $h_{j,n}(u_*|q) = \frac{H_{j,n}(u_*|q)}{u_*}$, tangent to $H_{j,n}(u|q)$ at some $0 < u_* < q$, and $H_{j,n}(u|q)$ on the right which ultimately transforms into a line. If moreover $h_{j,n}(q-|q) \le \frac{1-H_{j,n}(q-|q)}{1-q}$, then $\underline{H}_{j,n,0}(u|q) = H_{j,n}(u|q)$ for all $u_* \le u < q$. The last part of the minorant is the line connecting $(q, H_{j,n}(q-|q))$ with $(1,1)$. Otherwise $\underline{H}_{j,n,0}(u|q)$ should transform into a line at some $u_* < u_{**} < q$ determined by the tangency condition $h_{j,n}(u_{**}|q)(1-u_{u**}) = 1 - H_{j,n}(u_{**}|q)$. Note that in the last case it is admitted that $H_{j,n}(q-|q) \ge q$ but necessarily $H_{j,n}(u_{**}|q) < u_{**}$.

Once we determine the greatest convex minorants, we further proceed in a standard way. The bound amounts to n multiplied by the square root of the integral of the squared derivative of the minorant decreased by 1. The standardized quantile function of the distribution attaining the bound is proportional to the greatest convex minorant derivative decreased by 1. The last step of the proof consists in removing the distributions whose left-end support points differ from 0. Detailed calculations are left to the reader. □

Establishing lower bounds for (1.3) does not make sense, because when we consider random variables $X_1, ..., X_n$ taking on very large values, and we may get $X_k\eta_k = 0$ as the results of not reaching records in some trials, would make (1.3) negative and arbitrarily small. We illustrate the phenomenon in the following example.

Example. Suppose that X_k, $k = 1, ..., n$, are uniformly distributed on the interval $[m, m+1]$. They have the distribution function $F(x) = x - m$, $m < x < m+1$, and quantile function $F^{-1}(q) = m + q$, $0 < q < 1$. Applying (4.4) we calculate

$$\begin{aligned}
\mathbb{E}\left(\sum_{k=1}^{n} X_k\eta_k \middle| M_j = m+q\right) &= \frac{\sum_{k=2}^{j}\frac{1}{k}}{-\ln(1-q)}\int_m^{m+q}\frac{y\,dy}{1-y+m} + q + m + \int_{m+q}^{m+1} y\sum_{k=0}^{n-j-1}(y-m)^k\,dy \\
&= \frac{\sum_{k=2}^{j}\frac{1}{k}}{-\ln(1-q)}[-(m+1)\ln(1-q) - q] + q + m + \sum_{k=2}^{n-j+1}\frac{1-q^k}{k} + m\sum_{k=1}^{n-j}\frac{1-q^k}{k} \\
&= m\left[1 + \sum_{k=2}^{j}\frac{1}{k} + \sum_{k=1}^{n-j}\frac{1-q^k}{k}\right] + q + \sum_{k=2}^{j}\left[1 - \frac{q}{-\ln(1-q)}\right] + \sum_{k=2}^{n-j+1}\frac{1-q^k}{k}.
\end{aligned}$$

Since $\mathbb{E}X_1 = m + \frac{1}{2}$ and $\operatorname{Var} X_1 = \frac{1}{12}$, we have

$$\begin{aligned}
\frac{1}{\sigma}\mathbb{E}\left(\sum_{k=1}^{n} X_k\eta_k - n\mu \middle| M_j = m+q\right) &= 12m\left[1 + \sum_{k=2}^{j}\frac{1}{k} + \sum_{k=1}^{n-j}\frac{1-q^k}{k} - n\right] \\
&\quad + 12\left[q + \sum_{k=2}^{j}\left[1 - \frac{q}{-\ln(1-q)}\right] + \sum_{k=2}^{n-j+1}\frac{1-q^k}{k} - \frac{n}{2}\right]. \qquad (4.17)
\end{aligned}$$

Putting $m = 0$, we obtain the standardized conditional expectation for the standard uniform variables. However, when m increases to $+\infty$, then (4.17) tends to $-\infty$, because the factor at m is strictly negative.

We would avoid obtaining trivial lower bounds in (1.3) if we replaced $X_k\eta_k = 0$ by a quantity connected with the distribution of random variables, e.g. by the mean or a quantile of F of a positive order.

5. NUMERICAL EVALUATIONS

Here we present numerical upper bounds on the conditional expectations of sums of sample maxima and records described analytically in Sections 2–4 for $n = 10$, $j = 1, 5, 8$ and 10, and quantile orders $q = 0.1..., (0.1), ..., 0.9$. We also include two extreme cases $q = 0.05$ and 0.99. For comparison, we present numerical results for $n = 20$ and $j = 10$ as well. We do not evaluate numerically respective lower bounds because they have simple analytic forms.

The bounds strongly depend on the number n of summands. Therefore instead of bounds $nA_{j,n}(q)$, $nB_{j,n}(q)$, and $nC_{j,n}(q)$ on the expectations of the total sums, we present in Tables 1–5 the average bounds $A_{j,n}(q)$, $B_{j,n}(q)$, and $C_{j,n}(q)$ determined per each particular summand. Each numerical bound is accompanied by the reference to a particular part of the theorem which provides the tools for calculating it. This allows the reader to realize the shape of the parent distribution which attains the corresponding bound. For instance, the average upper bounds $A_{5,10}(q)$ are determined with use of Theorem 2.1. For $q = 0.05$ the quantile function is first a constant (which generates a jump of height $u_{**} > q$), and than is a curve linearly transforming $f_{5,10}(u|0.05)$ (see (2.7)). For $q = 0.1, ..., 0.4$ the extreme quantile function is first increasing, then constant, and again increasing. When $q = 0.1, 0.2, 0.3$ the transition from the curve to the horizontal line occurs at q (see Theorem 2.1(iva)), but for $q = 0.4$ it happens at some $u_* < q$ (see Theorem 2.1(ivb)). For $q \geq 0.5$ the conditions of Theorem 2.1(iib) hold which implies that the distribution functions attaining the respective bounds are continuous on some intervals, and have jumps of size $1 - q$ at their right-end points.

All the average bounds presented in Tables 1–5 are increasing with respect to q. This is easily justifiable: the greater is the extreme j-th variable, the greater is the expectation of the sum of n analogous observations. When we fix n and q, we observe that the bounds decrease when j increases. It has a clear explanation as well. E.g., when we assume that $M_{j_1} = x$ we may suspect that $\sum_{k=1}^{n} M_k$ is greater than in the case $M_{j_2} = x$ for some $j_2 > j_1$, because in the latter case the maximum equal to x appears later than in the former one. We note that $C_{j,n}(q) = 0$ except for about 10% upper quantile orders q. Trivial zero bounds $A_{j,n}(q)$ and $B_{j,n}(q)$ for the sums of maxima and records, respectively, appear only for relatively small q and large j.

By definition $\sum_{k=1}^{n} X_k\eta_k \leq \sum_{k=1}^{n} M_k \leq \sum_{k=1}^{n} R_k$ for any random sequence $X_1, ..., X_n, ...$. The corresponding relations for the bounds $C_{j,n}(q) < A_{j,n}(q) < B_{j,n}(q)$ are preserved and their values are significantly different when j is small with respect to n. When j is equal or close to n, the latter inequality is reversed, though. For $j = n$ it is justified by the following arguments.

Table 1: Average upper bounds for $n = 10$, and $j = 1$.

q	T2	$A_{j,n}(q)$	T5	$B_{j,n}(q)$	T7	$C_{j,n}(q)$
0.05	(ia)	0.99901	(i)	23.19053	(ia)	0
0.1	(ia)	1.00446	(i)	24.48122	(ia)	0
0.2	(ia)	1.01964	(i)	27.54620	(ia)	0
0.3	(ia)	1.04398	(i)	31.48623	(ia)	0
0.4	(ia)	1.08488	(i)	36.73885	(ia)	0
0.5	(ia)	1.15691	(i)	44.09161	(ia)	0
0.6	(ia)	1.28895	(i)	55.11962	(ia)	0
0.7	(ia)	1.53655	(i)	73.49811	(ia)	0
0.8	(ib)	2.00000	(i)	110.25284	(ia)	0
0.9	(ib)	3.00000	(i)	220.51248	(ib)	0.24836
0.99	(ib)	9.94987	(i)	2205.14721	(ib)	0.99321

Table 2: Average upper bounds for $n = 10$, and $j = 5$.

q	T1	$A_{j,n}(q)$	T4	$B_{j,n}(q)$	T7	$C_{j,n}(q)$
0.05	(iii)	0.11265	(i)	1.53182	(iia)	0
0.1	(iva)	0.11312	(i)	1.62854	(iia)	0
0.2	(iva)	0.16859	(i)	1.85718	(iia)	0
0.3	(iva)	0.28233	(i)	2.14943	(iia)	0
0.4	(ivb)	0.43379	(i)	2.53691	(iia)	0
0.5	(iib)	0.61390	(iia)	3.09867	(iia)	0
0.6	(iib)	0.82506	(iia)	3.92617	(iia)	0
0.7	(iib)	1.09660	(iia)	5.29056	(iia)	0
0.8	(iib)	1.50351	(iib)	7.86104	(iia)	0
0.9	(iib)	2.33534	(iib)	15.77310	(iib)	0.15107
0.99	(iib)	7.94034	(iib)	157.76816	(iic)	0.94803

Table 3: Average upper bounds for $n = 20$, and $j = 10$.

q	T1	$A_{j,n}(q)$	T4	$B_{j,n}(q)$	T7	$C_{j,n}(q)$
0.05	(iii)	0.38885	(i)	22.61002	(iia)	0
0.1	(iva)	0.39153	(i)	23.86739	(iia)	0
0.2	(iva)	0.41974	(i)	26.85346	(iia)	0
0.3	(iva)	0.47203	(i)	30.69239	(iia)	0
0.4	(iva)	0.54734	(i)	35.81061	(iia)	0
0.5	(iva)	0.65396	(i)	42.97568	(iia)	0
0.6	(ivb)	0.81552	(iia)	53.72374	(iia)	0
0.7	(ivb)	1.06348	(iia)	71.63557	(iia)	0
0.8	(iib)	1.45459	(iia)	107.45727	(iia)	0
0.9	(iib)	2.25974	(iia)	214.91881	(iia)	0
0.99	(iib)	7.69114	(iib)	2149.144709	(iic)	0.43245

There are no future maxima and records after the j-th one. Conditionally on $R_n = x$, the previous record values $R_1, ..., R_{n-1}$ are distributed as ordered i.i.d. random variables from the right-truncated at x parent distribution (cf. Lemma 3.1). The distributions of M_k, $k = 1, ..., n-1$, under the condition $M_n = x$ are the mixtures of maxima from k independent observations from the right-truncated baseline distribution and an atom at x (see Lemma 2.1). This implies $\mathbb{E}\big(\sum_{k=1}^{n} R_k \big| R_n = x\big) < \mathbb{E}\big(\sum_{k=1}^{n} M_k \big| M_n = x\big)$ for any parent distribution.

Numerical calculations show that the reversed inequality $B_{j,n}(q) < A_{j,n}(q)$ holds for $j = n-1$ and all q as well. Then the distribution $\mathcal{L}(R_n|R_{n-1} = x)$ is just the left-truncated parent distribution at x, and this does not affect much the whole sum. Table 4 shows that for $n = 10$, $j = n - 2 = 8$ the reversed inequalities $B_{8,10}(q) < A_{8,10}(q)$ are satisfied merely for some central q.

Table 4: Average upper bounds for $n = 10$, and $j = 8$.

q	T1	$A_{j,n}(q)$	T4	$B_{j,n}(q)$	T7	$C_{j,n}(q)$
0.05	(i)	0	(i)	0.01450	(iia)	0
0.1	(iia)	0.00644	(i)	0.01874	(iia)	0
0.2	(iia)	0.10956	(i)	0.03141	(iia)	0
0.3	(iia)	0.21969	(i)	0.05288	(iia)	0
0.4	(iia)	0.33303	(i)	0.08963	(iia)	0
0.5	(iia)	0.45780	(i)	0.15357	(iia)	0
0.6	(iia)	0.60936	(iia)	0.59290	(iia)	0
0.7	(iib)	0.82368	(iia)	0.96188	(iia)	0
0.8	(iib)	1.16140	(iia)	1.56008	(iia)	0
0.9	(iib)	1.85340	(iia)	3.22165	(iib)	0.06500
0.99	(iib)	6.43747	(iib)	16.06139	(iic)	0.92619

Table 5: Average upper bounds for $n = 10$, and $j = 10$.

q	T2	$A_{j,n}(q)$	T5	$B_{j,n}(q)$	T7	$C_{j,n}(q)$
0.05	(iia)	0	(iia)	0	(iia)	0
0.1	(iib)	0.00448	(iia)	0	(iia)	0
0.2	(iib)	0.10267	(iia)	0	(iia)	0
0.3	(iib)	0.20796	(iia)	0	(iia)	0
0.4	(iib)	0.31432	(iia)	0	(iia)	0
0.5	(iib)	0.42761	(iia)	0	(iia)	0
0.6	(iib)	0.55702	(iib)	0.00138	(iia)	0
0.7	(iib)	0.72122	(iib)	0.08846	(iia)	0
0.8	(iic)	0.98150	(iib)	0.25087	(iia)	0
0.9	(iic)	1.55748	(iib)	0.54681	(iia)	0
0.99	(iic)	5.44190	(iib)	1.91034	(iic)	0.91287

We finally focus on Tables 2 and 3 which contain results for $(j, n) = (5, 10)$ and $(10, 20)$. One could expect that the average bounds are similar if the proportion j/n is preserved. This actually happens in the case of sums of maxima. We see that $A_{5,10}(q) < A_{10,20}(q)$ for small q, and the opposite holds for large q. In the former case, when $M_j = F^{-1}(q)$ is relatively small, it is a great chance that the total sum of maxima shall increase more when we observe 10 future i.i.d. observations rather than 5. This chance decreases when M_j is close to the right end-point of the support. Much the same observation concerns $C_{j,n}(q)$, but the difference is not visible for small q, for which $C_{5,10}(q) = C_{10,20}(q) = 0$. The average bounds $B_{j,n}(q)$ for the sums of record values behave quite differently: $B_{5,10}(q)$ are much less than $B_{10,20}(q)$, and the latter are rather close to $B_{1,10}(q)$ (see Table 1). This shows that the average bounds $B_{j,n}(q)$ depend rather on the differences $n - j$, i.e., on the number of future records.

ACKNOWLEDGMENTS

The research of the second author has been partially supported by PUT under grant 0211/SBAD/0121. The authors thank the associate editor for valuable comments which allowed them to improve the presentation of the paper.

REFERENCES

[1] Ahmadi, J. and Balakrishnan, N. (2010). Prediction of order statistics and record values from two independent sequences, *Statistics*, **44**, 417–430.

[2] Arnold, B.C.; Balakrishnan, N. and Nagaraja, H.N. (1992). *A First Course in Order Statistics*, Wiley, New York.

[3] Arnold, B.C.; Balakrishnan, N. and Nagaraja, H.N. (1998). *Records*, Wiley, New York.

[4] Asgharzadeh, A.; Ahmadi, J.; Ganji, Z.M. and Valiollahi, R. (2012). Reconstruction of the past failure times for the proportional reversed hazard rate model, *J. Stat. Comput. Simul.*, **82**, 475–489.

[5] Balakrishnan, N.; Doostparast, M. and Ahmadi, J. (2009). Reconstruction of past records, *Metrika*, **70**, 89–109.

[6] Bel'kov, I.V. and Nevzorov, V.B. (2018). On a problem on the optimal choice of record values, *Vestnik St. Petersburg Univ. Math.*, **51**, 107–113.

[7] Bel'kov, I.V. and Nevzorov, V.B. (2020). On a problem of the optimal choice of record values, *J. Math. Sci. (N.Y.)*, **244**, 718–722.

[8] Chow, Y.S.; Robbins, H. and Siegmund, D. (1971). *Great Expectations: The Theory of Optimal Stopping*, Houghton Mifflin Co., Boston, Mass.

[9] David, H.A. and Nagaraja, H.N. (2003). *Order Statistics*, 3rd ed., Wiley, Hoboken, NJ.

[10] Freeman, P.R. (1983). The secretary problem and its extensions: a review, *Int. Stat. Rev.*, **51**, 189–206.

[11] Gilbert, J.P. and Mosteller, F. (1966). Recognizing the maximum of a sequence, *J. Am. Stat. Assoc.*, **61**, 35–73.

[12] Grau Ribas, J.M. (2019). A new look at the returning secretary problem, *J. Comb. Optim.*, **37**, 1216–1236.

[13] Gumbel, E.J. (1954). The maxima of the mean largest value and of the range, *Ann. Math. Statist.*, **25**, 76–84.

[14] Hartley, H.O. and David, H.A. (1954). Universal bounds for mean range and extreme observation, *Ann. Math. Statist.*, **25**, 85–99.

[15] Khatib, B. and Ahmadi, J. (2014). Best linear unbiased and invariant reconstructors for the past records, *Bull. Malays. Math. Sci. Soc.*, **37**(2), 1017–1028.

[16] Khatib, B.; Ahmadi, J. and Razmkhah, M. (2014). Reconstruction of the past lower record values in a proportional reversed hazard rate model, *Statistics*, **48**, 421–435.

[17] Klimczak, M. (2006). Prediction of k-th records, *Statist. Probab. Lett.*, **76**, 117–127.

[18] Klimczak, M. and Rychlik, T. (2005). Reconstruction of previous failure times and records, *Metrika*, **61**, 277–290.

[19] Kuchta, M. (2017). Iterated full information secretary problem, *Math. Methods Oper. Res.*, **86**, 277–292.

[20] MirMostafaee, S.M.T.K. and Ahmadi, J. (2011). Point prediction of future order statistics from an exponential distribution, *Statist. Probab. Lett.*, **81**, 360–370.

[21] Moriguti, S. (1953). A modification of Schwarz's inequality with applications to distributions, *Ann. Math. Statist.*, **24**, 107–113.

[22] Nagaraja, H.N. (1978). On the expected values of record values, *Austral. J. Statist.*, **20**, 176–182.

[23] Nevzorov, V.B. (2001). *Records: Mathematical Theory*, Translations of Mathematical Monographs, **194**, AMS, Providence, RI.

[24] Nevzorov, V.B.; Savinova, V. and Stepanov, A. (2022). Linear prediction of sequential minima, *Comm. Statist. Theory Methods*, **51**, 5446–5454.

[25] Nevzorov, V.B. and Stepanov, A. (2021). On maximum of expected sums of sequential minima, *Comm. Statist. Theory Methods*, **50**, 1362–1369.

[26] Nevzorov, V.B. and Tovmasyan, S.A. (2014). On the maximal value of the expectation of record numbers, *Vestnik St. Petersburg Univ. Math.*, **47**(2), 64–67.

[27] Ramsey, D.M. (2016). A secretary problem with missing observations, *Math. Appl. (Warsaw)*, **44**, 149–165.

[28] Raqab, M.Z. and Balakrishnan, N. (2008). Prediction intervals for future records, *Statist. Probab. Lett.*, **78**, 1955–1963.

[29] Rychlik, T. (2002). Predictions of increments of order and record statistics in nonparametric families of distributions, *J. Statist. Theory Appl.*, **1**, 43–56.

[30] Samuels, S.M. (1991). *Secretary problems*. In "Handbook of Sequential Analysis" (B.K. Ghosh and P.K. Sen, Eds.), Marcel Dekker, New York, pp. 381–405.

[31] Volterman, W.; Davies, K.F.; Balakrishnan, N. and Ahmadi, J. (2014). Nonparametric prediction of future order statistics, *J. Stat. Comput. Simul.*, **84**, 683–695.

[32] Woryna, A. (2017). The solution of a generalized secretary problem via analytic expressions, *J. Comb. Optim.*, **33**, 1469–1491.

REVSTAT – Statistical Journal
Volume 21, Number 2, April 2023, 267–281
https://doi.org/10.57805/revstat.v21i2.401

Median Distance Model for Likert-Type Items in Contingency Table Analysis

Authors: Serpil Aktas Altunay ✉
– Department of Statistics, Hacettepe University, Ankara, Turkey
spxl@hacettepe.edu.tr

Ayfer Ezgi Yilmaz
– Department of Statistics, Hacettepe University, Ankara, Turkey
ezgiyilmaz@hacettepe.edu.tr

Received: April 2021 Revised: November 2021 Accepted: November 2021

Abstract:

- Likert-type items (questions) are a widely used scale in questionnaire design. The "neutral" or "undecided" option may lead to misinterpretation and confusion about the results. This paper proposes two novel log-linear models to measure how much accumulation of the neutral option over the contingency tables at any question levels. These models also test the odds that a respondent's level how far from the median. These models will help the researchers how to incorporate the neutral option in conceptual frameworks.

Keywords:

- *Likert-type items; association models; ordinal variables; contingency table; log-linear models.*

AMS Subject Classification:

- 49A05, 78B26.

✉ Corresponding author.

1. INTRODUCTION

Likert-type scales or formally ordinal scales are psychometric scales used when there is an order in responses and distances between categories are not quantitative [4, 14]. Likert scale is widely used in medical, education, and many disciplines in social sciences.

There is a difference between the terms of Likert-type items and Likert scales [20]. Likert items are the single questions that use some aspect of the original Likert response alternatives and several of them built a Likert scale [10]. In this study, Likert-type items are considered as a part of a scale or not.

Likert-type items are usually formed in five responses: "1: strongly disagree", "2: disagree", "3: neutral", "4: agree", "5: strongly agree". Similarly, a 7-point Likert scale includes seven responses such as; "1: strongly disagree", "2: disagree", "3: somewhat disagree", "4: neither agree nor disagree", "5: somewhat agree", "6: Agree", "7: strongly agree".

The attitudes change from mildly positive to mildly negative. The neutral option that is sometimes referred to as "neither agree nor disagree" or "undecided" on a Likert scale means that respondents are not willing to answer a particular question or have no idea.

With regard to the neutral point on the scale, we should be aware that neutral does not imply the midpoint between the two extreme-scale scores.

Those respondents who check the neutral option might mislead the results and the main point might not be achieved. Hence, the question "neutral responses will be omitted or how to handle with neutral questions?" matter. In some surveys that there is often no neutral category included in the middle of the scale [7]. Sometimes it is placed at the end of the scale, and sometimes it is eliminated directly. The neutral means is the median or mid-point and the median is the 50% sample distribution and it means 50% of the participants have neutral to agree with opinions in a 5-point Likert scale. If the median is 4, it means 50% of participants have a positive opinion. The ordinal structure and the existence of a neutral category should be considered to model the Likert items. Despite the independence of the two Likert-type items is analyzed with the chi-square test, it does not accept the ordinal structure of the items. Linear-by-linear association model and its special form uniform association model are used to analyze the association between the variables of a contingency table with ordered categories [1, 8]. There are many extensions of association models (e.g. [5, 6, 18, 21, 22]). Even though all these models consider the ordinal structure of the variables, they ignore the ambiguous nature of the neutral category and treat it as if the neutral category has the same structure as other categories. Truebner [19] showed that changes in respondents' characteristics do not affect median response with the exception of age. Even though the intervals between the categories should be regarded as subjectively equal, Oppenheim [15] states that "attitudes may be shaped more like concentric circles or overlapping ellipses or three-dimensional cloud formations, therefore, the model of the linear continuum or dimension is not always easy or appropriate".

2. MATERIAL AND METHOD

A contingency table summarizes information of two or higher dimensions random variables. An example of the contingency table is given in Table 1 for the first question (Q_1) and second question (Q_2) in a questionnaire.

Table 1: Two-way classification table for a 5-point Likert scale questions.

Q_1	Q_2					**Total**
	1	2	3	4	5	
1	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	n_{1+}
2	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}	n_{2+}
3	n_{31}	n_{32}	n_{33}	n_{34}	n_{35}	n_{3+}
4	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}	n_{4+}
5	n_{51}	n_{52}	n_{53}	n_{54}	n_{55}	n_{5+}
Total	n_{+1}	n_{+2}	n_{+3}	n_{+4}	n_{+5}	n

Consider a two-way table in which both the row and column variables have R categories (levels). R denotes the R-point Likert scale. In an $R \times R$ table, n_{ij}'s denote the cell frequencies for the ith row and jth column where $i = 1, ..., R$. n_{i+} and n_{+j} are the row and column totals, respectively, satisfying

$$\sum_{i=1}^{R} n_{i+} = \sum_{j=1}^{R} n_{+j} = n.$$

The goal of the log-linear analysis is to determine which categorical variables represent the data. Log-linear models do not distinguish between response and explanatory variables. All variables in a log-linear model are treated as responses.

The relationship between two or more variables is examined in analyzing contingency tables. We will refer to the variables in two-way contingency tables as "question". In a two-way $R \times R$ contingency table, let $\{\mu_{ij}\}$ be the expected values corresponding to the observed values. The independence model for any pair of items is commonly defined for the two questions in Equation (2.1).

$$\text{Log}(\mu_{ij}) = \lambda + \lambda_i^{Q_1} + \lambda_j^{Q_2} + \lambda_{ij}^{Q_1 Q_2}, \qquad i, j = 1, ..., R, \tag{2.1}$$

where λ is the intercept term (overall mean of the natural log of the expected values), $\lambda_i^{Q_1}$ is the main effect for question Q_1, $\lambda_j^{Q_2}$ is the main effect for question Q_2, and $\lambda_{ij}^{Q_1 Q_2}$ is the interaction term. The parameters are set to satisfy the following restrictions:

$$\sum_{i=1}^{R} \lambda_i^{Q_1} = \sum_{j=1}^{R} \lambda_j^{Q_2} = \sum_{i=1}^{R} \sum_{j=1}^{R} \lambda_{ij}^{Q_1 Q_2} = 0.$$

Because concluding that respondents are neutral might be inaccurate, we suggest two models that measure the variability around the neutral option, namely in the third group for the 5-point Likert-type and the fourth group for the 7-point Likert-type as shown in Figures 1–3.

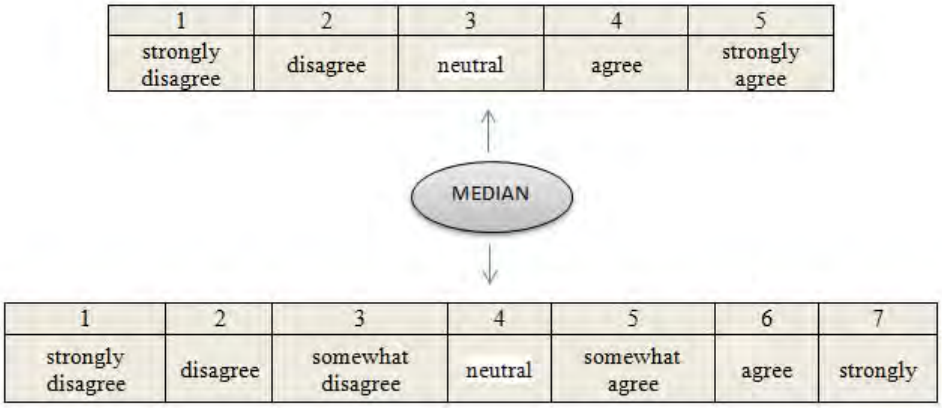

Figure 1: The position of the median in 5- and 7- point Likert scales.

Q_1	Q_2				
	1	2	3	4	5
1	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}
2	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}
3	n_{31}	n_{32}	n_{33}	n_{34}	n_{35}
4	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}
5	n_{51}	n_{52}	n_{53}	n_{54}	n_{55}

Figure 2: Variability around the median in a 5×5 table.

Q_1	Q_2						
	1	2	3	4	5	6	7
1	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	n_{16}	n_{17}
2	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}	n_{26}	n_{27}
3	n_{31}	n_{32}	n_{33}	n_{34}	n_{35}	n_{36}	n_{37}
4	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}	n_{46}	n_{47}
5	n_{51}	n_{52}	n_{53}	n_{54}	n_{55}	n_{56}	n_{57}
6	n_{61}	n_{62}	n_{63}	n_{64}	n_{65}	n_{66}	n_{67}
7	n_{71}	n_{72}	n_{73}	n_{74}	n_{75}	n_{76}	n_{77}

Figure 3: Variability around the median in a 7×7 table.

The median of an R categories is calculated as

$$m = \frac{R+1}{2},$$

and median cell implies that the cell falls into the (m, m). The median cell falls into the (3,3) cell for a 5×5 table, fall into the (4,4) cell for a 7×7 table.

We built two novel log-linear models taking the main effects (Q_1, Q_2), association parameter, and distance parameter. The simple model is the Median Distance (MD) model as

$$\text{Log}(\mu_{ij}) = \lambda + \lambda_i^{Q_1} + \lambda_j^{Q_2} + \delta_{ij}, \qquad i, j = 1, ..., R. \tag{2.2}$$

The parameter δ is the median distance parameter which is defined in Equation (2.3) and the method to identify the log-linear parameters involves fixing the parameters to zero for one category of Q_1 and Q_2, respectively. For an $R \times R$ table, the MD model has m median distance parameters:

$$\delta_{ij} = \begin{cases} \delta_1, & i = j = m \text{ (median cell)}, \\ \delta_2, & \text{one-step distance from the median cell}, \\ \delta_3, & \text{two-step distance from the median cell}, \\ \vdots & \vdots \\ \delta_{m-1}, & (m-2)\text{-step distance from the median cell}, \\ \delta_m, & (m-1)\text{-step distance from the median cell}. \end{cases} \tag{2.3}$$

For example, the light gray shaded area in Figure 2 represents one step from the midpoint, and the dark gray shaded area shows the two-step distance from the midpoint. The median distance parameters are set to satisfy the following restriction:

$$\sum_{i=1}^{m} \delta_i = 0.$$

This model has more $(m - 1 = (R - 1)/2)$ parameters than the independence model, the residual degrees of freedom under the MD model is

$$\begin{aligned} df &= R \times R - \left[1 - (R-1) + (R-1) + \left(\frac{R+1}{2} - 1\right)\right] \\ &= \frac{2R^2 - 5R + 3}{2}. \end{aligned}$$

The odds ratios matrix under the MD model for a 5×5 table is shown below

$$\begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \theta_{14} \\ \theta_{21} & \theta_{22} & \theta_{23} & \theta_{24} \\ \theta_{31} & \theta_{32} & \theta_{33} & \theta_{34} \\ \theta_{41} & \theta_{42} & \theta_{43} & \theta_{44} \end{bmatrix} = \exp \begin{bmatrix} \delta_2 - \delta_3 & 1 & 1 & \delta_3 - \delta_2 \\ 1 & \delta_1 - \delta_2 & \delta_2 - \delta_1 & 1 \\ 1 & \delta_2 - \delta_1 & \delta_1 - \delta_2 & 1 \\ \delta_3 - \delta_2 & 1 & 1 & \delta_2 - \delta_3 \end{bmatrix}$$

and for a 7×7 table is given as

$$\begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \theta_{14} & \theta_{15} & \theta_{16} \\ \theta_{21} & \theta_{22} & \theta_{23} & \theta_{24} & \theta_{25} & \theta_{26} \\ \theta_{31} & \theta_{32} & \theta_{33} & \theta_{34} & \theta_{35} & \theta_{36} \\ \theta_{41} & \theta_{42} & \theta_{43} & \theta_{44} & \theta_{45} & \theta_{46} \\ \theta_{51} & \theta_{52} & \theta_{53} & \theta_{54} & \theta_{55} & \theta_{56} \\ \theta_{61} & \theta_{62} & \theta_{63} & \theta_{64} & \theta_{65} & \theta_{66} \end{bmatrix} = \exp \begin{bmatrix} \delta_3 - \delta_4 & 1 & 1 & 1 & 1 & \delta_4 - \delta_3 \\ 1 & \delta_1 - \delta_3 & 1 & 1 & \delta_3 - \delta_2 & 1 \\ 1 & 1 & \delta_1 - \delta_2 & \delta_2 - \delta_1 & 1 & 1 \\ 1 & 1 & \delta_2 - \delta_1 & \delta_1 - \delta_2 & 1 & 1 \\ 1 & \delta_3 - \delta_2 & 1 & 1 & \delta_1 - \delta_3 & 1 \\ \delta_4 - \delta_3 & 1 & 1 & 1 & 1 & \delta_3 - \delta_4 \end{bmatrix}$$

When both the column and row variables of a two-dimensional table are ordinal, a simple log-linear model that utilizes the orderings of the rows and the columns is the linear-by-linear association model [1]. This ordinarily of the data needs an extra parameter

that gives the association of two ordinal variables. Hence, adding an association model to the MD model, the median distance + association (MDA) model is defined in a log-linear form as in Equation (2.4):

$$\text{Log}(\mu_{ij}) = \lambda + \lambda_i^{Q_1} + \lambda_j^{Q_2} + \beta u_{1i} u_{2j} + \delta_{ij}, \qquad i, j = 1, ..., R, \tag{2.4}$$

where β is the linear-by-linear association parameter and δ is the median distance parameter which is defined in Equation (2.3). The necessity of reflecting the ordinarily of the variables, assigning scores to the ordinal categories are fulfilled by the row and column scores, by u_{1i} and u_{2j} scores. The integer scores, meanly $u_{1i}, u_{2j} = 1, ..., R$ are the frequently used scores. This model has more one more parameter than the MA model, the residual degrees of freedom under the MDA model is

$$\begin{aligned} df &= R \times R - \left[1 - (R-1) + (R-1) + \left(\frac{R+1}{2} - 1\right) + 1\right] \\ &= \frac{2R^2 - 5R + 1}{2}. \end{aligned}$$

The matrix of odds ratios under the MDA model for a 5×5 table is

$$\begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \theta_{14} \\ \theta_{21} & \theta_{22} & \theta_{23} & \theta_{24} \\ \theta_{31} & \theta_{32} & \theta_{33} & \theta_{34} \\ \theta_{41} & \theta_{42} & \theta_{43} & \theta_{44} \end{bmatrix} = \exp \begin{bmatrix} \beta + \delta_2 - \delta_3 & \beta & \beta & \beta + \delta_3 - \delta_2 \\ \beta & \beta + \delta_1 - \delta_2 & \beta + \delta_2 - \delta_1 & \beta \\ \beta & \beta + \delta_2 - \delta_1 & \beta + \delta_1 - \delta_2 & \beta \\ \beta + \delta_3 - \delta_2 & \beta & \beta & \beta + \delta_2 - \delta_3 \end{bmatrix}.$$

The matrix of odds ratios under the MDA model for a 7×7 table is

$$\begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \theta_{14} & \theta_{15} & \theta_{16} \\ \theta_{21} & \theta_{22} & \theta_{23} & \theta_{24} & \theta_{25} & \theta_{26} \\ \theta_{31} & \theta_{32} & \theta_{33} & \theta_{34} & \theta_{35} & \theta_{36} \\ \theta_{41} & \theta_{42} & \theta_{43} & \theta_{44} & \theta_{45} & \theta_{46} \\ \theta_{51} & \theta_{52} & \theta_{53} & \theta_{54} & \theta_{55} & \theta_{56} \\ \theta_{61} & \theta_{62} & \theta_{63} & \theta_{64} & \theta_{65} & \theta_{66} \end{bmatrix} = \exp \begin{bmatrix} \beta + \delta_3 - \delta_4 & \beta & \beta & \beta & \beta & \beta + \delta_4 - \delta_3 \\ \beta & \beta + \delta_1 - \delta_3 & \beta & \beta & \beta + \delta_3 - \delta_2 & \beta \\ \beta & \beta & \beta + \delta_1 - \delta_2 & \beta + \delta_2 - \delta_1 & \beta + \beta & \beta \\ \beta & \beta & \beta + \delta_2 - \delta_1 & \beta + \delta_1 - \delta_2 & \beta & \beta \\ \beta & \beta + \delta_3 - \delta_2 & \beta & \beta & \beta + \delta_1 - \delta_3 & \beta \\ \beta + \delta_4 - \delta_3 & \beta & \beta & \beta & \beta & \beta + \delta_3 - \delta_4 \end{bmatrix}.$$

The goodness of fit hypothesis is tested by the likelihood ratio test statistic as

$$G^2 = 2 \sum_{i=1}^{R} \sum_{j=1}^{R} n_{ij} \log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right).$$

Under the null hypothesis is true, likelihood ratio statistic has an asymptotic chi-square distribution with associated degrees of freedom.

The design matrix of the MDA model for a 5×5 table is constructed as below. If we subtracted the last column from the design matrix the MDA model would turn into the

MA model. This implies that the MD model has one less parameter than the MDA model:

$$\log\begin{bmatrix}\mu_{11}\\ \mu_{12}\\ \mu_{13}\\ \mu_{14}\\ \mu_{15}\\ \mu_{21}\\ \mu_{22}\\ \mu_{23}\\ \mu_{24}\\ \mu_{25}\\ \mu_{31}\\ \mu_{32}\\ \mu_{33}\\ \mu_{34}\\ \mu_{35}\\ \mu_{41}\\ \mu_{42}\\ \mu_{43}\\ \mu_{44}\\ \mu_{45}\\ \mu_{51}\\ \mu_{52}\\ \mu_{53}\\ \mu_{54}\\ \mu_{55}\end{bmatrix} = \begin{bmatrix}1&1&0&0&0&1&0&0&0&-1&-1&1\\ 1&1&0&0&0&0&1&0&0&-1&-1&2\\ 1&1&0&0&0&0&0&1&0&-1&-1&3\\ 1&1&0&0&0&0&0&0&1&-1&-1&4\\ 1&1&0&0&0&-1&-1&-1&-1&-1&-1&5\\ 1&0&1&0&0&1&0&0&0&-1&-1&2\\ 1&0&1&0&0&0&1&0&0&0&1&4\\ 1&0&1&0&0&0&0&1&0&0&1&6\\ 1&0&1&0&0&0&0&0&1&0&1&8\\ 1&0&1&0&0&-1&-1&-1&-1&-1&-1&10\\ 1&0&0&1&0&1&0&0&0&-1&-1&3\\ 1&0&0&1&0&0&1&0&0&0&1&6\\ 1&0&0&1&0&0&0&1&0&1&0&9\\ 1&0&0&1&0&0&0&0&1&0&1&12\\ 1&0&0&1&0&-1&-1&-1&-1&-1&-1&15\\ 1&0&0&0&1&1&0&0&0&-1&-1&4\\ 1&0&0&0&1&0&1&0&0&0&1&8\\ 1&0&0&0&1&0&0&1&0&0&1&12\\ 1&0&0&0&1&0&0&0&1&0&1&16\\ 1&0&0&0&1&-1&-1&-1&-1&-1&-1&20\\ 1&-1&-1&-1&-1&-1&0&0&0&-1&-1&5\\ 1&-1&-1&-1&-1&0&-1&0&0&-1&-1&10\\ 1&-1&-1&-1&-1&0&0&-1&0&-1&-1&15\\ 1&-1&-1&-1&-1&0&0&0&-1&-1&-1&20\\ 1&-1&-1&-1&-1&-1&-1&-1&-1&-1&-1&25\end{bmatrix}\begin{bmatrix}\lambda\\ \lambda_1^{Q_1}\\ \lambda_2^{Q_1}\\ \lambda_3^{Q_1}\\ \lambda_4^{Q_1}\\ \lambda_1^{Q_2}\\ \lambda_2^{Q_2}\\ \lambda_3^{Q_2}\\ \lambda_4^{Q_2}\\ \delta_1\\ \delta_2\\ \beta\end{bmatrix}.$$

3. NUMERICAL EXAMPLES

In this section, we provide three data sets to illustrate the methods presented in this paper. Two of these data sets are artificial and one is real-life data. The observed frequencies in the artificial tables were generated so that the data set fits the model adequately, by adjusted according to the expected frequencies calculated under the models hold true. Models are applied to these numerical examples and the results are highlighted for the researchers to be able to understand and interpret the information more strategically and usefully. The models were analyzed using "General Log-linear models" in IBM SPSS 23 by entering the design matrix properly. In the design matrix, the δ and β parameters are defined as the covariates [12].

Example 1

An artificial 5×5 contingency table is given in Table 2 which displays for any two questions from a questionnaire, say Q_1 and Q_2.

The Independence, symmetry, quasi-symmetry, MD, and MDA models are applied to the data in Table 2 and the log-linear model results are summarized in Table 3 (see [1] and [3] for the details of symmetry and quasi-symmetry models). The quasi-symmetry, MD, and MDA models fit data (Table 3, $p>0.05$). The quasi-symmetry model implies that there is an agreement between Q_1 and Q_2.

Table 2: The frequencies (expected values) of a 5×5 table.

Q_1	Q_2					**Total**
	1	2	3	4	5	
1	8 (5.89)	6 (8.82)	9 (11.49)	11 (11.32)	13 (9.48)	47
2	15 (15.56)	51 (50.84)	65 (66.25)	67 (65.28)	25 (25.06)	223
3	18 (15.79)	51 (51.57)	150 (150.00)	67 (66.22)	23 (25.06)	309
4	13 (14.24)	46 (46.51)	59 (60.60)	61 (59.72)	25 (22.93)	204
5	5 (7.52)	15 (11.26)	20 (14.67)	11 (14.45)	9 (12.11)	60
Total	59	169	303	217	95	843

Table 3: Model results for the 5×5 table.

Model	G^2	***df***	***p*-value**	**AIC**	**BIC**
Independence	54.065	16	<0.001	—	—
Symmetry	25.347	10	0.005	—	—
Quasi-symmetry	7.118	6	0.310	−4.882	−33.304
MD	10.062	14	0.758	−17.938	−84.256
MDA	9.691	13	0.719	−16.309	−77.890

Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) [2, 17] are calculated for quasi-symmetry, MD, and MDA models to find the best fitting model to data. The MD model has the lowest AIC and BIC, it is considered as the best-fitted model. The parameter estimates under the MD model are summarized in Table 4.

Table 4: The parameter estimates under the MD model for the 5×5 table.

Parameter	**Estimate**	**Std. Error**	**Z**	***p*-value**	**95% CI**
Constant	3.282	0.320	10.253	<0.001	[2.654; 3.909]
$[Q_1 = 1]$	−0.244	0.195	−1.254	0.210	[−0.626; 0.138]
$[Q_1 = 2]$	0.728	0.218	3.336	0.001	[0.300; 1.155]
$[Q_1 = 3]$	0.742	0.220	3.374	0.001	[0.311; 1.173]
$[Q_1 = 4]$	0.639	0.219	2.915	0.004	[0.209; 1.068]
$[Q_1 = 5]$	$0^{(a)}$				
$[Q_2 = 1]$	−0.476	0.166	−2.874	0.004	[−0.801; −0.151]
$[Q_2 = 2]$	−0.073	0.226	−0.322	0.747	[−0.516; 0.370]
$[Q_2 = 3]$	0.192	0.225	0.853	0.394	[−0.249; 0.632]
$[Q_2 = 4]$	0.177	0.223	0.793	0.428	[−0.260; 0.615]
$[Q_2 = 5]$	$0^{(a)}$				
δ_1	0.795	0.119	6.683	<0.001	[0.562; 1.029]
δ_2	−0.008	0.096	−0.079	0.937	[−0.196; 0.181]

(a): This parameter is set to zero because it is redundant.

The δ_k parameters in Equation (2.2) have straightforward interpretations in terms of departures from the median category. The distance parameter estimates are $\hat{\delta}_1 = 0.795$, $\hat{\delta}_2 = -0.008$, and $\hat{\delta}_3 = 0 - [\hat{\delta}_1 + \hat{\delta}_2] = -0.787$. Odds ratios are calculated either taking the expected values in Table 2 or from the parameter estimates under the underlying model given in Table 4. For example for $\hat{\theta}_{11}$ is obtained as:

$$\hat{\theta}_{11} = \frac{5.89 \times 50.84}{8.82 \times 15.56} = \exp(\hat{\delta}_2 - \hat{\delta}_3) = 2.18.$$

This can be interpreted as: the respondent's response is 2.18 times more likely to fall into the neutral category than a category two-step away from the median category. The matrix of odds ratios:

$$\hat{\theta} = \begin{bmatrix} 2.18 & 1 & 1 & 0.46 \\ 1 & 2.23 & 0.45 & 1 \\ 1 & 0.45 & 2.23 & 1 \\ 0.46 & 1 & 1 & 2.18 \end{bmatrix}$$

$\exp(\hat{\delta}_1 - \hat{\delta}_2) = 2.23$ can be interpreted as: a respondent's response is 2.23 times more likely to fall into the neutral category than a category one-step away from the median category.

Example 2

Table 5 displays an artificial 7×7 contingency tables for any two questions from a questionnaire, say Q_1 and Q_2.

Table 5: The frequencies (expected values) of an hypothetical 7×7 table.

Q_1	Q_2							**Total**
	1	2	3	4	5	6	7	
1	5 (9.79)	15 (14.76)	18 (17.78)	24 (19.18)	22 (20.42)	13 (15.18)	9 (8.88)	106
2	16 (14.41)	27 (28.58)	32 (33.46)	29 (35.07)	45 (36.28)	28 (26.21)	8 (11.00)	125
3	17 (15.83)	21 (30.50)	75 (73.49)	80 (74.84)	82 (75.24)	21 (24.94)	9 (10.17)	305
4	20 (19.33)	45 (36.19)	87 (84.73)	95 (95.00)	70 (81.91)	27 (26.38)	10 (10.46)	354
5	21 (20.24)	40 (36.82)	82 (83.77)	80 (80.56)	75 (76.47)	25 (23.93)	8 (9.22)	331
6	19 (18.32)	32 (32.38)	35 (33.80)	31 (31.58)	25 (29.13)	21 (18.76)	8 (7.02)	171
7	10 (10.07)	12 (12.78)	11 (12.97)	9 (11.77)	11 (10.55)	7 (6.60)	8 (3.25)	68
Total	108	192	340	348	330	142	60	1520

The Independence, symmetry, quasi-symmetry, MD, and MDA models are applied to the data in Table 5 and the log-linear model results are summarized in Table 6. The symmetry,

quasi-symmetry, MD, and MDA models fit data (Table 6, p>0.05). The symmetry and quasi-symmetry model implies that there is an agreement between Q_1 and Q_2.

Table 6: Model results for the 7×7 table.

Model	G^2	*df*	*p*-value	AIC	BIC
Independence	87.455	36	<0.001	—	—
Symmetry	14.355	21	0.854	−27.645	−139.501
Quasi-symmetry	8.840	15	0.886	−21.160	−101.057
MD	33.818	33	0.428	−32.182	−207.955
MDA	25.648	32	0.779	−38.352	−208.799

AIC and BIC are calculated for symmetry, quasi-symmetry, MD, and MDA models. The MDA model has the lowest AIC and BIC, thus it is considered as the best-fitted model. The parameter estimates under the MDA model are summarized in Table 7.

Table 7: The parameter estimates under the MDA model for the 7×7 table.

Parameter	Estimate	Std. Error	Z	*p*-value	95% CI
Constant	3.219	0.505	6.369	<0.001	[2.228; 4.209]
$[Q_1=1]$	−0.200	0.275	−0.730	0.466	[−0.739; 0.338]
$[Q_1=2]$	0.215	0.292	0.734	0.463	[−0.358; 0.787]
$[Q_1=3]$	0.337	0.263	1.280	0.201	[−0.179; 0.853]
$[Q_1=4]$	0.566	0.245	2.308	0.021	[0.085; 1.046]
$[Q_1=5]$	0.640	0.232	2.761	0.006	[0.186; 1.095]
$[Q_1=6]$	0.569	0.233	2.446	0.014	[0.113; 1.025]
$[Q_1=7]$	$0^{(a)}$				
$[Q_2=1]$	−0.074	0.282	−0.263	0.793	[−0.628; 0.479]
$[Q_2=2]$	0.365	0.297	1.228	0.220	[−0.217; 0.947]
$[Q_2=3]$	0.580	0.267	2.172	0.030	[0.057; 1.103]
$[Q_2=4]$	0.684	0.249	2.749	0.006	[0.196; 1.172]
$[Q_2=5]$	0.775	0.235	3.298	0.001	[0.315; 1.236]
$[Q_2=6]$	0.508	0.238	2.129	0.033	[0.040; 0.975]
$[Q_2=7]$	$0^{(a)}$				
δ_1	0.545	0.120	4.541	<0.001	[0.310; 0.780]
δ_2	0.420	0.083	5.054	<0.001	[0.257; 0.583]
δ_3	−0.331	0.093	−3.537	<0.001	[−0.514; −0.147]
β	−0.029	0.010	−2.842	0.004	[−0.048; −0.009]

(a): This parameter is set to zero because it is redundant.

The negative value of β indicates that there is a negative relationship between Q_1 and Q_2 ($\hat{\beta}=-0.029$). The distance parameter estimates are $\hat{\delta}_1=0.545$, $\hat{\delta}_2=0.420$, $\hat{\delta}_3=-0.331$, and $\hat{\delta}_4=0-[\hat{\delta}_1+\hat{\delta}_2+\hat{\delta}_3]=-0.634$.

Odds ratios can be calculated over either the expected values in Table 5 or the parameter estimates in Table 7. For instance, $\hat{\theta}_{11}$ is calculated as

$$\hat{\theta}_{11}=\frac{9.79\times28.58}{14.76\times14.41}=\exp(\hat{\beta}+\hat{\delta}_3-\hat{\delta}_4)=1.32.$$

This can be interpreted as: the respondent's response is 1.32 times more likely to fall into the neutral category than a category three-step away from the median category, respectively. The matrix of odds ratios:

$$\hat{\theta} = \begin{bmatrix} 1.32 & 0.97 & 0.97 & 0.97 & 0.97 & 0.72 \\ 0.97 & 2.06 & 0.97 & 0.97 & 0.46 & 0.97 \\ 0.97 & 0.97 & 1.10 & 0.86 & 0.97 & 0.97 \\ 0.97 & 0.97 & 0.86 & 1.10 & 0.97 & 0.97 \\ 0.97 & 0.46 & 0.97 & 0.97 & 2.06 & 0.97 \\ 0.72 & 0.97 & 0.97 & 0.97 & 0.97 & 1.32 \end{bmatrix}$$

$\exp(\hat{\beta}+\hat{\delta}_1-\hat{\delta}_2) = 1.10$ can be interpreted as: a respondent's response is 1.10 times more likely to fall into the neutral category than a category one-step away from the median category. The respondent's response is $\exp(\hat{\beta}+\hat{\delta}_1-\hat{\delta}_3) = 2.06$ times more likely to fall into the neutral category than a category two-step away from the median category.

Real-Life Data

The study of hostel life data [16] is used to illustrate the proposed models. The project aims to measure the satisfaction level of the students towards facilities given in hostels. 5-point Likert items are used as: "1: very dissatisfied", "2: dissatisfied", "3: neutral", "4: satisfied", "5: very satisfied". Three items, "Overall Satisfaction about Hostel", "Management System of Mess", and "24 Hours Electricity" are selected. The answers of 184 students are given in Table 8.

Table 8: The study of hostel life data.

Overall Satisfaction	**Management System (24 Hours Electricity)**					**Total**
	1	2	3	4	5	
1	2 (0)	3 (2)	0 (5)	6 (5)	1 (0)	12
2	2 (3)	5 (3)	6 (8)	14 (14)	3 (2)	30
3	2 (1)	11 (13)	24 (10)	36 (35)	6 (20)	79
4	3 (2)	10 (3)	13 (12)	19 (20)	7 (15)	52
5	1 (0)	0 (0)	0 (2)	4 (3)	6 (6)	11
Total	10 (6)	29 (21)	43 (37)	79 (77)	23 (43)	184

The Independence, MD, and MDA models are applied to the overall satisfaction x management system of mess and overall satisfaction x 24 hour electricity tables. The log-linear model results are summarized in Table 9. For overall satisfaction x management system of mess table, both MD and MDA models fit the data well ($p>0.05$). For overall satisfaction x 24 hour electricity, only the MDA model fit the data well ($p>0.05$).

For overall satisfaction x management system of mess table, MD model has the lowest BIC and MDA model has the lowest AIC. We considered BIC. We follow the BIC results and decide that the MD model is the best-fitted model. The expected values under the best-fitted models are summarized in Table 10.

Table 9: Log-linear model results for hostel life data.

Table	Model	G^2	*df*	*p*-value	AIC	BIC
Overall satisfaction-Management system of mess	Independence	32.268	16	0.009	—	—
	MD	19.942	14	0.132	−8.058	−53.067
	MDA	17.750	13	0.167	−8.250	−50.044
Overall satisfaction-24 hours electricity	Independence	32.807	16	0.008	—	—
	MD	27.805	14	0.015	—	—
	MDA	16.426	13	0.227	−9.574	−51.368

Table 10: The expected values of hostel life data.

Overall Satisfaction	**Management System (24 Hours Electricity)**					**Total**
	1	2	3	4	5	
1	1.58 (1.07)	1.40 (2.58)	1.56 (4.00)	3.82 (3.37)	3.64 (0.98)	12
2	1.39 (1.58)	5.25 (4.65)	5.86 (9.39)	14.30 (10.27)	3.20 (4.12)	30
3	3.17 (2.60)	11.96 (9.93)	24.00 (10.00)	32.58 (37.11)	7.29 (19.37)	79
4	2.41 (0.69)	9.10 (3.41)	10.15 (11.66)	24.79 (21.58)	5.54 (14.65)	52
5	1.45 (0.06)	1.28 (0.44)	1.43 (1.94)	3.50 (4.67)	3.33 (3.89)	11
Total	10 (6)	29 (21)	43 (37)	79 (77)	23 (43)	184

The parameter estimates for overall satisfaction x management system of mess table under the MD model and overall satisfaction x 24 hours electricity table under the MDA model are summarized in Table 11 and Table 12, respectively.

Table 11: The parameter estimates under the MD model for overall satisfaction x management system of mess table.

Parameter	**Estimate**	**Std. Error**	**Z**	***p*-value**	**95% CI**
Constant	2.366	0.635	3.728	<0.001	[1.122; 3.610]
$[Q_1 = 1]$	0.087	0.417	0.208	0.835	[−0.731; 0.905]
$[Q_1 = 2]$	−0.041	0.468	−0.088	0.930	[−0.958; 0.876]
$[Q_1 = 3]$	0.782	0.448	1.746	0.081	[−0.096; 1.660]
$[Q_1 = 4]$	0.509	0.452	1.125	0.261	[−0.378; 1.396]
$[Q_1 = 5]$	$0^{(a)}$				
$[Q_2 = 1]$	−0.833	0.379	−2.199	0.028	[−1.575; −0.091]
$[Q_2 = 2]$	−0.953	0.468	−2.038	0.042	[−1.870; −0.036]
$[Q_2 = 3]$	−0.844	0.480	−1.761	0.078	[−1.784; 0.095]
$[Q_2 = 4]$	0.049	0.444	0.110	0.913	[−0.821; 0.919]
$[Q_2 = 5]$	$0^{(a)}$				
δ_1	0.875	0.272	3.214	0.001	[0.341; 1.408]
δ_2	0.287	0.208	1.381	0.167	[−0.120; 0.695]

(a): This parameter is set to zero because it is redundant.

The distance parameter estimates in Table 11 are $\hat{\delta}_1 = 0.875$, $\hat{\delta}_2 = 0.287$, and $\hat{\delta}_3 = 0 - [\hat{\delta}_1 + \hat{\delta}_2 + \hat{\delta}_3] = -1.162$. The odds ratios of overall satisfaction x management system of

mess table can be calculated by the expected values in Table 10 or by the parameter estimates in Table 11. For example, $\hat{\theta}_{11}$ is calculated as:

$$\hat{\theta}_{11} = \frac{1.58 \times 5.25}{1.40 \times 1.39} = \exp(\hat{\delta}_2 - \hat{\delta}_3) = 4.26.$$

This can be interpreted as: the student's response is 4.26 times more likely to fall into the neutral category than a category two-step away from the median category. The matrix of odds ratios for overall satisfaction x management system of mess table:

$$\hat{\theta} = \begin{bmatrix} 4.26 & 1 & 1 & 0.23 \\ 1 & 1.80 & 0.56 & 1 \\ 1 & 0.56 & 1.80 & 1 \\ 0.23 & 1 & 1 & 4.26 \end{bmatrix}$$

$\exp(\hat{\delta}_1 - \hat{\delta}_2) = 1.80$ can be interpreted as: a student's response is 1.80 times more likely to fall into the neutral category than a category one-step away from the median category.

Table 12: The parameter estimates under the MDA model for overall satisfaction x 24 hours electricity.

Parameter	Estimate	Std. Error	Z	*p*-value	95% CI
Constant	−5.581	2.209	−2.527	0.012	[−9.910; −1.252]
$[Q_1 = 1]$	3.882	1.239	3.132	0.002	[1.452; 6.311]
$[Q_1 = 2]$	4.005	1.133	3.536	<0.001	[1.785; 6.225]
$[Q_1 = 3]$	4.238	0.883	4.797	<0.001	[2.506; 5.969]
$[Q_1 = 4]$	2.643	0.633	4.176	<0.001	[1.403; 3.884]
$[Q_1 = 5]$	$0^{(a)}$				
$[Q_2 = 1]$	1.148	1.046	1.098	0.272	[−0.902; 3.198]
$[Q_2 = 2]$	1.760	1.006	1.750	0.080	[−0.211; 3.732]
$[Q_2 = 3]$	1.937	0.832	2.328	0.020	[0.306; 3.569]
$[Q_2 = 4]$	1.500	0.630	2.381	0.017	[0.265; 2.735]
$[Q_2 = 5]$	$0^{(a)}$				
δ_1	−0.660	0.316	−2.090	0.037	[−1.278; −0.041]
δ_2	0.300	0.231	1.297	0.195	[−0.153; 0.752]
β	0.263	0.083	3.177	0.001	[0.101; 0.425]

(a): This parameter is set to zero because it is redundant.

The distance parameter estimates in Table 12 are $\hat{\delta}_1 = -0.660$, $\hat{\delta}_2 = 0.300$, and $\hat{\delta}_3 = 0 - [\hat{\delta}_1 + \hat{\delta}_2 + \hat{\delta}_3] = -0.634$. Similarly, the odds ratios of overall satisfaction x 24 hour electricity table can be calculated either from the expected values in Table 10 or from the parameter estimates in Table 12. For the odds ratio $\hat{\theta}_{11}$, is obtained as:

$$\hat{\theta}_{11} = \frac{1.07 \times 4.65}{2.58 \times 1.58} = \exp(\hat{\beta} + \hat{\delta}_2 - \hat{\delta}_3) = 1.22.$$

The odds ratio can be interpreted as: the students' response is 1.22 times more likely to fall into the neutral category than a category two-step away from the median category. The matrix of odds ratios for overall satisfaction x 24 hours electricity table:

$$\hat{\theta} = \begin{bmatrix} 1.22 & 1.30 & 1.30 & 1.38 \\ 1.30 & 0.50 & 3.39 & 1.30 \\ 1.30 & 3.39 & 0.50 & 1.30 \\ 1.38 & 1.30 & 1.30 & 1.22 \end{bmatrix}$$

$1/\exp(\hat{\beta}+\hat{\delta}_1-\hat{\delta}_2)=2$ can be interpreted as: a respondent's response is 2 times more likely to fall into a category one-step away from the median category than the neutral category. The positive value of β means that there is a positive effect of 24 hour electricity on overall satisfaction about Hostel ($\hat{\beta}=0.263$).

4. CONCLUDING REMARKS

Attitudinal questions are a fundamental part of surveys in the social sciences. The items in a Likert scale are designed to measure respondent's attitudes to a particular question. Likert-type data is ordinal data, and a score is higher or lower than another. In any survey, if people feel that they really have no idea upon a question or feel that they are urged to make a choice, they choose the random or intentionally choose the neutral option. Neutral states that the respondent has neither a positive response nor a negative response. The researchers prefer to use a neutral category or midpoint so as to one side of which lay the favorable categories and to the other side the unfavorable categories. If the researcher does not set to a midpoint and respondents actually have a neutral opinion, they either tend to give a response that does not represent their actual attitude or avoid answering the question because the respondents sometimes tend to avoid using extreme categories. Essentially age and education are believed the two most relevant demographic factors which have been associated with a neutral option [13]. For instance, unlike the results that Harzing [11] showed that a higher neutral response for women than men, Grimm and Church [9] had found no gender effect.

The neutral point is the most difficult to locate and even more difficult to interpret. Moreover, the Likert scales tend to perform well with regard to a particular attitude of respondents that is in rough order. Assuming that we employ a 5-point or 7-point Likert scale and our questionnaire comprises a neutral option, with this regard we would mainly wish to know if there is any agglomeration in the neutral option. In fact, being in the neutral option would also imply that those users might be moved towards the satisfied group in some senses. This would cause a misinterpretation and deviates from the real context. Statistical modeling is a very essential part of data analysis. With this point of view, this paper proposes two log-linear models that take the ordinal information into account, besides the distance from the median category in Likert scale data. These models test whether the frequencies accumulate over the median group by subtracting the association. The distance parameters indicate that whether a subject is in favor to decide neutral, or measures how far a subject from the median. If the models hold true, the researcher will be able to draw conclusions from the evidence presented in the findings which are the results of the parameter estimates. It is noteworthy that the δ parameters and their associated odds ratios in the MDA model give evidence that how the frequencies in a two-way contingency table are distributed around the median category, moreover, how far the frequencies are from the median or midpoint. Interpretation of the log-odds coefficient gives the odds that a respondents' response falls in the median group than being an m-step distant from the midpoint category.

The models have a limitation that addresses the cognitive bias. As a consequence of cognitive bias, individuals make decisions according to their own perspectives, and therefore, cognitive biases may sometimes lead to inaccurate inferences or illogical interpretations. The impact of cognitive bias might be reduced by helping the participants to understand the consequences of the inference at the beginning.

REFERENCES

[1] Agresti, A. (2010). *Analysis of Ordinal Categorical Data*, John Wiley and Sons, New Jersey.

[2] Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**(6), 716–723.

[3] Aktaş Altunay, S.; Yilmaz, A.E.; Bahcecitapar, M. and Bakacak Karabenli, L. (2021). *SPSS ve R Uygulamalı Kategorik Veri Çözümlemesi*, Seçkin Yayıncılık, Ankara.

[4] Allen, I.E. and Seaman, C.A. (2007). Likert scales and data analyses, *Quality Progress*, **40**(7), 64–65.

[5] Altun, G. and Aktas, S. (2016). Asymmetry and skew-symmetry models for square contingency tables, *Turkiye Klinikleri Journal of Biostatistics*, **8**(2), 152–161.

[6] Bagheban, A.A. and Zayeri, F. (2010). A generalization of the uniform association model for assessing rater agreement in ordinal scales, *Journal of Applied Statistics*, **37**(8), 1265–1273.

[7] Barry, D. (2017). Do not use averages with likert scale data.
`https://bookdown.org/Rmadillo/likert/`

[8] Goodman, L.A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories, *Journal of the American Statistical Association*, **74**(367), 537–552.

[9] Grimm, S.D. and Church, A.T. (1999). Cross-cultural study of response biases in personality measures, *Journal of Research in Personality*, **33**(4), 415–441.

[10] Guerra, A.L.; Gidel, T. and Vezzetti, E. (2016). *Toward a common procedure using likert and likert-type scales in small groups comparative design observations*. In "Proceedings of International Design Conference", **33**(4), 23–32.

[11] Harzing, A.W. (2006). Response styles in cross-national survey research: a 26-country study, *International Journal of Cross Cultural Management*, **6**(2), 243–266.

[12] IBM Corp. Released (2015). *IBM SPSS Statistics for Windows*, Version 23.0, IBM Corp, New York.

[13] Knauper, B. (1999). The impact of age and education on response order effects in attitude measurement, *Public Opinion Quarterly*, **63**(3), 347–370.

[14] Likert, L. (1932). A technique for the measurement of attitudes, *Archives of Psychology*, **22**(140), 1–55.

[15] Oppenheim, A.N. (1992). *Questionnaire Design, Interviewing and Attitude Measurement*, Pinter Pub. Ltd., London.

[16] Sable, V. (2018). The study of hostel life (version 2).
`https://www.kaggle.com/vinayaksable/the-study-of-hostel-life`

[17] Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**(2), 461–464.

[18] Tomizawa, S. (1991). A model of uniform association plus two-diagonals parameter and its application to occupational mobility table data, *Statistical Paper*, **32**(1), 243–252.

[19] Truebner, M. (2019). Dynamics of "Neither Agree nor Disagree" answers in attitudinal questions, *Journal of Survey Statistics and Methodology*, **9**(1), 51–72.

[20] Tutz, G. (2020). Hierarchical models for the analysis of likert scales in regression and item response analysis, *International Statistical Review*, **89**(1), 18–35.

[21] Valet, F.; Guinot, C. and Mary, J.Y. (2007). Log-linear non-uniform association models for agreement between two ratings on an ordinal scale, *Statistics in Medicine*, **26**(3), 647–662.

[22] Yilmaz, A.E. and Saracbasi, T. (2015). Symmetric disagreement + exponential score association model in square tables, *Turkiye Klinikleri Journal of Biostatistics*, **7**(2), 96–102.

REVSTAT – Statistical Journal
Volume 21, Number 2, April 2023, 283–302
https://doi.org/10.57805/revstat.v21i2.400

Random Forests for Time Series

Authors: Benjamin Goehry ✉
– Laboratoire de Mathématiques d'Orsay, CNRS, Université Paris-Saclay, Faculté des Sciences d'Orsay, Bâtiment 307, 91405 Orsay, France
benjamin.goehry@gmail.com

Hui Yan
– EDF Lab,
7 bd Gaspard Monge, 91120 Palaiseau, France
hui.yan@edf.fr

Yannig Goude
– EDF Lab & Laboratoire de Mathématiques d'Orsay, CNRS, Université Paris-Saclay, Orsay, France
yannig.goude@edf.fr

Pascal Massart
– Laboratoire de Mathématiques d'Orsay, CNRS, Université Paris-Saclay, Orsay, France
pascal.massart@math.u-psud.fr

Jean-Michel Poggi
– University Paris & Laboratoire de Mathématiques d'Orsay, CNRS, Université Paris-Saclay, Orsay, France
jean-michel.poggi@math.u-psud.fr

Received: July 2020 Revised: November 2021 Accepted: November 2021

Abstract:

- Random forests are a powerful learning algorithm. However, when dealing with time series, the time-dependent structure is lost, assuming the observations are independent. We propose some variants of random forests for time series. The idea is to replace standard bootstrap with a dependent block bootstrap to subsample time series during tree construction. We present numerical experiments on electricity load forecasting. The first, at a disaggregated level and the second at a national level focusing on atypical periods. For both, we explore a heuristic for the choice of the block size. Additional experiments with generic time series data are also available.

Keywords:

- *block bootstrap; random forests; regression; time series.*

AMS Subject Classification:

- 62M10, 62P30.

✉ Corresponding author.

1. INTRODUCTION

Random forests were introduced in 2001 by Breiman in [1] and are since then one of the most popular algorithms in machine learning [2]. The popularity comes from the wide range of applications in which they are known to perform well on even high dimensional, are fast to compute and easy to tune. Successful applications can be cited: chemo-informatics [3], ecology [4, 5], 3D object recognition [6] and time series prediction [7, 8, 9, 10, 11].

Suppose that we have a random sequence $(X_t, Y_t)_{t\in\mathbb{Z}} \in \mathcal{X}\times\mathcal{Y}$ such that

$$Y_t = f(X_t) + \epsilon_t \tag{1.1}$$

and the error ϵ_t is such that $\mathbb{E}[\epsilon_t|X_t] = 0$. The purpose of random forests is to estimate, by only observing a training sample $\mathcal{D}_n = ((X_1, Y_1), ..., (X_n, Y_n))$, the regression function

$$\forall\, x \in \mathcal{X}, \qquad f(x) = \mathbb{E}[Y_t|X_t = x].$$

Random forests can be related to two main sources, regression trees [12] and bagging [13]. Regression trees are constructed by a recursive partitioning of the input space based on some criterion to estimate the regression function f. At each step of the tree construction, a split is selected (a variable and a location on the variable) based on the evaluation of the criterion among all the admissible splits based on all the variables. The cell is cut in two on the selected split and the previous step is reiterated on the new cells. A tree is then a piecewise constant decomposition of the input space. A binary tree can be associated to the input space partitioning. Each node corresponds to a test matching how the input space was cut. An illustration is given in Figure 1 of a partitioning in the two-dimensional space and its associated binary tree. The principle of bagging (short form of bootstrap aggregating) is to create M randomly generated training sets by randomly sampling α_n observations with or without replacement from the set $\mathcal{D}_n$ and to construct on each set a predictor. Once the predictors are constructed, the bagging prediction for a new observation x is an aggregation, generally the empirical mean, of the predictions given by the M predictors for the point x. This procedure aims to improve stability and accuracy of the base predictor. In the context of random forests the predictors are regression trees. In order to explain the random forest procedure we then have to explicit the construction of one tree.

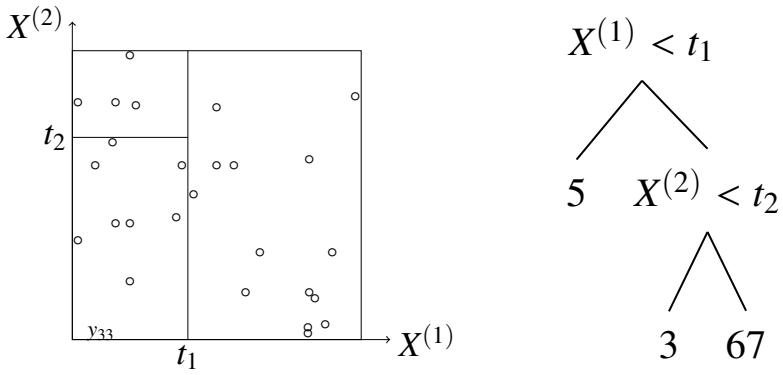

Figure 1: A partitioning of $[0,1]^2$ and the associated binary tree.

The first step is the bootstrap/subsampling: α_n points are selected with or without replacement among the n realisations. Then a tree is constructed based on these α_n selected points. At each node of the tree the best split (the variable and the location on this variable) is determined by minimising the intra-node variance. This is commonly called the CART criterion introduced in [12]. Instead of minimising this criterion among all the admissible splits based on all the variables the choice of inputs is restricted to a random subset of fixed size m_{try}. This procedure is then iterated on each node produced after binary splitting until stopping conditions are met. The first stopping rule is when the variance in a node is equal to zero. Since this is rarely the case a second condition is that the number of observations in a node must be greater than a given threshold.

Even if the theoretical settings of random forests was until recently restricted to the i.i.d. case, a theoretical study extending it to the time-dependent case is proposed in [14]. In addition, applications on time series could be found, as previously cited, in [7, 10], in electricity load forecasting [8], [9], [11].

The bootstrap step determines which observations are chosen to construct a tree. The original bootstrap which we call standard (or i.i.d.) bootstrap from [15] consists of randomly drawing α_n observations among the n with or without replacement. Note that we use here an abuse of language, the bootstrap is standardly defined as drawing n observations among the n observations with replacement. The goal of this bootstrap is to replicate the distribution of $\mathcal{D}_n$. However, this is adapted to the case of independent and identically distributed observations. When the data has an underlying dependence structure as for time series the i.i.d. hypothesis is not verified anymore and using the standard bootstrap destroys the dependence structure. We illustrate this phenomenon for a dataset from [16] which is described in Section 3.1. We observe in Figure 2 the original load over the month of January. Using the standard bootstrap we obtain the series in Figure 3 and immediately note that the structure we had in the original series is all gone. By contrast, using a moving block bootstrap, described in Section 2, using a block length of 24 hours we recover similar patterns as in the original series of Figure 4.

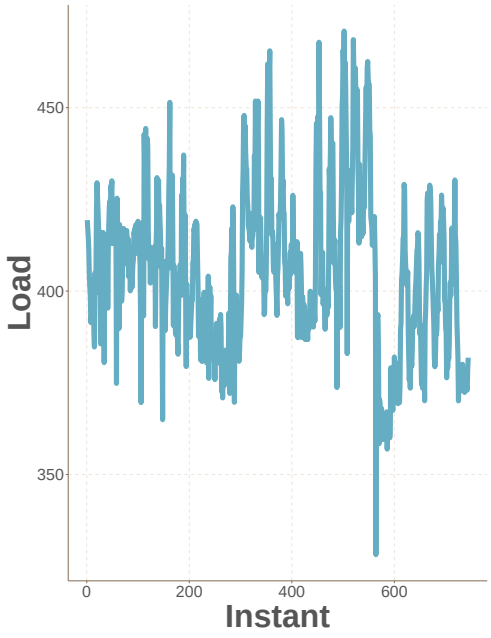

Figure 2: Original load hourly sampled.

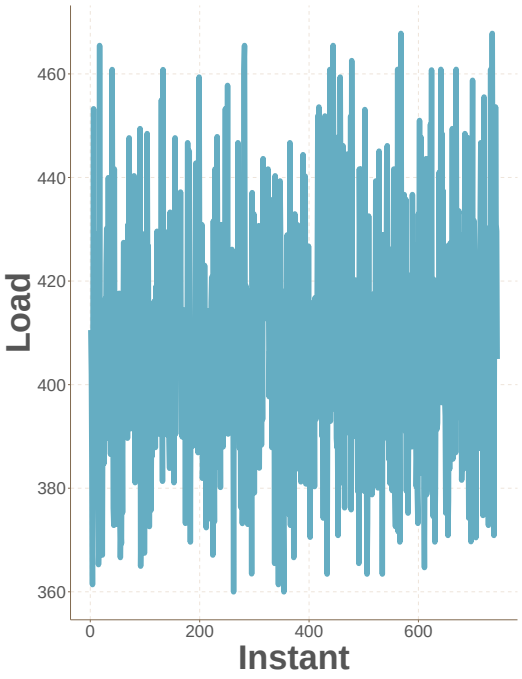

Figure 3: Bootstrapped load.

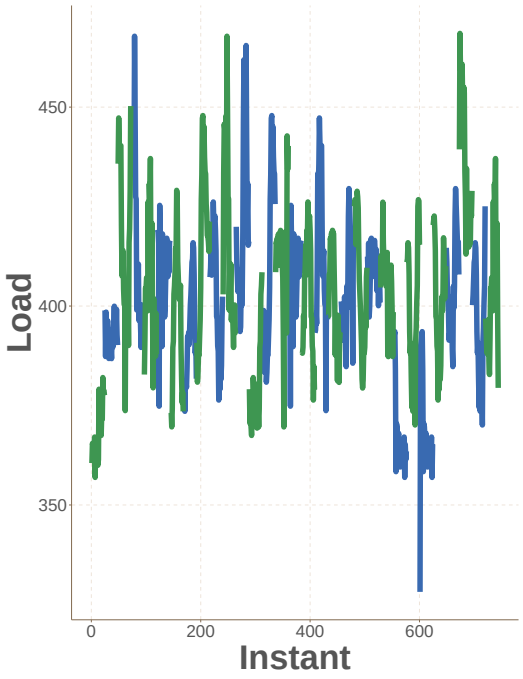

Figure 4: Block bootstrapped load with block size of 24h.

We list here a few papers using blocks bootstrap in the forecasting literature. The first one is [17] in which they use a sieve bootstrap to perform bagging with exponential smoothing models. They use exponential smoothing to decompose the data, then fit an autoregressive model to the residuals, and generate new residuals from this AR process. Finally, they fit the exponential smoothing model that was used for decomposition to all bootstrapped series. Another work is from [18] who propose a method of bagging which is as follows. After applying a Box-Cox transformation to the data, the series is decomposed into trend, seasonal and remainder components. The remainder component is then bootstrapped using the moving block bootstrap, defined in Section 2, the trend and seasonal components are added back, and the Box-Cox transformation is inverted. For each one of these bootstrapped time series, a model among several exponential smoothing models is chosen, using the bias-corrected AIC. Then, point forecasts are calculated using all the different models and the resulting forecasts are combined using the median. A companion paper [19] explores experimentally the value of bagging for time series forecasting. More generally, we refer to the special issue presented in [20] for more details about the recent developments in bootstraps methods for dependent data.

Our strategy is mainly motivated by the results on random forests in the time-dependent case in [14], proven using a block decomposition on the entries $(X_i, Y_i)_{1\leq i\leq n}$. The proofs rely on a lemma from [21] that shows that the blocks are close to being independent, under the condition that the block length is well-chosen. But it should be noted that after obtaining the bootstrap sample, the procedure to build a tree is unchanged and flipping the data after bootstrap will not change the resulting tree. The data are, in that sense and at this stage, considered to be exchangeable since the splitting criterion is unchanged and since it does not take into account the time dependence between the observations. We then try to make the data, before this stage, as much compatible with the underlying independence hypothesis.

A typical example of weak dependence is the m-dependent case, for which considering block bootstrap of length at least m allows to recover exchangeability. In the general weak dependence case, it is reasonable to consider that performing block bootstrap with a suitably chosen block-length could make the data more compatible with the exchangeable hypothesis. The aim of this work is to show that, based on the theoretical work in [14], the forecasting performance could be improved by replacing the bootstrap step by what we call block bootstrap variants, to subsample time series during the tree construction phase and thereby keep the dependence structure. This intuition is supported by the experiment reported in Appendix 2 (with time shuffling) illustrating that preserving the temporal structure is, at least empirically, beneficial.

Since random forests were already introduced in this introduction. The next section presents the different block bootstrap variants, the new algorithm and a new way to compute the variable importance. We then present two numerical experiments. The first one is based on an application to load forecasting of a building from the dataset described in [16] and see how the variants may perform. The second one on the French national forecasting problem and explore a heuristic on the choice of the new parameter.

2. RANDOM FORESTS FOR TIME SERIES

2.1. Block bootstrap variants

Non-overlapping block bootstrap. A first variant is found in [22]: the *non-overlapping block bootstrap*. The idea is to construct a number of non-overlapping blocks and then to draw uniformly, with replacement, among the constructed blocks. More precisely, let l_n be the size of a block and $B \geq 1$ the greatest integer such that $l_n B \leq n$. The blocks are then constructed in the following way

$$B_b = \big((X_{(b-1)l_n+1}, Y_{(b-1)l_n+1}), ..., (X_{bl_n}, Y_{bl_n}) \big), \qquad b = 1, ..., B.$$

The bootstrap set $\mathcal{D}_n^\star$ is then obtained by drawing K blocks, $(B_1^\star, ..., B_K^\star)$, uniformly with replacement in the collection of non-overlapping blocks$(B_b)_{1 \leq b \leq B}$ for a suitably chosen K.

Moving block bootstrap. [23] and [24] introduced the so-called *moving block bootstrap*. The idea is, instead of picking randomly one observation among the n observations as for the standard bootstrap, the moving block bootstrap pick randomly a block of l_n consecutive observations. Repeating this step and concatenating all the selected blocks, we get a new time series with a preserved structure at least in each block. More precisely, let us denote by $B_{i,l_n} = ((X_i, Y_i), ..., (X_{i+l_n-1}, Y_{i+l_n-1}))$ the block of size l_n beginning with the observation (X_i, Y_i) for $i \in \{1, ..., n - l + 1\}$. The procedure then consists to draw randomly K indices $(I_j)_{1 \leq k \leq K}$ uniformly on the set $\{1, ..., n - l_n + 1\}$ and associate one block to each index, $(B_{I_k})_{1 \leq k \leq K}$. The bootstrap set is then defined as $\mathcal{D}_n^\star = (B_{I_1}, ..., B_{I_K})$.

Circular block bootstrap. When studying the moving block bootstrap we can note that less weight is given to the endpoints of the time series which also leads in theory to non negligible bias when computing the mean. A way to correct this issue is given in [25]

introducing the so-called *circular block bootstrap*. The idea is to wrap the time series writing $X_i := X_{i_n}$ where $i_n = i \bmod n$, $X_0 := X_n$ and then use the same procedure as in the moving block bootstrap where the index I is drawn uniformly on the set $\{1, ..., n\}$ instead.

Note that in each above variant, taking $l_n = 1$ we recover the standard bootstrap of [15]. For a given number of selected observations in each tree α_n the number of blocks K is such that $K = \frac{\alpha_n}{l_n}$.

2.2. Proposed random forest for time series

Our proposition in order to incorporate the dependence structure is by replacing the first step for the construction of a random tree in the random forest building procedure, namely replacing the standard bootstrap step with one of the block bootstrap variants recalled in Section 2.1.

Note that the proposed algorithm only considers the dependence during the bootstrapping phase, directly on the entries $(X_i, Y_i)_{1 \le i \le n}$. Once the bootstrap sample is drawn the splitting is done as in the independent case. The adapted algorithm is found in Algorithm 1 underlining the modification with respect to the original random forest procedure.

input: $((X_1, Y_1), \ldots, (X_n, Y_n))$
parameters: $M, \alpha_n, m_{try}, \tau_n, \underline{l_n}$
stopping criteria: the variance in the node is zero or the number of observations in a node is below the threshold τ_n
for $j \leftarrow 1$ *to* M **do**
  Construct the jth tree:
  - Draw $\alpha_n \le n$ observations using a block bootstrap variant with parameter l_n.
  - Repeat recursively on each resulting node the following steps until a stopping criterion is met:
    - At each node, select randomly m_{try} variables
    - Select the best split using the variance criterion among the previously chosen variables.
    - Cut according to the chosen split.

end
output for a new observation x: mean of the M predictions given by the trees for x.

Algorithm 1: Random forest for time series.

Note that here, we consider the bootstrap directly on the entries $(X_i, Y_i)_{1 \le i \le n}$, and thus keeping the black box design of the random forests. Even if the time series nature of the data is forgotten after the bootstrap step, it should be noted that to include the time as a dependent variable could provide an indirect way to weakly take into account, at some extent,

the temporal nature of the data. Works on blocks bootstraps in the forecasting literature presented in Section 1 use generally the block bootstrap on the residuals after removing trends and seasonality. However, using such a procedure in our experiments (by bootstrapping the residuals of a pilot random forest) led to worse performance and further explain our approach.

2.3. Block permutation importance

Random forests can be used to rank with respect to a decreasing order of importance the variables. One way to measure the significance of a variable is the *Mean Decrease Accuracy* introduced in [1] which stems from the idea that if a variable is not important, then permuting its value should not change the prediction accuracy.

For each tree, we have access to the so-called *out-of-bag* observations denoted by OOB_m, composed of the observations not included in the bootstrap sample $\mathcal{D}_n^m$ used to construct the m-th tree. The OOB_m sample can then be used to estimate the out-of-bag error denoted by $errOOB_m$. In order to compute the importance of the variable $X^{(j)}$, the values of the j-th variable are randomly permuted in the OOB sample and we compute for each tree an out-of-bag error estimation for the permuted observations. The importance of the variable $X^{(j)}$ is then obtained by averaging the difference between the out-of-bag error before and after permutation. More formally, if, for the m-th tree, we denote by $\widetilde{errOOB}_m^j$ the OOB_m sample's error when the j-th variable is permuted, then the importance of the variable $X^{(j)}$ is defined by

$$VI\Big(X^{(j)}\Big) = \frac{1}{M}\sum_{m=1}^{M}\Big(\widetilde{errOOB}_m^j - errOOB_m\Big).$$

The higher the increase in the prediction error after the permutation of the j-th variable in the out-of-bag observations, the more important the variable is. However, if the permutation of $X^{(j)}$ does not change much the error prediction then the importance of the considered variable is small.

In the case of dependent observations we are faced with the same issue as in the construction of the random forests, namely the permutation of variable in the out-of-bag observations does not preserve the dependence structure. In the case where block instead of standard bootstrap is used in the random forest we introduce a new variable importance computation: the *block (permutation) variable importance*. However, using a block bootstrap variant does not necessarily lead to a out-of-bag observations with constant number of consecutive observations but we solve this issue in the following. Let us first suppose that the out-of-bag observations can be separated in blocks of size l_n and denote by B_m^* the blocks in the out-of-observations for the m-th tree. In order to compute the importance of the j-th variable, the permutation of the considered variable is done by only permuting the blocks in B_m^* and preserving the structure in each block. We can then compute a block permuted out-of-bag error estimation for the j-th variable denoted by $\overline{errOOB_m^j}$. The block variable importance for the j-th variable is then defined by

$$VI\Big(X^{(j)}\Big) = \frac{1}{M}\sum_{m=1}^{M}\Big(\overline{errOOB_m^j} - errOOB_m\Big).$$

The out-of-bag observations stemming from the block bootstrap with parameter l_n are not necessarily composed of blocks of the size l_n. In order to obtain an OOB sample which has the same block size as in the construction of the random forest we adapt the obtained out-of-bag observations to get a new set of blocks of out-of-bag observations as follows. The three following cases are exclusive. First, if a block of consecutive observations in the out-of-bag observations is of the right length l_n we add it to the block out-of-bag observations. Second, if the length is larger than l_n and less than $2l_n$ we draw a random subset of consecutive observations of length l_n. Finally, if a block of consecutive observations in the out-of-observations has a length less than l_n then the block is not kept. Then the block out-of-bag observations is composed of the kept block observations of length l_n and satisfies the conditions to compute the block permutation variable importance as previously defined.

3. NUMERICAL EXPERIMENTS

We consider two experiments in this work. One regarding the performance the variants may attain on a real world application of load forecasting, at a disaggregated level, on one of the building dataset from [16], which is composed of different building loads with hourly observations. The other regarding the choice of the block length parameter, this time on the French national load forecasting problem, at a more aggregated level but focusing on atypical periods.

In the following experiments, the results are obtained over 50 runs. The parameters of the random forest are set to default except for the m_{try} parameter which is optimised on a validation set and the block size parameter for which we carry out an in-depth analysis in Section 3.2.

We run the experiments by implementing the extra features we propose in this paper as an extension of the R package *ranger* [26], and thus inherit the availability in both C++ and R. Our R package *rangerts* is freely available from the github repository `https://github.com/hyanworkspace/rangerts`. Additional experiments with time series data are performed and the results can be found in the same github repository as our modified R package, omitted here for brevity reasons.

3.1. First load forecasting application: On the performance and variable importance

This experiment is based on the so-called building loads, a collection of 507 whole buildings electrical meters made publicly available. We refer to the paper [16] for a complete description of the collection. We consider one specific building in the building data genome project called *UnivLab Patrick*. This building belongs to the college laboratory category located in the New York time zone and has an area of around 7054 square meters. We have access to its electricity load from the 1st January 2015 to the 31th December 2015 with a sampling rate of one observation per hour. The weekly profile is found in Figure 5. We see a clear daily trend as well as a clear distinction between the week and the end of the week due

to less activity. We also have access to exogenous variables: the temperature as well as to the schedule of the building, indicating if a day is ordinary, a break or a holiday. We decompose the year in three parts: the training set is composed of the observations from the 1st January to the 31st October, the validation set corresponds to the month of November and the test set corresponds to the month of December.

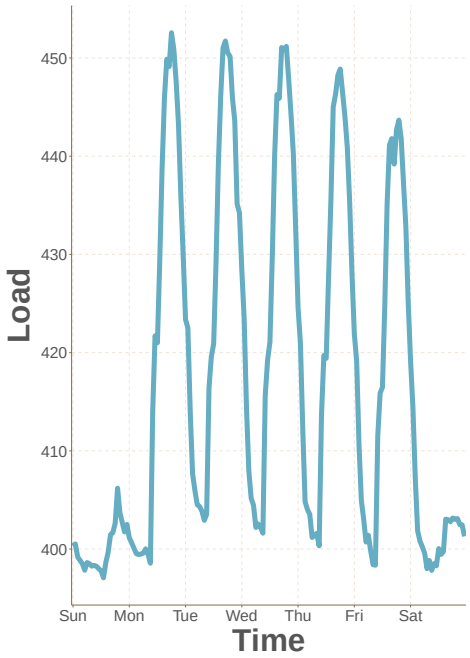

Figure 5: Weekly profile hourly sampled of the UnivLab Patrick dataset.

Let us denote by Y_t the system load of the building at hour t. In this experiment, we aim to forecast at a horizon of 24 hours. Based on the weekly profile, having hourly sampled observations, the chosen model is inspired by [27] in which they also considered random forests with a similar model for the same kind of problem. This results in the model described in (1.1) with X_t of the form

$$X_t = (Y_{t-24}, Y_{t-168}, \text{Temp}_t, \text{Schedule}_t, \text{Hour}_t, \text{InstantWeek}_t, \text{DayType}_t, \text{Time}_t) \tag{3.1}$$

where:

- Temp_t corresponds to the temperature at instant t;
- Schedule_t take three values: Regular, Break, Holiday;
- Hour_t corresponds to the hour of the day at instant t;
- InstantWeek_t corresponds to the hour in the month;
- DayType_t corresponds to the day of the week;
- Time_t corresponds to the day of the year divided by 366.

The selected value for m_{try} according to the best performance on the validation set for the standard random forest is $m_{try} = 2$. For this parameter we computed the different variants varying the block size parameters multiple of 6 hours up to 90 hours. We first optimise the

performances on the validation set, looking for the best block size value minimising the RMSE and then plug it in for the test set. The performance are resumed in Figure 6. For the sake of comparison, the baseline $Y_t = Y_{t-24}$ has a RMSE of 19.43 on the test set. We observe an improvement for the three variants with an improvement up to 11% for the mean RMSE compared to the standard random forest. We also show the evolution of the performance according to the block size parameter in Figure 7. We can find the same kind of figures for each m_{try} from 1 to 8 in Appendix 1 from Figures 15 to 22. We observe for the three variants a similar pattern in the evolution of the performance, namely a decrease for which the three variants performs better than the standard random forest and then an increase. We note that, even if the performance get worse when the block size is large, we also have a large window for which the performance is far better for these three variants with an optimal block size parameter of around 24 hours also corresponding to the forecasting horizon and the main seasonality of the data.

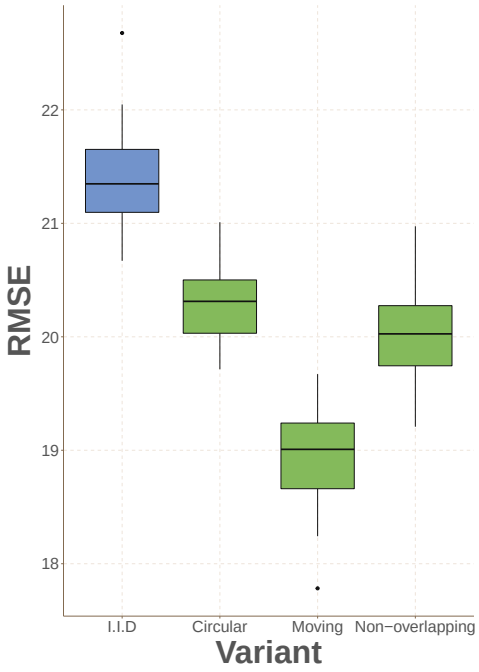

Figure 6: Performance of the different variants for $m_{try} = 2$, evaluated on the month of December of the UnivLab Patrick dataset.

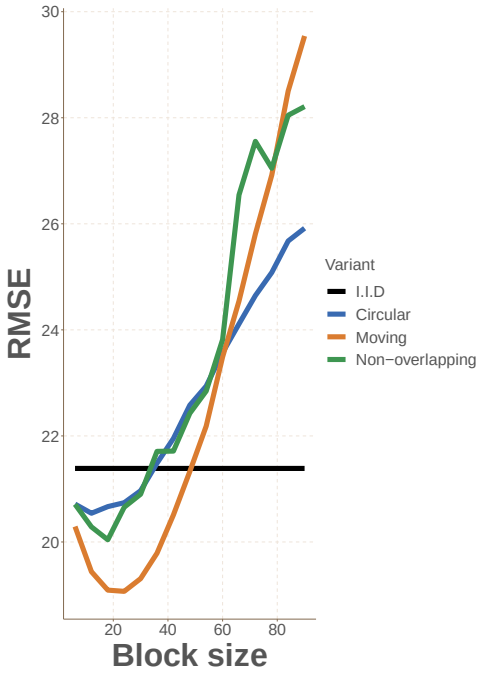

Figure 7: Performance of the variants for $m_{try} = 2$ when the block size changes, evaluated on the month of December of the UnivLab Patrick dataset.

One may wonder if the block bootstrap mechanism really helps to take into account time dependence or if it is another underlying mechanism. In order to illustrate this point, we shuffled the instances in the training set. If it was another mechanism at play, we would have the same results as before. The results after shuffling the training set can be found in Appendix 2. We can clearly see that once the training set does not have the dependence structure, using the block bootstrap variants has basically the same behaviour as the standard random forests, regardless of the block length, and thus further confirms that the block bootstrap random forests take into account the dependence structure.

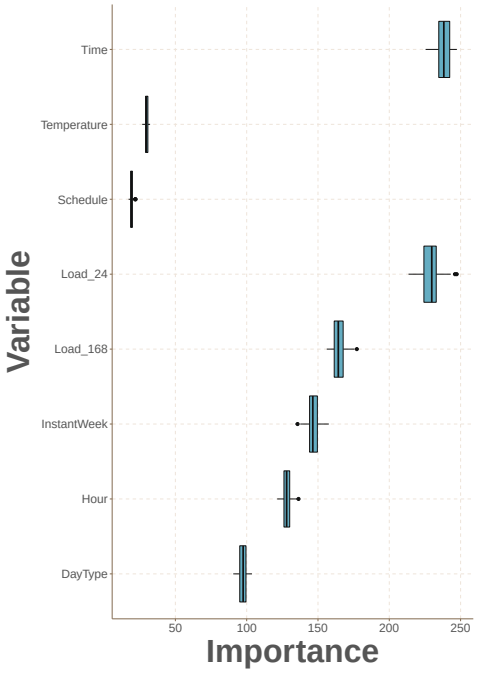

Figure 8: Variable importance moving bootstrap variant under the standard permutation on the UnivLab Patrick dataset.

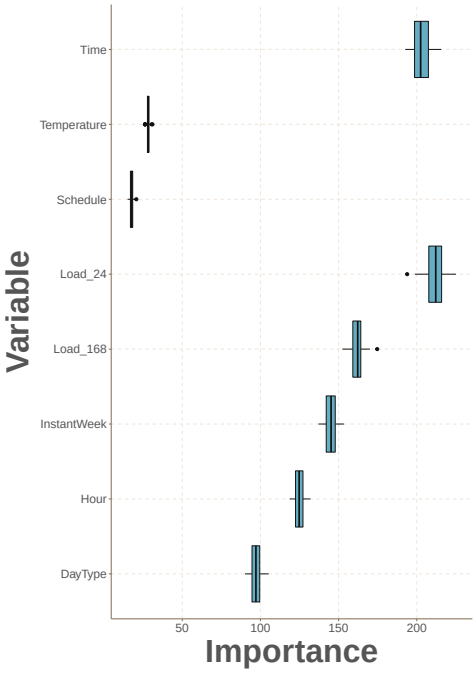

Figure 9: Block moving bootstrap variant importance with block size of 24h on the UnivLab Patrick dataset.

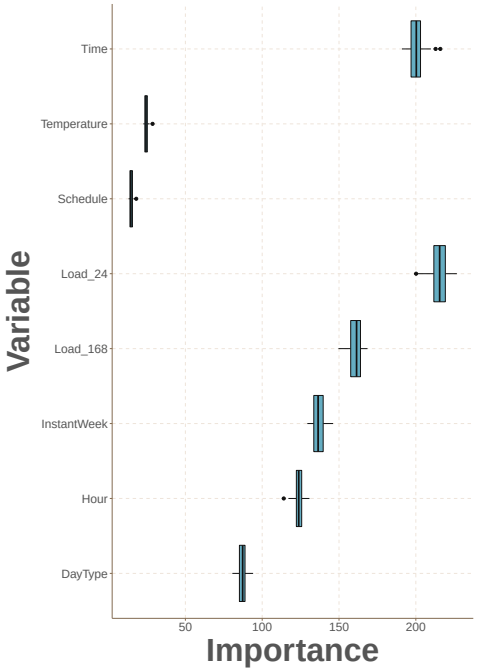

Figure 10: Variable importance non-overlapping variant under the standard permutation on the UnivLab Patrick dataset.

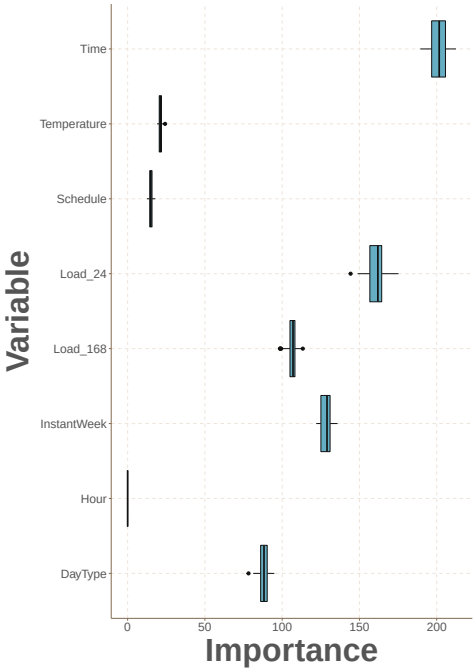

Figure 11: Block non-overlapping variant importance with block size of 24h on the UnivLab Patrick dataset.

Computing the variable importance for blocks of size 24 hours we obtain Figures 8 to 11. We observe that the difference between the standard variable importance and the block variable importance is essentially noticeable for the non-overlapping block bootstrap variant.

The most evident difference is for the variable *Hour* for which the importance is set to zero using the block variable importance. Since the blocks are of length 24 hours and always beginning at the same time, permuting the blocks will not change the out-of-bag error since each permutation is replaced by an identical copy and thus the output from this procedure for the variable *Hour*.

3.2. Second load forecasting application: On the block length choice

We discuss here the choice of the block length parameter, found in every block bootstrap variant. In the previous experiment, we notice that the optimal choice for the block length was 24 hours, corresponding to the daily step and seasonality in the dataset. However, the last experiment is done by optimising the block length on the validation set error. It would be interesting to choose this parameter more wisely in order to avoid unnecessary computations and we think that it should be proportional to the (minimal) seasonality in the dataset. The block bootstrap aims to build blocks that preserve the dependency in them but that the blocks are independent to a certain extent. In the case of seasonal trends, the intuition would consequently be to choose blocks correlated to basic seasonal components. We illustrate this with another dataset, on the French national load with goal to forecast at a 24 hours horizon as well, having a longer span of time and thus having more stable results.

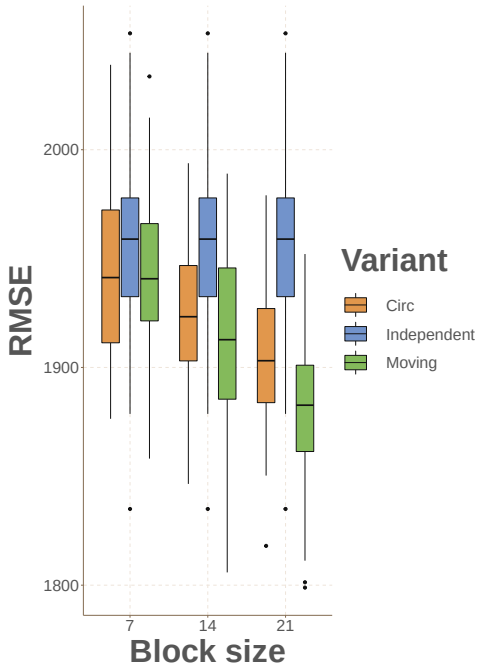

Figure 12: Performances evaluated on April 2016 on the French load forecasting problem of the different variants for three block length values.

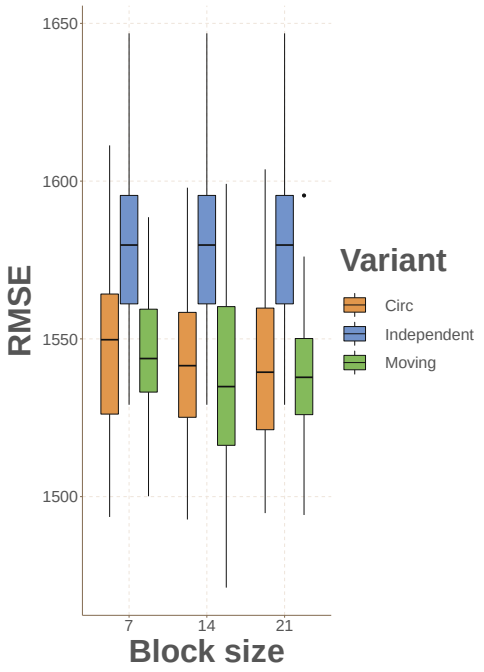

Figure 13: Performances evaluated on October 2016 on the French load forecasting problem of the different variants for three block length values.

We consider the French electricity load of the year 2015 as the training set with a sampling rate of one observation per day at noon. The test set for this experiment are the months April and October of the year 2016, corresponding to the transition between summer and

winter season, a particularly difficult period to forecast. We observed in various experiments that the random forests for time series variants work the best when it is "difficult" to forecast. This typically corresponds to the shoulder seasons in the load forecasting field. We use here the model described in (3.1) as well without the variables *Hour* and *InstantWeek*. Since the observations are daily occurrences, the minimal seasonality would be the week. Hence, we consider three values for the block length parameter: 7, 14 and 21 days. The selected value for m_{try} is 3 corresponding to the worst case scenario, in the sense that for another value of m_{try} the block bootstrap variants are doing better than shown in this example. Note that for this example we removed the non-overlapping block bootstrap variant. We have found that this variant needs more observations to get consistent results, providing less diversity in the trees due to its construction.

The results are found, respectively for April and October 2016, in Figures 12 and 13. We observe that, for both months, we have a consistent improvement of the performance in comparison to the standard random forest for each choice of block length. We even note significant improvement in the performance when taking twice or thrice the seasonality for April. However, taking larger values than these would lead to a diversity problem in the trees as mentioned before and thus have less consistent performance. This concludes that the heuristic for the block length parameter choice would be to take the smallest seasonality up to a multiplying factor of two or three.

3.3. Supplementary experiments

Further experiments are carried out with two forecasting competition data sets: quarterly and monthly series from M3 [28] (2184 series, 756 quarterly data and 1428 monthly data) and M4 [29] (4151 series, 402 quarterly data and 3749 monthly data) competitions to assess the performance of the proposed variants. Our main objective here is to compare the performance of the standard random forest and the block-bootstrap variants extensively on general time series data, instead of accessing how competitive the random forest algorithm itself is for these two data sets. Note that both stationary and non-stationary data are included in the data set whereas random forest cannot extrapolate and thus performs poorly on non-stationary data comparing to other baseline time series methods in the literature such as ARIMA models. The metrics we use for evaluation here are the normalized RMSE (NRMSE $= \frac{RMSE}{\sigma(serie_{train})}$, where σ is the standard deviation of the training part of the series), and the normalized difference in MAPE (NdMAPE $= \frac{\Delta(MAPE)}{MAPE_{i.i.d.}}$) where $\Delta(MAPE) = MAPE_{i.i.d.} - MAPE_{variant}$. Higher values indicate better results with variants.

As described in (1.1), let us denote by Y_t the series to be predicted at step t. Unlike the load forecasting application, only the frequency and time features are used. For monthly data, the frequency feature ranges from 1 to 12, which corresponds to the month. For quarterly, this feature is thus 1 to 4, and 1 stands for the first quarter. By regressing on time, we aim at estimating the trend and the seasonality components of each series. Including lags as explanatory variable would be a natural choice in time series forecasting tasks, here we choose not to do that to stay as far as possible from the exchangeability of the data.

We keep all other hyper parameters of the random forest identical to the standard i.i.d. version to compare the obtained results with those from the block bootstrap variants.

The only hyper parameter remains to be tuned is thus the block size. To be able to choose the block size automatically, we propose to set a general auto-correlation threshold for all series, to determine for each of them, the largest lag as the block length.

Better performance is achieved as shown in Figure 14 with the moving block variant on the monthly series (the same for the M4 data set). A Wilcoxon signed rank test confirms the gain with respect to the standard i.i.d. forest. We also observe in Table 1 that in general, higher auto-correlation thresholds lead to better results.

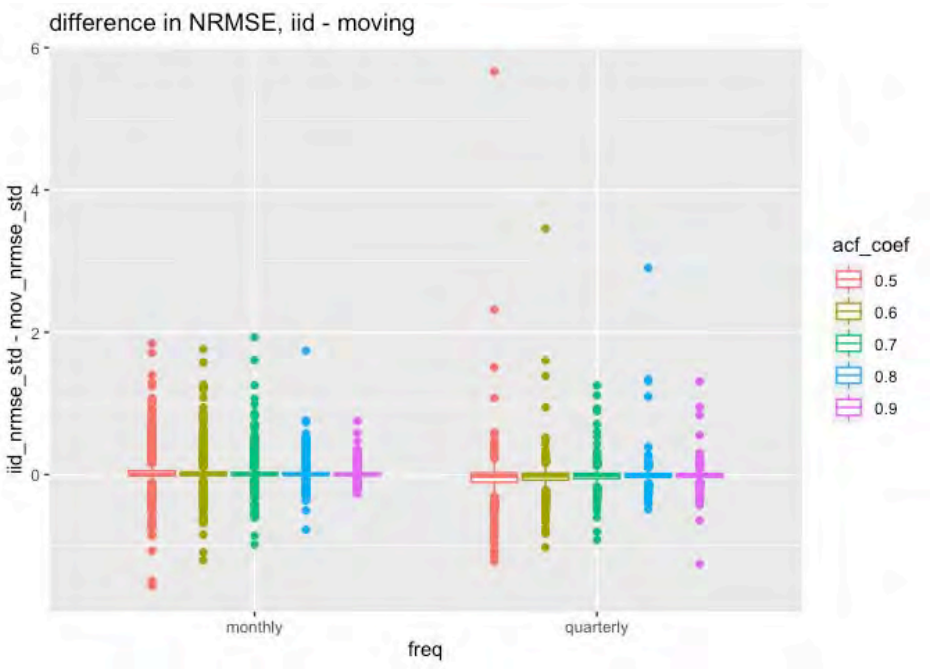

Figure 14: Difference in NRMSE of the standard random forest (i.i.d.) and the moving block variant (moving), for monthly and quarterly data, with different auto-correlation threshold values from 0.5 to 0.9, from the M3 data set.

Table 1: The percentage of cases where the block bootstrap variant outperforms the i.i.d. in terms of NdMAPE.

acf_coef	M3	M4
0.5	0.581	0.488
0.6	0.567	0.496
0.7	0.589	0.505
0.8	0.586	0.515
0.9	0.572	0.515

We choose to present our major results with a restricted number of graphs and statistics to conserve space. All the codes and other supporting materials can be found in the same GitHub repository as our implemented variants under the sub-directory benchmark_Mcomp.

4. CONCLUSION AND PERSPECTIVES

We introduced a new variant of random forests taking into account the temporal dependency of the observations and showed that we can improve significantly the performance on forecasting tasks when choosing the right block length. A variant of the variable importance based on the block bootstrap mechanism is also introduced. The non-overlapping variant seems to be mistaken regarding the importance of the variables, forgetting some variables fundamental to the forecasting problem as the hour variable in our first application, and thus we do not advise to use this variant for this purpose. However, both moving and circular variants seem to perform much better than the standard random forests when the block length is well-chosen, and we showed that a good heuristic for the block length choice is correlated to a multiple of the smallest seasonality.

This work is mainly methodological, a first perspective would be to prove theoretical results on the random forests variants under time-dependent observations hypotheses. Consistency of random forests is proven under stationary and $\beta-$mixing hypotheses in [14] when trees are not fully grown and the observations are subsampled. The previously cited works regarding the block bootstrap as [22, 23, 24, 25] also show consistency of some estimators, generally under less restrictive hypotheses. It would be interesting to prove similar results on the variants by adapting and combining the previous proof techniques.

We have performed a detailed study on one specific field of application and an automatic extensive study was conducted on the time series of the M3 and M4 competitions. We illustrated the potential value of the random forests variants. We also showed that it could be useful to develop an adaptive and automatic way to choose the block length parameter. Finally, it could be interesting to explore more deeply under which conditions (input variables, etc.) the variants work, going well beyond the scope of this paper.

APPENDIX 1

Performance of the variants for each given m_{try} from 1 to 8, when the block size changes, evaluated on the month of December of the UnivLab Patrick dataset can be found from Figures 15 to 22.

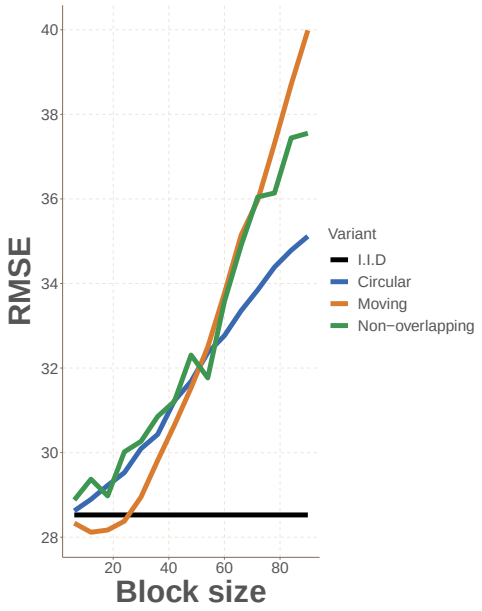

Figure 15: Performance of the variants for $m_{try}=1$ when the block size changes, evaluated on the month of December of the UnivLab Patrick dataset.

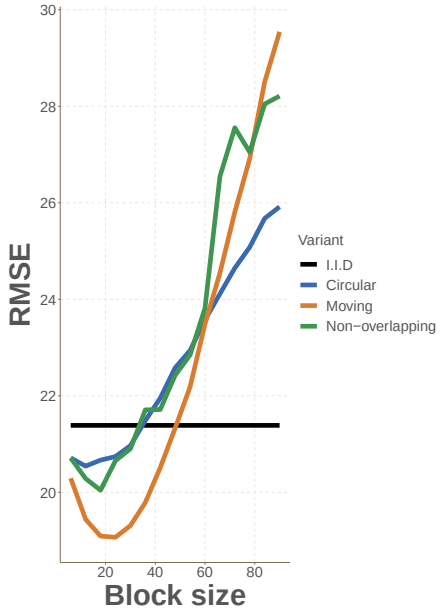

Figure 16: Performance of the variants for $m_{try}=2$ when the block size changes, evaluated on the month of December of the UnivLab Patrick dataset.

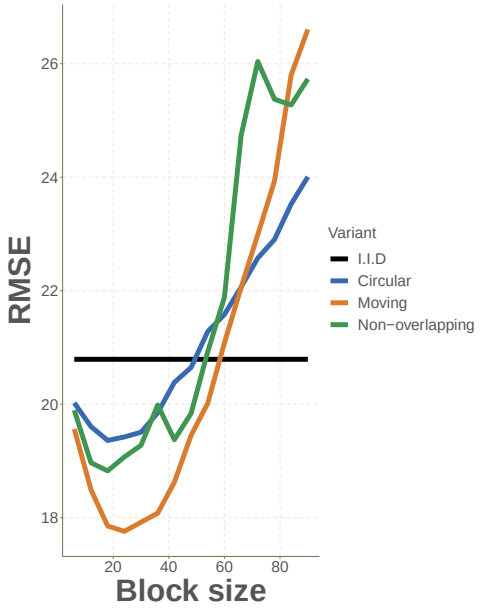

Figure 17: Performance of the variants for $m_{try}=3$ when the block size changes, evaluated on the month of December of the UnivLab Patrick dataset.

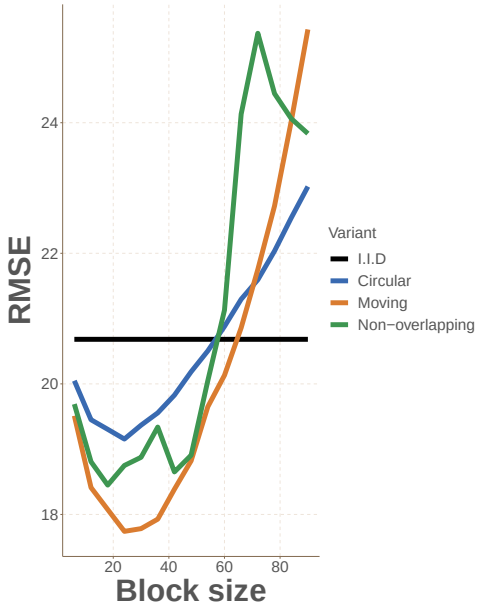

Figure 18: Performance of the variants for $m_{try}=4$ when the block size changes, evaluated on the month of December of the UnivLab Patrick dataset.

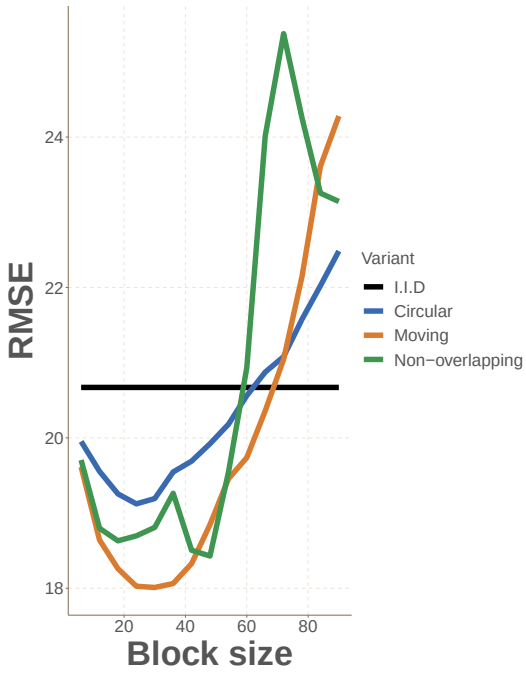

Figure 19: Performance of the variants for $m_{try} = 5$ when the block size changes, evaluated on the month of December of the UnivLab Patrick dataset.

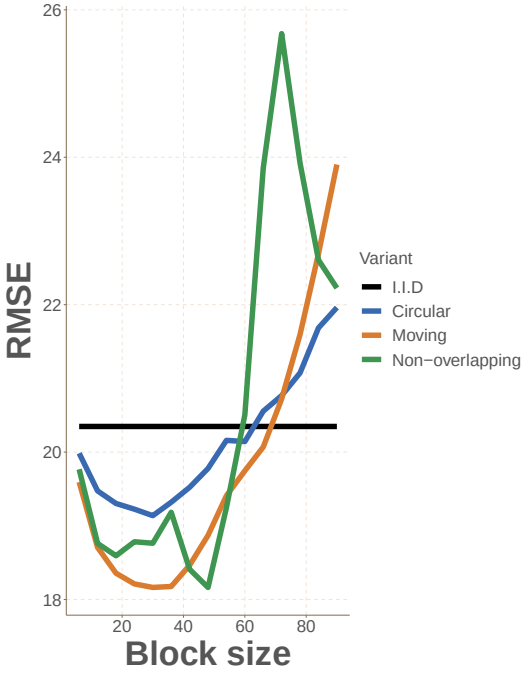

Figure 20: Performance of the variants for $m_{try} = 6$ when the block size changes, evaluated on the month of December of the UnivLab Patrick dataset.

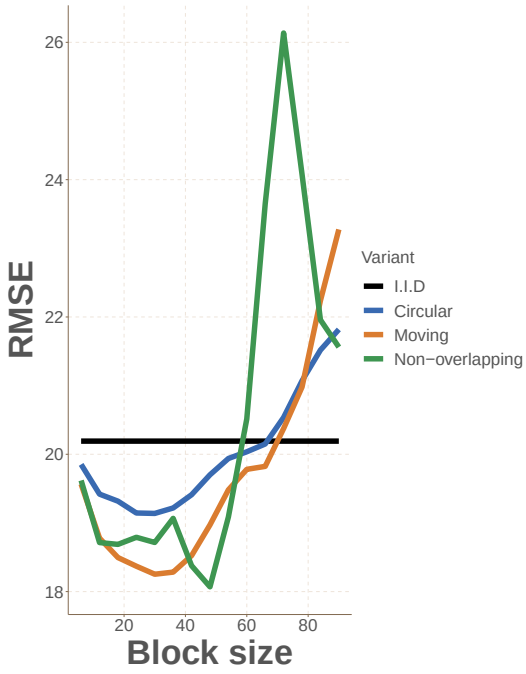

Figure 21: Performance of the variants for $m_{try} = 7$ when the block size changes, evaluated on the month of December of the UnivLab Patrick dataset.

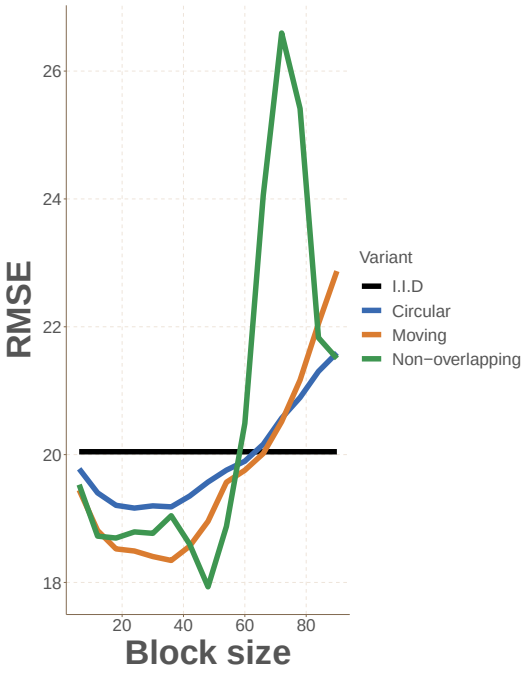

Figure 22: Performance of the variants for $m_{try} = 8$ when the block size changes, evaluated on the month of December of the UnivLab Patrick dataset.

APPENDIX 2

Performance of the variants, when the observations in the training set are shuffled beforehand, for m_{try} equal to 1 and 2, when the block size changes, evaluated on the month of December of the UnivLab Patrick dataset can be found from Figures 23 to 24. We have similar results for m_{try} from 3 to 8.

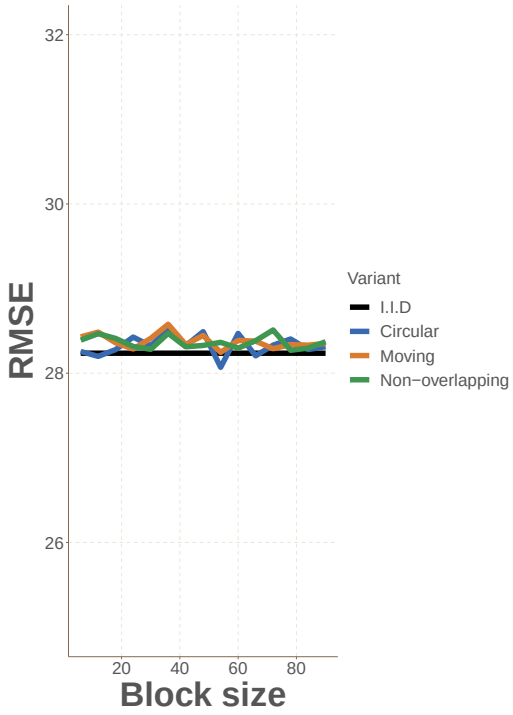

Figure 23: Performance of the variants when training set is shuffled for $m_{try}=1$ when the block size changes, evaluated on the month of December of the UnivLab Patrick dataset.

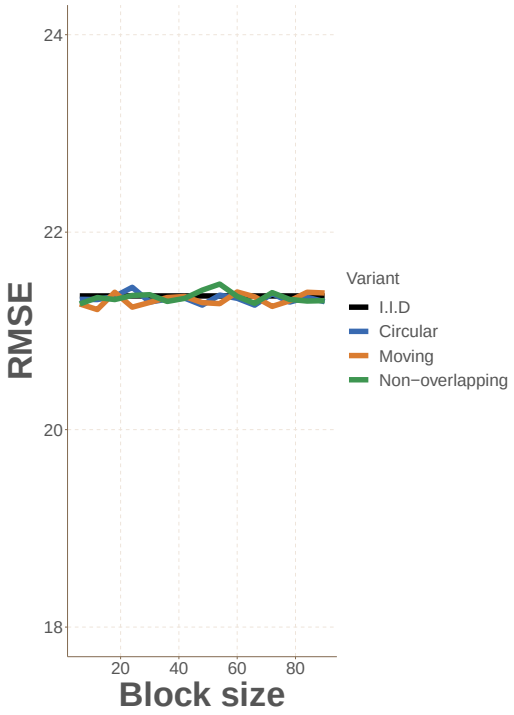

Figure 24: Performance of the variants when training set is shuffled for $m_{try}=2$ when the block size changes, evaluated on the month of December of the UnivLab Patrick dataset.

ACKNOWLEDGMENTS

The authors would like to thank the reviewer for their helpful comments and suggestions.

REFERENCES

[1] Breiman, L. (2001). Random forests, *Machine Learning*, **45**(1), 5–32.

[2] Fernández-Delgado, M.; Cernadas, E.; Barro, S. and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?, *The Journal of Machine Learning Research*, **15**(1), 3133–3181.

[3] Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P. and Feuston, B.P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling, *Journal of Chemical Information and Computer Sciences*, **43**(6), 1947–1958.

[4] Cutler, D.R.; Edwards, T.C.; Beard, K.H.; Cutler, A.; Hess, K.T.; Gibson, J. and Lawler, J.J. (2007). Random forests for classification in ecology, *Ecology*, **88**(11), 2783–2792.

[5] Prasad, A.M.; Iverson, L.R. and Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction, *Ecosystems*, **9**(2), 181–199.

[6] Shotton, J.; Sharp, T.; Kipman, A.; Fitzgibbon, A.; Finocchio, M.; Blake, A.; Cook, M. and Moore, R. (2013). Real-time human pose recognition in parts from single depth images, *Communications of the ACM*, **56**(1), 116–124.

[7] Kane, M.J.; Price, N.; Scotch, M. and Rabinowitz, P. (2014). Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks, *BMC Bioinformatics*, **15**(1), 276.

[8] Dudek, G. (2015). *Short-term load forecasting using random forests*. In "Intelligent Systems' 2014", Springer, pp. 821–828.

[9] Lahouar, A. and Ben Hadj Slama, J. (2015). *Random forests model for one day ahead load forecasting*. In "IREC2015 – The Sixth International Renewable Energy Congress", 1–6.

[10] Fischer, A.; Montuelle, L.; Mougeot, M. and Picard, D. (2017). Statistical learning for wind power: a modeling and stability study towards forecasting, *Wind Energy*, **20**(12), 2037–2047.

[11] Moon, J.; Kim, Y.; Son, M. and Hwang, E. (2018). Hybrid short-term load forecasting scheme using random forest and multilayer perceptron, *Energies*, **11**, 3283.

[12] Breiman, L.; Friedman, J.; Stone, C.J. and Olshen, R.A. (1984). *Classification and Regression Trees*, The Wadsworth and Brooks–Cole Statistics-Probability Series, Taylor & Francis.

[13] Breiman, L. (1996). Bagging predictors, *Machine Learning*, **24**(2), 123–140.

[14] Goehry, B. (2020). Random forests for time-dependent processes, *ESAIM: PS*, **24**, 801–826.

[15] Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Ann. Statist.*, **7**(1), 1–26.

[16] Miller, C. and Meggers, F. (2017). The Building Data Genome Project: an open, public data set from non-residential building electrical meters, *Energy Procedia*, **122**, 439–444.

[17] Cordeiro, C. and Neves, M. (2009). Forecasting time series with BOOT. EXPOS procedure, *REVSTAT – Statistical Journal*, **7**(2), 135–149.

[18] Bergmeir, C.; Hyndman, R.J. and Benítez, J.M. (2016). Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation, *International Journal of Forecasting*, **32**(2), 303–312.

[19] Petropoulos, F.; Hyndman, R.J. and Bergmeir, C. (2018). Exploring the sources of uncertainty: why does bagging for time series forecasting work?, *European Journal of Operational Research*, **268**(2), 545–554.

[20] Cavaliere, G.; Politis, D.N. and Rahbek, A. (2015). Recent developments in bootstrap methods for dependent data, *Journal of Time Series Analysis*, **36**(3), 269–271.

[21] Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences, *The Annals of Probability*, **22**(1), 94–116.

[22] Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence, *Ann. Statist.*, **14**(3), 1171–1179.

[23] Kunsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations, *Ann. Statist.*, **17**(3), 1217–1241.

[24] Liu, R.Y. and Singh, K. (1992). *Moving blocks jackknife and bootstrap capture weak dependence*. In "Exploring the Limits of Bootstrap (East Lansing, MI, 1990), Wiley Ser. Probab. Math. Statist.", Wiley, New York, pp. 225–248.

[25] Politis, D.N. and Romano, J.P. (1992). *A circular block-resampling procedure for stationary data*. In "Exploring the Limits of Bootstrap (East Lansing, MI, 1990), Wiley Ser. Probab. Math. Statist.", Wiley, New York, pp. 263–270.

[26] Wright, M. and Ziegler, A. (2017). ranger: a fast implementation of random forests for high dimensional data in C++ and R, *Journal of Statistical Software, Articles*, **77**, 1–17.

[27] Goehry, B.; Goude, Y.; Massart, P. and Poggi, J.-M. (2019). Aggregation of multi-scale experts for bottom-up load forecasting, *IEEE Transactions on Smart Grid*, **11**(3), 1895–1904.

[28] Makridakis, S. and Hibon, M. (2000). The M3-Competition: results, conclusions and implications, *International Journal of Forecasting*, (6), 451–476.

[29] Makridakis, S.; Spiliotis, E. and Assimakopoulos, V. (2018). The M4-Competition: results, findings, conclusion and way forward, *International Journal of Forecasting*, (34), 802–808.

REVSTAT-Statistical journal

Aims and Scope

The aim of REVSTAT-Statistical Journal is to publish articles of high scientific content, developing Statistical Science focused on innovative theory, methods, and applications in different areas of knowledge. Important survey/review contributing to Probability and Statistics advancement is also welcome.

Background

Statistics Portugal started in 1996 the publication of the scientific statistical journal *Revista de Estatística*, in Portuguese, a quarterly publication whose goal was the publication of papers containing original research results, and application studies, namely in the economic, social and demographic fields. Statistics Portugal was aware of how vital statistical culture is in understanding most phenomena in the present-day world, and of its responsibilities in disseminating statistical knowledge.

In 1998 it was decided to publish papers in English. This step has been taken to achieve a larger diffusion, and to encourage foreign contributors to submit their work. At the time, the editorial board was mainly composed by Portuguese university professors, and this has been the first step aimed at changing the character of *Revista de Estatística* from a national to an international scientific journal. In 2001, the *Revista de Estatística* published a three volumes special issue containing extended abstracts of the invited and contributed papers presented at the 23rd European Meeting of Statisticians (EMS). During the EMS 2001, its editor-in-chief invited several international participants to join the editorial staff.

In 2003 the name changed to REVSTAT-Statistical Journal, published in English, with a prestigious international editorial board, hoping to become one more place where scientists may feel proud of publishing their research results.

Editorial policy

REVSTAT-Statistical Journal is an open access peer-reviewed journal published quarterly, in English, by Statistics Portugal.

The editorial policy of REVSTAT is mainly placed on the originality and importance of the research. The whole submission and review processes for REVSTAT are conducted exclusively online on the journal's webpage revstat.ine.pt based in Open Journal System (OJS). The only working language allowed is English. Authors intending to submit any work must register, login and follow the guidelines.

There are no fees for publishing accepted manuscripts that will be made available in open access.

All articles consistent with REVSTAT aims and scope will undergo scientific evaluation by at least two reviewers, one from the Editorial Board and another external. Authors can suggest an editor or reviewer who is expert on the paper subject providing her/his complete information, namely: name, affiliation, email and, if possible, personal URL or ORCID number.

All published works are Open Access (CC BY 4.0) which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. Also, in the context of archiving policy, REVSTAT is a *blue* journal welcoming authors to deposit their works in other scientific repositories regarding the use of the published edition and providing its source.

Journal prints may be ordered at expenses of the author(s), and prior to publication.

Abstract and Indexing services

REVSTAT-Statistical Journal is covered *by Journal Citation Reports - JCR (Clarivate); DOAJ-Directory Of Open Access Journals; Current Index to Statistics; Google Scholar; Mathematical Reviews® (MathSciNet®); Zentralblatt für Mathematic; Scimago Journal & Country Rank; Scopus*

Author guidelines

The whole submission and review processes for REVSTAT are conducted exclusively online on the journal's webpage https://revstat.ine.pt/ based in Open Journal System (OJS). Authors intending to submit any work must *register, login* and follow the indications choosing *Submissions*.

REVSTAT - Statistical Journal adopts the COPE guidelines on publication ethics.

Work presentation

- the only working language is English;
- the first page should include the name, ORCID iD (optional), Institution, country, and mail-address of the author(s);
- a summary of fewer than one hundred words, followed by a maximum of six keywords and the MSC 2020 subject classification should be included also in the first page;
- manuscripts should be typed only in black, in double-spacing, with a left margin of at least 3 cm, with numbered lines, and a maximum of 25 pages;
- the title should be with no more than 120 characters (with spaces);
- figures must be a minimum of 300dpi and will be reproduced online as in the original work, however, authors should take into account that the printed version is always in black and grey tones;
- authors are encouraged to submit articles using LaTeX which macros are available at *REVSTAT style*;
- citations in text should be included in the text by name and year in parentheses, as in the following examples: § article title in lowercase (Author 1980); § This

theorem was proved later by AuthorB and AuthorC (1990); § This subject has been widely addressed (AuthorA 1990; AuthorB et al. 1995; AuthorA and AuthorB 1998).

- references should be listed in alphabetical order of the author's scientific surname at the end of the article;
- acknowledgments of people, grants or funds should be placed in a short section before the References title page. Note that religious beliefs, ethnic background, citizenship and political orientations of the author(s) are not allowed in the text;
- authors are welcome to suggest one of the Editors or Associate Editors or yet other reviewer expert on the subject providing a complete information, namely: name, affiliation, email and personal URL or ORCID number in the Comments for the Editor (submission form).

Accepted papers

After final revision and acceptance of an article for publication, authors are requested to provide the corresponding LaTeX file, as in REVSTAT style.

Supplementary files may be included and submitted separately in .tiff, .gif, .jpg, .png .eps, .ps or .pdf format. These supplementary files may be published online along with an article, containing data, programming code, extra figures, or extra proofs, etc; however, REVSTAT is not responsible for any supporting information supplied by the author(s).

Copyright Notice

Upon acceptance of an article, the author(s) will be asked to transfer copyright of the article to the publisher, Statistics Portugal, in order to ensure the widest possible dissemination of information.

According to REVSTAT's *archiving policy*, after assigning the copyright form, authors may cite and use limited excerpts (figures, tables, etc.) of their works accepted/published in REVSTAT in other publications and may deposit only the published edition in scientific repositories providing its source as REVSTAT while the original place of publication. The Executive Editor of the Journal must be notified in writing in advance.

EDITORIAL BOARD 2019-2023

Editor-in-Chief

Isabel FRAGA ALVES, University of Lisbon, Portugal

Co-Editor

Giovani L. SILVA, University of Lisbon, Portugal

Associate Editors

Marília ANTUNES, University of Lisbon, Portugal
Barry ARNOLD, University of California, USA
Narayanaswamy BALAKRISHNAN, McMaster University, Canada
Jan BEIRLANT, Katholieke Universiteit Leuven, Belgium
Graciela BOENTE, University of Buenos Aires, Argentina
Paula BRITO, University of Porto, Portugal
Valérie CHAVEZ-DEMOULIN, University of Lausanne, Switzerland
David CONESA, University of Valencia, Spain
Charmaine DEAN, University of Waterloo, Canada
Fernanda FIGUEIREDO, University of Porto, Portugal
Jorge Milhazes FREITAS, University of Porto, Portugal
Alan GELFAND, Duke University, USA
Stéphane GIRARD, Inria Grenoble Rhône-Alpes, France
Marie KRATZ, ESSEC Business School, France
Victor LEIVA, Pontificia Universidad Católica de Valparaíso, Chile
Artur LEMONTE, Federal University of Rio Grande do Norte, Brazil
Shuangzhe LIU, University of Canberra, Australia
Maria Nazaré MENDES-LOPES, University of Coimbra, Portugal
Fernando MOURA, Federal University of Rio de Janeiro, Brazil
John NOLAN, American University, USA
Paulo Eduardo OLIVEIRA, University of Coimbra, Portugal
Pedro OLIVEIRA, University of Porto, Portugal
Carlos Daniel PAULINO, University of Lisbon, Portugal
Arthur PEWSEY, University of Extremadura, Spain
Gilbert SAPORTA, Conservatoire National des Arts et Métiers, France
Alexandra M. SCHMIDT, McGill University, Canada
Manuel SCOTTO, University of Lisbon, Portugal
Lisete SOUSA, University of Lisbon, Portugal
Milan STEHLÍK, University of Valparaíso, Chile and LIT-JK University Linz, Austria
María Dolores UGARTE, Public University of Navarre, Spain

Executive Editor

Olga Bessa MENDES, Statistics Portugal