

REVSTAT

Statistical Journal

vol. 20 - no. 3 - July 2022

Special Issue

Celebrating 20 years of publication



REVSTAT-Statistical Journal, vol.20, n. 3 (July 2022)

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

Continues: Revista de Estatística = ISSN 0873-4275.

ISSN 1645-6726 ; e-ISSN 2183-0371

Editorial Board (2019-2023)

Editor-in-Chief – *Isabel FRAGA ALVES*

Co-Editor – *Giovani L. SILVA*

Associate Editors

Marília ANTUNES

Barry ARNOLD

Narayanaswamy BALAKRISHNAN

Jan BEIRLANT

Graciela BOENTE

Paula BRITO

Valérie CHAVEZ-DEMOULIN

David CONESA

Charmaine DEAN

Fernanda FIGUEIREDO

Jorge Milhazes FREITAS

Alan GELFAND

Stéphane GIRARD

Marie KRATZ

Victor LEIVA

Artur LEMONTE

Shuangzhe LIU

Maria Nazaré MENDES-LOPES

Fernando MOURA

John NOLAN

Paulo Eduardo OLIVEIRA

Pedro OLIVEIRA

Carlos Daniel PAULINO

Arthur PEWSEY

Gilbert SAPORTA

Alexandra M. SCHMIDT

Manuel SCOTTO

Julio SINGER

Lisete SOUSA

Milan STEHLÍK

María Dolores UGARTE

Executive Editor – *José A. PINTO MARTINS*

Assistant Editors – *José Cordeiro | Olga Bessa Mendes*

Publisher – *Statistics Portugal*

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

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



Creative Commons Attribution 4.0 International (CC BY 4.0)

© Statistics Portugal, Lisbon. Portugal, 2022

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

Editorial

Celebrating 20 years of publication | *REVSTAT-Statistical Journal*

This year we complete 20 years of publication of *REVSTAT-Statistical Journal*.

Statistics Portugal ([INE](#)), aware of its responsibilities in disseminating statistical knowledge and intending to fill a gap of scientific publication in Statistics in 1996, started the publication of [Revista de Estatística](#) in Portuguese, a quarterly publication whose goal was the publication of papers containing original statistical methodological results, and its applications, namely in the economic, social and demographic fields.

A decisive step was taken in 1998, publishing papers also in English, aiming to change the character of *Revista de Estatística* from a national to an international scientific journal, the *Statistical Review*. Afterwards, during the EMS 2001, the 23rd European Meeting of Statisticians, its Editor-in-Chief Ivette Gomes invited several international top researcher participants to join the editorial board for the becoming new international journal and in 2003 the name changed to [REVSTAT-Statistical Journal](#).

The first issue of REVSTAT, [Vol1\(1\)](#), was published in November 2003 and consisted of four papers in different topics, namely in random-effects log-linear model, extremal index of sub-sampled periodic sequences, in lifetime models and multiple correspondence analysis.

The year of 2010 was the first electronic [JCR](#) (Journal Citation Reports) year. Since then, an impressive increasing number of submissions has occurred. The year of 2015 was the first one with more than one hundred of submissions. It should be mentioned that, in the period 2015-2021, an average yearly number of around 150 submissions have been received in our *REVSTAT*, with a mean number of 25 published papers per year.

The rate of acceptance of the last three years is around 20%.



Ivette Gomes has acted, heroically, till the end of 2018, as Editor-in-Chief of REVSTAT, improving the visibility and impact of REVSTAT, serving the science community with dedication and commitment, setting up and shaping the journal to ensure it has maintained its place among the international benchmark journals in the field of Statistics. The importance of obtaining high quality reviews and timeliness in publish-decisions, has been a major concern under Ivette's Editorship, assisted by the Co-Editor Antónia Amaral Turkman.

Since 2019, the current Editor-in-Chief Isabel Fraga Alves and Co-Editor Giovani L. Silva have been trying to maintain the high level of the journal, acting with the utmost scientific integrity, and promoting some special issues too, spreading the scientific communications of some thematic meetings. One of the keystones that we propose ourselves was the creation of a new [journal platform](#), making possible not only the electronic submission and refereeing process for our [Associate Editors](#), but also making true the aim of DOI attribution and in a near future the Open Access (OA) registration in [DOAJ](#), although REVSTAT has been OA since its birth. This goal is now partially accomplished, made possible due to our teamwork with INE and represents a continuous work in progress. We assume that these recent years during covid-19 pandemic were very difficult, and represented a big challenge, not easy to accomplish by the two Editors, who must answer to their normal professional academic duties, apart from this *pro bono* job.

At last, we should enhance the fact that a successful journal depends on the efficiency and competence of the teamwork with its editorial board and its reviewers. We would like to take this opportunity to thank all Associate Editors, Executive Editors, Reviewers, Authors, and the readers of REVSTAT for their support. In the future, it is our continuous goal to improve the profile of the REVSTAT Journal, next to the objectives of its readers, answering not only to the theoretical, methodological, and applied topics but also to the nowadays bigdata paradigm.

July 14, 2022

Isabel Fraga Alves
Giovani L. Silva

INDEX

Bias Reduced Peaks over Threshold Tail Estimation

Jan Beirlant, Gaonyalelwe Maribe, Philippe Naveau and Andréhette Verster 277

A Note on the Right Truncated Weibull Distribution and the Minimum of Power Function Distributions

Pedro Jodrá 305

Asymptotic Confidence Intervals for the Difference and the Ratio of the Weighted Kappa Coefficients of Two Diagnostic Tests Subject to a Paired Design

José Antonio Roldán-Nofuentes and Saad Bouh Sidaty-Regad 309

On Construction of Bernstein-Bézier Type Bivariate Archimedean Copula

Selim Orhun Susam and Burcu Hudaverdi 337

Wavelet Estimation of Regression Derivatives for Biased and Negatively Associated Data

Junke Kou and Christophe Chesneau 353

Approximation Results for the Sums of Independent Random Variables

Pratima Eknath Kadu 373

Modeling Heavy-Tailed Bounded Data by the Trapezoidal Beta Distribution with Applications

Jorge I. Figueroa-Zúñiga, Sebastián Niklitschek-Soto, Víctor Leiva and Shuangzhe Liu 387

Bias Reduced Peaks over Threshold Tail Estimation

Authors: JAN BEIRLANT  

– Department of Mathematics, LStat and LRisk, KU Leuven,
Belgium

– Department of Mathematical Statistics and Actuarial Science, Free State University,
South Africa

jan.beirlant@kuleuven.be

GAONYALELWE MARIBE 

– Department of Statistics, University of Pretoria,
South Africa

g.maribe@up.ac.za

PHILIPPE NAVEAU 

– Laboratoire des Sciences du Climat et de l'Environnement, CNRS, Université Paris-Saclay,
France

Philippe.Naveau@lsce.ipsl.fr

ANDRÉHETTE VERSTER 

– Department of Mathematical Statistics and Actuarial Science, Free State University,
South Africa

verstera@ufs.ac.za

Received: September 2019

Revised: February 2020

Accepted: March 2020

Abstract:


- Bias reduction in tail estimation has mainly been performed in case of Pareto-type models; see for instance Drees (1996) [11], Peng (1998) [20], Feuerverger and Hall (1999) [14], Beirlant *et al.* (1999 [3], 2002 [4]), Gomes and Martins (2002) [16] and Caeiro *et al.* (2005 [9], 2009 [10]). In that context, Beirlant *et al.* (2009) [7] and Papastathopoulos and Tawn (2013) [19] constructed distributional models that are based on second order rates of convergence for distributions of peaks over thresholds (POT). Such approach also allows to connect the tail and the bulk of the distribution. Bias reduction for all max-domains of attractions, i.e. without restricting to the Pareto-type case, received much less attention up to now. Here we extend the second-order refined POT approach started in Beirlant *et al.* (2009) [7] providing a bias reduction technique for the classical generalized Pareto (GP) approximation for POTs. We consider parametric and nonparametric modelling of the second order component.

Keywords:

- *peaks over threshold; generalized Pareto distribution; tail estimation; mixture models.*

AMS Subject Classification:

- 62G32, 62F10, 62F15, 62J07.

 Corresponding author: Jan Beirlant, KU Leuven, Department of Mathematics, Celestijnenlaan 200B, 3001 Heverlee, Belgium; E-mail: jan.beirlant@kuleuven.be.

1. INTRODUCTION

Extreme value (EV) methodology starts from the assumption that the distribution of the available sample X_1, X_2, \dots, X_n belongs to the domain of attraction of a generalized extreme value distribution, i.e. there exists sequences $(b_n)_n$ and $(a_n > 0)_n$ such that as $n \rightarrow \infty$

$$(1.1) \quad \frac{\max(X_1, X_2, \dots, X_n) - b_n}{a_n} \rightarrow_d Y_\xi,$$

where $\mathbb{P}(Y_\xi \leq y) = \exp(-(1 + \xi y)^{-1/\xi})$, for some $\xi \in \mathbb{R}$ with $1 + \xi y > 0$. The parameter ξ is termed the extreme value index (EVI). It is well-known (see e.g. Beirlant *et al.*, 2004 [5], and de Haan and Ferreira, 2006 [17]) that (1.1) is equivalent to the existence of a positive function $t \mapsto \sigma_t$, such that

$$(1.2) \quad \mathbb{P}\left(\frac{X-t}{\sigma_t} > y | X > t\right) = \frac{\bar{F}(t + y\sigma_t)}{\bar{F}(t)} \xrightarrow{t \rightarrow x_+} \bar{H}_\xi^{GP}(y) = (1 + \xi y)^{-1/\xi},$$

where $\bar{F}(x) = \mathbb{P}(X > x)$ and x_+ denotes the endpoint of the distribution of X . The conditional distribution of $X - t$ given $X > t$ is called the peaks over threshold (POT) distribution, while \bar{H}_ξ^{GP} is the survival function of the generalized Pareto distribution (GPD).

Estimation of ξ and tail quantities such as return periods is then based on fitting a GPD to the observed excesses $X - t$ given $X > t$. The main difficulty in such an EV application is the choice of the threshold t . Most often, the threshold t is chosen as one of the top data points $X_{n-k,n}$ for some $k \in \{1, 2, \dots, n\}$ where $X_{1,n} \leq X_{2,n} \leq \dots \leq X_{n,n}$ denotes the ordered sample. The parameters (ξ, σ) are then estimated by fitting the GPD $H_\xi^{GP}(\frac{y}{\sigma})$ to the spacings $X_{n,n} - X_{n-k,n}, \dots, X_{n-k+1,n} - X_{n-k,n}$.

The limit result in (1.2) requires t to be chosen as large as possible (or, equivalently, k as small as possible) for the bias in the estimation of ξ and other tail parameters to be limited. However, in order to limit the estimation variance, t should be as small as possible, i.e. the number of data points k used in the estimation should be as large as possible. Several adaptive procedures for choosing t or k have been proposed, but mainly in the Pareto-type case with $\xi > 0$, i.e. when

$$(1.3) \quad \bar{F}(x) = x^{-1/\xi} \ell(x),$$

for some slowly varying function ℓ , i.e. satisfying $\frac{\ell(yt)}{\ell(t)} \rightarrow 1$ as $t \rightarrow \infty$, for every $y > 1$. One then typically assumes a second-order specification of (1.3) of the type

$$(1.4) \quad \frac{\ell(yt)}{\ell(t)} - 1 = \delta_t \left(y^{-\beta} - 1 \right),$$

where $\delta_t = \delta(t) = t^{-\beta} \tilde{\ell}(t)$, with $\beta > 0$ and $\tilde{\ell}$ slowly varying at infinity.

As an alternative, bias reduction techniques have been proposed in the Pareto-type case $\xi > 0$, among others in Feuerverger and Hall (1999) [14], Beirlant *et al.* (1999 [3], 2002 [4]) and Gomes and Martins (2002) [16]. However while the bias is reduced, the variance is increased. In Caeiro *et al.* (2005 [9], 2009 [10]) methods are proposed to limit the variance of bias-reduced estimators assuming a third-order slow variation model. These methods focus

on the distribution of the log-spacings of high order statistics. Other construction methods for asymptotically unbiased estimators of $\xi > 0$ were introduced in Peng (1998) [20] and Drees (1996) [11].

Another approach consists of proposing penultimate limit distributions. In case $\xi > 0$, Beirlant *et al.* (2009) [7] proposed an extension of the Pareto distribution (EPD) to approximate the tail probability of the POT distribution $\mathbb{P}\left(\frac{X}{t} > y | X > t\right)$ as $t \rightarrow \infty$:

$$(1.5) \quad \bar{H}_{\xi, \delta, \rho}^{EP}(y) = 1 - H_{\xi, \delta, \rho}^{EP}(y) = y^{-1/\xi} \left(1 + \delta_t \left((y^{-1/\xi})^{-\rho} - 1\right)\right), \quad y > 1,$$

with δ_t satisfying $\delta_t \downarrow 0$ as $t \rightarrow \infty$ and $\rho = -\beta\xi$. In the literature, the second order parameter ρ typically is estimated externally with a different sequence of extreme order statistics than with ξ and δ , or it is given an appropriate 'canonical' value such as -1 . We suppress the notation ρ from the extended distribution notation.

Fitting the extended Pareto distribution $H_{\xi, \sigma}^{EP}$ to the relative excesses $\left\{\frac{X_{n-j+1, n}}{X_{n-k, n}}, j = 1, \dots, k\right\}$ leads to estimates of ξ that are more stable as a function of k compared to the original ML estimator derived by Hill (1975) [18]

$$\hat{\xi}_{k, n}^H = \frac{1}{k} \sum_{j=1}^k \log \frac{X_{n-j+1, n}}{X_{n-k, n}},$$

which is obtained by fitting the Pareto distribution $H_{\xi, 0}^{EP}$. Denoting the maximum likelihood estimators of ξ by $\hat{\xi}_k^{EP}$, it can indeed be shown under the assumption that the EP model for the excesses X/t is correct and that ρ is estimated consistently, that the asymptotic bias of $\hat{\xi}_k^{EP}$ is 0 as long as $k(k/n)^{-2\rho} \rightarrow \lambda \geq 0$ as $k, n \rightarrow \infty$, while the asymptotic bias of $\hat{\xi}_{k, n}^H$ is only 0 when $k(k/n)^{-2\rho} \rightarrow 0$. On the other hand, the asymptotic variance of $\hat{\xi}_k^{EP}$ equals $\left(\frac{1-\rho}{\rho}\right)^2 \frac{\xi^2}{k}$, where $\frac{\xi^2}{k}$ is the asymptotic variance of $\hat{\xi}_{k, n}^H$.

In case of a real-valued EVI, for the selection of an appropriate threshold or the construction of bias-reduced methods, only a few methods are available. Dupuis (1999) [12] suggested a robust model validation mechanism to guide the threshold selection, assigning weights between 0 and 1 to each data point where a high weight means that the point should be retained since a GPD model is fitting it well. However, thresholding is required at the level of the weights and hence the method cannot be used in an unsupervised manner. Buitendag *et al.* (2019) [8] present a ridge regression method to reduce the bias of the generalized Hill estimator proposed in Beirlant *et al.* (2005) [6].

In this paper we concentrate on bias reduction when fitting the GPD to the distribution of POTs $X - t | X > t$ using maximum likelihood estimation. We hence extend the second-order refined POT approach based on $\bar{H}_{\xi, \delta}^{EP}$ from (1.5) to all max-domains of attraction. Here the corresponding basic second order regular variation theory can be found in Theorem 2.3.8 in de Haan and Ferreira (2006) [17] stating that

$$(1.6) \quad \lim_{t \rightarrow x_+} \frac{\mathbb{P}(X - t > y\sigma_t | X > t) - (1 + \xi y)^{-1/\xi}}{\delta(t)} = (1 + \xi y)^{-1-1/\xi} \Psi_{\xi, \tilde{\rho}}((1 + \xi y)^{1/\xi}),$$

with $\delta(t) \rightarrow 0$ as $t \rightarrow x_+$ and $\Psi_{\xi, \tilde{\rho}}(x) = \frac{1}{\tilde{\rho}} \left(\frac{x^{\xi + \tilde{\rho}} - 1}{\xi + \tilde{\rho}} - \frac{x^{\xi} - 1}{\xi}\right)$ which for the cases $\xi = 0$ and $\tilde{\rho} = 0$ is understood to be equal to the limit as $\xi \rightarrow 0$ and $\tilde{\rho} \rightarrow 0$. We further allow more flexible second-order models than the ones arising from second-order regular variation theory such

as in (1.6) using non-parametric modelling of the second-order component and the flexible semiparametric GP modelling introduced in Tencaliec *et al.* (2019) [21]. This newly proposed method can also be applied to the specific case of Pareto-type distributions.

In the next section we propose our extended GPD models, and detail the estimation methods. Some basic asymptotic results are provided in Section 3. In the final section we discuss simulation results and some practical case studies.

2. TRANSFORMED AND EXTENDED GPD MODELS

In this paper we propose to approximate the POT distribution with an extended GPD model with survival function

$$(\mathcal{E}) : \quad \bar{F}_t^{EGP}(y) = \bar{H}_\xi^{GP}\left(\frac{y}{\sigma}\right) \left\{ 1 + \delta_t B_\eta \left(\bar{H}_\xi^{GP}\left(\frac{y}{\sigma}\right) \right) \right\},$$

where

- $\delta_t = \delta(t) \rightarrow 0$ as $t \rightarrow x_+$,
- $B_\eta(1) = 0$ and $\lim_{u \rightarrow 0} u^{1-\epsilon} B_\eta(u) = 0$ for every $0 < \epsilon < 1$,
- B_η is twice continuously differentiable.

Here the parameter η represents a second order nuisance parameter. For negative δ -values one needs $\delta_t > \left\{ \min_u \left(1 - \frac{d}{du} (u B_\eta(u)) \right) \right\}^{-1}$ to obtain a valid distribution.

Note that this model is a transformation model $G_t \left(\bar{H}_\xi^{GP}\left(\frac{y}{\sigma}\right) \right)$ where the transformation function $G_t : (0, 1) \rightarrow (0, 1), u \mapsto u(1 + \delta_t B_\eta(u))$ satisfies $\frac{G_t(u)}{u} \rightarrow 1$ as $t \rightarrow \infty$ for every $u \in (0, 1)$ as follows from (1.2).

Also, model (\mathcal{E}) generalizes the EPD model (1.5) replacing the Pareto survival function $y^{-1/\xi}$ ($\xi > 0$) by the GPD survival function \bar{H}_ξ^{GP} ($\xi \in \mathbb{R}$), and considering a general function $B_\eta(u)$.

We here detail *a parametric and non-parametric estimation procedure* for (ξ, σ) under (\mathcal{E}) based on excesses $Y_{j,k} = X_{n-j+1,n} - X_{n-k,n}$ ($j = 1, \dots, k$), while considering external estimation of the parameters in the B_η component of the model. In this we use the reparametrization (ξ, τ) with $\tau = \xi/\sigma$. Modelling the distribution of the exceedances Y with model (\mathcal{E}) leads to maximum likelihood estimators based on the excesses $Y_{j,k} = X_{n-j+1,n} - X_{n-k,n}$ ($j = 1, \dots, k$):

$$(2.1) \quad (\hat{\xi}_k^E, \hat{\tau}_k^E, \hat{\delta}_k^E) = \operatorname{argmax} \left\{ \sum_{j=1}^k \log \left(1 + \delta_k b_\eta \left((1 + \tau Y_{j,k})^{-1/\xi} \right) \right) + \sum_{j=1}^k \log \left\{ \frac{\tau}{\xi} (1 + \tau Y_{j,k})^{-1-1/\xi} \right\} \right\}$$

with $b_\eta(u) = \frac{d}{du} (u B_\eta(u))$ for a given choice of B_η .

Estimates of small tail probabilities $\mathbb{P}(X > c)$ are then obtained through

$$\hat{\mathbb{P}}_k^E(X > c) = \frac{k}{n} \bar{H}_{\hat{\xi}_k^E}^{GP} \left(\frac{\hat{\tau}_k^E}{\hat{\xi}_k^E} (c - X_{n-k,n}) \right) \left(1 + \hat{\delta}_k^E \hat{B}_\eta \left(\bar{H}_{\hat{\xi}_k^E}^{GP} \left(\frac{\hat{\tau}_k^E}{\hat{\xi}_k^E} (c - X_{n-k,n}) \right) \right) \right).$$

A general approach to choose the parameters contained in the B_η component can be to minimize the variance of the obtained estimates of ξ over $k = 2, \dots, n$. See also the simulation Section 4.

A parametric approach (Ep). The second-order result (1.6) leads to the parametric choice $B_{\xi, \tilde{\rho}}(u) = \frac{u^\xi}{\tilde{\rho}} \left(\frac{u^{-\xi-\tilde{\rho}-1}}{\xi+\tilde{\rho}} - \frac{u^{-\xi-1}}{\xi} \right)$ in case $\xi + \tilde{\rho} \neq 0$ and $\xi \neq 0$.

Model (\mathcal{E}) allows for bias reduction in the estimation of (ξ, τ) under the assumption that the corresponding second-order model (1.6) is correct for the POTs $X - t | X > t$. Note that here the B_η component contains two parameters ξ and $\tilde{\rho}$. So in this component ξ and $\tilde{\rho}$ will be substituted with an external value.

Here

$$b_\eta(u) = u^{-\tilde{\rho}} \left(\frac{1 - \tilde{\rho}}{\tilde{\rho}(\xi + \tilde{\rho})} \right) + u^\xi \left(\frac{1 + \xi}{\xi(\xi + \tilde{\rho})} \right) - \frac{1}{\xi\tilde{\rho}},$$

in which the classical estimator of ξ (with $\delta_k = 0$), or an appropriate value ξ_0 , is used to substitute ξ . A consistent estimator of $\tilde{\rho}$ is provided in Fraga Alves *et al.* (2003) [15]. Another option is to choose $(\xi_0, \tilde{\rho})$ minimizing the variance in the plot of the resulting estimates of ξ as a function of k .

A non-parametric approach ($E\tilde{p}$). In practice a particular distribution probably follows laws of nature, environment or business and does not have to follow the second-order regular variation assumptions as in (1.6). A non-parametric approximation of $u \mapsto uB_\eta(u)$ can be obtained from an estimator \hat{G}_{t_*} of G_{t_*} , or equivalently \hat{G}_{k_*} of G_{k_*} , of the transformation $G_t(u) = u(1 + \delta_t B_\eta(u))$ ($u \in (0, 1)$) at some particular t_* or k_* . Indeed, using $\hat{G}_{k_*}^{(m)}(u) - u$ as an approximation of $u \mapsto \delta_{k_*} u B_\eta(u)$, and reparametrizing δ_k by δ_k / δ_{k_*} , we obtain $\hat{b}_{\eta, k_*}(u) = -1 + \frac{d}{du} \hat{G}_{k_*}^{(m)}(u)$ as an estimator of b_η .

For any t , an estimator \hat{G}_t of G_t can be obtained using the Bernstein polynomial algorithm from Tencaliec *et al.* (2019) [21]. The Bernstein approximation of order m of a continuous distribution function G on $[0, 1]$ is given by

$$G^{(m)}(u) = \sum_{j=0}^m G \left(\frac{j}{m} \right) \binom{m}{j} u^j (1-u)^{m-j}, \quad u \in [0, 1].$$

As in Babu *et al.* (2002) [2] one then replaces the unknown distribution function G itself with the empirical distribution function \hat{G}_n of the available data in order to obtain a smooth estimator of G :

$$\hat{G}_n^{(m)}(u) = \sum_{j=0}^m \hat{G}_n \left(\frac{j}{m} \right) \binom{m}{j} u^j (1-u)^{m-j}.$$

Note that G_t is the distribution function of $\bar{H}_\xi^{GP}(Y/\sigma)$. Hence, in the present application, data from G_t are only available after imputing a value for (ξ, τ) . This then leads to the iterative algorithm from Tencaliec *et al.* (2019) [21], which is applied to every threshold t , or every number of top k data:

- (i) Set starting values $(\hat{\xi}_k^{(0)}, \hat{\tau}_k^{(0)})$. Here one can use $(\hat{\xi}_k^{ML}, \hat{\tau}_k^{ML})$ from using $G_t(u) = u$.
- (ii) Iterate for $r = 0, 1, \dots$ until the difference in log-likelihood taken in $(\hat{\xi}_k^{(r)}, \hat{\tau}_k^{(r)})$ and $(\hat{\xi}_k^{(r+1)}, \hat{\tau}_k^{(r+1)})$ is smaller than a prescribed small value:
 - (a) Given $(\hat{\xi}_k^{(r)}, \hat{\tau}_k^{(r)})$ construct rv's $\hat{Z}_{j,k} = \left(1 + \hat{\tau}_k^{(r)} Y_{j,k}\right)^{-1/\hat{\xi}_k^{(r)}}$;
 - (b) Construct Bernstein approximation based on $\hat{Z}_{j,k}$ ($1 \leq j \leq k$)

$$\hat{G}_k^{(m)}(u) = \sum_{j=0}^m \hat{G}_k \left(\frac{j}{m}\right) \binom{m}{j} u^j (1-u)^{m-j}$$

with \hat{G}_k the empirical distribution function of $\hat{Z}_{j,k}$;

- (c) Obtain new estimates $(\hat{\xi}_k^{(r+1)}, \hat{\tau}_k^{(r+1)})$ with ML:

$$(\hat{\xi}_k^{(r+1)}, \hat{\tau}_k^{(r+1)}) = \operatorname{argmax} \left\{ \sum_{j=1}^k \log \left\{ \hat{g}_k^{(m)} \left((1 + \tau \hat{Z}_{j,k})^{-1/\xi} \right) \right\} + \sum_{j=1}^k \log \left\{ \frac{\tau}{\xi} (1 + \tau \hat{Z}_{j,k})^{-1-1/\xi} \right\} \right\}$$

with $\hat{g}_k^{(m)}$ denoting the derivative of $\hat{G}_k^{(m)}$.

As noted in Tencaliec *et al.* (2019) [21] a theoretical study of these estimates is difficult and has not been established.

Remark 2.1. The estimation methods described above of course can be rewritten for the specific case of Pareto-type distributions where the distribution of POTs $Y = \frac{X}{t} | X > t$ are approximated by transformed Pareto distributions. The model (\mathcal{E}) is then rephrased as

$$(\mathcal{E}^+) : \quad \bar{F}_t^E(y) = \bar{H}_\xi^P(y) \{1 + \delta_t B_\eta(\bar{H}_\xi^P(y))\}.$$

The likelihood estimation method, now based on the exceedances $Y_{j,k} = X_{n-j+1,n} / X_{n-k,n}$ ($j = 1, \dots, k$), is then adapted to

$$(2.2) \quad (\hat{\xi}_k^{E+}, \hat{\delta}_k^{E+}) = \operatorname{argmax} \left\{ \sum_{j=1}^k \log \left(1 + \delta_k b_\eta(Y_{j,k}^{-1/\xi}) \right) + \sum_{j=1}^k \log \left\{ \frac{1}{\xi} (Y_{j,k})^{-1-1/\xi} \right\} \right\}.$$

Note that the (Ep^+) approach using the parametric version $B_\eta(u) = u^{-\rho} - 1$ for a particular fixed $\rho < 0$ equals the EPD method from Beirlant *et al.* (2009) [7], while $(E\bar{p}^+)$ is new. Estimators of tail probabilities are then given by

$$\hat{\mathbb{P}}_k^{E+}(X > c) = \frac{k}{n} \bar{H}_{\hat{\xi}_k^{E+}}^P \left(c / X_{n-k,n} \right) \left(1 + \hat{\delta}_k^{E+} \hat{B}_\eta \left(\bar{H}_{\hat{\xi}_k^{E+}}^P \left(c / X_{n-k,n} \right) \right) \right).$$

3. BASIC ASYMPTOTICS UNDER MODEL (\mathcal{E})

In this section we discuss the asymptotic properties of the maximum likelihood estimators solving (2.1) and (2.2). To this end, as in Beirlant *et al.* (2009) [7], we develop the likelihood equations up to linear terms in δ_k since $\delta_k \rightarrow 0$ with decreasing value of k .

Below we set $\bar{H}_\theta(y) = (1 + \tau y)^{-1/\xi}$ when using extended GPD modelling, while $\bar{H}_\theta(y) = y^{-1/\xi}$ when using extended Pareto modelling under $\xi > 0$.

Extended Pareto POT modelling. The likelihood problem (2.2) was already considered in Beirlant *et al.* (2009) [7] in case of parametric modelling for B_η . We here propose a more general treatment. The limit statements in the derivation can be obtained using the methods from Beirlant *et al.* (2009) [7]. Denoting the log-likelihood function in (2.2) by ℓ , the likelihood equations are given by

$$(3.1) \quad \begin{cases} \frac{\partial}{\partial \xi} \ell = -\frac{k}{\xi} + \frac{1}{\xi^2} \sum_{j=1}^k \log Y_{j,k} + \frac{\delta_k}{\xi^2} \sum_{j=1}^k \frac{b'_\eta(\bar{H}_\theta(Y_{j,k})) \bar{H}_\theta(Y_{j,k}) \log Y_{j,k}}{1 + \delta_k b_\eta(\bar{H}_\theta(Y_{j,k}))}, \\ \frac{\partial}{\partial \delta_k} \ell = \sum_{j=1}^k b_\eta(\bar{H}_\theta(Y_{j,k})) - \delta_k \sum_{j=1}^k b_\eta^2(\bar{H}_\theta(Y_{j,k})). \end{cases}$$

Extended Generalized Pareto POT modelling. The likelihood equations following from (2.1) up to linear terms in δ_k are now given by

$$\begin{cases} \frac{\partial}{\partial \xi} \ell = -\frac{k}{\xi} + \frac{1}{\xi^2} \sum_{j=1}^k \log(1 + \tau Y_{j,k}) + \frac{\delta_k}{\xi^2} \sum_{j=1}^k b'_\eta(\bar{H}_\theta(Y_{j,k})) \bar{H}_\theta(Y_{j,k}) \log(1 + \tau Y_{j,k}), \\ \frac{\partial}{\partial \tau} \ell = \frac{k}{\xi \tau} \left\{ -1 + (1 + \xi) \frac{1}{k} \sum_{j=1}^k \frac{1}{1 + \tau Y_{j,k}} - \frac{\delta_k}{k} \sum_{j=1}^k b'_\eta(\bar{H}_\theta(Y_{j,k})) (\tau Y_{j,k}) (1 + \tau Y_{j,k})^{-1-1/\xi} \right\}, \\ \frac{\partial}{\partial \delta_k} \ell = \sum_{j=1}^k b_\eta(\bar{H}_\theta(Y_{j,k})) - \delta_k \sum_{j=1}^k b_\eta^2(\bar{H}_\theta(Y_{j,k})), \end{cases}$$

from which

$$(3.2) \quad \begin{cases} \hat{\delta}_k = \frac{\sum_{j=1}^k b_\eta(\bar{H}_{\hat{\theta}_k}(Y_{j,k}))}{\sum_{j=1}^k b_\eta^2(\bar{H}_{\hat{\theta}_k}(Y_{j,k}))}, \\ \frac{1}{k} \sum_{j=1}^k \log(1 + \hat{\tau}_k Y_{j,k}) = \hat{\xi}_k - \frac{\hat{\delta}_k}{k} \sum_{j=1}^k b'_\eta(\bar{H}_{\hat{\theta}_k}(Y_{j,k})) \bar{H}_{\hat{\theta}_k}(Y_{j,k}) \log(1 + \hat{\tau}_k Y_{j,k}), \\ \frac{1}{k} \sum_{j=1}^k \frac{1}{1 + \hat{\tau}_k Y_{j,k}} = \frac{1}{1 + \hat{\xi}_k} + \frac{\hat{\delta}_k}{1 + \hat{\xi}_k} \left\{ \frac{1}{k} \sum_{j=1}^k b'_\eta(\bar{H}_{\hat{\theta}_k}(Y_{j,k})) \bar{H}_{\hat{\theta}_k}(Y_{j,k}) \right. \\ \left. - \frac{1}{k} \sum_{j=1}^k b'_\eta(\bar{H}_{\hat{\theta}_k}(Y_{j,k})) \bar{H}_{\hat{\theta}_k}(Y_{j,k}) \frac{1}{1 + \hat{\tau}_k Y_{j,k}} \right\}. \end{cases}$$

Under the extended model we now state the asymptotic distribution of the estimators $(\hat{\xi}_k^E, \hat{\tau}_k^E)$ and $\hat{\xi}_k^{E+}$. To this end let Q denote the quantile function of F , and let $U(x) = Q(1 - x^{-1})$ denote the corresponding tail quantile function. Model (E) assumption can be rephrased in terms of U :

$$(\tilde{\mathcal{E}}) : \frac{U(vx) - U(v)}{\sigma_{U(v)}} - h_\xi(x) \rightarrow_{v \rightarrow \infty} x^\xi B_\eta(1/x),$$

where $h_\xi(x) = (x^\xi - 1)/\xi$ and $\delta(U)$ regularly varying with index $\tilde{\rho} < 0$. Moreover in the mathematical derivations one needs the extra condition that for every $\epsilon, \nu > 0$, and v, vx sufficiently large

$$(\tilde{\mathcal{E}}_2) : \left| \frac{\frac{U(vx) - U(v)}{\sigma_{U(v)}} - h_\xi(x)}{\delta(U(v))} - x^\xi B_\eta(1/x) \right| \leq \epsilon x^\xi |B_\eta(1/x)| \max\{x^\nu, x^{-\nu}\}.$$

Similarly, (\mathcal{E}^+) is rewritten as

$$(\tilde{\mathcal{E}}^+) : \frac{\frac{U(vx)}{U(v)} - x^\xi}{\xi \delta(U(v))} \xrightarrow{v \rightarrow \infty} x^\xi B_\eta(1/x).$$

The analogue of $(\tilde{\mathcal{E}}_2)$ in this specific case is given by

$$(\tilde{\mathcal{E}}_2^+) : \left| \frac{\frac{U(vx)}{U(v)} - x^\xi}{\xi \delta(U(v))} - x^\xi B_\eta(1/x) \right| \leq \epsilon x^\xi |B_\eta(1/x)| \max\{x^\nu, x^{-\nu}\},$$

with $\delta(U)$ regularly varying with index $\rho < 0$.

Finally, in the expression of the asymptotic variances we use

$$Eb_\eta^2 = \int_0^1 b_\eta^2(u) du, \quad EB_\eta = \int_0^1 B_\eta(u) du, \quad EC_\eta = \int_0^1 u^\xi B_\eta(u) du.$$

The proof of the next theorem is outlined in the [Appendix](#). It allows to construct confidence intervals for the estimators of ξ obtained under the extended models.

Theorem 3.1. *Let $k = k_n$ be a sequence such that $k, n \rightarrow \infty$ and $k/n \rightarrow 0$ such that $\sqrt{k}\delta(U(n/k)) \rightarrow \lambda \in \mathbb{R}$. Moreover assume that in (2.1) and (2.2), B_η is substituted by a consistent estimator as $n \rightarrow \infty$. Then:*

- i. When $\xi > -1/2$ with $(\tilde{\mathcal{E}}_2)$

$$\begin{aligned} & \left(\sqrt{k}(\hat{\xi}_k^E - \xi), \sqrt{k}\left(\frac{\hat{\tau}_k^E}{\tau} - 1\right) \right) \rightarrow_d \mathcal{N}_2(\mathbf{0}, \Sigma) \\ \Sigma &= \frac{\xi^2}{D} \begin{pmatrix} \frac{1}{(1+\xi)^2(1+2\xi)} - \frac{(EC_\eta)^2}{Eb_\eta^2} & \frac{1}{\xi(1+\xi)^3} - \frac{EB_\eta EC_\eta}{\xi(1+\xi)Eb_\eta^2} \\ \frac{1}{\xi(1+\xi)^3} - \frac{EB_\eta EC_\eta}{\xi(1+\xi)Eb_\eta^2} & \frac{1}{\xi^2(1+\xi)^2} \left(1 - \frac{(EB_\eta)^2}{Eb_\eta^2}\right) \end{pmatrix}, \end{aligned}$$

where

$$D = \left(\frac{1}{(1+\xi)^2(1+2\xi)} - \frac{(EC_\eta)^2}{Eb_\eta^2} \right) \left(1 - \frac{(EB_\eta)^2}{Eb_\eta^2} \right) - \left(\frac{1}{(1+\xi)^2} - \frac{EB_\eta EC_\eta}{Eb_\eta^2} \right)^2;$$

- ii. When $\xi > 0$ with $(\tilde{\mathcal{E}}_2^+)$

$$\begin{aligned} & \left(\sqrt{k}(\hat{\xi}_k^{E^+} - \xi), \sqrt{k}(\hat{\delta}_k^{E^+} - \delta_k) \right) \rightarrow_d \mathcal{N}_2(\mathbf{0}, \Sigma^+), \\ \Sigma^+ &= \frac{1}{Eb_\eta^2 - (EB_\eta)^2} \begin{pmatrix} \xi^2 Eb_\eta^2 & -\xi EB_\eta \\ -\xi EB_\eta & 1 \end{pmatrix}. \end{aligned}$$

Remark 3.1. The asymptotic variance of $\hat{\xi}_k^{E+}$ is larger than the asymptotic variance ξ^2 of the Hill estimator $\hat{\xi}_{k,n}^H$. Indeed,

$$\begin{aligned} (EB_\eta)^2 &= \left(\int_0^1 \log(1/u) b_\eta(u) du \right)^2 \\ &= \left(\int_0^1 (\log(1/u) - 1) b_\eta(u) du \right)^2 \\ &\leq \left(\int_0^1 (\log(1/u) - 1)^2 du \right) \left(\int_0^1 b_\eta^2(u) du \right) \\ &= (Eb_\eta^2), \end{aligned}$$

where the above inequality follows using the Cauchy-Schwarz inequality.

Similarly, one can show that

$$(EC_\eta)^2 = \xi^{-2} \left(\int_0^1 (u^\xi - \frac{1}{1+\xi}) b_\eta du \right)^2 \leq \frac{1}{(1+2\xi)(1+\xi)^2} (Eb_\eta^2).$$

The asymptotic variance of $\hat{\xi}_k^E$ equals

$$\frac{(1+\xi)^2}{k} \frac{1 - (1+\xi)^2(1+2\xi)(EC_\eta)^2 / (Eb_\eta^2)}{1 - \frac{(1+\xi)^4(1+2\xi)}{\xi^2} (Eb_\eta^2)^{-1} [(EC_\eta)^2 - 2\frac{(EC_\eta)(EB_\eta)}{(1+\xi)^2} + \frac{(EB_\eta)^2}{(1+\xi)^2(1+2\xi)}]}$$

which can be shown to be larger than the asymptotic variance $(1+\xi)^2/k$ of the classical GPD maximum likelihood estimator. In the parametric case with $B_\eta(u) = \frac{u^\xi}{\rho} \left(\frac{u^{-\xi-\tilde{\rho}}-1}{\xi+\tilde{\rho}} - \frac{u^{-\xi}-1}{\xi} \right)$, one obtains $EB_\eta = (1+\xi)^{-1}(1-\tilde{\rho})^{-1}$, $EC_\eta = (1+\xi)^{-1}(1+2\xi)^{-1}(\xi-\tilde{\rho}+1)^{-1}$ and $Eb_\eta^2 = 2(1+2\xi)^{-1}(1-2\tilde{\rho})^{-1}(\xi-\tilde{\rho}+1)^{-1}$. It then follows that the asymptotic variance of $\hat{\xi}_k^E$ equals $\frac{(1+\xi)^2}{k} \left(\frac{1-\tilde{\rho}}{\tilde{\rho}} \right)^2$.

In case $\xi > 0$ with $B_\eta(u) = u^{-\rho} - 1$, the asymptotic variance of $\hat{\xi}_k^{E+}$ is given by $\frac{\xi^2}{k} \left(\frac{1-\rho}{\rho} \right)^2$ as already found in Beirlant *et al.* (2009) [7].

Finally, an asymptotic representation of $\sqrt{k}(\hat{\delta}_k^E - \delta_k)$ can be found at the end of the proof of Theorem 3.1 in the Appendix.

In the case studies in the next section, asymptotic confidence intervals based on Theorem 3.1 can be added to the analysis.

Remark 3.2. Since in model (\mathcal{E}) the B_η factor is multiplied by δ_t , the asymptotic distribution of tail estimators based on (\mathcal{E}) will not depend on the asymptotic distribution of the estimator of B_η . As in Beirlant *et al.* (2009) [7] when using the EPD model in the Pareto-type setting, one can rely in the parametric approach on consistent estimators of the nuisance parameter η using a larger proportion k_* of the data. Alternatively, one can also consider different values of η in the parametric approach, and of (k_*, m) in the non-parametric setting, and search for values of this nuisance parameter which stabilizes the plots of the EVI estimates as a function of k using the minimum variance principle for the estimates as a function of k . Clearly one loses the asymptotic unbiasedness in Theorem 3.1 if B_η is not consistently estimated. For the moment no proof is available to show that the estimators of the parameters in the second order component B_η through the minimum variance principle are consistent. Note that the estimator of $\tilde{\rho}$ presented in Fraga Alves *et al.* (2003) [15] has been shown to be consistent.

As becomes clear from the simulation results, in many instances the extreme value index estimators are not very sensitive to such a misspecification, especially in the non-parametric approach leading to $E\bar{p}$ and $E\bar{p}^+$, and the proposed estimators can still outperform the classical maximum likelihood estimators based on the first order approximations of the POT distributions.

4. SIMULATIONS AND CASE STUDIES

Simulation results and practical cases are proposed in a Shinyapp written in R:

<https://phdshinygao.shinyapps.io/ExtendedModels/>

Under *Simulations* one finds simulation results with sample sizes $n = 200$ for different distributions from each max-domain of attraction. The bias and MSE for the different estimators are plotted as a function of the number of exceedances k . Using the notation from the preceding sections one has a choice to apply the technique with \bar{H}_θ equal to the GPD, respectively the simple Pareto distribution (only when $\xi > 0$).

Sliders are provided for the following parameters:

- in case of GPD modelling: $\tilde{\rho}$ in Ep , and (k_*, m) in $E\bar{p}$ estimation,
- in case of Pareto modelling: ρ in Ep^+ , and (k_*, m) in $E\bar{p}^+$ estimation.

Again one can indicate to choose these parameters so as to minimize the variance of $\hat{\xi}_k$ over $k = 2, \dots, n$. The value of ξ in the parametric function $B_{\xi, \tilde{\rho}}$ in Ep is imputed with the classical GPD-ML estimator at the given value of k .

Also bias and RMSE plots of the corresponding tail probability estimates of $p = \mathbb{P}(X > c)$ are given, where c is chosen so that these probabilities equal $p = 0.005$ or $p = 0.003$. Here the bias, respectively RMSE, are expressed as the average, respectively the average of squared values, of $\log(p/\hat{p})$.

One can also change the vertical scale of the plots, smooth the figures by taking moving averages of a certain number of estimates. Finally one can download the figures in pdf.

While on the above link, several other distributions are used and sliders are provided for the different parameters ρ , $\tilde{\rho}$, and (k_*, m) , we collect here the resulting figures for estimation of ξ and estimating 0.003 tail probabilities, when using the minimum variance principle for all parameters, in case of the following subset of models:

- The *Burr*(τ, λ) distribution with $\bar{F}(x) = (1 + x^\tau)^{-\lambda}$ for $x > 0$ with $\tau = 1$ and $\lambda = 2$, so that $\xi = \frac{1}{\tau\lambda} = \frac{1}{2}$ and $\rho = \tilde{\rho} = -\frac{1}{\lambda} = -\frac{1}{2}$.
- The *Fréchet*(2) distribution with $\bar{F}(x) = 1 - \exp(-x^{-2})$ for $x > 0$, so that $\xi = \frac{1}{2}$ and $\rho = \tilde{\rho} = -1$.
- The *standard normal distribution* with $\xi = 0$ and $\tilde{\rho} = 0$.
- The *Exponential distribution* with $\bar{F}(x) = e^{-\lambda x}$ for $x > 0$, so that $\xi = 0$ and $\tilde{\rho} = 0$.

- The Reversed Burr distribution with $\bar{F}(x) = (1 + (1 - x)^{-\tau})^{-\lambda}$ for $x < 1$ with $\tau = 5$ and $\lambda = 1$, so that $\xi = -1/(\tau\lambda) = -\frac{1}{5}$ with $\tilde{\rho} = -1/\lambda = -1$.
- The extreme value Weibull distribution with $\bar{F}(x) = 1 - e^{-(1-x)^\alpha}$ for $x < 1$ with $\alpha = 4$, so that $\xi = -\frac{1}{4}$ with $\tilde{\rho} = -1$.

We also compare the bias and RMSE results for $\hat{\xi}_k^E$ with those of the ridge regression estimator presented in Buitendag *et al.* (2019) [8]. This regression method is constructed on the basis of a regression model of the type

$$Y_j = \xi + b_{n,k} \left(\frac{j}{k+1} \right)^{-\tilde{\rho}}, \quad j = 1, \dots, k,$$

where

$$Y_j = (j+1) \left(\log \frac{X_{n-j,n} \hat{\xi}_{j,n}^H}{X_{n-j-1,n} \hat{\xi}_{j+1,n}^H} - \log \left(1 + \frac{1}{j} \right) + \frac{1}{j} \right), \quad j = 1, \dots, n-1.$$

In case $\xi > 0$, the results for $\hat{\xi}_k^{E+}$ are also compared with the corrected Hill method presented in Caeiro *et al.* (2005) [9] and (2009) [10], also based on regression representations of top order statistics $X_{n-j+1,n}$, and which have been shown to have asymptotic bias 0 while keeping the same asymptotic variance ξ^2/k as the Hill estimator $\hat{\xi}_{k,n}^H$ under a third-order slow variation model.

In general the minimum variance principle works well, though in some cases some improved results can be obtained by choosing specific values of the parameters ρ , $\tilde{\rho}$, and (k_*, m) . This is mainly the case for the Pareto-type models when using $E\bar{p}$, such as for the Fréchet distribution. Also, in case of tail probability estimation using Ep for cases with $\xi < 0$ particular choices of the corresponding parameters lead to improvements over the minimum variance principle.

Overall the Ep approach yields the best results, both in estimation of ξ and tail probabilities. The improvement over the classical GPD maximum likelihood approach is smaller for $E\bar{p}$, and in case of situations where the second order parameter $\tilde{\rho}$ equals 0 then $E\bar{p}$ basically equals the ML estimators. Note that when $\tilde{\rho} = 0$ the conditions of the main theorem are not met, in which case the GPD and the bias reductions are known to exhibit a large bias. This is typically the case when $\xi = 0$. This is also known to be the case using simple Pareto modelling when $\rho = 0$.

The proposed methods compare well with the ridge regression method. One exception is the Fréchet distribution (see Figure 3) in which the ridge regression method offers exceptionally good results.

In case of simple Pareto modelling for $\xi > 0$ cases (see Figures 2 and 4) the Ep^+ and $E\bar{p}^+$ approaches yield serious improvements over the Hill estimator, with small bias for Ep^+ and $E\bar{p}^+$, while the parametric approach Ep^+ naturally exhibits the best RMSE. The results obtained with proposed methods are comparable with the CH estimator (see Figures 2 and 4).

Under *Applications* the app also offers the analysis of some case studies, some of which are discussed here in more detail. We use Belgian car insurance claim ultimates of a Belgian car insurance portfolio discussed in Albrecher *et al.* (2017) [1], and lifetime data discussed in Einmahl *et al.* (2019) [13]. We then present estimates of ξ , σ and tail probabilities $\mathbb{P}(X > x_{n,n})$ with $x_{n,n}$ denoting the largest observation, so that the estimated probability is supposed to be close to $1/n$. An option is provided in the Shinyapp to construct asymptotic confidence intervals for ξ for the Ep and Ep^+ based estimates of ξ , on the basis of Theorem 3.1.

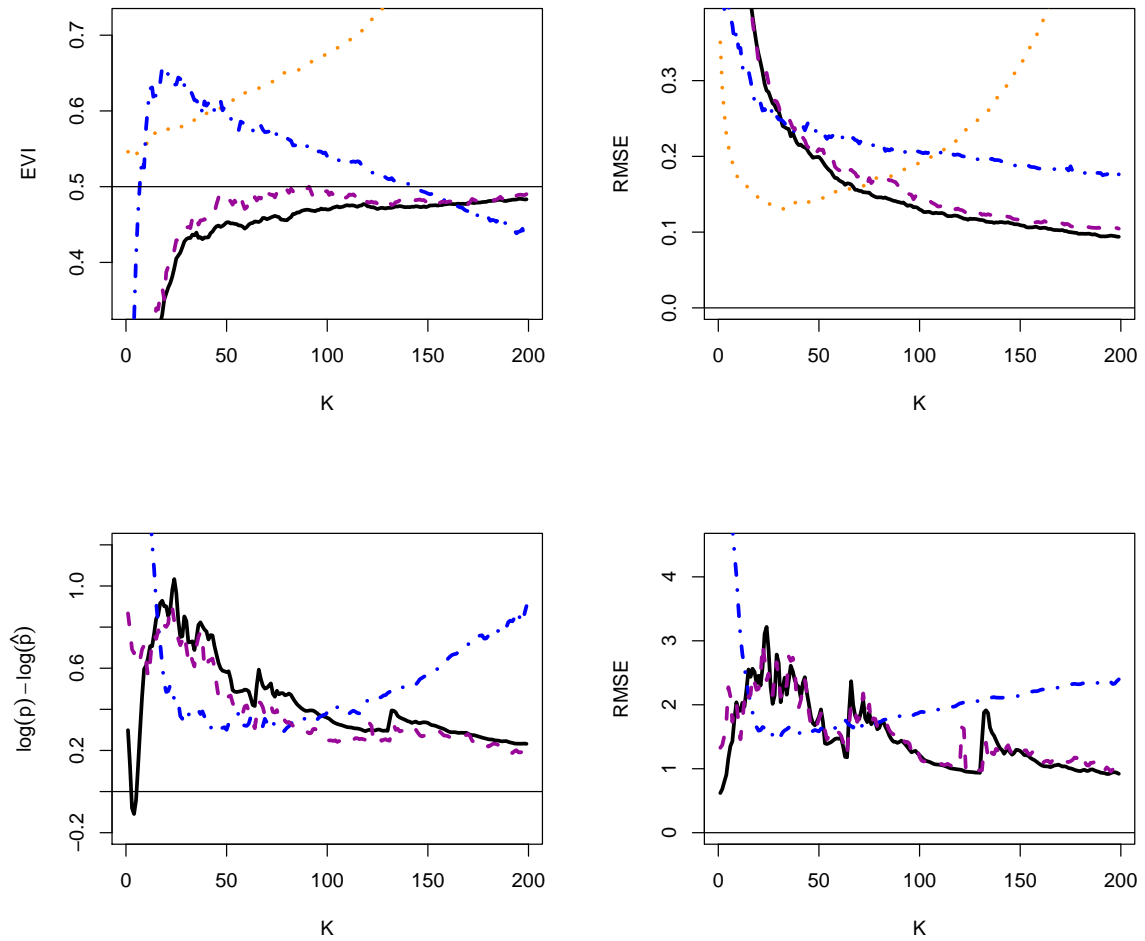


Figure 1: Burr distribution with $\xi = 0.5$ and $\rho = -0.5$. Estimation of ξ (top) and tail probability (bottom) using minimum variance principle, bias (left), RMSE (right): GPD-ML (full line), $E\hat{p}$ (dash-dotted), $E\bar{p}$ (dashed) and ridge regression estimator (dotted).

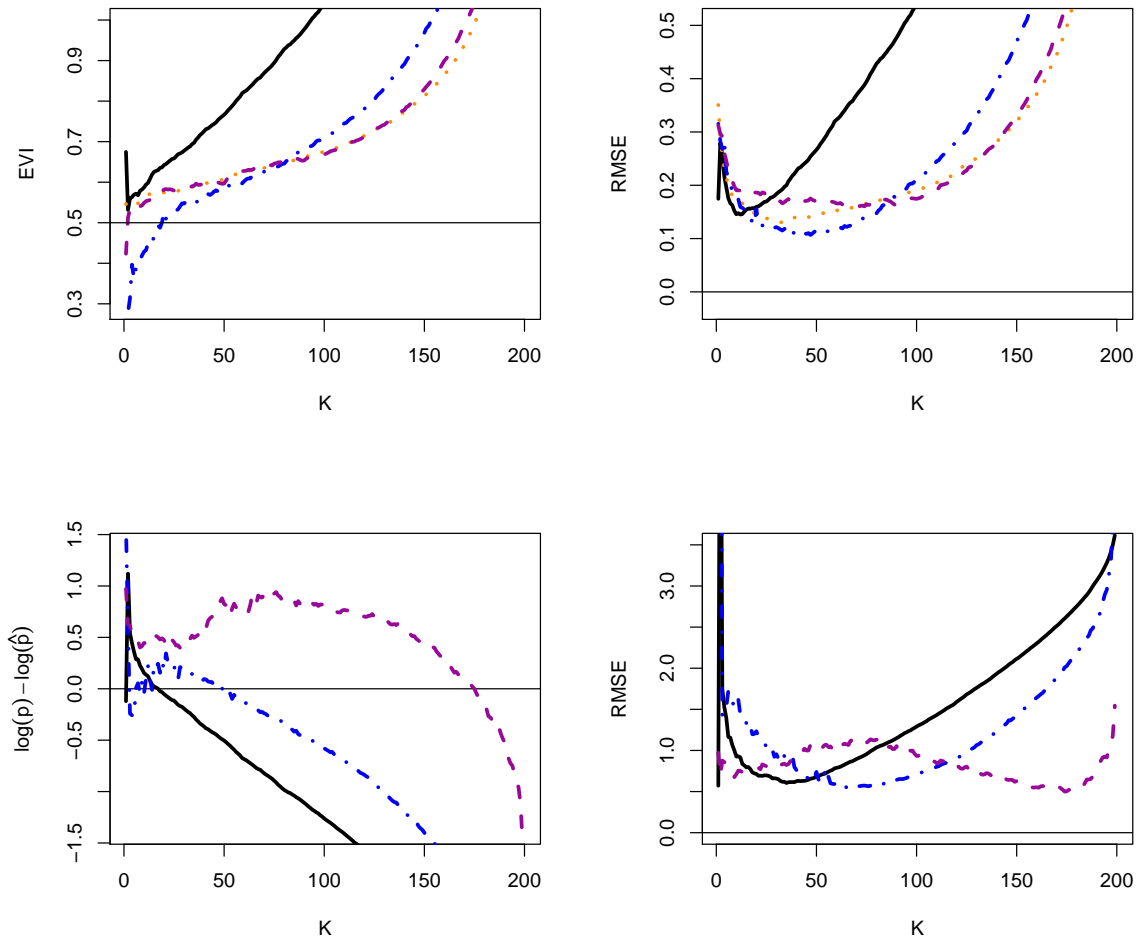


Figure 2: Burr distribution with $\xi = 0.5$ and $\rho = -0.5$. Estimation of ξ (top) and tail probability (bottom) using minimum variance principle, bias (left), RMSE (right): Pareto-ML (full line), $E p^+$ (dash-dotted), $E \bar{p}^+$ (dashed) and corrected Hill estimator (dotted).

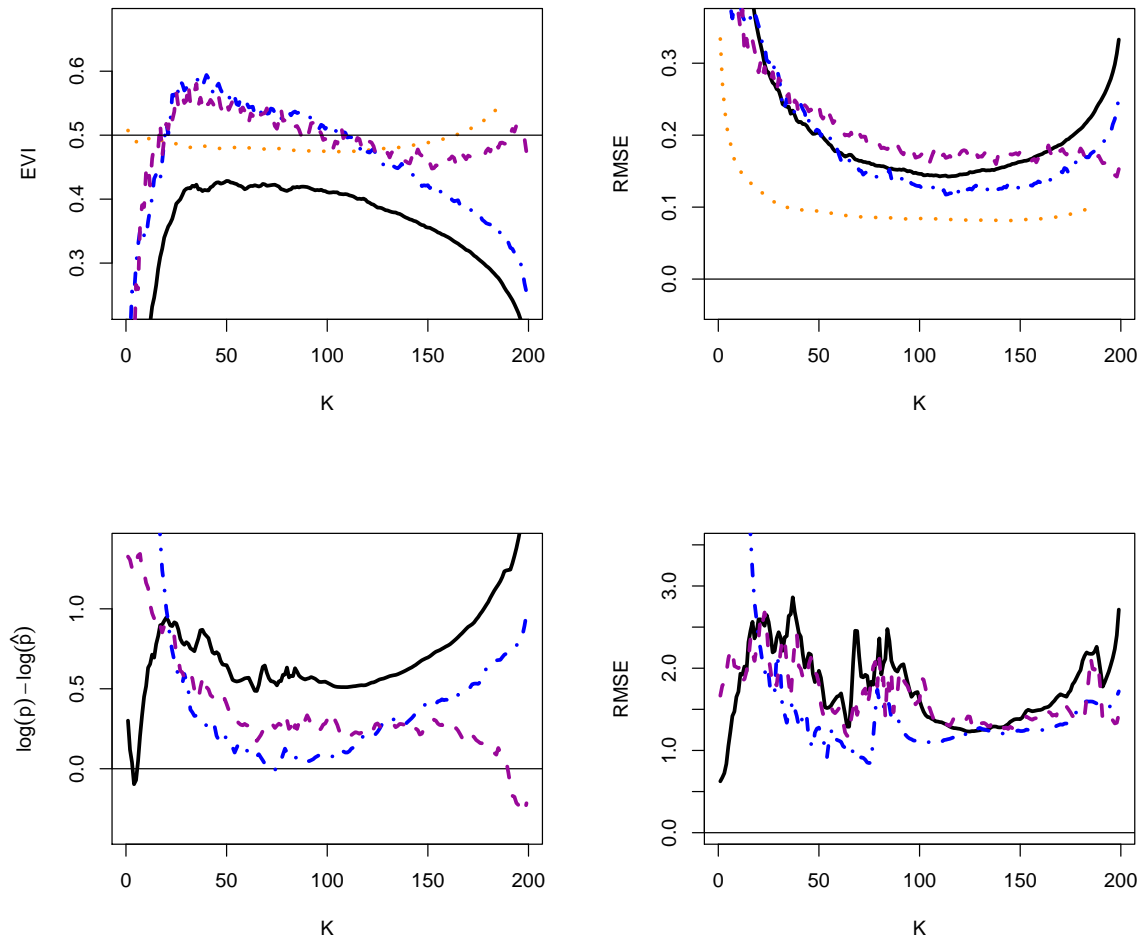


Figure 3: Fréchet distribution with $\xi = 0.5$. Estimation of ξ (top) and tail probability (bottom), bias (left), RMSE (right): GPD-ML (full line), $E\hat{p}$ with $\rho = -2$ (dash-dotted), $E\bar{p}$ with $(k_*, m) = (190, 150)$ (dashed), and ridge regression estimator (dotted).

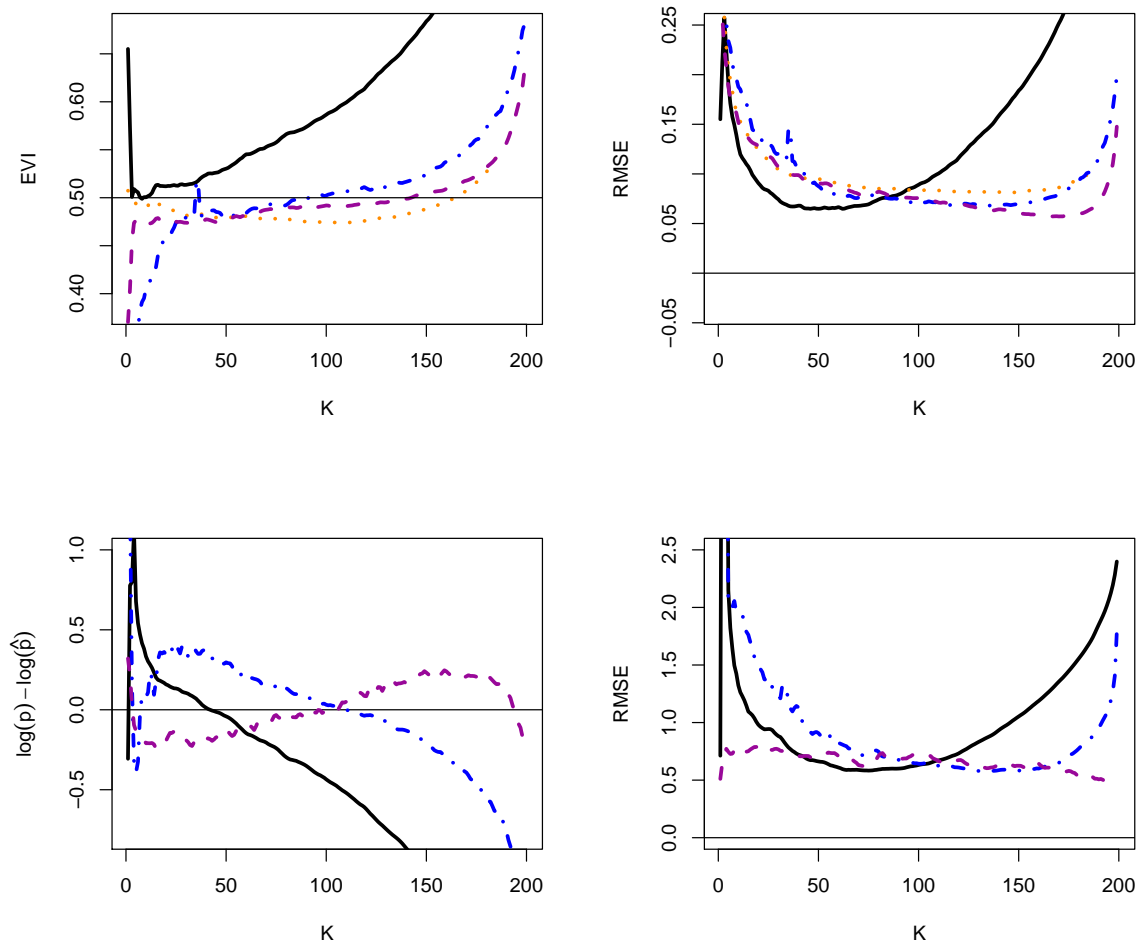


Figure 4: Fréchet distribution with $\xi = 0.5$. Estimation of ξ (top) and tail probability (bottom) using minimum variance principle, bias (left), RMSE (right): Pareto-ML (full line), Ep^+ (dash-dotted), $E\bar{p}^+$ (dashed) and corrected Hill estimator (dotted).

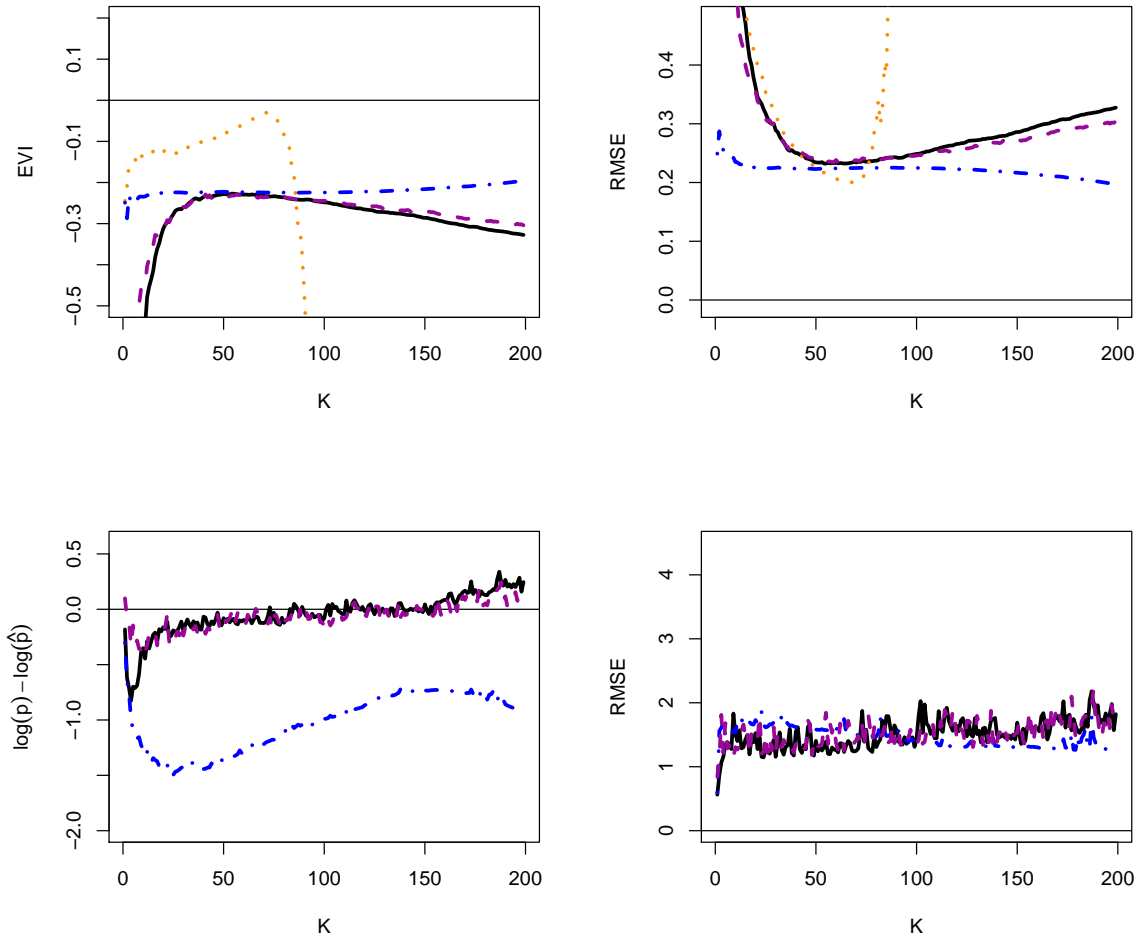


Figure 5: Standard normal distribution ($\xi = 0$ and $\tilde{\rho} = 0$). Estimation of ξ (top) and tail probability (bottom) using minimum variance principle, bias (left), RMSE (right): GPD-ML (full line), E_p (dash-dotted), $E_{\tilde{p}}$ (dashed) and ridge regression estimator (dotted).

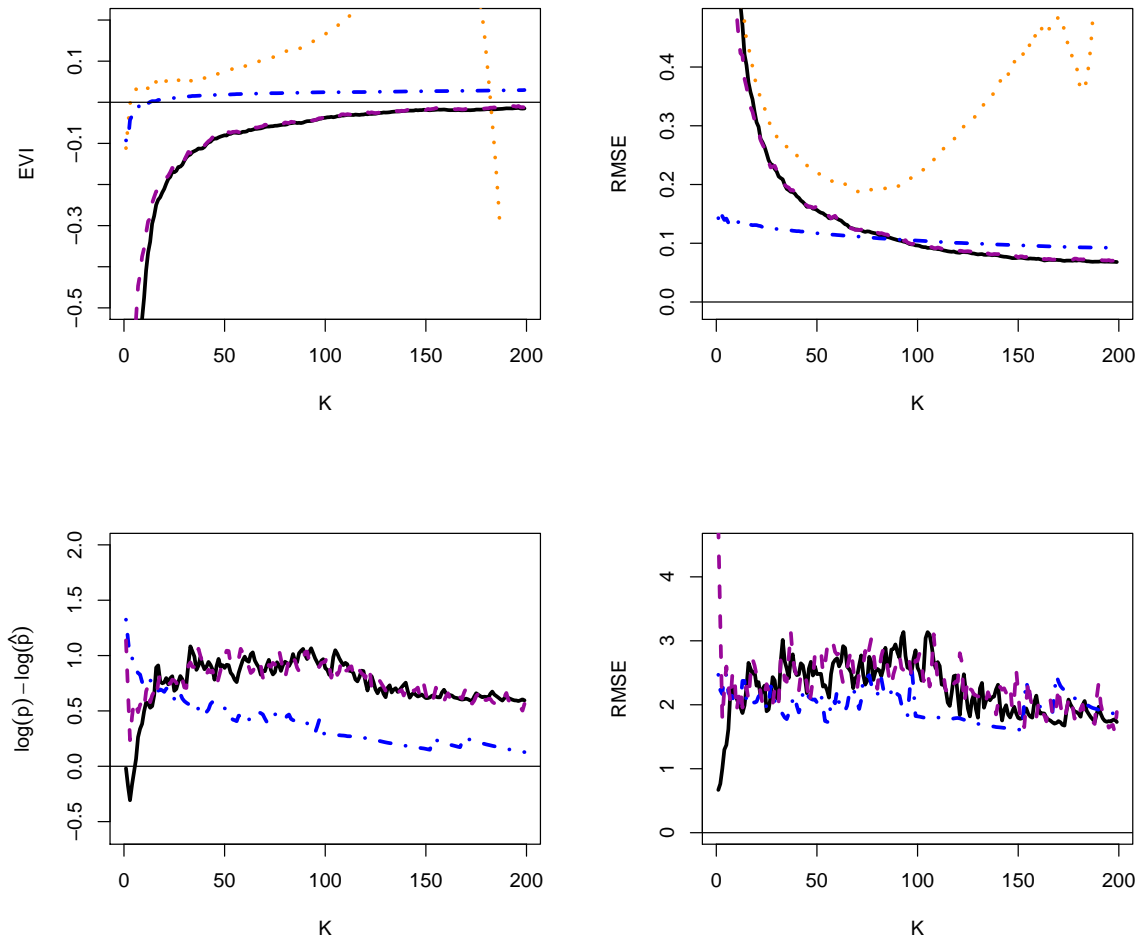


Figure 6: The exponential distribution ($\xi = 0$ and $\tilde{\rho} = 0$). Estimation of ξ (top) and tail probability (bottom) using minimum variance principle, bias (left), RMSE (right): GPD-ML (full line), $E\hat{p}$ (dash-dotted), $E\bar{p}$ (dashed) and ridge regression estimator (dotted).

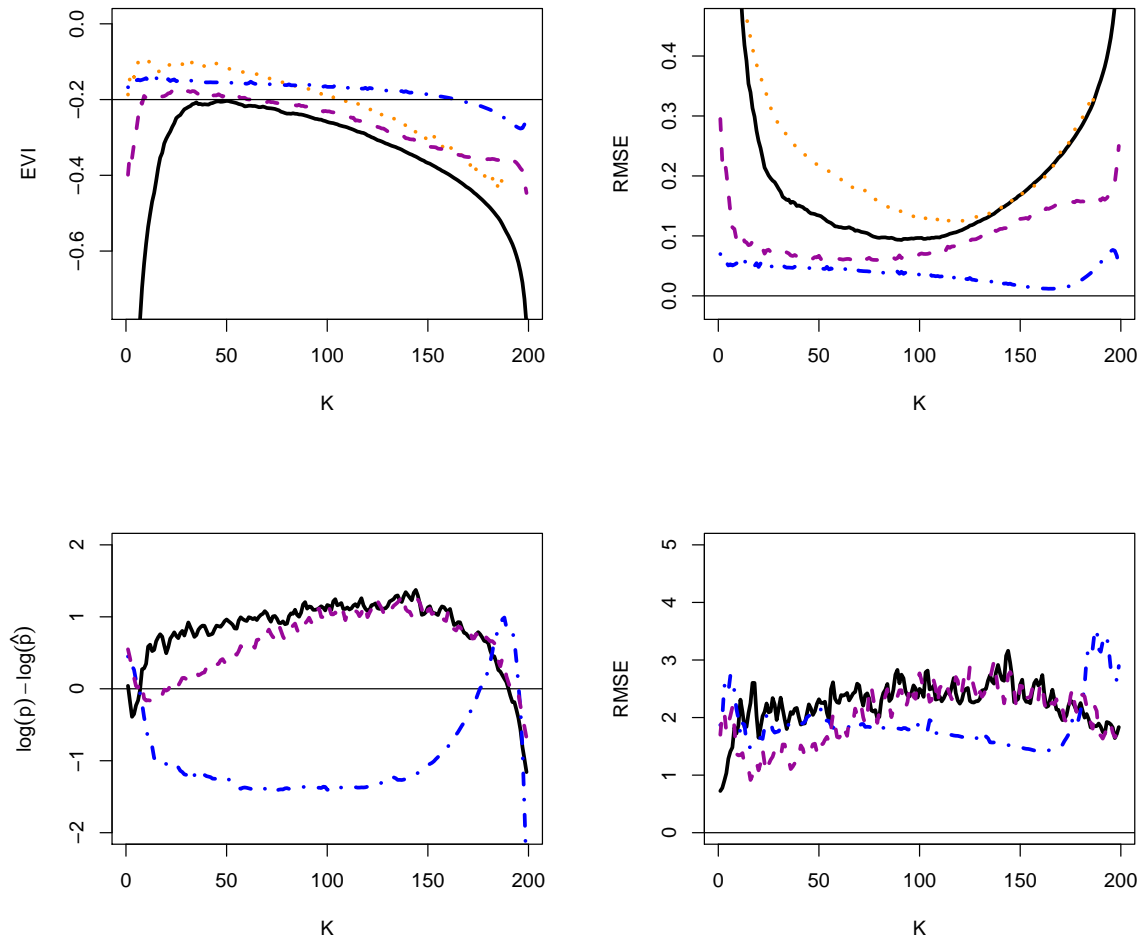


Figure 7: Reversed Burr distribution ($\xi = -0.2$ and $\tilde{\rho} = -1$). Estimation of ξ (top) and tail probability (bottom) using minimum variance principle, bias (left), RMSE (right): GPD-ML (full line), $E\hat{p}$ (dash-dotted), $E\tilde{p}$ (dashed) and ridge regression estimator (dotted).

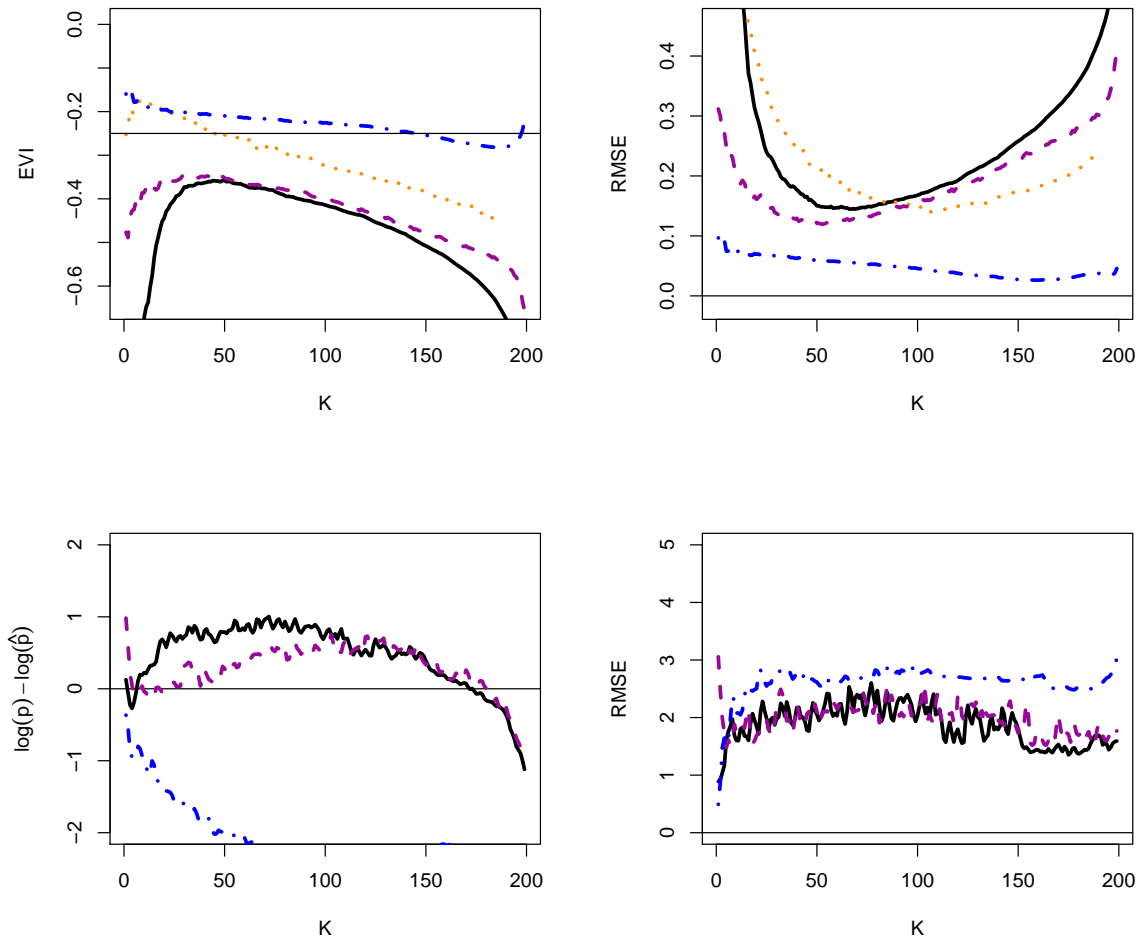


Figure 8: Extreme value Weibull distribution ($\xi = -0.25$ and $\tilde{\rho} = -1$). Estimation of ξ (top) and tail probability (bottom) using minimum variance principle, bias (left), RMSE (right): GPD-ML (full line), $E\hat{p}$ (dash-dotted), $E\bar{p}$ (dashed) and ridge regression estimator (dotted).

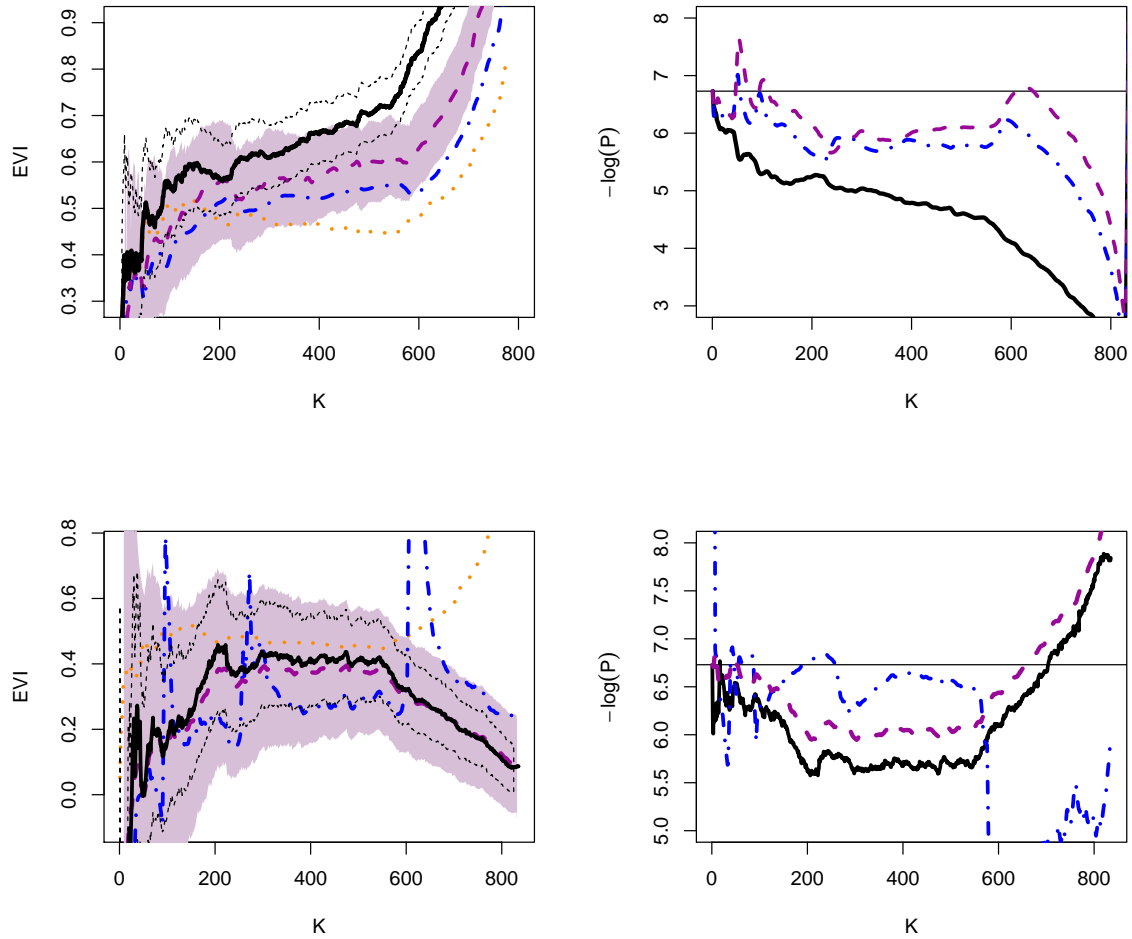


Figure 9: Ultimates of Belgian car insurance claims: estimation of ξ with asymptotic confidence intervals (left), tail probability estimation at maximum observation (right), Pareto-based analysis (top) and GPD-based analysis (bottom): classical ML estimation (full line with dotted confidence intervals), E_p (dashed with shaded confidence intervals) and $E_{\bar{p}}$ (dash-dotted). CH (top left) and ridge regression (bottom left) estimators are indicated by dotted lines.

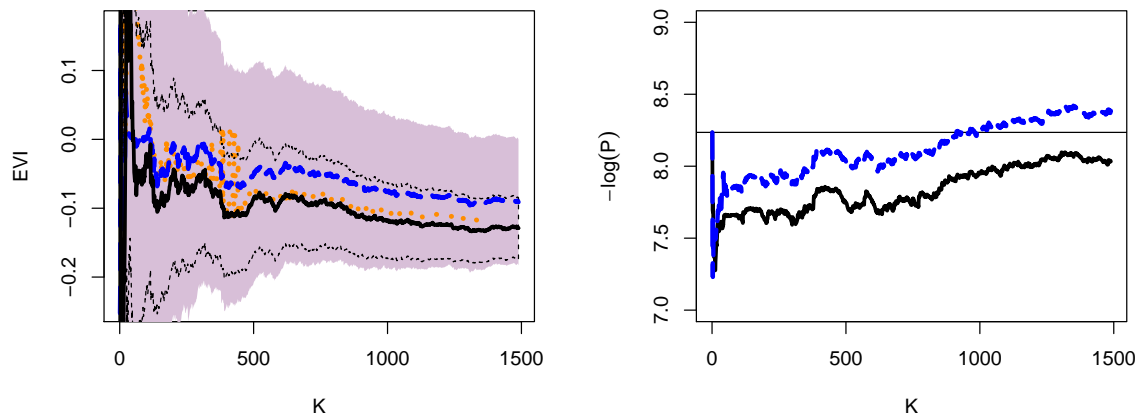


Figure 10: Lifetime data from the Netherlands, female persons who died in 1986. Left: estimation of ξ with asymptotic confidence intervals for classical ML estimation (full line with dotted confidence intervals), Ep (dashed with shaded confidence intervals, $\tilde{\rho} = -0.5$) and ridge regression (dotted). Right: tail probability estimation at maximum observation for classical ML estimation (full line) and Ep (dashed).

In actuarial statistics, Pareto-type modelling is customary in case of car insurance claim modelling. So here we provide both the plots of $\hat{\xi}_{k,n}^H$, Ep^+ , $E\bar{p}^+$ and the CH estimator (see top left in Figure 9), as well as the GPD-ML, Ep , $E\bar{p}$ and ridge regression estimator (bottom left in Figure 9), and the corresponding tail probability estimates at the right hand side. Under the Pareto approach, confining oneself to $\xi > 0$, the level 0.4 clearly appears for the EVI both using Ep^+ and $E\bar{p}^+$ when using the minimum variance principle. The CH estimator also shows a stable area around the value 0.5. The tail probability estimates of $\mathbb{P}(X > x_{n,n})$ are close to $1/n$ for almost all k values while the plot of the classical estimates is difficult to interpret.

With GPD based modelling two EVI levels are visible, around 0.2 and 0.4, of which the lower level is more clearly indicated when using $E\bar{p}$ with $k_* = 427$ and $m = 25$ as shown in Figure 9, bottom left. The ridge estimator is stable at the value 0.4. The corresponding tail probability estimates based on $E\bar{p}$ are also stable at the value $1/n$ for a long k range.

In Einmahl *et al.* (2019) [13] the life spans are studied for Dutch males and females reaching age 92 years and higher, considering their age at death. For every year, from 1986 till 2015, the life spans of this subgroup were analyzed. The authors decided to use $k = 1500$ for every year when using the classical GPD-ML estimators, and found an EVI estimate $\hat{\xi}$ between -0.1 and -0.15 for females, while for males a value around -0.15 is common over the whole period. Here we restrict ourselves to the female data from 1986. The results of Ep with asymptotic confidence intervals as discussed in Remark 3.1 with $\tilde{\rho} = -0.5$ are shown in Figure 10 (left). While the classical GPD-ML estimates decrease with increasing k from 1 to 1500, the Ep estimates show a more stable plot at a negative ξ value which is rather between -0.05 and -0.1 . The ridge regression method shows a similar value for $k \leq 500$. The corresponding tail probability estimates for a larger k indicate a value closer to the tail probability estimate $1/n$ based on the empirical distribution function, in contrast to the classical GPD approach.

5. CONCLUSIONS

In this contribution we have constructed bias reduced estimators of tail parameters extending the classical POT method. The bias can be modelled parametrically (for instance based on second order regular variation theory), or non-parametrically using Bernstein polynomial approximations. A basic asymptotic limit theorem is provided for the estimators of the extreme value parameters which allows to compute asymptotic confidence intervals. A shinyapp has been constructed with which the characteristics and the effectiveness of the proposed methods are illustrated through simulations and practical case studies. From this it follows that within the proposed methods it is always possible to improve upon the classical POT method both in bias and RMSE. This approach can also be used as a data analytic tool to enhance an extreme value analysis.

A. APPENDIX

In this section we provide details concerning the proof of Theorem 3.1.

Asymptotic distribution of $\hat{\xi}_k^{E+}$.

From (3.1) we obtain up to linear terms in δ_k that (denoting $\hat{\xi}_k$ for $\hat{\xi}_k^{E+}$)

$$\begin{cases} \hat{\delta}_k = \frac{\sum_{j=1}^k b_\eta(Y_{j,k}^{-1/\hat{\xi}_k})}{\sum_{j=1}^k b_\eta^2(Y_{j,k}^{-1/\hat{\xi}_k})}, \\ \hat{\xi}_k = \hat{\xi}_{k,n}^H + \hat{\delta}_k B_k^{(1)}, \end{cases}$$

with $B_k^{(1)} = \frac{1}{k} \sum_{j=1}^k b'_\eta(Y_{j,k}^{-1/\hat{\xi}_k}) Y_{j,k}^{-1/\hat{\xi}_k} \log Y_{j,k}$. As $k, n \rightarrow \infty$ and $k/n \rightarrow 0$ we have $B_k^{(1)} \rightarrow_p -\xi \int_0^1 b'_\eta(u) u \log u du = -\xi EB_\eta$.

Using a Taylor expansion on the numerator of the right hand side of the first equation leads to

$$\frac{1}{k} \sum_{j=1}^k b_\eta(Y_{j,k}^{-1/\hat{\xi}_k}) = \frac{1}{k} \sum_{j=1}^k b_\eta(Y_{j,k}^{-1/\xi}) - (\hat{\xi}_k - \xi) \xi^{-1} (EB_\eta) (1 + o_p(1)),$$

so that, with $\frac{1}{k} \sum_{j=1}^k b_\eta^2(Y_{j,k}^{-1/\hat{\xi}_k}) \rightarrow_p Eb_\eta^2$, up to lower order terms

$$\hat{\delta}_k = \frac{1}{Eb_\eta^2} \frac{1}{k} \sum_{j=1}^k b_\eta(Y_{j,k}^{-1/\xi}) - (\hat{\xi}_k - \xi) \xi^{-1} \frac{EB_\eta}{Eb_\eta^2} (1 + o_p(1)).$$

Hence, inserting this expansion into $\hat{\xi}_k = \hat{\xi}_{k,n}^H + \hat{\delta}_k B_k^{(1)}$, finally leads to

$$\begin{aligned} \sqrt{k}(\hat{\xi}_k - \xi)(1 + o_p(1)) &= \frac{Eb_\eta^2}{Eb_\eta^2 - (EB_\eta)^2} \sqrt{k} (\hat{\xi}_{k,n}^H - \xi) \\ &\quad - \frac{\xi EB_\eta}{Eb_\eta^2 - (EB_\eta)^2} \sqrt{k} \left(\frac{1}{k} \sum_{j=1}^k b_\eta(Y_{j,k}^{-1/\xi}) \right) \\ &= \frac{Eb_\eta^2}{Eb_\eta^2 - (EB_\eta)^2} \sqrt{k} (\hat{\xi}_{k,n}^H - \xi - \xi \delta_k EB_\eta) \\ &\quad - \frac{\xi EB_\eta}{Eb_\eta^2 - (EB_\eta)^2} \sqrt{k} \left(\frac{1}{k} \sum_{j=1}^k b_\eta(Y_{j,k}^{-1/\xi}) - \delta_k Eb_\eta^2 \right), \end{aligned}$$

with $\delta_k = \delta(U(n/k))$. We now show that this final expression is a linear combination of two zero centered statistics (up to the required accuracy) which is asymptotically normal with the stated asymptotic variance. To this end let $Z_{n-k,n} \leq Z_{n-k+1,n} \leq \dots \leq Z_{n,n}$ denote the top $k + 1$ order statistics of a sample of size n from the standard Pareto distribution with

distribution function $z \mapsto z^{-1}$, $z > 1$. Then from $(\tilde{\mathcal{E}}_2^+)$

$$\begin{aligned} \hat{\xi}_{k,n}^H &= \frac{1}{k} \sum_{j=1}^k (\log U(Z_{n-j+1,n}) - \log U(Z_{n-k,n})) \\ &= \frac{1}{k} \sum_{j=1}^k \log \left\{ \left(\frac{Z_{n-j+1,n}}{Z_{n-k,n}} \right)^\xi \left[1 + \xi \delta(U(Z_{n-k,n})) B_\eta \left(\frac{Z_{n-k,n}}{Z_{n-j+1,n}} \right) \right. \right. \\ &\quad \left. \left. + o_p(1) |\delta(U(Z_{n-k,n}))| B_\eta \left(\frac{Z_{n-k,n}}{Z_{n-j+1,n}} \right) \left| \left(\frac{Z_{n-j+1,n}}{Z_{n-k,n}} \right)^\epsilon \right| \right] \right\} \\ &= \xi \frac{1}{k} \sum_{j=1}^k \log \frac{Z_{n-j+1,n}}{Z_{n-k,n}} + \xi \delta(U(Z_{n-k,n})) \frac{1}{k} \sum_{j=1}^k B_\eta \left(\frac{Z_{n-k,n}}{Z_{n-j+1,n}} \right) \\ &\quad + o_p(1) |\delta(U(Z_{n-k,n}))| \frac{1}{k} \sum_{j=1}^k \left| B_\eta \left(\frac{Z_{n-k,n}}{Z_{n-j+1,n}} \right) \right| \left(\frac{Z_{n-j+1,n}}{Z_{n-k,n}} \right)^\epsilon. \end{aligned}$$

Now $\log Z_{n-j+1,n} - \log Z_{n-k,n} =_d E_{k-j+1,k}$, the $(k-j+1)$ -th smallest value from a standard exponential sample E_1, \dots, E_k of size k , so that $\frac{1}{k} \sum_{j=1}^k \log \frac{Z_{n-j+1,n}}{Z_{n-k,n}} =_d \frac{1}{k} \sum_{j=1}^k E_j$ and $\frac{1}{k} \sum_{j=1}^k B_\eta \left(\frac{Z_{n-k,n}}{Z_{n-j+1,n}} \right) =_d \frac{1}{k} \sum_{j=1}^k B_\eta(e^{-E_j}) =_d \frac{1}{k} \sum_{j=1}^k B_\eta(U_j)$ where U_1, \dots, U_k is a uniform $(0,1)$ sample. Hence, since $\delta(U(Z_{n-k,n}))/\delta(U(n/k)) \rightarrow_p 1$ and $\frac{1}{k} \sum_{j=1}^k B_\eta(U_j) \rightarrow_p EB_\eta$, we have that $\hat{\xi}_{k,n}^H - \xi - \xi \delta_k EB_\eta$ is asymptotically equivalent to $\frac{1}{k} \sum_{j=1}^k \xi(E_j - 1)$ as $\sqrt{k} \delta_k \rightarrow \lambda$. Similarly

$$\begin{aligned} \frac{1}{k} \sum_{j=1}^k b_\eta(Y_{j,k}^{-1/\xi}) &= \frac{1}{k} \sum_{j=1}^k b_\eta \left(\left[\frac{U \left(\frac{Z_{n-j+1,n}}{Z_{n-k,n}} \frac{Z_{n-k,n}}{Z_{n-k,n}} \right)}{U(Z_{n-k,n})} \right]^{-1/\xi} \right) \\ &= \frac{1}{k} \sum_{j=1}^k b_\eta \left(\left(\frac{Z_{n-j+1,n}}{Z_{n-k,n}} \right)^{-1} \left[1 + \xi \delta(U(Z_{n-k,n})) B_\eta \left(\frac{Z_{n-k,n}}{Z_{n-j+1,n}} \right) \right. \right. \\ &\quad \left. \left. + o_p(1) |\delta(U(Z_{n-k,n}))| B_\eta \left(\frac{Z_{n-k,n}}{Z_{n-j+1,n}} \right) \left| \left(\frac{Z_{n-j+1,n}}{Z_{n-k,n}} \right)^\epsilon \right| \right]^{-1/\xi} \right) \\ &= \frac{1}{k} \sum_{j=1}^k b_\eta \left(\left(\frac{Z_{n-j+1,n}}{Z_{n-k,n}} \right)^{-1} \left[1 - \delta(U(Z_{n-k,n})) B_\eta \left(\frac{Z_{n-k,n}}{Z_{n-j+1,n}} \right) \right. \right. \\ &\quad \left. \left. + o_p(1) |\delta(U(Z_{n-k,n}))| B_\eta \left(\frac{Z_{n-k,n}}{Z_{n-j+1,n}} \right) \left| \left(\frac{Z_{n-j+1,n}}{Z_{n-k,n}} \right)^\epsilon \right| \right] \right) \\ &= \frac{1}{k} \sum_{j=1}^k b_\eta(e^{-E_j}) \\ &\quad - \delta(U(Z_{n-k,n})) \frac{1}{k} \sum_{j=1}^k b'_\eta \left(\frac{Z_{n-k,n}}{Z_{n-j+1,n}} \right) B_\eta \left(\frac{Z_{n-k,n}}{Z_{n-j+1,n}} \right) \left(\frac{Z_{n-k,n}}{Z_{n-j+1,n}} \right) (1 + o_p(1)). \end{aligned}$$

Since $\delta(U(Z_{n-k,n}))/\delta_k \rightarrow_p 1$ and $\frac{1}{k} \sum_{j=1}^k b'_\eta \left(\frac{Z_{n-k,n}}{Z_{n-j+1,n}} \right) B_\eta \left(\frac{Z_{n-k,n}}{Z_{n-j+1,n}} \right) \left(\frac{Z_{n-k,n}}{Z_{n-j+1,n}} \right) \rightarrow_p -Eb_\eta^2$ it follows that $\frac{1}{k} \sum_{j=1}^k b_\eta(Y_{j,k}^{-1/\xi}) - \delta_k Eb_\eta^2$ is asymptotically equivalent to $\frac{1}{k} \sum_{j=1}^k b_\eta(e^{-E_j}) =_d \frac{1}{k} \sum_{j=1}^k b_\eta(U_j)$ as $\sqrt{k} \delta_k \rightarrow \lambda$, which is centered at 0 since $E(b_\eta(U)) = 0$. The results incorporating $\hat{\delta}_k^{E+}$ follow similarly.

Asymptotic distribution of $\hat{\xi}_k^E$.

This derivation follows similar lines starting from (3.2):

$$\left\{ \begin{array}{l} \frac{1}{k} \sum_{j=1}^k b'_\eta(\bar{H}_{\hat{\theta}_k}(Y_{j,k})) \bar{H}_{\hat{\theta}_k}(Y_{j,k}) \log(1 + \hat{\tau}_k Y_{j,k}) \rightarrow_p -\xi EB_\eta, \\ \frac{1}{k} \sum_{j=1}^k b_\eta^2(\bar{H}_{\hat{\theta}_k}(Y_{j,k})) \rightarrow_p Eb_\eta^2, \\ \frac{1}{k} \sum_{j=1}^k b'_\eta(\bar{H}_{\hat{\theta}_k}(Y_{j,k})) \bar{H}_{\hat{\theta}_k}(Y_{j,k}) \rightarrow_p b_\eta(1), \\ \frac{1}{k} \sum_{j=1}^k b'_\eta(\bar{H}_{\hat{\theta}_k}(Y_{j,k})) \bar{H}_{\hat{\theta}_k}(Y_{j,k}) \frac{1}{1 + \hat{\tau}_k Y_{j,k}} \rightarrow_p \xi(1 + \xi) EC_\eta + b_\eta(1), \end{array} \right.$$

as $k, n \rightarrow \infty$ and $k/n \rightarrow \infty$, so that the system of equations is asymptotically equivalent to

$$\left\{ \begin{array}{l} \hat{\delta}_k = \frac{\frac{1}{k} \sum_{j=1}^k b_\eta(\bar{H}_{\hat{\theta}_k}(Y_{j,k}))}{Eb_\eta^2}, \\ \frac{1}{k} \sum_{j=1}^k \log(1 + \hat{\tau}_k Y_{j,k}) = \hat{\xi}_k + \hat{\xi}_k \hat{\delta}_k EB_\eta, \\ \frac{1}{k} \sum_{j=1}^k \frac{1}{1 + \hat{\tau}_k Y_{j,k}} = \frac{1}{1 + \hat{\xi}_k} - \hat{\xi}_k \hat{\delta}_k EC_\eta. \end{array} \right.$$

Using a Taylor expansion on the numerator of the right hand side of the first equation leads to

$$\hat{\delta}_k Eb_\eta^2 = \frac{1}{k} \sum_{j=1}^k b_\eta(\bar{H}_\theta(Y_{j,k})) - \frac{EB_\eta}{\xi} (\hat{\xi}_k - \xi) + (1 + \xi) EC_\eta \left(\frac{\hat{\tau}_k}{\tau} - 1 \right).$$

Imputing this in the second and third equation in ξ and τ , and expanding these equations linearly around the correct values (ξ, τ) , while using, as $k, n \rightarrow \infty$ and $k/n \rightarrow 0$

$$\frac{1}{k} \sum_{j=1}^k \frac{\tau Y_{j,k}}{1 + \tau Y_{j,k}} \rightarrow_p \frac{\xi}{1 + \xi} \quad \text{and} \quad \frac{1}{k} \sum_{j=1}^k \frac{\tau Y_{j,k}}{(1 + \tau Y_{j,k})^2} \rightarrow_p \frac{\xi}{(1 + \xi)(1 + 2\xi)},$$

leads to the linearized equations

$$(A.1) \left\{ \begin{array}{l} \left(\hat{\xi}_k - \xi \right) \left(-1 + \frac{(EB_\eta)^2}{Eb_\eta^2} \right) + \left(\frac{\hat{\tau}_k}{\tau} - 1 \right) \left(\frac{\xi}{1 + \xi} - \xi(1 + \xi) \frac{EB_\eta EC_\eta}{Eb_\eta^2} \right) \\ \quad = - \left(\frac{1}{k} \sum_{j=1}^k \log(1 + \tau Y_{j,k}) - \xi \right) + \frac{\xi EB_\eta}{Eb_\eta^2} \frac{1}{k} \sum_{j=1}^k b_\eta(\bar{H}_\theta(Y_{j,k})), \\ \left(\hat{\xi}_k - \xi \right) \left(\frac{1}{(1 + \xi)^2} - \frac{EB_\eta EC_\eta}{Eb_\eta^2} \right) + \left(\frac{\hat{\tau}_k}{\tau} - 1 \right) \left(-\frac{\xi}{(1 + \xi)(1 + 2\xi)} + \xi(1 + \xi) \frac{(EC_\eta)^2}{Eb_\eta^2} \right) \\ \quad = - \left(\frac{1}{k} \sum_{j=1}^k \frac{1}{1 + \tau Y_{j,k}} - \frac{1}{1 + \xi} \right) - \frac{\xi EC_\eta}{Eb_\eta^2} \frac{1}{k} \sum_{j=1}^k b_\eta(\bar{H}_\theta(Y_{j,k})). \end{array} \right.$$

Using similar derivations as in the case $\hat{\xi}_k^{E+}$, it follows that the right hand sides in (A.1) can be rewritten as a linear combination of two zero centered statistics from which the asymptotic normality of $\left(\sqrt{k}(\hat{\xi}_k^E - \xi), \sqrt{k}(\frac{\hat{\tau}_k^E}{\tau} - 1)\right)$ can be obtained, as stated in Theorem 3.1:

$$\left\{ \begin{aligned} & \left(\hat{\xi}_k - \xi \right) \left(-1 + \frac{(EB_\eta)^2}{Eb_\eta^2} \right) + \left(\frac{\hat{\tau}_k}{\tau} - 1 \right) \left(\frac{\xi}{1 + \xi} - \xi(1 + \xi) \frac{EB_\eta EC_\eta}{Eb_\eta^2} \right) \\ &= - \left(\frac{1}{k} \sum_{j=1}^k \log(1 + \tau Y_{j,k}) - \xi - \xi \delta_k EB_\eta \right) + \frac{\xi EB_\eta}{Eb_\eta^2} \left(\frac{1}{k} \sum_{j=1}^k b_\eta(\bar{H}_\theta(Y_{j,k})) - \delta_k Eb_\eta^2 \right), \\ & \left(\hat{\xi}_k - \xi \right) \left(\frac{1}{(1 + \xi)^2} - \frac{EB_\eta EC_\eta}{Eb_\eta^2} \right) + \left(\frac{\hat{\tau}_k}{\tau} - 1 \right) \left(-\frac{\xi}{(1 + \xi)(1 + 2\xi)} + \xi(1 + \xi) \frac{(EC_\eta)^2}{Eb_\eta^2} \right) \\ &= - \left(\frac{1}{k} \sum_{j=1}^k \frac{1}{1 + \tau Y_{j,k}} - \frac{1}{1 + \xi} + \xi \delta_k EC_\eta \right) - \frac{\xi EC_\eta}{Eb_\eta^2} \left(\frac{1}{k} \sum_{j=1}^k b_\eta(\bar{H}_\theta(Y_{j,k})) - \delta_k Eb_\eta^2 \right). \end{aligned} \right.$$

We hence obtain the following asymptotic representation

$$\left(\hat{\xi}_k^E - \xi, \frac{\hat{\tau}_k^E}{\tau} - 1 \right)^\top = W^{-1} \begin{pmatrix} -1 & 0 & \xi \frac{EB_\eta}{Eb_\eta^2} \\ 0 & -1 & -\xi \frac{EC_\eta}{Eb_\eta^2} \end{pmatrix} \left(U_k^{(1)}, U_k^{(2)}, U_k^{(3)} \right)^\top$$

where

$$W = \begin{pmatrix} -1 + \frac{(EB_\eta)^2}{Eb_\eta^2} & \frac{\xi}{1 + \xi} - \xi(1 + \xi) \frac{EB_\eta EC_\eta}{Eb_\eta^2} \\ \frac{1}{(1 + \xi)^2} - \frac{EB_\eta EC_\eta}{Eb_\eta^2} & -\frac{\xi}{(1 + \xi)(1 + 2\xi)} + \xi(1 + \xi) \frac{(EC_\eta)^2}{Eb_\eta^2} \end{pmatrix},$$

and

$$\sqrt{k} \left(U_k^{(1)}, U_k^{(2)}, U_k^{(3)} \right)^\top := \begin{pmatrix} \frac{1}{k} \sum_{j=1}^k \log(1 + \tau Y_{j,k}) - \xi - \xi \delta_k EB_\eta \\ \frac{1}{k} \sum_{j=1}^k \frac{1}{1 + \tau Y_{j,k}} - \frac{1}{1 + \xi} + \xi \delta_k EC_\eta \\ \frac{1}{k} \sum_{j=1}^k b_\eta(\bar{H}_\theta(Y_{j,k})) - \delta_k Eb_\eta^2 \end{pmatrix}$$

is asymptotically normal with variance-covariance matrix

$$\Sigma_U = \begin{pmatrix} \xi^2 & -\xi^2(1 + \xi)^{-2} & \xi EB_\eta \\ -\xi^2(1 + \xi)^{-2} & \xi^2(1 + \xi)^{-2}(1 + 2\xi)^{-1} & -\xi EC_\eta \\ \xi EB_\eta & -\xi EC_\eta & Eb_\eta^2 \end{pmatrix}.$$

Concerning $\hat{\delta}_k^E$ we find the following representation:

$$(Eb_\eta^2)\sqrt{k} \left(\hat{\delta}_k^E - \delta_k \right) = \begin{pmatrix} (0 \ 0 \ 1) + (-EB_\eta/\xi \ (1 + \xi)EC_\eta)W^{-1} \begin{pmatrix} -1 & 0 & \xi \frac{EB_\eta}{Eb_\eta^2} \\ 0 & -1 & -\xi \frac{EC_\eta}{Eb_\eta^2} \end{pmatrix} \end{pmatrix} \begin{pmatrix} U_k^{(1)} \\ U_k^{(2)} \\ U_k^{(3)} \end{pmatrix}.$$

ACKNOWLEDGMENTS

The authors want to thank the referees for their constructive suggestions.

The research of G. Maribe was supported wholly/in part by the National Research Foundation of South Africa (Grant Number 102628) and the DST-NRF Centre of Excellence in Mathematical and Statistical Sciences (COE-Mass). The Grantholder acknowledges that opinions, findings and conclusions or recommendations expressed in any publication generated by the NRF supported research is that of the author(s), and that the NRF accepts no liability whatsoever in this regard.


Part of P. Naveau's work was supported by the European DAMOCLES-COST-ACTION on compound events, and also benefited from French national programs, in particular FRAISE-LEFE/INSU, MELODY-ANR, and ANR-11-IDEX-0004 – 17-EURE-0006.

REFERENCES

- [1] ALBRECHER, H.; BEIRLANT, J. and TEUGELS, J. (2017). *Reinsurance: Actuarial and Statistical Aspects*, Wiley.
- [2] BABU, G.J.; CANTY, A.J. and CHAUBEY, Y.P. (2002). Application of Bernstein Polynomials for smooth estimation of a distribution and density function, *Journal of Statistical Planning and Inference*, **105**, 377–392.
- [3] BEIRLANT, J.; DIERCKX, G.; GOEGEBEUR, Y. and MATTHYS, G. (1999). Tail index estimation and an exponential regression model, *Extremes*, **2**, 157–180.
- [4] BEIRLANT, J.; DIERCKX, G.; GUILLOU, A. and STARICA, C. (2002). On exponential representations of log-spacings of extreme order statistics, *Extremes*, **5**, 257–180.
- [5] BEIRLANT, J.; GOEGEBEUR, Y.; TEUGELS, J. and SEGERS, J. (2004). *Statistics of Extremes: Theory and Applications*, Wiley, UK.
- [6] BEIRLANT, J.; DIERCKX, G.; MATTHYS, G. and GUILLOU, A. (2005). Estimation of the extreme value index and regression on generalized quantile plots, *Bernoulli*, **11**, 949–970.
- [7] BEIRLANT, J.; JOOSSENS, E. and SEGERS, J. (2009). Second-order refined peaks-over-threshold modelling for heavy-tailed distributions, *Journal of Statistical Planning and Inference*, **139**(8), 2800–2815.
- [8] BUITENDAG, S.; BEIRLANT, J. and DE WET, T. (2019). Ridge regression estimators for the extreme value index, *Extremes*, **22**, 271–292.
- [9] CAEIRO, F.; GOMES, M.I. and PESTANA, D. (2005). Direct reduction of bias of the classical Hill estimator, *REVSTAT*, **3**(2), 113–136.
- [10] CAEIRO, F.; GOMES, M.I. and RODRIGUES, L.H. (2009). Reduced-bias tail index estimators under a third order framework, *Communications in Statistics – Theory and Methods*, **38**(7), 1019–1040.
- [11] DREES, H. (1996). Refined Pickands estimators with bias correction, *Communications in Statistics – Theory and Methods*, **25**, 837–851.
- [12] DUPUIS, D. (1999). Exceedances over high thresholds: a guide to threshold selection, *Extremes*, **1**, 251–261.
- [13] EINMAHL, J.; EINMAHL, J.H.J. and HAAN, L. DE (2019). Limits to human life span through extreme value theory, *Journal of the American Statistical Association*, **114**(527), 1075–1080.

- [14] FEUERVERGER, A. and HALL, P. (1999). Estimating a tail exponent by modelling departure from a Pareto distribution, *Ann. Statist.*, **27**, 760–781.
- [15] FRAGA ALVES, M.I.; DE HAAN, L. and LIN, T. (2003). Estimation of the parameter controlling the speed of convergence in extreme value theory, *Math. Methods Statist.*, **12**, 155–176.
- [16] GOMES, M.I. and MARTINS, M.J. (2002). Asymptotically unbiased estimators of the tail index based on external estimation of the second order parameter, *Extremes*, **5**, 5–31.
- [17] HAAN, L. DE and FERREIRA, A. (2006). *Extreme Value Theory: an Introduction*, Springer Science and Business Media, LLC, New York.
- [18] HILL, B.M. (1975). A simple general approach about the tail of a distribution, *Annals of Statistics*, **3**, 1163–1174.
- [19] PAPASTATHOPOULOS, I. and TAWN, J. (2013). Extended generalized Pareto models for tail estimation, *Journal of Statistical Planning and Inference*, **143**, 131–143.
- [20] PENG, L. (1998). Asymptotically unbiased estimator for the extreme-value index, *Statist. Prob. Lett.*, **38**, 107–115.
- [21] TENCALIEC, P.; FAVRE, A.-C.; NAVEAU, P. and PRIEUR, C. (2019). Flexible semiparametric generalized Pareto modeling of the entire range of rainfall amount, *Environmetrics*, <https://doi.org/10.1002/env.2582>.

A Note on the Right Truncated Weibull Distribution and the Minimum of Power Function Distributions

Author: PEDRO JODRÁ 
– Departamento de Métodos Estadísticos, EINA, Universidad de Zaragoza,
Zaragoza, Spain
pjodra@unizar.es

Received: February 2020

Accepted: March 2020

Abstract:

- In this note, the right truncated Weibull distribution is derived as the distribution of the minimum of a random number of independent and identically distributed random variables. Specifically, the independent random variables have a common power function distribution and the random number has a zero-truncated Poisson distribution.

Keywords:

- *right truncated Weibull distribution; power function distribution; truncated Poisson distribution; minimum of random variables.*

AMS Subject Classification:

- 60E05, 62E15, 62N05.

1. INTRODUCTION

The Weibull distribution is one of the most popular probability models, both from a theoretical and practical viewpoint, and it has been successfully used to model lifetime and failure data in a wide variety of areas. Prabhakar *et al.* [11] and Rinne [13] are two excellent monograph books that review the history, theory and applications of the Weibull distribution.

To be more precise, let X be a random variable having a two-parameter Weibull distribution, that is, its cumulative distribution function (cdf) is given by

$$(1.1) \quad F_X(x; \alpha, \beta) = 1 - \exp(-\alpha x^\beta), \quad x > 0,$$

where $\alpha > 0$ and $\beta > 0$ are the scale and shape parameters, respectively. Note that the domain of the Weibull model is the positive real line. However, there are many real situations in which the data take values in a bounded interval and then a truncated distribution may be preferred. In this note, the attention will be focussed on the Weibull distribution truncated to the interval $(0, c)$, $c > 0$, which is commonly referred to as the right –or upper– truncated Weibull (RTW) distribution. The cdf of a random variable Y having a RTW distribution on $(0, c)$ is easily deduced from (1.1), namely,

$$(1.2) \quad \begin{aligned} F_Y(y; \alpha, \beta, c) &= P(X \leq y | X \leq c) = \frac{F_X(y; \alpha, \beta) - F_X(0; \alpha, \beta)}{F_X(c; \alpha, \beta) - F_X(0; \alpha, \beta)} \\ &= \frac{1 - \exp(-\alpha y^\beta)}{1 - \exp(-\alpha c^\beta)}, \quad 0 < y < c, \end{aligned}$$

where $\alpha > 0$ and $\beta > 0$. Statistical properties concerning the RTW model can be found in Martínez and Quintana [7], McEwen and Parresol [8], Rao [12], Wingo [16] and Zhang and Xie [18], among others.

On the other hand, let Z be a random variable having a power function (PF) distribution on the interval $(0, c)$, that is, its cdf is given by

$$F_Z(z; \beta, c) = \left(\frac{z}{c}\right)^\beta, \quad 0 < z < c,$$

where $\beta > 0$ is a shape parameter. Recall that the PF distribution is obtained by inverting the Pareto distribution. Statistical properties of the PF distribution can be found in Forbes *et al.* [3, Chapter 36] and Johnson *et al.* [6, Chapter 20]. A detailed review of research concerning the PF law is given in Tahir *et al.* [15]. Practical applications in different areas can also be found in Ferreira and Andrade [2] (queuing theory), Meniconi and Barry [9] (electrical component reliability) and Wu *et al.* [17] (economics and finance), among others.

There exists a well-known relationship between the non-truncated Weibull distribution and the PF distribution. If a random variable Z follows a PF distribution on $(0, 1)$ with shape parameter $\alpha > 0$, then the random variable $(-\log Z)^{1/\beta}$ has a Weibull distribution with cdf (1.1). The aim of this note is to present a non-trivial connection between the distributions RTW and PF. In the next section, it is shown that the RTW model can be derived as the distribution of the minimum of a positive random number N of independent and identically distributed (iid) random variables having a common PF distribution. Specifically, the random number N follows a zero-truncated Poisson distribution.

Before going further, it is interesting to point out that families of distributions derived as the minimum of a positive random number N of iid random variables are common in statistical applications. For example, this stochastic representation arises in reliability analysis of series systems, in which the failure of the system is due to the presence of an unknown number of independent components of the same kind and it is assumed that the system fails if at least one component fails. Some of those families of distributions are listed in Nadarajah *et al.* [10] and some applications can be found in Silva *et al.* [14]. In addition, Bobotas and Koutras [1] have also studied the special case where N is a non-negative random number with $P(N=0) > 0$.

2. MAIN RESULT

Let N be a random variable having a zero-truncated Poisson distribution with parameter $\lambda > 0$. The probability mass function of N is given by

$$(2.1) \quad P(N=n) = \frac{\lambda^n \exp(-\lambda)}{(1 - \exp(-\lambda)) n!}, \quad n = 1, 2, \dots$$

The following result provides a relationship between the RTW and the minimum of iid PF distributions. The zero-truncated Poisson distribution plays a crucial role.

Proposition 2.1. *For any $c > 0$, let Z_1, \dots, Z_N be iid random variables having a PF distribution on the interval $(0, c)$ with shape parameter $\beta > 0$. For any $\alpha > 0$, let N be a random variable having a zero-truncated Poisson distribution with parameter $\lambda = \alpha c^\beta$. Then, the random variable $T = \min\{Z_1, \dots, Z_N\}$ has a RTW distribution on the interval $(0, c)$.*

Proof: For any $n = 1, 2, \dots, c > 0$ and $\beta > 0$, the conditional cdf of the random variable $T|N=n$ is given by

$$F_{T|N=n}(t; \beta, c) = 1 - \prod_{i=1}^n (1 - F_{Z_i}(t; \beta, c)) = 1 - \left(1 - \left(\frac{t}{c}\right)^\beta\right)^n, \quad 0 < t < c.$$

From the above equation together with (2.1), for any $\alpha > 0$ the marginal cdf of T is obtained as follows:

$$\begin{aligned} F_T(t; \alpha, \beta, c) &= \sum_{n=1}^{\infty} P(T \leq t, N=n) = \sum_{n=1}^{\infty} F_{T|N=n}(t; \beta, c) P(N=n) \\ &= \sum_{n=1}^{\infty} \left[1 - \left(1 - \left(\frac{t}{c}\right)^\beta\right)^n\right] \frac{(\alpha c^\beta)^n \exp(-\alpha c^\beta)}{(1 - \exp(-\alpha c^\beta)) n!} \\ &= \frac{1 - \exp(-\alpha t^\beta)}{1 - \exp(-\alpha c^\beta)}, \quad 0 < t < c, \end{aligned}$$

which taking into account (1.2) implies the desired result. \square

To conclude, it is interesting to note that by taking the minimum of a random number N of iid PF random variables on the unit interval $(0, 1)$, Jodrá [4] and Jodrá and Jiménez-Gamero [5] have introduced two new probability distributions depending on if N follows a shifted Poisson distribution or a zero-truncated geometric distribution, respectively. Surprisingly, the well-studied RTW distribution is obtained if the random number N has a zero-truncated Poisson distribution.



ACKNOWLEDGMENTS


Research in this paper has been partially funded by Diputación General de Aragón –Grupo E24-17R– and ERDF funds.

REFERENCES

- [1] BOBOTAS, P. and KOUTRAS, M.V. (2019). Distributions of the minimum and the maximum of a random number of random variables, *Statistics and Probability Letters*, **146**, 57–64.
- [2] FERREIRA, M.A. and ANDRADE, M. (2011). The $M/G/\infty$ queue busy period distribution exponentiality, *Journal of Applied Mathematics*, **4**(3), 249–260.
- [3] FORBES, C.; EVANS, M.; HASTINGS, N. and PEACOCK, B. (2011). *Statistical Distributions*, fourth edition, John Wiley & Sons, Inc., Hoboken, New Jersey.
- [4] JODRÁ, P. (2020). A bounded distribution derived from the shifted Gompertz law, *Journal of King Saud University – Science*, **32**, 523–536.
- [5] JODRÁ, P. and JIMÉNEZ-GAMERO, M.D. A quantile regression model for bounded responses based on the exponential-geometric distribution, *REVSTAT Statistical Journal*, **18**(4), 415–436.
- [6] JOHNSON, N.L.; KOTZ, S. and BALAKRISHNAN, N. (1994). *Continuous Univariate Distributions*, Volume 1, second edition, John Wiley & Sons, Inc., New York.
- [7] MARTÍNEZ, S. and QUINTANA, F. (1991). On a test for generalized upper truncated Weibull distributions, *Statistics and Probability Letters*, **12**(4), 273–279.
- [8] MCEWEN, R.P. and PARRESOL, B.R. (1991). Moment expressions and summary statistics for the complete and truncated Weibull distribution, *Communications in Statistics – Theory and Methods*, **20**(4), 1361–1372.
- [9] MENICONI, M. and BARRY, D.M. (1996). The power function distribution: a useful and simple distribution to assess electrical component reliability, *Microelectronics Reliability*, **36**(9), 1207–1212.
- [10] NADARAJAH, S.; POPOVIĆ, B.V. and RISTIĆ, M.M. (2013). Compounding: an R package for computing continuous distributions obtained by compounding a continuous and a discrete distribution, *Computational Statistics*, **28**(3), 977–992.
- [11] PRABHAKAR, D.N.; XIE, M. and JIANG, R. (2004). *Weibull Models*, Wiley Series in Probability and Statistics, Wiley-Interscience, John Wiley & Sons, Inc., Hoboken, New Jersey.
- [12] RAO, A.S.R.S. (2006). A note on derivation of the generating function for the right truncated Rayleigh distribution, *Applied Mathematics Letters*, **19**(8), 789–794.
- [13] RINNE, H. (2009). *The Weibull Distribution. A Handbook*, CRC Press, Boca Raton.
- [14] SILVA, R.B.; BOURGUIGNON, M.; DIAS, C.R.B. and CORDEIRO, G.M. (2013). The compound class of extended Weibull power series distributions, *Computational Statistics and Data Analysis*, **58**, 352–367.
- [15] TAHIR, M.H.; ALIZADEH, M.; MANSOOR, M.; GAUSS, M.C. and ZUBAIR, M. (2016). The Weibull-power function distribution with applications, *Hacettepe Journal of Mathematics and Statistics*, **45**(1), 245–265.
- [16] WINGO, D.R. (1988). Methods for fitting the right-truncated Weibull distribution to life-test and survival data, *Biometrical Journal*, **30**(5), 545–551.
- [17] WU, Z.; KAZAZ, B.; WEBSTER, S. and YANG, K.K. (2012). Ordering, pricing, and lead-time quotation under lead-time and demand uncertainty, *Production and Operations Management*, **21**, 576–589.
- [18] ZHANG, T. and XIE, M. (2011). On the upper truncated Weibull distribution and its reliability implications, *Reliability Engineering and System Safety*, **96**, 194–200.

Asymptotic Confidence Intervals for the Difference and the Ratio of the Weighted Kappa Coefficients of Two Diagnostic Tests Subject to a Paired Design

Authors: JOSÉ ANTONIO ROLDÁN-NOFUENTES  
– Statistics (Biostatistics), University of Granada,
Spain
jaroldan@ugr.es

SAAD BOUH SIDATY-REGAD 
– Public Health and Epidemiology, University of Nouakchott,
Mauritania
sidaty_saad@yahoo.com

Received: January 2019

Revised: February 2019

Accepted: April 2020

Abstract:

- The weighted kappa coefficient of a binary diagnostic test is a measure of the beyond-chance agreement between the diagnostic test and the gold standard, and depends on the sensitivity and specificity of the diagnostic test, on the disease prevalence and on the relative importance between the false negatives and the false positives. This article studies the comparison of the weighted kappa coefficients of two binary diagnostic tests subject to a paired design through confidence intervals. Three asymptotic confidence intervals are studied for the difference between the parameters and five other intervals for the ratio. Simulation experiments were carried out to study the coverage probabilities and the average lengths of the intervals, giving some general rules for application. A method is also proposed to calculate the sample size necessary to compare the two weighted kappa coefficients through a confidence interval. A program in R has been written to solve the problem studied and it is available as supplementary material. The results were applied to a real example of the diagnosis of malaria.

Keywords:

- *binary diagnostic test; paired design; weighted kappa coefficient.*

AMS Subject Classification:

- 62P10, 6207.

1. INTRODUCTION

A diagnostic test is medical test that is applied to an individual in order to determine the presence or absence of a disease. When the result of a diagnostic test is positive (indicating the presence of the disease) or negative (indicating its absence), the diagnostic test is called a binary diagnostic test (BDT) and its accuracy is measured in terms of two fundamental parameters: sensitivity and specificity. Sensitivity (Se) is the probability of the BDT result being positive when the individual has the disease, and specificity (Sp) is the probability of the BDT result being negative when the individual does not have the disease. Sensitivity is also called true positive fraction (TPF) and specificity is also called true negative fraction (TNF), verifying that $TPF = 1 - FNF$ and that $TNF = 1 - FPF$, where FNF (FPF) is the false negative (positive) fraction. The accuracy of a BDT is assessed in relation to a gold standard (GS), which is a medical test that objectively determines whether or not an individual has the disease. When considering the losses of an erroneous classification with the BDT, the performance of the BDT is measured in terms of the weighted kappa coefficient (Kraemer *et al.*, 1990 [7]; Kraemer, 1992 [8]; Kraemer *et al.*, 2002 [9]). The weighted kappa coefficient depends on the Se and Sp of the BDT, on the disease prevalence (p) and on the relative importance between the false negatives and the false positives (weighting index c). The weighted kappa coefficient is a measure of the beyond-chance agreement between the BDT and the GS.

Furthermore, the comparison of the performance of two BDTs is an important topic in the study of Statistical Methods for Diagnosis in Medicine. The comparison of two BDTs can be made subject to two types of sample designs: unpaired design and paired design. In the book by Pepe (2003) [13] we can see a broad discussion about both types of sample designs. Summing up, subject to an unpaired design each individual is tested with a single BDT, whereas subject to a paired design each individual is tested with the two BDTs. Consequently, unpaired design consists of applying a BDT to a sample of n_1 individuals and the other BDT to another sample of n_2 individuals; paired design consists of applying both BDTs to all of the individuals of a sample sized n . The comparative studies based on a paired design are more efficient from a statistical point of view than the studies based on an unpaired design, since it minimizes the impact of the between-individual variability. Therefore, in this article we focus on paired design. Subject to this type of design, Bloch (1997) [3] has studied an asymptotic hypothesis test to compare the weighted kappa coefficients of two BDTs. Nevertheless, if the hypothesis test is significant, this method does not allow us to assess how much bigger one weighted kappa coefficient is compared to another one, and it is necessary to estimate this effect through confidence intervals (CIs). Thus, the objective of our study is to compare the weighted kappa coefficients of two BDTs through CIs. Frequentist and Bayesian CIs have been studied for the difference and for the ratio of the two weighted kappa coefficients. If a CI for the difference (ratio) does not contain the zero (one) value, then we reject the equality between the two weighted kappa coefficients and we estimate how much bigger one coefficient is than another one. Consequently, our study is an extension of the Bloch method to the situation of the CIs. We have also dealt with the problem of calculating the sample size to compare the two parameters through a CI.

The manuscript is structured in the following way. In Section 2, we explain the weighted kappa coefficient of a BDT and we relate the comparison of the weighted kappa coefficients of two BDTs with the relative true (false) positive fraction of the two BDTs.

Section 3 summarizes the Bloch method and we propose CIs for the difference and the ratio of the weighted kappa coefficients of two BDTs subject to a paired design. In Section 4, simulation experiments are carried out to study the asymptotic behaviour of the proposed CIs, and some general rules of application are given. In Section 5, we propose a method to calculate the sample size necessary to compare the two weighted kappa coefficients through a CI. In Section 6, a programme written in R is presented to solve the problems posed in this manuscript. In Section 7, the results were applied to a real example on the diagnosis of malaria, and in Section 8 the results are discussed.

2. WEIGHTED KAPPA COEFFICIENT

Let us consider a BDT that is assessed in relation to a GS. Let L (L') the loss which occurs when for a diseased (non-diseased) individual the BDT gives a negative (positive) result. Therefore, the loss L (L') is associated with a false negative (positive). If an individual (with or without the disease) is correctly diagnosed by the BDT then $L = L' = 0$. Let D be the variable that models the result of the GS: $D = 1$ when an individual has the disease and $D = 0$ when this is not the case. Let $p = P(D = 1)$ be the prevalence of the disease and $q = 1 - p$. Let T be the random variable that models the result of the BDT: $T = 1$ when the result of the BDT is positive and $T = 0$ when the result is negative. Table 1 shows the losses and the probabilities associated with the assessment of a BDT in relation to a GS, and the probabilities when the BDT and the GS are independent, i.e. when $P(T = i|D = j) = P(T = i)$. Multiplying each loss in the 2×2 table by its corresponding probability and adding up all the terms, we find $p(1 - Se)L + q(1 - Sp)L'$, a term that is defined as expected loss. Therefore, the expected loss is the loss that occurs when erroneously classifying with the BDT an individual with or without the disease.

Table 1: Losses and probabilities.

Losses (Probabilities)			
	$T = 1$	$T = 0$	Total
$D = 1$	0 (pSe)	L ($p(1 - Se)$)	L (p)
$D = 0$	L' ($q(1 - Sp)$)	0 (qSp)	L' (q)
Total	L' ($Q = pSe + q(1 - Sp)$)	L ($1 - Q = p(1 - Se) + qSp$)	$L + L'$ (1)

Probabilities when the BDT and the GS are independent			
	$T = 1$	$T = 0$	Total
$D = 1$	pQ	$p(1 - Q)$	p
$D = 0$	qQ	$q(1 - Q)$	q
Total	Q	$1 - Q$	1

Moreover, if the BDT and the GS are independent, multiplying each loss by its corresponding probability (subject to the independence between the BDT and the GS) and adding up all of the terms we find $p[p(1 - Se) + qSp]L + q[pSe + q(1 - Sp)]L'$, a term that is defined as random loss.

Therefore, the random loss is the loss that occurs when the BDT and the GS are independent. The independence between the BDT and the GS is equivalent to the Youden index of the BDT being equal to zero i.e. $Se + Sp - 1$, and is also equivalent to the expected loss being equal to the random loss. In terms of expected and random losses, the weighted kappa coefficient of a BDT is defined as

$$\kappa = \frac{\text{Random loss} - \text{Expected loss}}{\text{Random loss}}.$$

Substituting in this equation each loss with its expression, the weighted kappa coefficient of a BDT is expressed (Kraemer *et al.*, 1990 [7]; Kraemer, 1992 [8]; Kraemer *et al.*, 2002 [9]) as

$$(2.1) \quad \kappa(c) = \frac{pqY}{p(1-Q)c + qQ(1-c)},$$

where $Y = Se + Sp - 1$ is the Youden index, $Q = pSe + q(1 - Sp)$ is the probability that the BDT result is positive, and $c = L/(L + L')$ is the weighting index. The weighting index c is a measure of the relative importance between the false negatives and the false positives. For example, let us consider the diagnosis of breast cancer using as a diagnostic mammography test. If the mammography test is positive in a woman that does not have cancer (false positive), the woman will be given a biopsy that will give a negative result. The loss L' is determined from the economic costs of the diagnosis and also from the risk, stress, anxiety, etc., caused to the woman. If the mammography test is negative in a woman who has breast cancer (false negative), the woman may be diagnosed at a later stage, but the cancer may spread, and the possibility of the treatment being successful will have diminished. The loss L is determined from these considerations. The losses L and L' are measured in terms of economic costs and also from risks, stress, etc., which is why in practice their values cannot be determined. Therefore, as loss L (L') cannot be determined, L (L') is substituted by the importance that a false negative (positive) has for the clinician. The value of the weighting index c will depend therefore on the relative importance between a false negative and a false positive. If the clinician is more concerned about false negatives, as in a screening test, then $0.5 < c \leq 1$. If the clinician has greater concerns about false positives, as it is the situation in which the BDT is used as a definitive test prior to a treatment that involves a risk for the individual (e.g., a definitive test prior to a surgical operation), then $0 \leq c < 0.5$. The index c is equal to 0.5 when the clinician considers that the false negatives and the false positives have the same importance, in which case $\kappa(0.5)$ is the Cohen kappa coefficient. Weighting index c quantifies the relative importance between a false negative and a false positive, but it is not a measure that quantifies how much bigger the proportion of false negatives is compared to the false positives. If $c = 0$ then

$$(2.2) \quad \kappa(0) = \frac{Sp - (1 - Q)}{Q} = \frac{p(1 - FNF - FPF)}{p(1 - FNF) + qFPF},$$

which is the chance corrected specificity according to the kappa model. If $c = 1$ then

$$(2.3) \quad \kappa(1) = \frac{Se - Q}{1 - Q} = \frac{q(1 - FNF - FPF)}{pFNF + q(1 - FPF)},$$

which is the chance corrected sensitivity according to the kappa model. A low (high) value of $\kappa(1)$ will indicate that the value of FNF is high (low), and a low (high) value of $\kappa(0)$ will indicate that the value of FPF is high (low). The weighted kappa coefficient can be written as

$$(2.4) \quad \kappa(c) = \frac{pc(1-Q)\kappa(1) + q(1-c)Q\kappa(0)}{p(1-Q)c + qQ(1-c)},$$

which is a weighted average of $\kappa(0)$ and $\kappa(1)$. Therefore, the weighted kappa coefficient is a measure that considers the proportion of false negatives (FNF) and the proportion of false positives (FPF). Moreover, for a set value of the c index and of the accuracy (Se and Sp) of the BDT, the weighted kappa coefficient strongly depends on the disease prevalence among the population being studied, and its value increases when the disease prevalence increases. The weighted kappa coefficient is a measure of the beyond-chance agreement between the BDT and the GS. The properties of the kappa coefficient can be seen in the manuscripts of Kraemer *et al.* (2002) [9], Roldán-Nofuentes *et al.* (2009) [15] and of Roldán-Nofuentes and Amro (2018) [16].

When comparing the accuracies of two BDTs, Pepe (2003) [13] recommends using the parameters $rTPF_{12} = \frac{Se_1}{Se_2}$ and $rFPF_{12} = \frac{FPF_1}{FPF_2}$, where $FPF_h = 1 - Sp_h$, with $h = 1, 2$. If $rTPF_{12} > 1$ then the sensitivity of Test 1 is greater than that of Test 2, and if $rFPF_{12} > 1$ then the FPF of Test 1 is greater than that of Test 2 (the specificity of Test 2 is greater than that of Test 1). The comparison of the weighted kappa coefficients of two BDTs can be related to the previous measures, and these have an important effect on the comparison of $\kappa_1(c)$ and $\kappa_2(c)$. From now onwards, it is considered that $0 < Se_h < 1$, $0 < Sp_h < 1$ and $0 < p < 1$, with $h = 1, 2$. Let us consider the subindexes i and j , in such a way that if $i = 1$ ($i = 2$) then $j = 2$ ($j = 1$). It is obvious that if $rTPF_{ij} = rFPF_{ij} = 1$ then $Se_1 = Se_2$ and $Sp_1 = Sp_2$, and that therefore $\kappa_1(c) = \kappa_2(c)$ with $0 \leq c \leq 1$. Let

$$(2.5) \quad c' = \frac{(1-p)[Se_2(1-Sp_1) - Se_1(1-Sp_2)]}{p(Se_1 - Se_2) + (1-Sp_1)(Se_2 - p) - (1-Sp_2)(Se_1 - p)}.$$

In terms of $rTPF_{ij}$ and $rFPF_{ij}$ the following rules are verified to compare $\kappa_1(c)$ and $\kappa_2(c)$:

- a) If $rTPF_{ij} \geq 1$ and $rFPF_{ij} < 1$, or $rTPF_{ij} > 1$ and $rFPF_{ij} \leq 1$, then $\kappa_i(c) > \kappa_j(c)$ for $0 \leq c \leq 1$.
- b) If $rTPF_{ij} > 1$ and $rFPF_{ij} > 1$, then:
 - b.1) $\kappa_i(c) > \kappa_j(c)$ if $0 < c' < c \leq 1$;
 - b.2) $\kappa_i(c) < \kappa_j(c)$ if $0 \leq c < c' < 1$;
 - b.3) $\kappa_i(c) = \kappa_j(c)$ if $c = c'$, with $0 < c' < 1$;
 - b.4) $\kappa_i(c) > \kappa_j(c)$ for $0 \leq c \leq 1$ if $c' < 0$ (or $c' > 1$) and $rTPF_{ij} > rFPF_{ij} > 1$;
 - b.5) $\kappa_i(c) < \kappa_j(c)$ for $0 \leq c \leq 1$ if $c' < 0$ (or $c' > 1$) and $rFPF_{ij} > rTPF_{ij} > 1$.
- c) If $rTPF_{ij} < 1$ and $rFPF_{ij} < 1$, then:
 - c.1) $\kappa_i(c) > \kappa_j(c)$ if $0 \leq c < c' < 1$;
 - c.2) $\kappa_i(c) < \kappa_j(c)$ if $0 < c' < c \leq 1$;
 - c.3) $\kappa_i(c) = \kappa_j(c)$ if $c = c'$, with $0 < c' < 1$;
 - c.4) $\kappa_i(c) > \kappa_j(c)$ for $0 \leq c \leq 1$ if $c' < 0$ (or $c' > 1$) and $rTPF_{ij} > rFPF_{ij} > 1$;
 - c.5) $\kappa_i(c) < \kappa_j(c)$ for $0 \leq c \leq 1$ if $c' < 0$ (or $c' > 1$) and $rFPF_{ij} > rTPF_{ij} > 1$.

The demonstrations can be seen in the Appendix A of the supplementary material. Regarding c' , this is obtained solving the equation $\kappa_1(c) - \kappa_2(c) = 0$ in c . The graphs in Figure 1 show how $\kappa_1(c)$ (on a continuous line) and $\kappa_2(c)$ (on a dotted line) vary depending on the weighting index c , taking as prevalence $p = \{5\%, 25\%, 50\%, 75\%\}$, for $Se_1 = 0.80$, $Sp_1 = 0.95$, $Se_2 = 0.90$ and $Sp_2 = 0.85$. These graphs correspond to the case in which $rTPF_{12} < 1$ and $rFPF_{12} < 1$, and therefore $\kappa_1(c) > \kappa_2(c)$ when $c < c'$, and $\kappa_2(c) > \kappa_1(c)$ when $c > c'$, and c' is equal to 0.95 when $p = 5\%$, 0.75 when $p = 25\%$, 0.50 when $p = 50\%$ and 0.25 when $p = 75\%$. If the clinician considers that a false positive is 1.5 times more important than a false negative, then $c = 0.4$ and $\kappa_1(c) > \kappa_2(c)$ in the population with $p = \{5\%, 25\%, 50\%\}$ and $\kappa_2(c) > \kappa_1(c)$ in the population with $p = 75\%$. If in the population with $p = 75\%$ the clinician has a greater concern about a false positive than a false negative ($0 \leq c < 0.5$), then $\kappa_1(c) > \kappa_2(c)$ if $0 \leq c < 0.25$ and $\kappa_2(c) > \kappa_1(c)$ if $0.25 < c < 0.5$; in the populations with $p = \{5\%, 25\%, 50\%\}$, $\kappa_1(c) > \kappa_2(c)$ when $0 \leq c < 0.5$.

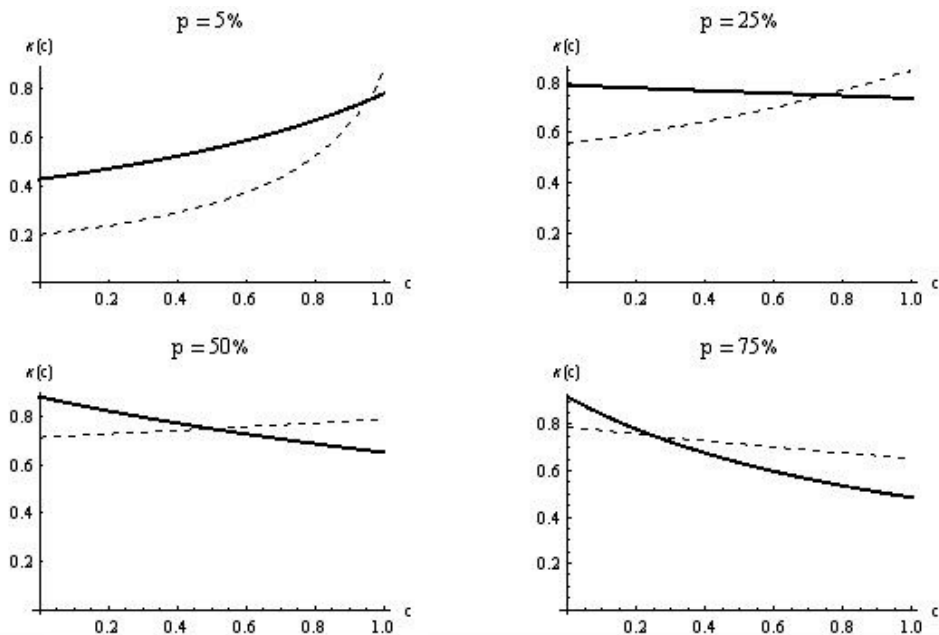


Figure 1: Weighted kappa coefficients with $rTPF_{12} < 1$ and $rFPF_{12} < 1$.

We will now study the comparison of the weighted kappa coefficients of two BDTs through CIs subject to a paired design.

3. CONFIDENCE INTERVALS

Let us consider two BDTs which are assessed in relation to the same GS. Let T_1 and T_2 be the random binary variables that model the results of each BDT respectively. Let Se_h and Sp_h be the sensitivity and specificity of the h -th BDT, with $h = 1, 2$. Table 2 (Observed frequencies) shows the frequencies that are obtained when both BDTs and the GS are applied to all the individuals in a random sample sized n . The frequencies s_{ij} and r_{ij} are the product

of a multinomial distribution whose probabilities are also shown in Table 2 (Theoretical probabilities), where $p_{ij} = P(D = 1, T_1 = i, T_2 = j)$ and $q_{ij} = P(D = 0, T_1 = i, T_2 = j)$, with $i, j = 0, 1$. Applying the Vacek (1985) [17] conditional dependency model, the probabilities p_{ij} and q_{ij} are written as

$$(3.1) \quad p_{ij} = p \left[Se_1^i (1 - Se_1)^{1-i} Se_2^j (1 - Se_2)^{1-j} + \delta_{ij} \epsilon_1 \right]$$

and

$$(3.2) \quad q_{ij} = q \left[Sp_1^{1-i} (1 - Sp_1)^i Sp_2^{1-j} (1 - Sp_2)^j + \delta_{ij} \epsilon_0 \right],$$

where ϵ_1 (ϵ_0) is the covariance or dependence factor between the two BDTs when $D = 1$ ($D = 0$), $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = -1$ if $i \neq j$, with $i, j = 0, 1$. It is verified that

$$0 \leq \epsilon_1 \leq \text{Min} \{Se_1 (1 - Se_2), Se_2 (1 - Se_1)\}$$

and

$$0 \leq \epsilon_0 \leq \text{Min} \{Sp_1 (1 - Sp_2), Sp_2 (1 - Sp_1)\}.$$

If $\epsilon_1 = \epsilon_0 = 0$ then the two BDTs are conditionally independent on the disease. In practice, the assumption of conditional independence is not realistic, and so $\epsilon_1 > 0$ and/or $\epsilon_0 > 0$. Let $\pi = (p_{11}, p_{10}, p_{01}, p_{00}, q_{11}, q_{10}, q_{01}, q_{00})^T$ be the vector of probabilities of the multinomial distribution, and it is verified that $p = \sum_{i,j=0}^1 p_{ij}$ and $q = 1 - p = \sum_{i,j=0}^1 q_{ij}$. The maximum likelihood estimators of these probabilities are $\hat{p}_{ij} = s_{ij}/n$ and $\hat{q}_{ij} = r_{ij}/n$.

The rules given in Section 2 about the effect of $rTPF$ and $rFPF$ on the comparison of $\kappa_1(c)$ and $\kappa_2(c)$ are theoretical rules that can be applied to the estimators, but they cannot guarantee that one weighted kappa coefficient will be higher than another. This question should be studied through hypothesis tests and confidence intervals. The Bloch method to compare the weighted kappa coefficients of two BDTs subject to a paired design is summarized below, and different CIs are proposed to compare these parameters subject to the same type of sample design.

Table 2: Observed frequencies and theoretical probabilities subject to a paired design.

Observed frequencies					
	$T_1 = 1$		$T_1 = 0$		
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	Total
$D = 1$	s_{11}	s_{10}	s_{01}	s_{00}	s
$D = 0$	r_{11}	r_{10}	r_{01}	r_{00}	r
Total	$s_{11} + r_{11}$	$s_{10} + r_{10}$	$s_{01} + r_{01}$	$s_{00} + r_{00}$	n

Theoretical probabilities					
	$T_1 = 1$		$T_1 = 0$		
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	Total
$D = 1$	p_{11}	p_{10}	p_{01}	p_{00}	p
$D = 0$	q_{11}	q_{10}	q_{01}	q_{00}	q
Total	$p_{11} + q_{11}$	$p_{10} + q_{10}$	$p_{01} + q_{01}$	$p_{00} + q_{00}$	1

3.1. Hypothesis test

Bloch (1997) [3] studied the comparison of the weighted kappa coefficients of two BDTs subject to a paired design. In terms of probabilities (3.1) and (3.2), the weighted kappa coefficient of Test 1 is

$$\kappa_1(c) = \frac{(p_{11} + p_{10})(q_{01} + q_{00}) - (p_{01} + p_{00})(q_{10} + q_{11})}{pc \sum_{k=0}^1 (p_{0k} + q_{0k}) + q(1-c) \sum_{k=0}^1 (p_{1k} + q_{1k})},$$

and that of Test 2 is

$$\kappa_2(c) = \frac{(p_{11} + p_{01})(q_{10} + q_{00}) - (p_{10} + p_{00})(q_{01} + q_{11})}{pc \sum_{k=0}^1 (p_{k0} + q_{k0}) + q(1-c) \sum_{k=0}^1 (p_{k1} + q_{k1})}.$$

Substituting in the previous expressions the parameters by their estimators, the estimators of the weighted kappa coefficients are

$$(3.3) \quad \hat{\kappa}_1(c) = \frac{(s_{11} + s_{10})(r_{01} + r_{00}) - (s_{01} + s_{00})(r_{10} + r_{11})}{sc \sum_{k=0}^1 (s_{0k} + r_{0k}) + r(1-c) \sum_{k=0}^1 (s_{1k} + r_{1k})}$$

and

$$(3.4) \quad \hat{\kappa}_2(c) = \frac{(s_{11} + s_{01})(r_{10} + r_{00}) - (s_{10} + s_{00})(r_{01} + r_{11})}{sc \sum_{k=0}^1 (s_{k0} + r_{k0}) + r(1-c) \sum_{k=0}^1 (s_{k1} + r_{k1})}.$$

Their variances-covariance are obtained applying the delta method (see the Appendix B of the supplementary material). Subject to paired design, the covariance between the two sensitivities and between the two specificities are given by $\text{Cov}(\hat{S}e_1, \hat{S}e_2) = \frac{\epsilon_1}{np}$ and $\text{Cov}(\hat{S}p_1, \hat{S}p_2) = \frac{\epsilon_0}{nq}$ respectively (Appendix B of the supplementary material), where ϵ_1 and ϵ_0 are the covariances between the two BDTs when $D = 1$ and $D = 0$ respectively. These covariances also affect the covariances between the two weighted kappa coefficients, just as can be seen in the expressions given in the Appendix B of the supplementary material. Finally, the statistic for the hypothesis test $H_0 : \kappa_1(c) = \kappa_2(c)$ vs $H_0 : \kappa_1(c) \neq \kappa_2(c)$ is

$$(3.5) \quad z = \frac{\hat{\kappa}_1(c) - \hat{\kappa}_2(c)}{\sqrt{\widehat{\text{Var}}[\hat{\kappa}_1(c)] + \widehat{\text{Var}}[\hat{\kappa}_2(c)] - 2\widehat{\text{Cov}}[\hat{\kappa}_1(c), \hat{\kappa}_2(c)]}} \xrightarrow{n \rightarrow \infty} N(0, 1).$$

3.2. Confidence intervals

When two parameters are compared, the interest is generally focused on studying the difference or the ratio between them. We then compare the weighted kappa coefficients of two BDTs through CIs for the difference $\delta = \kappa_1(c) - \kappa_2(c)$ and for the ratio $\theta = \frac{\kappa_1(c)}{\kappa_2(c)}$. Through the CIs: a) the two weighted kappa coefficients are compared, in such a way that if a CI for the difference (ratio) does not contain the zero (one) value, then we reject the equality between

the weighted kappa coefficients; and b) we estimate (if the two weighted kappa coefficients are different) how much bigger one weighted kappa coefficient is than the other. Firstly, three CIs are proposed for the difference of the two weighted kappa coefficients, and secondly five CIs are proposed for the ratio.

3.2.1. CIs for the difference

For the difference of the two weighted kappa coefficients we propose the Wald, bootstrap and Bayesian CIs.

Wald CI. Based on the asymptotic normality of the estimator of $\delta = \kappa_1(c) - \kappa_2(c)$, i.e. $\hat{\delta} \rightarrow N[\delta, \text{Var}(\delta)]$ when the sample size n is large, the Wald CI for the difference δ is very easy to obtain inverting the test statistic proposed by Bloch (1997) [3], therefore

$$(3.6) \quad \delta \in \hat{\kappa}_1(c) - \hat{\kappa}_2(c) \pm z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}[\hat{\kappa}_1(c)] + \widehat{\text{Var}}[\hat{\kappa}_2(c)] - 2\widehat{\text{Cov}}[\hat{\kappa}_1(c), \hat{\kappa}_2(c)]},$$

where $z_{1-\alpha/2}$ is the $100(1 - \alpha/2)$ -th percentile of the standard normal distribution.

Bootstrap CI. The bootstrap CI is calculated generating B random samples with replacement from the sample of n individuals. In each sample with replacement, we calculate the estimators of the weighted kappa coefficients and the difference between them, i.e. $\hat{\kappa}_{i1B}(c)$, $\hat{\kappa}_{i2B}(c)$ and $\hat{\delta}_{iB} = \hat{\kappa}_{i1B}(c) - \hat{\kappa}_{i2B}(c)$, with $i = 1, \dots, B$. Then, based on the B differences calculated, the average difference is estimated as $\hat{\delta}_B = \frac{1}{B} \sum_{i=1}^B \hat{\delta}_{iB}$. Assuming that the bootstrap statistic $\hat{\delta}_B$ can be transformed to a normal distribution, the bias-corrected bootstrap CI (Efron and Tibshirani, 1993 [5]) for δ is calculated in the following way. Let $A = \#(\hat{\delta}_{iB} < \hat{\delta})$ be the number of bootstrap estimators $\hat{\delta}_{iB}$ that are lower than the maximum likelihood estimator $\hat{\delta} = \hat{\kappa}_1(c) - \hat{\kappa}_2(c)$, and let $\hat{z}_0 = \Phi^{-1}(A/B)$, where $\Phi^{-1}(\cdot)$ is the inverse function of the standard normal cumulative distribution function. Let $\alpha_1 = \Phi(2\hat{z}_0 - z_{1-\alpha/2})$ and $\alpha_2 = \Phi(2\hat{z}_0 + z_{1-\alpha/2})$, then the bias-corrected bootstrap CI is $(\hat{\delta}_B^{(\alpha_1)}, \hat{\delta}_B^{(\alpha_2)})$, where $\hat{\delta}_B^{(\alpha_j)}$ is the j -th quantile of the distribution of the B bootstrap estimations of δ .

Bayesian CI. The problem is now approached from a Bayesian perspective. The number of individuals with the disease (s) is the product of a binomial distribution with parameters n and p , i.e. $s \rightarrow B(n, p)$. Conditioning on the individuals with the disease, i.e. conditioning on $D = 1$, it is verified that

$$(3.7) \quad s_{11} + s_{10} \rightarrow B(s, Se_1) \text{ and } s_{11} + s_{01} \rightarrow B(s, Se_2).$$

The number of individuals without the disease (r) is the product of a binomial distribution with parameters n and q , i.e. $r \rightarrow B(n, q)$, with $q = 1 - p$. Conditioning on the individuals without the disease ($D = 0$), it is verified that

$$(3.8) \quad r_{01} + r_{00} \rightarrow B(r, Sp_1) \text{ and } r_{10} + r_{00} \rightarrow B(r, Sp_2).$$

Considering the marginal distributions of each BDT, the estimators of the sensitivity and the specificity of the Test 1, $\hat{Se}_1 = \frac{s_{11} + s_{10}}{s}$ and $\hat{Sp}_1 = \frac{r_{01} + r_{00}}{r}$, and of the Test 2, $\hat{Se}_2 = \frac{s_{11} + s_{01}}{s}$

and $\hat{S}p_2 = \frac{r_{10}+r_{00}}{r}$, are estimators of binomial proportions. In a similar way, considering the marginal distribution of the GS, the estimator of the disease prevalence, $\hat{p} = \frac{s}{n}$, is also the estimator of a binomial proportion. Therefore, for these estimators we propose conjugate beta prior distributions, which are the appropriate distributions for the binomial distributions involved, i.e.

$$(3.9) \quad \hat{S}e_h \rightarrow \text{Beta}(\alpha_{Se_h}, \beta_{Se_h}), \hat{S}p_h \rightarrow \text{Beta}(\alpha_{Sp_h}, \beta_{Sp_h}) \text{ and } \hat{p} \rightarrow \text{Beta}(\alpha_p, \beta_p).$$

Let $\mathbf{v} = (s_{11}, s_{10}, s_{01}, s, r_{11}, r_{10}, r_{01}, r)$ be the vector of observed frequencies, with $s_{00} = s - s_{11} - s_{10} - s_{01}$, $r = n - s$ and $r_{00} = r - r_{11} - r_{10} - r_{01}$. Then the posteriori distributions for the estimators of the sensitivities, of the specificities and of the prevalence are:

$$(3.10) \quad \begin{aligned} \hat{S}e_1 | \mathbf{v} &\rightarrow \text{Beta}(s_{11} + s_{10} + \alpha_{Se_1}, s - s_{11} - s_{10} + \beta_{Se_1}), \\ \hat{S}e_2 | \mathbf{v} &\rightarrow \text{Beta}(s_{11} + s_{01} + \alpha_{Se_2}, s - s_{11} - s_{01} + \beta_{Se_2}), \\ \hat{S}p_1 | \mathbf{v} &\rightarrow \text{Beta}(r_{01} + r_{00} + \alpha_{Sp_1}, r - r_{01} - r_{00} + \beta_{Sp_1}), \\ \hat{S}p_2 | \mathbf{v} &\rightarrow \text{Beta}(r_{10} + r_{00} + \alpha_{Sp_2}, r - r_{10} - r_{00} + \beta_{Sp_2}), \\ \hat{p} | \mathbf{v} &\rightarrow \text{Beta}(s + \alpha_p, r + \beta_p). \end{aligned}$$

Once we have defined all distributions, the posteriori distribution for the weighted kappa coefficient of each BDT, and for the difference between them, can be approximated applying the Monte Carlo method. This method consists of generating M values of the posteriori distributions given in equations (3.10). In the m -th iteration, the values generated for sensitivity $\hat{S}e_h^{(m)}$ and specificity $\hat{S}p_h^{(m)}$ of each BDT, and for the prevalence $\hat{p}^{(m)}$, are plugged in the equations

$$(3.11) \quad \hat{\kappa}_h^{(m)}(c) = \frac{\hat{p}^{(m)} \hat{q}^{(m)} (\hat{S}e_h^{(m)} + \hat{S}p_h^{(m)} - 1)}{\hat{p}^{(m)} (1 - \hat{Q}_h^{(m)}) c + \hat{q}^{(m)} \hat{Q}_h^{(m)} (1 - c)}, \quad h = 1, 2,$$

where $\hat{Q}_h^{(m)} = \hat{p}^{(m)} \hat{S}e_h^{(m)} + \hat{q}^{(m)} (1 - \hat{S}p_h^{(m)})$. We then calculate the difference between the two weighted kappa coefficients in the m -th iteration: $\hat{\delta}^{(m)} = \hat{\kappa}_1^{(m)}(c) - \hat{\kappa}_2^{(m)}(c)$. As the estimator of the average difference of the weighted kappa coefficients, we calculate the average of the M estimations of difference, i.e. $\hat{\delta} = \frac{1}{M} \sum_{m=1}^M \hat{\delta}^{(m)}$. Once the Monte Carlo method is applied, based on the M values $\hat{\delta}^{(m)}$ we propose the calculation of a CI based on quantiles, i.e. the $100(1 - \alpha)\%$ CI for δ is

$$(3.12) \quad (q_{\alpha/2}, q_{1-\alpha/2}),$$

where q_γ is the γ -th quantile of the distribution of the M values $\hat{\delta}^{(m)}$.

3.2.2. CIs for the ratio

We propose five CIs for the ratio of the two weighted kappa coefficients: Wald, logarithmic, Fieller, bootstrap and Bayesian CIs.

Wald CI. Assuming the asymptotic normality of the estimator of $\theta = \kappa_1(c)/\kappa_2(c)$, i.e. $\hat{\theta} \rightarrow N[\theta, \text{Var}(\theta)]$ when the sample size n is large, the Wald CI for θ is

$$(3.13) \quad \theta \in \hat{\theta} \pm z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{\theta})},$$

where $\widehat{\text{Var}}(\hat{\theta})$ is obtained applying the delta method (Agresti, 2002 [1]), and whose expression is

$$\widehat{\text{Var}}(\hat{\theta}) \approx \frac{\hat{\kappa}_2^2(c) \widehat{\text{Var}}[\hat{\kappa}_1(c)] + \hat{\kappa}_1^2(c) \widehat{\text{Var}}[\hat{\kappa}_2(c)] - 2\hat{\kappa}_1(c) \hat{\kappa}_2(c) \widehat{\text{Cov}}[\hat{\kappa}_1(c), \hat{\kappa}_2(c)]}{\hat{\kappa}_2^4(c)}.$$

Expressions of the variances-covariance can be seen in the Appendix B of the supplementary material.

Logarithmic CI. Assuming the asymptotic normality of the Napierian logarithm of the $\hat{\theta}$, i.e. $\ln(\hat{\theta}) \rightarrow N(\ln(\theta), \text{Var}[\ln(\theta)])$ when the sample size n is large, an asymptotic CI for $\ln(\theta)$ is

$$\ln(\theta) \in \ln(\hat{\theta}) \pm z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}[\ln(\hat{\theta})]}.$$

Taking exponential, the logarithmic CI for θ is

$$(3.14) \quad \theta \in \hat{\theta} \times \exp\left\{\pm z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}[\ln(\hat{\theta})]}\right\},$$

where $\widehat{\text{Var}}[\ln(\hat{\theta})]$ is obtained applying the delta method (see the Appendix B of the supplementary material), i.e.

$$\widehat{\text{Var}}[\ln(\hat{\theta})] \approx \frac{\widehat{\text{Var}}[\hat{\kappa}_1(c)]}{\hat{\kappa}_1^2(c)} + \frac{\widehat{\text{Var}}[\hat{\kappa}_2(c)]}{\hat{\kappa}_2^2(c)} - \frac{2 \widehat{\text{Cov}}[\hat{\kappa}_1(c), \hat{\kappa}_2(c)]}{\hat{\kappa}_1(c) \hat{\kappa}_2(c)}.$$

Fieller CI. The Fieller method (1940) [6] is a classic method to obtain a CI for the ratio of two parameters. This method requires us to assume that the estimators are distributed according to a normal bivariate distribution, i.e. $(\hat{\kappa}_1(c), \hat{\kappa}_2(c))^T \rightarrow N[\boldsymbol{\kappa}(c), \sum_{\boldsymbol{\kappa}(c)}]$ when the sample size n is large, where

$$\boldsymbol{\kappa}(c) = (\kappa_1(c), \kappa_2(c))^T$$

and

$$\sum_{\boldsymbol{\kappa}(c)} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} \text{Var}[\kappa_1(c)] & \text{Cov}[\kappa_1(c), \kappa_2(c)] \\ \text{Cov}[\kappa_1(c), \kappa_2(c)] & \text{Var}[\kappa_2(c)] \end{pmatrix}.$$

Applying the Fieller method it is verified that

$$\hat{\kappa}_1(c) - \theta \hat{\kappa}_2(c) \xrightarrow[n \rightarrow \infty]{} N(0, \sigma_{11} - 2\theta\sigma_{12} + \theta^2\sigma_{22}).$$

The Fieller CI is obtained by searching for the set of values for that satisfy the inequality

$$\frac{[\hat{\kappa}_1(c) - \theta \hat{\kappa}_2(c)]^2}{\hat{\sigma}_{11} - 2\theta \hat{\sigma}_{12} + \theta^2 \hat{\sigma}_{22}} < z_{1-\alpha/2}^2.$$

Finally, the Fieller CI for $\theta = \kappa_1(c)/\kappa_2(c)$ is

$$(3.15) \quad \theta \in \frac{\hat{\omega}_{12} \pm \sqrt{\hat{\omega}_{12}^2 - \hat{\omega}_{11}\hat{\omega}_{22}}}{\hat{\omega}_{22}},$$

where $\hat{\omega}_{ij} = \hat{\kappa}_i(c) \times \hat{\kappa}_j(c) - \hat{\sigma}_{ij} z_{1-\alpha/2}^2$ with $i, j = 1, 2$, and verifying that $\hat{\omega}_{12} = \hat{\omega}_{21}$. This interval is valid when $\hat{\omega}_{12}^2 > \hat{\omega}_{11}\hat{\omega}_{22}$ and $\hat{\omega}_{22} \neq 0$.

Bootstrap CI. The bootstrap CI for θ is calculated in a similar way to that of the bootstrap interval explained in Section 3.1 but considering θ instead of δ . In each sample with replacement obtained we calculate the estimators of the weighted kappa coefficients and the ratio between them, i.e. $\hat{\kappa}_{i1B}(c)$, $\hat{\kappa}_{i2B}(c)$ and $\hat{\theta}_{iB} = \hat{\kappa}_{i1B}(c)/\hat{\kappa}_{i2B}(c)$, with $i = 1, \dots, B$. Then, based on the B ratios calculated we estimate the average ratio as $\hat{\theta}_B = \frac{1}{B} \sum_{i=1}^B \hat{\theta}_{iB}$. Assuming that the statistic $\hat{\theta}_B$ can be transformed to a normal distribution, the bias-corrected bootstrap CI (Efron and Tibshirani, 1993 [5]) for θ is obtained in a similar way to how the bootstrap CI for δ is calculated, considering now that $A = \#(\hat{\theta}_{iB} < \hat{\theta})$. Finally, the bias-corrected bootstrap CI is $(\hat{\theta}_B^{(\alpha_1)}, \hat{\theta}_B^{(\alpha_2)})$, where $\hat{\theta}_B^{(\alpha_j)}$ is the j -th quantile of the distribution of the B bootstrap estimations of θ .

Bayesian CI. The Bayesian CI for θ is also calculated in a similar way to that of the bayesian CI presented in Section 3.1. Considering the same distributions given in equations (3.9) and (3.10), in the m -th iteration of the Monte Carlo method we calculate the ratio $\hat{\theta}^{(m)} = \hat{\kappa}_1^{(m)}(c)/\hat{\kappa}_2^{(m)}(c)$ and as an estimator we calculate $\hat{\theta} = \frac{1}{M} \sum_{m=1}^M \hat{\theta}^{(m)}$. Finally, based on the M values $\hat{\theta}^{(m)}$ we calculate the CI based on quantiles.

The five previous CIs are for the ratio $\theta = \kappa_1(c)/\kappa_2(c)$. If we want to calculate the CI for the ratio $\kappa_2(c)/\kappa_1(c)$ ($= \theta' = 1/\theta$), then the logarithmic, Fieller, bootstrap and Bayesian CIs are obtained by calculating the inverse of each boundary of the corresponding CI for $\theta = \kappa_1(c)/\kappa_2(c)$. Nevertheless, the Wald CI for θ' is obtained from the Wald CI for θ dividing each boundary by $\hat{\theta}^2$, i.e. if (L_θ, U_θ) is the Wald CI for $\theta = \kappa_1(c)/\kappa_2(c)$ then the Wald CI for $\theta' = \kappa_2(c)/\kappa_1(c)$ is $(L_\theta/\hat{\theta}^2, U_\theta/\hat{\theta}^2)$.

4. SIMULATION EXPERIMENTS

Monte Carlo simulation experiments were carried out to study the coverage probability (CP) and the average length (AL) of each of the CIs presented in Section 3.2. For this purpose, we generated $N = 10,000$ random samples with multinomial distribution sized $n = \{25, 50, 100, 200, 300, 400, 500, 1000\}$. The random samples were generated setting the values of the weighted kappa coefficients, following these Steps:

1. For the disease prevalence, we took the values $p = \{5\%, 10\%, 25\%, 50\%\}$.
2. For the weighting index, we took a small, intermediate and high value:
 $c = \{0.1, 0.5, 0.9\}$.
3. As values of the weighted kappa coefficients with $c = 0$ and $c = 1$, we took the following values: $\kappa_h(0), \kappa_h(1) = \{0.01, 0.02, \dots, 0.98, 0.99\}$.
4. Next, using all of the values set previously, we calculated the sensitivity and the specificity of each diagnostic test solving the equations

$$Se_h = \frac{[q\kappa_h(0) + p]\kappa_h(1)}{q\kappa_h(0) + p\kappa_h(1)} \text{ and } Sp_h = \frac{[p\kappa_h(1) + q]\kappa_h(0)}{q\kappa_h(0) + p\kappa_h(1)},$$

considering, quite logically, only those cases in which the Youden index is higher than 0, i.e. $Y_h = Se_h + Sp_h - 1 > 0$.

5. The values of $\kappa_h(c)$ were calculated applying the equation

$$\kappa_h(c) = \frac{pc(1 - Q_h)\kappa_h(1) + q(1 - c)Q_h\kappa_h(0)}{pc(1 - Q_h) + q(1 - c)Q_h},$$

where $Q_h = pSe_h + q(1 - Sp_h)$.

6. As values of the weighted kappa coefficients we considered $\kappa_h(c) = \{0.2, 0.4, 0.6, 0.8\}$, and from these we calculated δ and θ . In order to be able to compare the coverage probabilities of the CIs for δ and for θ , $\kappa_1(c)$ and $\kappa_2(c)$ must be the same for δ and θ .

Following the idea of Cicchetti (2001) [4], simulations were carried out for values of $\kappa_h(c)$ with different levels of significance: poor ($\kappa_h(c) < 0.40$), fair ($0.40 \leq \kappa_h(c) \leq 0.59$), good ($0.60 \leq \kappa_h(c) \leq 0.74$) and excellent ($0.75 \leq \kappa_h(c) \leq 1$). As values of the dependence factors ε_1 and ε_0 we took intermediate values (50% of the maximum value of each ε_i) and high values (80% of the maximum value of each ε_i), i.e. $\varepsilon_1 = f \times \text{Min}\{Se_1(1 - Se_2), Se_2(1 - Se_1)\}$ and $\varepsilon_0 = f \times \text{Min}\{Sp_1(1 - Sp_2), Sp_2(1 - Sp_1)\}$, where $f = \{0.50, 0.80\}$. Probabilities of the multinomial distributions, equations (3.1) and (3.2), were calculated from values of the weighted kappa coefficients, and not setting the values of the sensitivities and specificities. In each scenario considered, for each one of the N random samples we calculated all the CIs proposed in Section 3.2. For the bayesian CIs we considered as prior distribution a Beta(1, 1) distribution for all of the estimators (sensitivities, specificities and prevalence). This distribution is a non-informative distribution and is flat for all possible values of each sensitivity, specificity and prevalence, and has a minimum impact on each posteriori distribution. For the bootstrap method, for each one of the N random samples we also generated $B = 2,000$ samples with replacement; and for the Bayesian method, for each one of the N random samples we also generated another $M = 10,000$. Moreover, the simulation experiments were designed in such a way that in all of the random samples generated we can estimate the weighted kappa coefficients and their variances-covariance, in order to be able to calculate all of the intervals proposed in Section 3.2. As the confidence level, we took 95%.

The comparison of the asymptotic behaviour of the CIs was made following a similar procedure to that used by other authors (Price and Bonett, 2004 [14]; Martín-Andrés and Alvarez-Hernández, 2014a [10], 2014b [11]; Montero-Alonso and Roldán-Nofuentes, 2019 [12]). This procedure consists of determining if the CI “fails” for a confidence of 95%, which happens if the CI has a $CP \leq 93\%$. The selection of the CI with the best asymptotic behaviour (for the difference and for the ratio) was made following the following Steps: 1) Choose the CIs with the least failures ($CP > 93\%$), and 2) Choose the CIs which are the most accurate, i.e. those which have the lowest AL. In the Appendix C of the supplementary material this method is justified.

4.1. CIs for the difference δ

Tables 3 and 4 show some of the results obtained (CPs and ALs) for $\delta = \{-0.6, -0.4, -0.2, 0\}$, indicating in each case the scenarios ($\kappa_h(c)$, Se_h , Sp_h and p) in which these values were obtained, and for intermediate values of the dependence factors ε_1 and ε_0 . These tables indicate the failures in bold type and it was considered that $\kappa_1(c) \leq \kappa_2(c)$.

Table 3: Coverage probabilities (CPs) and average lengths (ALs) of the CIs for the difference δ of the two weighted kappa coefficients (I).

$\kappa_1(0.1) = 0.2, \kappa_2(0.1) = 0.8, \delta = -0.6$ $Se_1 = 0.484, Sp_1 = 0.684, Se_2 = 0.852, Sp_2 = 0.911$ $\epsilon_1 = 0.0359, \epsilon_0 = 0.0306, p = 50\%$						
n	Wald		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL
25	0.335	0.866	0	0.643	0.287	0.923
50	0.737	0.646	0.038	0.589	0.762	0.690
100	0.912	0.470	0.750	0.473	0.937	0.501
200	0.958	0.337	0.952	0.354	0.968	0.364
300	0.972	0.276	0.980	0.295	0.982	0.301
400	0.960	0.239	0.969	0.258	0.971	0.262
500	0.955	0.214	0.972	0.231	0.975	0.236
1000	0.937	0.152	0.963	0.164	0.965	0.168

$\kappa_1(0.9) = 0.2, \kappa_2(0.9) = 0.8, \delta = -0.6$ $Se_1 = 0.28, Sp_1 = 0.92, Se_2 = 0.82, Sp_2 = 0.98$ $\epsilon_1 = 0.0252, \epsilon_0 = 0.0092, p = 10\%$						
n	Wald		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL
25	0.114	0.999	0	0.651	0.033	0.987
50	0.566	0.863	0	0.640	0.280	0.838
100	0.760	0.682	0.031	0.614	0.600	0.667
200	0.885	0.503	0.487	0.490	0.815	0.503
300	0.934	0.411	0.733	0.402	0.886	0.418
400	0.935	0.354	0.823	0.347	0.903	0.365
500	0.947	0.314	0.892	0.309	0.937	0.326
1000	0.947	0.220	0.938	0.218	0.947	0.233

$\kappa_1(0.1) = 0.4, \kappa_2(0.1) = 0.8, \delta = -0.4$ $Se_1 = 0.804, Sp_1 = 0.887, Se_2 = 0.82, Sp_2 = 0.98$ $\epsilon_1 = 0.0723, \epsilon_0 = 0.0089, p = 10\%$						
n	Wald		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL
25	0.847	0.812	0.473	0.671	0.920	0.899
50	0.856	0.715	0.602	0.608	0.910	0.764
100	0.924	0.534	0.847	0.528	0.953	0.580
200	0.968	0.373	0.955	0.423	0.978	0.426
300	0.957	0.302	0.986	0.367	0.976	0.369
400	0.951	0.261	0.992	0.313	0.978	0.315
500	0.955	0.232	0.994	0.259	0.979	0.262
1000	0.941	0.164	0.994	0.202	0.967	0.204

$\kappa_1(0.5) = 0.4, \kappa_2(0.5) = 0.8, \delta = -0.4$ $Se_1 = 0.76, Sp_1 = 0.72, Se_2 = 0.85, Sp_2 = 0.95$ $\epsilon_1 = 0.0570, \epsilon_0 = 0.0180, p = 25\%$						
n	Wald		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL
25	0.894	0.810	0.004	0.613	0.962	0.858
50	0.935	0.580	0.516	0.516	0.961	0.641
100	0.945	0.397	0.824	0.379	0.970	0.458
200	0.946	0.275	0.928	0.271	0.971	0.320
300	0.952	0.221	0.934	0.220	0.974	0.259
400	0.940	0.191	0.938	0.192	0.963	0.224
500	0.948	0.171	0.942	0.170	0.979	0.200
1000	0.945	0.120	0.944	0.119	0.979	0.140

Table 4: Coverage probabilities (CPs) and average lengths (ALs) of the CIs for the difference δ of the two weighted kappa coefficients (II).

$\kappa_1(0.9) = 0.6, \kappa_2(0.9) = 0.8, \delta = -0.2$ $Se_1 = 0.62, Sp_1 = 0.98, Se_2 = 0.911, Sp_2 = 0.937$ $\varepsilon_1 = 0.0277, \varepsilon_0 = 0.0094, p = 5\%$						
n	Wald		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL
25	1	1.009	0.757	0.724	1	1.018
50	0.996	0.913	0.829	0.659	0.999	0.916
100	0.993	0.823	0.928	0.580	0.998	0.801
200	0.934	0.642	0.763	0.535	0.986	0.649
300	0.922	0.533	0.745	0.483	0.964	0.551
400	0.941	0.456	0.794	0.434	0.971	0.481
500	0.933	0.404	0.799	0.393	0.962	0.430
1000	0.948	0.282	0.913	0.282	0.967	0.305

$\kappa_1(0.1) = 0.6, \kappa_2(0.1) = 0.8, \delta = -0.2$ $Se_1 = 0.195, Sp_1 = 0.995, Se_2 = 0.477, Sp_2 = 0.987$ $\varepsilon_1 = 0.0509, \varepsilon_0 = 0.0026, p = 25\%$						
n	Wald		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL
25	1	0.928	1.000	0.644	1	0.981
50	0.999	0.787	1.000	0.613	1	0.866
100	0.994	0.604	0.999	0.581	0.999	0.692
200	0.985	0.429	0.997	0.464	0.998	0.505
300	0.981	0.347	0.991	0.393	0.994	0.411
400	0.973	0.297	0.986	0.346	0.992	0.352
500	0.967	0.263	0.984	0.311	0.989	0.311
1000	0.957	0.182	0.988	0.222	0.987	0.213

$\kappa_1(0.5) = 0.4, \kappa_2(0.5) = 0.4, \delta = 0$ $Se_1 = 0.76, Sp_1 = 0.72, Se_2 = 0.40, Sp_2 = 0.943$ $\varepsilon_1 = 0.0480, \varepsilon_0 = 0.0206, p = 25\%$						
n	Wald		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL
25	0.990	0.811	0.988	0.624	0.999	0.826
50	0.978	0.683	0.998	0.598	0.994	0.691
100	0.962	0.499	0.967	0.466	0.985	0.522
200	0.955	0.353	0.963	0.340	0.981	0.381
300	0.944	0.288	0.943	0.280	0.965	0.314
400	0.960	0.250	0.962	0.244	0.980	0.274
500	0.946	0.223	0.945	0.219	0.966	0.246
1000	0.951	0.158	0.951	0.155	0.972	0.175

$\kappa_1(0.9) = 0.4, \kappa_2(0.9) = 0.4, \delta = 0$ $Se_1 = 0.943, Sp_1 = 0.229, Se_2 = 0.70, Sp_2 = 0.70$ $\varepsilon_1 = 0.0200, \varepsilon_0 = 0.0343, p = 50\%$						
n	Wald		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL
25	1	0.936	1	0.735	1	0.950
50	0.997	0.788	0.997	0.717	1	0.786
100	0.992	0.602	0.982	0.578	0.997	0.617
200	0.980	0.435	0.981	0.432	0.990	0.461
300	0.959	0.356	0.965	0.358	0.973	0.382
400	0.951	0.307	0.958	0.311	0.972	0.332
500	0.956	0.274	0.958	0.278	0.969	0.297
1000	0.956	0.193	0.958	0.196	0.970	0.210

If it is considered that $\kappa_1(c) > \kappa_2(c)$, the CPs are the same and the conclusions too. From the results, the following conclusions are obtained:

- a) Wald CI. For $\delta = \{-0.6, -0.4\}$ the Wald CI fails for a small ($n \leq 50$) and a moderate sample size ($n = 100$), and for a large sample size ($n \geq 200$) the Wald CI does not fail. For $\delta = \{-0.2, 0\}$ the Wald CI does not fail.
- b) Bootstrap CI. In very general terms, for $\delta = \{-0.6, -0.4\}$ this CI fails when $n \leq 100$, and for $n \geq 200$ this interval does not fail. For $\delta = -0.2$ this CI fails for almost all the sample sizes, and for $\delta = 0$ does not fail. When this CI does not fail, the AL is slightly lower than the Wald CI for $\delta = \{-0.2, 0\}$, and slightly higher for $\delta = \{-0.6, -0.4\}$ and $n \geq 200$.
- c) Bayesian CI. In very general terms, for $\delta = \{-0.6, -0.4\}$ this CI fails when $n \leq 50$, whereas for $n \geq 100$ this CI does not fail. For $\delta = \{-0.2, 0\}$ this CI does not fail. Regarding the AL, in the situations in which it does not fail, the AL is slightly higher than the ALs of the Wald CI and of the bootstrap CI.

Similar conclusions are obtained when the dependence factors take high values. Therefore, regarding the effect of the dependence factors ϵ_i on the asymptotic behaviour of the CIs, in general terms they do not have a clear effect on the CPs of the CIs.

4.2. CIs for the ratio θ

Tables 5 and 6 show some of the results obtained for $\theta = \{0.25, 0.50, 0.75, 1\}$, considering the same scenarios as in Tables 3 and 4. As in the case of the previous CIs, it was considered that $\kappa_1(c) \leq \kappa_2(c)$, and the same conclusions are obtained if $\kappa_1(c) > \kappa_2(c)$. From the results, the following conclusions are obtained:

- a) Wald CI. The Wald CI fails when $\theta = 0.25$ and the sample size is small ($n \leq 50$) or moderate ($n = 100$), and this CI does not fail for the rest of the values of θ and sample sizes.
- b) Logarithmic CI. This CI fails when $\theta = \{0.25, 0.50\}$ and $n \leq 200 - 300$ depending on the value of θ . For $\theta = 0.75$ this CI fails for some large sample sizes, and for $\theta = 1$ it does not fail. This CI fails more than the Wald CI, and in the situations in which it does not fail, its AL is slightly higher than that of the Wald CI.
- c) Fieller CI. This CI fails when $\theta = \{0.25, 0.5\}$ and $n \leq 50$, and it does not fail for the rest of the values of θ and sample sizes. In general terms, when there are no failures, its AL is similar to that of the Wald and logarithmic CIs.
- d) Bootstrap CI. This CI has numerous failures when $\theta = \{0.25, 0.50, 0.75\}$, whereas for $\theta = 1$ it does not fail. When $\theta = 1$, its AL is greater than that of the Wald and logarithmic CIs, especially when $n \leq 400$, and its AL is also slightly lower than that of the Fieller CI.
- e) Bayesian CI. This CI only fails when $\theta = 0.25$ and $n \leq 50$. When this CI does not fail, its AL is, in general terms, somewhat larger than that of the rest of the CIs.

Similar conclusions are obtained when the dependence factors take high values. Therefore, regarding the effect of the dependence factors on the CIs, in general terms they do not have a clear effect on the CPs of the CIs.

Table 5: Coverage probabilities (CPs) and average lengths (ALs) of the CIs for the ratio θ of the two weighted kappa coefficients (I).

$\kappa_1(0.1) = 0.2, \kappa_2(0.1) = 0.8, \theta = 0.25$ $Se_1 = 0.484, Sp_1 = 0.684, Se_2 = 0.852, Sp_2 = 0.911$ $\epsilon_1 = 0.0359, \epsilon_0 = 0.0306, p = 50\%$										
n	Wald		Logarit.		Fieller		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
25	0.823	1.351	0.088	1.517	0.700	1.950	0.368	2.260	0.884	2.704
50	0.837	0.803	0.532	0.886	0.828	0.851	0.634	0.882	0.905	0.965
100	0.931	0.551	0.832	0.608	0.942	0.565	0.889	0.569	0.954	0.585
200	0.957	0.389	0.920	0.422	0.962	0.392	0.952	0.388	0.970	0.402
300	0.970	0.318	0.933	0.340	0.974	0.319	0.969	0.316	0.984	0.328
400	0.960	0.277	0.936	0.293	0.967	0.278	0.962	0.276	0.976	0.285
500	0.957	0.248	0.944	0.260	0.967	0.248	0.969	0.247	0.975	0.256
1000	0.945	0.175	0.963	0.179	0.944	0.176	0.943	0.175	0.953	0.182

$\kappa_1(0.9) = 0.2, \kappa_2(0.9) = 0.8, \theta = 0.25$ $Se_1 = 0.28, Sp_1 = 0.92, Se_2 = 0.82, Sp_2 = 0.98$ $\epsilon_1 = 0.0252, \epsilon_0 = 0.0092, p = 10\%$										
n	Wald		Logarit.		Fieller		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
25	0.885	1.760	0.002	2.029	0.566	3.567	0.011	3.175	0.866	3.851
50	0.916	1.249	0.259	1.415	0.765	1.660	0.040	1.722	0.767	1.816
100	0.936	0.846	0.636	0.947	0.884	0.939	0.363	1.048	0.843	0.986
200	0.958	0.560	0.835	0.617	0.945	0.581	0.807	0.607	0.932	0.594
300	0.967	0.440	0.900	0.479	0.960	0.450	0.902	0.456	0.948	0.459
400	0.965	0.373	0.931	0.402	0.959	0.379	0.932	0.380	0.943	0.387
500	0.971	0.327	0.936	0.349	0.971	0.331	0.942	0.330	0.960	0.339
1000	0.950	0.227	0.941	0.235	0.950	0.228	0.949	0.227	0.955	0.234

$\kappa_1(0.1) = 0.4, \kappa_2(0.1) = 0.8, \theta = 0.5$ $Se_1 = 0.804, Sp_1 = 0.887, Se_2 = 0.82, Sp_2 = 0.98$ $\epsilon_1 = 0.0723, \epsilon_0 = 0.0089, p = 10\%$										
n	Wald		Logarit.		Fieller		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
25	0.918	1.141	0.835	1.259	0.893	2.824	0.543	1.157	0.906	2.310
50	0.959	1.021	0.859	1.119	0.939	1.518	0.897	1.140	0.978	1.710
100	0.961	0.619	0.922	0.655	0.949	0.693	0.880	0.670	0.975	0.828
200	0.962	0.395	0.947	0.406	0.959	0.409	0.914	0.400	0.977	0.470
300	0.955	0.315	0.951	0.320	0.956	0.321	0.928	0.312	0.976	0.363
400	0.953	0.271	0.949	0.274	0.952	0.274	0.935	0.265	0.975	0.308
500	0.951	0.240	0.950	0.242	0.953	0.242	0.932	0.234	0.971	0.271
1000	0.939	0.169	0.943	0.170	0.939	0.170	0.934	0.163	0.963	0.189

$\kappa_1(0.5) = 0.4, \kappa_2(0.5) = 0.8, \theta = 0.5$ $Se_1 = 0.76, Sp_1 = 0.72, Se_2 = 0.85, Sp_2 = 0.95$ $\epsilon_1 = 0.0570, \epsilon_0 = 0.0180, p = 25\%$										
n	Wald		Logarit.		Fieller		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
25	0.997	1.328	0.918	1.493	0.966	2.222	0.901	2.463	0.999	2.825
50	0.983	0.780	0.924	0.848	0.966	0.855	0.925	0.894	0.995	1.057
100	0.977	0.488	0.957	0.510	0.969	0.501	0.952	0.498	0.990	0.586
200	0.958	0.323	0.956	0.329	0.957	0.327	0.940	0.320	0.981	0.372
300	0.958	0.257	0.954	0.260	0.957	0.259	0.945	0.252	0.978	0.292
400	0.948	0.221	0.947	0.222	0.948	0.221	0.936	0.215	0.966	0.249
500	0.954	0.196	0.953	0.197	0.954	0.196	0.943	0.190	0.972	0.220
1000	0.944	0.137	0.951	0.137	0.945	0.137	0.933	0.132	0.968	0.152

Table 6: Coverage probabilities (CPs) and average lengths (ALs) of the CIs for the ratio θ of the two weighted kappa coefficients (II).

$\kappa_1(0.9) = 0.6, \kappa_2(0.9) = 0.8, \theta = 0.75$ $Se_1 = 0.62, Sp_1 = 0.98, Se_2 = 0.911, Sp_2 = 0.936$ $\epsilon_1 = 0.0277, \epsilon_0 = 0.0094, p = 5\%$										
n	Wald		Logarit.		Fieller		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
25	1	1.514	1	1.679	1	2.689	0.999	2.578	1	3.538
50	0.999	1.409	0.994	1.487	0.993	1.972	0.979	2.311	1	2.392
100	0.999	1.323	0.993	1.451	0.993	1.899	0.975	1.425	1	1.980
200	0.971	0.909	0.933	0.965	0.940	1.037	0.965	0.998	0.991	1.173
300	0.946	0.709	0.916	0.738	0.939	0.767	0.958	0.784	0.973	0.854
400	0.955	0.583	0.933	0.599	0.944	0.601	0.959	0.620	0.977	0.679
500	0.943	0.506	0.925	0.516	0.931	0.516	0.961	0.551	0.969	0.579
1000	0.947	0.341	0.945	0.344	0.943	0.344	0.969	0.375	0.969	0.377
$\kappa_1(0.1) = 0.6, \kappa_2(0.1) = 0.8, \theta = 0.75$ $Se_1 = 0.195, Sp_1 = 0.995, Se_2 = 0.477, Sp_2 = 0.987$ $\epsilon_1 = 0.0509, \epsilon_0 = 0.0026, p = 25\%$										
n	Wald		Logarit.		Fieller		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
25	1	1.687	1	1.924	1	4.747	1	2.676	1	4.561
50	1	1.266	1	1.400	1	2.837	1	1.609	1	2.308
100	0.999	0.865	0.997	0.923	0.997	0.946	0.998	0.945	1	1.188
200	0.992	0.565	0.990	0.583	0.986	0.579	0.975	0.618	0.997	0.700
300	0.971	0.444	0.990	0.452	0.976	0.449	0.958	0.493	0.992	0.536
400	0.971	0.375	0.985	0.380	0.972	0.378	0.960	0.420	0.989	0.448
500	0.966	0.328	0.976	0.331	0.971	0.331	0.964	0.371	0.987	0.390
1000	0.955	0.223	0.965	0.224	0.960	0.224	0.976	0.255	0.986	0.258
$\kappa_1(0.5) = 0.4, \kappa_2(0.5) = 0.4, \theta = 1$ $Se_1 = 0.76, Sp_1 = 0.72, Se_2 = 0.40, Sp_2 = 0.943$ $\epsilon_1 = 0.0480, \epsilon_0 = 0.0206, p = 25\%$										
n	Wald		Logarit.		Fieller		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
25	0.979	1.627	0.999	1.835	0.990	5.762	0.977	2.244	0.999	3.650
50	0.953	1.525	0.991	1.708	0.977	3.028	0.981	2.173	0.995	2.728
100	0.941	1.350	0.983	1.467	0.962	2.342	0.956	1.703	0.984	2.051
200	0.953	0.972	0.971	1.014	0.955	1.212	0.960	1.091	0.979	1.251
300	0.950	0.770	0.953	0.790	0.944	0.851	0.941	0.825	0.965	0.931
400	0.955	0.658	0.969	0.670	0.960	0.705	0.959	0.694	0.980	0.776
500	0.951	0.582	0.954	0.590	0.947	0.612	0.943	0.607	0.965	0.678
1000	0.952	0.403	0.955	0.406	0.951	0.413	0.950	0.410	0.972	0.458
$\kappa_1(0.9) = 0.4, \kappa_2(0.9) = 0.4, \theta = 1$ $Se_1 = 0.943, Sp_1 = 0.229, Se_2 = 0.70, Sp_2 = 0.70$ $\epsilon_1 = 0.0200, \epsilon_0 = 0.0343, p = 50\%$										
n	Wald		Logarit.		Fieller		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
25	1	1.857	1	2.233	1	4.483	1	2.595	1	4.216
50	0.999	1.762	0.999	2.134	0.997	3.455	0.979	1.943	1	3.294
100	0.995	1.685	0.997	1.876	0.992	2.338	0.974	1.770	0.997	2.396
200	0.983	1.195	0.988	1.278	0.980	1.345	0.980	1.268	0.990	1.445
300	0.964	0.943	0.982	0.986	0.959	1.003	0.965	0.989	0.971	1.093
400	0.957	0.803	0.976	0.828	0.951	0.838	0.957	0.839	0.971	0.913
500	0.954	0.709	0.970	0.726	0.956	0.733	0.960	0.739	0.970	0.801
1000	0.956	0.491	0.964	0.496	0.956	0.499	0.959	0.505	0.969	0.545

4.3. CIs with a small sample

The results of the simulation experiments have shown that the CIs may fail when the sample size is small ($n = 25 - 50$). A classic solution to this problem is adding the correction 0.5 to each observed frequency, as is frequent in the analysis of 2×2 tables. To assess this procedure, the same simulation experiments as before were carried out for $n = \{25, 50, 100\}$ adding the value 0.5 to all of the observed frequencies s_{ij} and r_{ij} . Table 7 shows some of the results obtained for the CIs for the ratio θ . The results for the difference δ are not shown since, although this method improves the CP of the CIs, these intervals continue to fail when they failed without adding the correction. The results for $n = 100$ are not shown either, since these are very similar to those obtained without adding the correction. As conclusions, in general terms, it holds that: a) the Wald CI for θ does not fail, its CP is 100% or very close to 100%, and its AL is lower than the rest of the intervals when these do not fail; b) the logarithmic, Fieller, Bootstrap and Bayesian CIs may continue to fail when $\theta = 0.25$. Consequently, when the sample size is small one must use the Wald CI for θ adding the value 0.5 to all of the observed frequencies.

Table 7: Coverage probabilities (CPs) and average lengths (ALs) of the CIs for θ with small samples.

$\kappa_1(0.9) = 0.2, \kappa_2(0.9) = 0.8, \theta = 0.25$ $Se_1 = 0.28, Sp_1 = 0.92, Se_2 = 0.82, Sp_2 = 0.98$ $\epsilon_1 = 0.0252, \epsilon_0 = 0.00092, p = 10\%$										
n	Wald		Logarit.		Fieller		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
25	0.999	1.808	0.008	1.960	0.653	3.014	0.145	2.150	0.783	3.531
50	0.940	1.287	0.262	1.464	0.768	1.710	0.556	1.440	0.768	1.813
$\kappa_1(0.5) = 0.4, \kappa_2(0.5) = 0.8, \theta = 0.5$ $Se_1 = 0.76, Sp_1 = 0.72, Se_2 = 0.85, Sp_2 = 0.95$ $\epsilon_1 = 0.0570, \epsilon_0 = 0.0180, p = 25\%$										
n	Wald		Logarit.		Fieller		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
25	1	1.458	0.961	1.659	0.984	2.332	0.940	1.897	1	3.118
50	0.992	0.836	0.960	0.913	0.982	0.932	0.962	0.869	0.997	1.141
$\kappa_1(0.9) = 0.6, \kappa_2(0.9) = 0.8, \theta = 0.75$ $Se_1 = 0.62, Sp_1 = 0.98, Se_2 = 0.911, Sp_2 = 0.936$ $\epsilon_1 = 0.0277, \epsilon_0 = 0.0094, p = 5\%$										
n	Wald		Logarit.		Fieller		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
25	1	1.812	1	2.073	1	3.554	1	2.425	1	4.053
50	1	1.593	1	1.789	1	2.564	0.999	2.067	1	2.682
$\kappa_1(0.9) = 0.4, \kappa_2(0.9) = 0.4, \theta = 1$ $Se_1 = 0.943, Sp_1 = 0.229, Se_2 = 0.70, Sp_2 = 0.70$ $\epsilon_1 = 0.0200, \epsilon_0 = 0.0343, p = 50\%$										
n	Wald		Logarit.		Fieller		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
25	1	1.896	1	2.140	1	4.727	1	2.571	1	4.234
50	1	1.798	1	1.991	1	3.211	1	2.418	1	3.242

4.4. Rules of application

The CIs for the difference and for the ratio of the two weighted kappa coefficients compare both parameters, and therefore we can decide which method is preferable to make this comparison. Once we have studied the coverage probabilities and the average lengths of the CIs for $\delta = \kappa_1(c) - \kappa_2(c)$ and for $\theta = \kappa_1(c)/\kappa_2(c)$, from the results obtained some general rules of application can be given for the CIs in terms of sample size. These rules are based on the failures and on the coverage probabilities, since the average lengths of the CIs for the difference and for the ratio cannot be compared as they are different intervals. In terms of sample size n :

- a) If n is small ($n < 100$), use the Wald CI for θ increasing the frequencies s_{ij} and r_{ij} in 0.5.
- b) If $100 \leq n \leq 400$, use the Wald CI for the ratio θ without adding 0.5.
- c) If $n \geq 500$, use any of the CIs (for the difference or for the ratio) proposed in Section 3.2 without adding 0.5.

In general terms, if the sample size is small, the Wald CI calculated adding 0.5 to each observed frequency does not fail. In this situation, its AL increases in relation to the Wald CI without adding 0.5, but its CP also increases meaning that the interval does not fail. When $100 \leq n \leq 400$ the CI that behaves best (fewest failures and its CP shows better fluctuations around 95%) is the Wald CI for the ratio θ . When the sample size is very large ($n \geq 500$), there is no important difference between the asymptotic behaviour of the proposed CIs, and therefore any one of them can be used. When the sample size is small, ($n \leq 50$) the CIs may fail, especially when the difference between the two weighted kappa coefficients is not small.

5. SAMPLE SIZE

The determination of the sample size to compare parameters of two BDTs is a topic of interest. We then propose a method to calculate the sample size to estimate the ratio θ between two weighted kappa coefficients with a precision ϕ and a confidence $100(1 - \alpha)\%$. This method is based on the Wald CI for θ , which is, in general terms, the interval with the best asymptotic behaviour. Furthermore, this method requires a pilot sample (or another previous study) from which we calculate estimations of all of the parameters (Se_h , Sp_h , ϵ_1 , ϵ_0 and p , and consequently of $\kappa_h(c)$) and the Wald CI for θ . If the pilot sample size is not small and the Wald CI for θ calculated from this sample contains the value 1, it makes no sense to determine the sample size necessary to estimate how much bigger one weighted kappa coefficient is than the other one, as the equality between both is not rejected. Nevertheless, if the pilot sample is small and the Wald CI (adding 0.5) contains the value 1, it may be useful to calculate the sample size to estimate the ratio θ . In this situation, the Wald CI (adding 0.5) will be very wide (as the pilot sample is small) and may contain the value 1 even if $\kappa_1(c)$ and $\kappa_2(c)$ are different. Let us consider that $\kappa_2(c) \geq \kappa_1(c)$ and therefore $\theta \leq 1$, and let ϕ be the precision set by the researcher. As it has been assumed that $\theta \leq 1$, then ϕ must be lower than one, and if we want to have a high level of precision then ϕ must be a small value.

On the other and, based on the asymptotic normality of $\hat{\theta} = \hat{\kappa}_1(c)/\hat{\kappa}_2(c)$ it is verified that $\hat{\theta} \in \theta \pm z_{1-\alpha/2} \sqrt{\text{Var}(\hat{\theta})}$, i.e. the probability of obtaining an estimator $\hat{\theta}$ is in this interval with a probability $100(1 - \alpha)\%$. Setting a precision ϕ , we can then calculate the sample size n from

$$(5.1) \quad \phi = z_{1-\alpha/2} \sqrt{\text{Var}(\hat{\theta})},$$

where

$$\text{Var}(\hat{\theta}) \approx \frac{\kappa_2^2(c) \text{Var}[\hat{\kappa}_1(c)] + \kappa_1^2(c) \text{Var}[\hat{\kappa}_2(c)] - 2\kappa_1(c) \kappa_2(c) \text{Cov}[\hat{\kappa}_1(c), \hat{\kappa}_2(c)]}{\kappa_2^4(c)}.$$

In the Appendix B of the supplementary material, we can see how this expression is obtained. This variance depends on the weighted kappa coefficients and on their respective variances and covariance. Furthermore, the variances $\text{Var}[\hat{\kappa}_h(c)]$ and the covariance $\text{Cov}[\hat{\kappa}_1(c), \hat{\kappa}_2(c)]$ (their expressions can be seen in the Appendix B of the supplementary material) depend, among other parameters, on the sample size n . Consequently, it is possible to use this relation to calculate the sample size to estimate the ratio θ . Substituting in the equation of $\text{Var}(\hat{\theta})$ the variances and the covariance with its respective expressions, substituting the parameters with their estimators and clearing n in equation (5.1), it is obtained that

$$(5.2) \quad n = \frac{z_{1-\alpha/2}^2 \hat{\theta}^2}{\phi^2 \hat{p}^3 \hat{q}^3} \left\{ \sum_{h=1}^2 \left[\frac{\hat{a}_{h1}^2 \hat{S}e_h (1 - \hat{S}e_h) \hat{q} + \hat{a}_{h2}^2 \hat{S}p_h (1 - \hat{S}p_h) \hat{p} + \hat{a}_{h3}^2 \hat{p}^2 \hat{q}^2}{\hat{Y}_h^2} \right] - \frac{2}{\hat{Y}_1 \hat{Y}_2} [\hat{a}_{11} \hat{a}_{21} \hat{\epsilon}_1 \hat{q} + \hat{a}_{12} \hat{a}_{22} \hat{\epsilon}_0 \hat{p} + \hat{a}_{13} \hat{a}_{23} \hat{p}^2 \hat{q}^2] \right\},$$

where $\hat{a}_{h1} = \hat{p}\hat{q} - \hat{p}(\hat{q} - c)\hat{\kappa}_h(c)$, $\hat{a}_{h2} = \hat{a}_{h1} + (\hat{q} - c)\hat{\kappa}_h(c)$ and $\hat{a}_{h3} = (1 - 2\hat{p})\hat{Y}_h - [(1 - c - 2\hat{p})\hat{Y}_h + \hat{S}p_h + c - 1]\hat{\kappa}_h(c)$, with $h = 1, 2$. This method requires us to know $\hat{S}e_h$, $\hat{S}p_h$, $\hat{\epsilon}_1$, $\hat{\epsilon}_0$ and \hat{p} (and therefore $\hat{\kappa}_h(c)$), for example obtained from a pilot sample or from previous studies. The procedure to calculate the sample size consists of the following Steps:

1. Take pilot samples sized n' (in general terms, $n' \geq 100$ to be able to calculate the Wald CI without adding 0.5 or use the Wald CI adding 0.5 to the frequencies if n is small), and from this sample calculate $\hat{S}e_h$, $\hat{S}p_h$, $\hat{\epsilon}_1$, $\hat{\epsilon}_0$, \hat{p} and $\hat{\kappa}_h(c)$, and then calculate the Wald CI for θ . If the Wald CI calculated has a precision ϕ , i.e. if $\frac{\text{Upper limit} - \text{Lower limit}}{2} \leq \phi$, then with the pilot sample the precision has been reached and the process has finished (θ has been estimated with a precision ϕ to a confidence $100(1 - \alpha)\%$); if this is not the case, go to the following Step.
2. From the estimations obtained in Step 1, calculate the new sample size n applying equation (5.2).
3. Take the sample of n individuals ($n - n'$ is added to the pilot sample), and from the new sample we calculate $\hat{S}e_h$, $\hat{S}p_h$, $\hat{\epsilon}_1$, $\hat{\epsilon}_0$, \hat{p} , $\hat{\kappa}_h(c)$ and the Wald CI for θ . If the Wald CI calculated has a precision ϕ , then with the new sample the precision has been reached and the process has finished. If the Wald CI does not have the required precision, then this new sample is considered as a pilot sample and the process starts again at Step 1. In this situation, the new sample has a size n calculated in Step 2, i.e. we add $n - n'$ individuals to the initial pilot sample (sized n'). Therefore, the process starts again at Step 1 considering the new sample as the pilot sample and from this sample we calculate the values of the estimators and the Wald CI.

The method to calculate the sample size is an iterative method which depends on the pilot sample and which does not guarantee that θ will be estimated with the required precision. Each time that the previous process (Steps 1–3) is repeated, we calculate (starting from an initial sample) the new sample size to estimate θ , i.e. we calculate the number of individuals that must be added to the initial sample to obtain a new sample. Therefore, this process adjusts the size of the initial pilot sample, adding (in each iteration of the process: Steps 1–3) the number of individuals necessary to obtain the right sample size to estimate θ with the precision required. The programme in R described in the Section 6 allows us to calculate the sample size to estimate θ .

If the Wald CI for θ is higher than one, the BDTs can always be permuted and θ will then be lower than one. Another alternative consists of setting a value for a precision ϕ' , in a similar way to the previous situation when $\theta \leq 1$, and then apply the equation (5.2) with $\phi = \hat{\theta}^2 \phi'$, where $\hat{\theta} = \hat{\kappa}_1(c)/\hat{\kappa}_2(c) \leq 1$. This is due to the fact that if (L_θ, U_θ) is the Wald CI for $\theta = \kappa_1(c)/\kappa_2(c) \leq 1$ then the Wald CI for $\theta' = 1/\theta = \kappa_2(c)/\kappa_1(c)$ is $(L_\theta/\hat{\theta}^2, U_\theta/\hat{\theta}^2)$. It is easy to check that the calculated value of the sample size n is the same both if $\theta \leq 1$ (with precision ϕ) and if $\theta > 1$ (with precision $\phi = \hat{\theta}^2 \phi'$).

Simulation experiments were carried out to study the effect that the pilot sample has on the calculation of the sample size. These experiments consisted of generating $N = 10,000$ random samples of multinomial distributions considering the same scenarios as those given in Tables 5 and 6. The equation of the sample size depends on the values of the estimators, which in turn depend on the pilot sample. Consequently, the pilot sample may have an effect on the sample size calculated. To study this effect, the simulation experiments consisted of the following Steps:

1. Calculate the sample size n from the values of the parameters set in the different scenarios considered. Therefore, equation (5.2) was applied using the values of the parameters (instead of their estimators).
2. Generate the N multinomial random samples sized n calculating the probabilities from equations (3.1) and (3.2), using the values of the previous parameters, and as ε_i we considered low values (25%), intermediate values (50%) and high values (80%). From each one of the N random samples, $\hat{S}e_h, \hat{S}p_h, \hat{\varepsilon}_1, \hat{\varepsilon}_0$ and \hat{p} (and therefore $\hat{\kappa}_h(c)$) were calculated, and then we calculated the sample size n'_i applying equation (5.2).
3. For each scenario, the average sample size and the relative bias were calculated, i.e. $\bar{n} = \sum n'_i/N$ and $RB(n') = (\bar{n} - n)/n$.

Table 8 shows some of the results obtained. The relative biases are very small, which indicates that the equation of the calculation of the sample size provides robust values, and therefore the choice of the pilot sample does not have an important effect on the calculation of the sample size.

Table 8: Effect of the pilot sample on the sample size.

$\kappa_1(0.1) = 0.2 \quad \kappa_2(0.1) = 0.8 \quad \theta = 0.25$ $Se_1 = 0.484 \quad Sp_1 = 0.684 \quad Se_2 = 0.852 \quad Sp_2 = 0.911 \quad p = 50\%$						
	$\epsilon_1 = 0.0179 \quad \epsilon_0 = 0.0153$		$\epsilon_1 = 0.0359 \quad \epsilon_0 = 0.0306$		$\epsilon_1 = 0.0574 \quad \epsilon_0 = 0.0489$	
	$\phi = 0.05$	$\phi = 0.10$	$\phi = 0.05$	$\phi = 0.10$	$\phi = 0.05$	$\phi = 0.10$
Sample size	3170	793	3066	767	2942	736
Average sample size	3173	795	3068	769	2946	738
Relative bias (%)	0.095	0.252	0.065	0.261	0.136	0.272
$\kappa_1(0.9) = 0.2 \quad \kappa_2(0.9) = 0.8 \quad \theta = 0.25$ $Se_1 = 0.28 \quad Sp_1 = 0.92 \quad Se_2 = 0.82 \quad Sp_2 = 0.98 \quad p = 10\%$						
	$\epsilon_1 = 0.0126 \quad \epsilon_0 = 0.0046$		$\epsilon_1 = 0.0252 \quad \epsilon_0 = 0.0092$		$\epsilon_1 = 0.0403 \quad \epsilon_0 = 0.0147$	
	$\phi = 0.05$	$\phi = 0.10$	$\phi = 0.05$	$\phi = 0.10$	$\phi = 0.05$	$\phi = 0.10$
Sample size	5104	1276	4947	1237	4758	1190
Average sample size	5113	1287	4948	1246	4759	1218
Relative bias (%)	0.18	0.83	0.02	0.73	0.02	2.35

6. PROGRAMME `citwkc`

A programme has been written in R and called “`citwkc`” (Confidence Intervals for Two Weighted Kappa Coefficients) which allows us to calculate the CIs proposed in Section 3 and the sample size proposed in Section 5. The programme runs with the command

$$\text{citwkc}(s_{11}, s_{10}, s_{01}, s_{00}, r_{11}, r_{10}, r_{01}, r_{00}, \text{cindex}, \text{preci} = 0, \text{conf} = 0.95),$$

where *cindex* is the weighting index, *preci* is the precision that is needed to calculate the sample size and *conf* is the level of confidence (by default 95%). By default *preci* = 0, and the programme does not calculate the sample size, and only calculates it when *preci* > 0. In this situation (*preci* > 0), the programme checks if it is necessary to calculate the sample size. The programme checks that the values of the frequencies and of the parameters are viable (e.g. that there are no negative values, frequencies with decimals, etc.), and also checks that it is possible to estimate all of the parameters and their variances-covariances. For the intervals obtained applying the bootstrap method, 2,000 samples with replacement are generated, and for the Bayesian intervals 10,000 random samples are generated. The results obtained on running the programme are saved in file called “`Results_citwkc.txt`” in the same folders from where the programme is run. The program is available for free at URL:

<https://www.ugr.es/local/bioest/software/cmd.php?seccion=mdb>

7. APPLICATION

The results obtained have been applied to the study by Batwala *et al.* (2010) [2] on the diagnosis of malaria. Batwala *et al.* have applied the Expert Microscopy Test and the HRP2-Based Rapid Diagnostic Test to a sample of 300 individuals using the PCR as the GS. The observed frequencies of this study are shown in Table 9, where the T_1 models the result of the Expert Microscopy Test, T_2 models the result of the HRP2-Based Rapid Diagnostic Test and D models the result of the PCR. In this example, $\hat{S}e_1 = 46.07\%$, $\hat{S}p_1 = 97.16\%$, $\hat{S}e_2 = 91.01\%$ and $\hat{S}p_2 = 86.26\%$, and therefore $r\widehat{TPF}_{12} = 0.506$ and $r\widehat{FPF}_{12} = 0.207$. Applying the equation (2.5) it holds that $c' = 0.1902$. As $r\widehat{TPF}_{12} < 1$ and $r\widehat{FPF}_{12} < 1$, applying the rule c) given in Section 2, it holds that $\hat{\kappa}_1(c) > \hat{\kappa}_2(c)$ for $0 \leq c < 0.1902$ and that $\hat{\kappa}_1(c) < \hat{\kappa}_2(c)$ for $0.1902 < c \leq 1$. Applying the rules given in Section 4, as $n = 300 < 400$ then it is necessary to use the Wald CI for the ratio θ . Table 10 shows the values of $\hat{\kappa}_h(c)$, $\hat{\delta}$, $\hat{\theta}$ and the 95% CIs for θ when $c = \{0.1, 0.1902, 0.2, \dots, 0.8, 0.9\}$. The results were obtained running the programme “citwkc” with the command “citwkc (41, 0, 40, 8, 5, 1, 24, 181, c)” taking $c = \{0.1, 0.1902, 0.2, \dots, 0.8, 0.9\}$.

Table 9: Observed frequencies of the study of Batwala *et al.*

Frequencies					
	$T_1 = 1$		$T_1 = 0$		
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	Total
$D = 1$	41	0	40	8	89
$D = 0$	5	1	24	181	211
Total	46	1	64	189	300

Table 10: CIs for the ratio $\theta = \kappa_1(c)/\kappa_2(c)$.

c	$\hat{\kappa}_1(c)$	$\hat{\kappa}_2(c)$	$\hat{\delta}$	Wald	Logarithmic	Fieller	Bootstrap	Bayesian
0.1	0.726	0.642	1.131	0.925 , 1.335	0.943 , 1.355	0.940 , 1.357	0.926 , 1.344	0.883 , 1.393
0.1902	0.659	0.659	1	0.811 , 1.189	0.828 , 1.208	0.823 , 1.206	0.817 , 1.204	0.776 , 1.234
0.2	0.653	0.661	0.988	0.800 , 1.174	0.817 , 1.194	0.812 , 1.192	0.808 , 1.192	0.766 , 1.219
0.3	0.593	0.681	0.871	0.695 , 1.046	0.711 , 1.065	0.704 , 1.059	0.701 , 1.065	0.673 , 1.083
0.4	0.543	0.701	0.775	0.609 , 0.939	0.625 , 0.958	0.615 , 0.948	0.615 , 0.952	0.593 , 0.971
0.5	0.501	0.723	0.693	0.537 , 0.847	0.553 , 0.866	0.541 , 0.854	0.541 , 0.857	0.525 , 0.877
0.6	0.464	0.747	0.621	0.476 , 0.768	0.492 , 0.786	0.479 , 0.772	0.481 , 0.776	0.468 , 0.799
0.7	0.433	0.772	0.561	0.425 , 0.698	0.440 , 0.716	0.426 , 0.701	0.430 , 0.707	0.418 , 0.727
0.8	0.406	0.799	0.508	0.380 , 0.637	0.395 , 0.654	0.381 , 0.639	0.384 , 0.644	0.375 , 0.667
0.9	0.382	0.827	0.462	0.341, 0.582	0.356 , 0.599	0.342 , 0.584	0.347 , 0.594	0.339 , 0.611

For $c = \{0.1, 0.1902, 0.2, 0.3\}$, the Wald CI for θ contains the value 1, and therefore in these cases we do not reject the equality of the weighted kappa coefficients of the Expert Microscopy Test and of the HRP2-Based Rapid Diagnostic Test. Therefore, when the clinician

considers that a false positive is 9, 4 or 2.33 times more important than a false negative, we do not reject the equality between the weighted kappa coefficients of the Expert Microscopy Test and of the HRP2-Based Rapid Diagnostic Test in the population studied. The rest of the intervals for θ also contain the value 1.

For $c = \{0.4, 0.5, \dots, 0.8, 0.9\}$, the Wald CI θ does not contain the value 1, and therefore in all of these cases we reject the equality of the weighted kappa coefficients of the Expert Microscopy Test and of the HRP2-Based Rapid Diagnostic Test in the population studied. Therefore, the clinician considers that $0.5 < c \leq 0.9$, i.e. a false negative is more important than a false positive (as happens in the situation in which the diagnostic tests are applied as screening tests), the weighted kappa coefficient of the HRP2-Based Rapid Diagnostic Test is significantly greater than the weighted kappa coefficient of the Expert Microscopy Test in the population studied. The same conclusion is obtained when the clinician considers that a false positive and a false negative have the same importance ($c = 0.5$). If the clinician considers that a false positive is 1.5 times greater than a false negative (i.e. $c = 0.4$), then the same conclusion is obtained. The rest of the CIs for θ do not contain the value 1. For example, considering $c = 0.9$, it is concluded that in the population being studied the beyond-chance agreement between the HRP2-Based Rapid Diagnostic Test and the PCR is, with a confidence of 95%, a value between 1.72 ($1/0.582 \approx 1.72$) and 2.94 ($1/0.341 \approx 2.94$) times greater than the beyond-chance agreement between the Expert Microscopy Test and the PCR.

In order to illustrate the method to calculate the sample size presented in Section 5 we will consider that $c = 0.9$, and therefore that the two BDTs are applied as a screening test. In this situation, the 95% Wald CI for θ is (0.341, 0.582), and the precision is 0.1205. As an example, we will consider that the clinician wishes to estimate the ratio between the two weighted kappa coefficients with a precision $\phi = 0.10$. As with the sample of 300 individuals the desired precision ($\phi = 0.10 < 0.1205$) was not achieved, then using this sample as a pilot sample and running the programme “citwkc” with the command “citwkc (41, 0, 40, 8, 5, 1, 24, 181, 0.9, 0.1)” it holds that $n = 435$. Therefore, to the sample pilot of 300 individuals we must add 135 more. Once the new sample has been taken, it is necessary to check that the precision $\phi = 0.10$ is verified.

8. DISCUSSION

The weighted kappa coefficient of a BDT is a measure of the beyond-chance agreement between the BDT and the GS, and depends on the sensitivity and specificity of the BDT, on the disease prevalence and on the weighting index. The weighted kappa coefficient is a parameter that is used to assess and compare the performance of BDTs. In this article, we have studied the comparison of the weighted kappa coefficients of two BDTs through confidence intervals when the sample design is paired. Three intervals have been studied for the difference of the two weighted kappa coefficients and five more intervals for the ratio of the two parameters. All the intervals studied are asymptotic and simulation experiments have been carried out to study their coverage probabilities and average lengths subject to different scenarios and for different sample sizes. Based on the results of the simulation experiments, some general rules of application have been given. When the sample size is moderate ($n = 100$) or large ($n = 200 - 400$) it is preferable to compare the two weighted

kappa coefficients through an interval for the ratio, and when the sample size is very large ($n \geq 500$) the two weighted kappa coefficients can be compared through the difference or the ratio. When the sample size is small ($n \leq 50$), the interval with the best behaviour is the Wald CI for the ratio θ adding 0.5 to all of the observed frequencies. Adding 0.5 to all of the frequencies does not improve the behaviour of the intervals for the difference δ , since these continue to fail when they failed without adding the value 0.5. This question may be due to the fact that the ratio $\hat{\theta}$ converges more quickly to the normal distribution than the difference $\hat{\delta}$. In the simulation experiments, the asymptotic behaviour of the Bayesian CIs has been studied using the Beta(1,1) distribution as prior distribution for all of the parameters. The choice of the values of the hyperparameters of the Beta distribution will depend on the previous information that the researcher has. If the researcher has some information and wants this information to have some weight in the data, then it is possible to use higher values of α and β , i.e. considering a Beta(α, β) distribution with $\alpha, \beta > 1$. The increase in α and β adds information and decreases the variance and, therefore, there is less uncertainty about the parameter. If the researcher does not want this information to have a great weight in the posteriori distribution, then the researcher chooses moderate values of α and β which are consistent with the information available, i.e. the average should be compatible with that information. To assess the effect that the Beta distribution has on the asymptotic behaviour of the Bayesian interval, we have carried out simulations (in a similar way to those carried out in Section 4) using as prior the distributions Beta(5,5) and Beta(25,25) for the Bayesian interval for $\theta = \frac{\kappa_1(c)}{\kappa_2(c)}$. These two distributions have the same average as the Beta(1,1) distribution but different variances. The first distribution has a moderate weight in the subsequent distribution and the second has an important weight. In general terms, the results obtained with the distribution Beta(5,5) are very similar to those obtained with the Beta(1,1) distribution. Regarding the Beta(25,25) distribution, there is no important difference in relation to the CPs obtained with the Beta(1,1), although for $\theta = \{0.25, 0.50\}$ the AL is slightly lower with the Beta(25,25), and when $\theta = \{0.75, 1\}$ the AL is slightly higher with the Beta(25,25). In general terms, when the Bayesian interval fails using the Beta(1,1) distribution then it also fails using the Beta(5,5) and the Beta(25,25). Furthermore, the Bayesian CI for $\theta = \kappa_1(c)/\kappa_2(c)$ with the Beta(5,5) and Beta(25,25), respectively, does not display a better CP than the Wald CI (when it does not fail), and therefore the Bayesian CI does not improve the asymptotic behaviour of the Wald CI. The application of the CIs requires the marginal frequencies s and r to be higher than zero. If the marginal frequency s (or r) is equal to zero, then it is not possible to estimate the weighted kappa coefficient of each BDT. Moreover, if a marginal frequency $s_{ij} + r_{ij}$ is equal to zero, then it is possible to calculate all of the CIs proposed; but not if two of these marginal frequencies are equal to zero. In this last situation, one of the weighted kappa coefficients (or both) is equal to zero, and the variance and the covariance are also equal to zero. If $s_{10} + r_{10} = s_{01} + r_{01} = 0$ then $\hat{\kappa}_1(c) = \hat{\kappa}_2(c)$ and $\widehat{\text{Var}}[\hat{\kappa}_1(c)] = \widehat{\text{Var}}[\hat{\kappa}_2(c)] = \text{Cov}[\hat{\kappa}_1(c), \hat{\kappa}_2(c)]$, and the frequentist intervals cannot be calculated. A solution to this problem is to add 0.5 to each observed frequency.

In this article, we have also proposed a method to calculate the sample size to estimate the ratio between the two weighted kappa coefficients with a determined precision and confidence. This method, based on the Wald CI for the ratio, is an iterative method, which starting from a pilot sample adds individuals to the sample until the CI has the set precision. From the initial sample we estimate a vector of parameters and in the second stage we calculate the sample size. Furthermore, the simulation experiments carried out to study

the robustness of the method to calculate the sample size have shown that the method has practical validity and the choice of the pilot sample has very little effect on this method.

When the two diagnostic tests are continuous, for each cut off point of each estimated ROC curve there will be a value of $\hat{S}e_h$ and of \widehat{FPF}_h (and therefore of $\hat{S}p_h = 1 - \widehat{FPF}_h$), with $h = 1, 2$. Once the clinician has set the value of the weighting index, $\hat{\kappa}_1(c)$ and $\hat{\kappa}_2(c)$ are calculated and therefore the CIs studied in Section 3 can be applied.

9. SUPPLEMENTARY MATERIAL

Appendices A, B and C are available as supplementary material of the manuscript in the URL:

<https://www.ugr.es/local/bioest/software/cmd.php?seccion=mb>

ACKNOWLEDGMENTS


This research was supported by the Spanish Ministry of Economy, Grant Number MTM2016-76938-P. We thank the referee, the Associate Editor, the Editor (Maria I. Fraga) and the Co-Editor (Giovani L. Silva) of REVSTAT Statistical Journal for their helpful comments that improved the quality of the paper.



REFERENCES

- [1] AGRESTI, A. (2002). *Categorical Data Analysis*, Wiley, New York.
- [2] BATWALA, V.; MAGNUSSEN, P. and NUWABA, F. (2010). Are rapid diagnostic tests more accurate in diagnosis of plasmodium falciparum malaria compared to microscopy at rural health centers?, *Malaria Journal*, **9**, 349.
- [3] BLOCH, D.A. (1997). Comparing two diagnostic tests against the same “gold standard” in the same sample, *Biometrics*, **53**, 73–85.
- [4] CICHETTI, D.V. (2001). The precision of reliability and validity estimates re-visited: distinguishing between clinical and statistical significance of sample size requirements, *Journal of Clinical and Experimental Neuropsychology*, **23**, 695–700.
- [5] EFRON, B. and TIBSHIRANI, R.J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York.
- [6] FIELLER, E.C. (1940). The biological standardization of insulin, *Journal of the Royal Statistical Society*, **7**, 1–64.
- [7] KRAEMER, H.C. and BLOCH, D.A. (1990). A note on case-control sampling to estimate kappa coefficients, *Biometrics*, **46**, 49–59.

- [8] KRAEMER, H.C. (1992). *Evaluating medical tests. Objective and quantitative guidelines*, Sage Publications, Newbury Park.
- [9] KRAEMER, H.C.; PERIYAKOIL, V.S. and NODA, A. (2002). Kappa coefficients in medical research, *Statistics in Medicine*, **2**, 2109–2129.
- [10] MARTÍN-ANDRÉS, A. and ALVAREZ-HERNÁNDEZ, M. (2014a). Two-tailed asymptotic inferences for a proportion, *Journal of Applied Statistics*, **41**, 1516–1529.
- [11] MARTÍN-ANDRÉS, A. and ALVAREZ-HERNÁNDEZ, M. (2014b). Two-tailed approximate confidence intervals for the ratio of proportions, *Statistics and Computing*, **24**, 65–75.
- [12] MONTERO-ALONSO, M.A. and ROLDÁN-NOFUENTES, J.A. (2019). Approximate confidence intervals for the likelihood ratios of a binary diagnostic test in the presence of partial disease verification, *Journal of Biopharmaceutical Statistics*, **29**, 56–81.
- [13] PEPE, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, New York.
- [14] PRICE, R.M. and BONETT, D.G. (2004). An improved confidence interval for a linear function of binomial proportions, *Computational Statistics and Data Analysis*, **45**, 449–456.
- [15] ROLDÁN-NOFUENTES, J.A.; LUNA DEL CASTILLO, J.D. and MONTERO-ALONSO, M.A. (2009). Confidence intervals of weighted kappa coefficient of a binary diagnostic test, *Communications in Statistics – Simulation and Computation*, **38**, 1562–1578.
- [16] ROLDÁN-NOFUENTES, J.A. and AMRO, R. (2018). Combination of the weighted kappa coefficients of two binary diagnostic tests, *Journal of Biopharmaceutical Statistics*, **28**, 909–926.
- [17] VACEK, P.M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests, *Biometrics*, **41**, 959–968.

On Construction of Bernstein-Bézier Type Bivariate Archimedean Copula

Authors: SELIM ORHUN SUSAM 
– Department of Statistics, Dokuz Eylul University,
Izmir, Turkey
orhun.susam@deu.edu.tr

BURCU HUDAVERDI  
– Department of Statistics, Dokuz Eylul University,
Izmir, Turkey
burcu.hudaverdi@deu.edu.tr

Received: February 2019

Revised: September 2019

Accepted: April 2020

Abstract:

- In this paper, a new class of bivariate multi-parameter Archimedean copula based on Kendall distribution using Bernstein-Bézier polynomials is introduced. The new class copula has flexible dependence properties depending on the polynomial degree and the control points. Some dependence characteristics such as Kendall's tau, upper tail and lower tail dependence of the new Archimedean copula class are derived. The simulation procedure based on these desired dependence characteristics is presented. Also, a parameter estimation process based on minimum Cramér-von-Mises distance is also given and its estimation performance is investigated through Monte Carlo simulation study.

Keywords:

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

AMS Subject Classification:

- 62G05, 60E05.

1. INTRODUCTION

Copula models are popular tools for describing multivariate data where the univariate distribution functions are combined to joint distribution function by Sklar's theorem (Sklar, 1959 [13]). Let X and Y be random variables with joint distribution function H and the marginal distribution functions F and G , respectively. Then, there exists a copula C such that $H(x, y) = C(F(x), G(y))$, for all x, y in \mathbb{R} . As an advantage of the copula models, the dependence structure can be modelled separately from the marginal distributions. If F and G are continuous, then C is unique. Otherwise, the copula C is uniquely determined on $\text{Ran}(F) \times \text{Ran}(G)$. There are various families of copulas. One of the most popular families is Archimedean copula family of which the dependence structure can be characterized by an univariate distribution function (Nelsen, 2006 [12], Section 4). The important feature that separates this class from the others is that it has a generator function φ which is used to construct an Archimedean copula.

Definition 1.1. A generator function φ is a continuous, strictly decreasing convex function defined from \mathbf{I} to $[0, \infty)$ such that $\varphi(1) = 0$. If $\varphi(0) = \infty$, then the generator is called as a strict generator. The pseudo inverse of φ is the function $\varphi^{[-1]}$, defined on $[0, \infty)$ to \mathbf{I} is given by

$$\varphi^{[-1]} = \begin{cases} \varphi^{-1}(t), & 0 \leq t \leq \varphi(0), \\ 0, & \varphi(0) \leq t < \infty. \end{cases}$$

A bivariate Archimedean copula with generator function $\varphi, C : \mathbf{I}^2 \rightarrow \mathbf{I}$ is defined by

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

where $u = F(x)$ and $v = G(y)$.

An Archimedean copula function can be reduced to an univariate distribution function through generator function. Genest *et al.* (1993) [8] showed that the function $\varphi(t)$ can be obtained by the univariate distribution function $K(t) = Pr(C(u, v) \leq t)$. Remarkably, there is a link between the function $\varphi(t)$ and $K(t)$ such as

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

$K(t)$ called as Kendall distribution function identifies the generator function $\varphi(t)$ and so the dependence structure of the Archimedean copula family. Dependence measures such as Kendall's tau, upper and lower tail dependence coefficients can be obtained by using Kendall distribution function. For a bivariate Archimedean copula with Kendall distribution function $K(t)$, Genest and MacKay (1986) [7] defined Kendall's Tau (τ) as

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

And also, Michiels *et al.* (2011) [10] defined lower λ_L and upper λ_U tail dependence as

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

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

and they investigated a general method for constructing bivariate Archimedean copula families using λ function. They worked with polynomials to construct multi-parameter copula families. Genest *et al.* (1998) [9] proposed several ways to generate bivariate Archimedean copula models via smooth transformations of existing generator function. Dimitrova *et al.* (2008) [4] defined an estimation method of Kendall distribution using B-spline functions. In addition, they defined sufficient conditions for the B-spline estimator to possess the properties of the Kendall distribution function. So, the function can be considered as a proper Kendall distribution function and associated with the multivariate Archimedean copula. Cooray (2018) [3] introduced two-parameter strict Archimedean generator function based on Clayton copula. Najjari *et al.* (2014) [11] constructed a new generator function $\varphi(t)$ using hyperbolic functions as generators of Archimedean copulas. The majority of the papers proposed some methods based on generator function φ for constructing a new Archimedean family of copulas. In this study, we propose constructing a multi-parameter Archimedean copula using Kendall distribution function $K(t)$. We use Bernstein-Bézier polynomials to create the new Archimedean class. Kendall's tau, lower and upper tail dependence coefficients are also obtained according to the polynomial degree and the control points. This new multi-parameter Archimedean copula family is contributed to the expansion of the existing Archimedean copula family.

The contribution of this study is two fold: First, a new Archimedean copula class based on Bernstein-Bézier polynomial is proposed. Different values of Kendall's tau (negative or positive), lower and upper tail dependence coefficients can be obtained by changing the polynomial degree and the control points, so the proposed class has flexible dependence structure. It is possible to create a new distribution function which has desirable dependence characteristics. This is quite useful in power analysis of goodness-of-fit test statistic. Second, an algorithm is proposed to create different distributions with the same dependence level by changing the control points for polynomial degree. Also, an estimation process based on minimizing Cramér-von Mises distance is presented and a Monte Carlo simulation study is employed to measure the performance of the parameter estimates.

The rest of the paper is organized as follows. In Section 2, Bernstein-Bézier type Archimedean copula is given and some dependence characteristics are investigated. A simulation procedure of this new class for different polynomial degrees is given in Section 3. Parameter estimation procedure which is based on minimum Cramér-von-Mises measure is given and parameter estimates are obtained in Section 4. And the last section is devoted to the conclusion.

2. BERNSTEIN BÉZIER TYPE BIVARIATE ARCHIMEDEAN COPULA

A Kendall distribution function $K(t)$ should satisfy the following properties (1–4) described in Nelsen (2006) [12]:

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

Let $K(m, \alpha; t)$ be a Bernstein-Bézier type Kendall distribution function with polynomial degree m and control points α defined as

$$(2.1) \quad K(m, \alpha; t) = \sum_{k=0}^m \alpha_k B_{k,m}(t)$$

where $B_{k,m}(t) = \binom{m}{k} t^k (1-t)^{m-k}$ for $t \in [0, 1]$.

Lemma 2.1. *A Bernstein-Bézier type Kendall distribution function $K(m, \alpha; t)$ satisfies the properties (1-4) if the following constraints hold:*

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

Proof: $K(m, \alpha, t = 0,) = \sum_{k=0}^m \alpha_k B_{k,m}(t = 0) = 0$ holds since $\alpha_0 = 0$. Similarly, $K(m, \alpha, t = 1) = \sum_{k=0}^m \alpha_k B_{k,m}(t = 1) = 1$ holds since $\alpha_m = 1$.

Also, $K(m, \alpha, t)' = m \sum_{k=0}^{m-1} (\alpha_{k+1} - \alpha_k) P_{k,m-1}(t) \geq 0$. See, Duncan (2005) [5]. So, $\alpha_0 = 0 < \alpha_1 < \alpha_2 < \dots < \alpha_m = 1$.

If the Bézier control points $\alpha_k > \frac{k}{m}, k = 1, \dots, m - 1$ where $\alpha_k = k/m + \epsilon_k$, then,

$$\begin{aligned} K(m, \alpha, t) &= \sum_{k=0}^m \alpha_k \binom{m}{k} t^k (1-t)^{m-k} \\ &= \sum_{k=0}^m \left(\frac{k}{m} + \epsilon_k\right) \binom{m}{k} t^k (1-t)^{m-k} \\ &= \sum_{k=0}^m \left(\frac{k}{m}\right) \binom{m}{k} t^k (1-t)^{m-k} + \sum_{k=0}^m (\epsilon_k) \binom{m}{k} t^k (1-t)^{m-k} \\ &= t \sum_{k=1}^m \binom{m-1}{k-1} t^{k-1} (1-t)^{m-k} + \sum_{k=0}^m (\epsilon_k) \binom{m}{k} t^k (1-t)^{m-k} \\ &= t \sum_{p=0}^{m-1} t^p (1-t)^{m-p-1} \binom{m-1}{p} + \sum_{k=0}^m (\epsilon_k) \binom{m}{k} t^k (1-t)^{m-k} \\ &= t + \sum_{k=0}^m (\epsilon_k) \binom{m}{k} t^k (1-t)^{m-k} > t. \end{aligned} \quad \square$$

We also obtain Kendall's tau, lower and upper tail dependence of the Bernstein-Bézier type Archimedean copula class using the following lemmas.

Lemma 2.2. *Kendall's tau for Bernstein-Bézier type Archimedean copula is obtained as*

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

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

Proof: τ is easily derived from equation $\tau = 3 - 4 \int_0^1 K(t) dt$. □

Lemma 2.3. *The lower tail λ_L and the upper tail λ_U dependence for Bernstein-Bézier type Archimedean copula are obtained by*

$$\begin{aligned}\lambda_L &= 2^{1-m\alpha_1}, \\ \lambda_U &= 2 - 2^{1-m(1-\alpha_{m-1})}.\end{aligned}$$

Proof: λ_U and λ_L are easily derived from equation $\lambda_L = 2^{\lim_{t \rightarrow 0^+} (t-K(t))'}$, $\lambda_U = 2 - 2^{\lim_{t \rightarrow 1^-} (t-K(t))'}$. □

It is seen that λ_L and λ_U are affected by only the control points α_1 and α_{m-1} , respectively. We can create Bernstein-Bézier type Archimedean copula using λ_L and λ_U , setting up the control points α_1 and α_{m-1} .

The following inequalities given in the next lemma provide an information for proper selection of λ_U and λ_L .

Lemma 2.4. *Let λ_L and λ_U be lower and upper tail dependence of Bernstein-Bézier type Archimedean copula with polynomial degree m . Then,*

$$1 > \lambda_L > \frac{2^{2-m}}{2 - \lambda_U}$$

holds for all values of polynomial degree m .

Proof: It can be proved using the inequality $\alpha_1 < \alpha_{m-1}$. Also, $0 < \lambda_U, \lambda_L < 1$, see Charpentier and Segers (2009) [2]. □

Suppose that the parameters α_k are defined as $\alpha_k > \frac{k}{m}$ for $k = 1, \dots, m - 1$, then $K(m, \alpha; t) > t$. See, Lemma 2.1. Also, we note that if the control points are selected as $\alpha_k \rightarrow \frac{k}{m}$, then the dependence coefficients $(\tau, \lambda_U, \lambda_L)$ approximate 1. In other words, the Bernstein-Bézier type Archimedean copula approximates comonotonic dependence when the control points are closely distributed uniform.

The Bernstein-Bézier type Archimedean copula with higher degree can represent various dependence forms. However, they may have some disadvantages:

1. As the degree increases, the complexity and therefore the processing time increase;
2. Because of the complexity, the curves of higher degree are more sensitive to round off errors.

As opposed to these disadvantages, we can combine several Bernstein-Bézier type Kendall distribution functions, mostly of degree three and four. We note that the Bernstein-Bézier polynomials are invariant under barycentric combinations (Farin (2001) [6], p. 61).

So, we obtain the following Bernstein-Bezier type Archimedean copulas for $\theta \in [0, 1]$:

$$\begin{aligned} K(m, \alpha; t) &= \sum_{k=0}^m (\theta\alpha_{1,k} + (1-\theta)\alpha_{2,k}) B_{k,m}(t) \\ &= \theta \sum_{k=0}^m \alpha_{1,k} B_{k,m}(t) + (1-\theta) \sum_{k=0}^m \alpha_{2,k} B_{k,m}(t) \\ &= \theta K(m, \alpha_{1,\cdot}; t) + (1-\theta) K(m, \alpha_{2,\cdot}; t). \end{aligned}$$

We can construct the weighted average of two Bernstein-Bézier Archimedean copulas either by taking the weighted average of corresponding points on the distribution, or by taking the weighted average of corresponding parameters α .

Dependence coefficients of two barycentric combinations of Bernstein-Bézier type Archimedean copula are given by

$$\begin{aligned} \tau &= 3 - 4 \sum_{k=0}^m \alpha_{2,k} \beta(k+1, m-k+1) \binom{m}{k} \\ &\quad + 4\theta \left(\sum_{k=0}^m (\alpha_{2,k} - \alpha_{1,k}) \beta(k+1, m-k+1) \binom{m}{k} \right), \\ \lambda_U &= 2 - 2^{1+\theta m \alpha_{1,m-1} + (1-\theta) m \alpha_{2,m-1-m}}, \\ \lambda_L &= 2^{1 - (\theta m \alpha_{1,1} + (1-\theta) m \alpha_{2,1})}. \end{aligned}$$

Note that if θ is selected as 1, then the classical Bernstein-Bézier type Archimedean copula is obtained.

3. SIMULATING DATA FROM BERNSTEIN BÉZIER TYPE ARCHIMEDEAN COPULA

In this section, data simulation from Bernstein-Bézier type Archimedean copula is given. Construction of a new distribution function which has desirable Kendall's tau and tail dependence coefficients are investigated.

The following procedure is used to create a distribution with the dependence characteristics represented by Kendall's tau and tail dependence coefficients:

1. The arbitrary value of the upper tail dependence λ_U is determined primarily.
2. λ_L is determined arbitrarily by using Lemma 2.4.
3. The value of Kendall's tau τ is determined for the distributions with polynomial degrees 2 and 3. For the distributions having polynomial degree $m \geq 4$, an interval of Kendall's tau is determined. Then, Kendall's tau is selected arbitrarily from this interval.
4. Bivariate data is simulated using the following algorithm. See, Nelsen (2006) [12].

The algorithm based on Michiels *et al.* (2011) [10] allows one to simulate $C(u, v)$ by Kendall distribution function $K(t)$ given as:

- Simulate uniformly distributed random pair (s, t) on $[0, 1]$.
- Set $w = K^{-1}(t)$.
- Set u such that $\int_w^u \frac{1}{t-K(t)} dt - \ln(s) = 0$.
- Set v such that $\int_w^v \frac{1}{t-K(t)} dt - \ln(1 - s) = 0$.

The range of the parameters and the dependence coefficients depending on the Bernstein-Bézier polynomial degree m are summarized in Table 1. It is observed that as the degree of the polynomial increases, the range of the dependence coefficients gets wider.

Table 1: Range of parameters and dependence coefficients.

m	α_0	α_1	α_2	α_3	α_4	α_5	τ	λ_U	λ_L
3	0	$(\frac{1}{3}, 1)$	$(\max(\frac{2}{3}, \alpha_1), 1)$	1	—	—	$(0, 1)$	$(0, 1)$	$(\frac{1}{4}, 1)$
4	0	$(\frac{1}{4}, 1)$	$(\max(\frac{3}{4}, \alpha_1), 1)$	$(\max(\frac{3}{4}, \alpha_2), 1)$	1	—	$(-0.2, 1)$	$(0, 1)$	$(\frac{1}{8}, 1)$
5	0	$(\frac{1}{5}, 1)$	$(\max(\frac{4}{5}, \alpha_1), 1)$	$(\max(\frac{3}{5}, \alpha_2), 1)$	$(\max(\frac{4}{5}, \alpha_3), 1)$	1	$(-0.33, 1)$	$(0, 1)$	$(\frac{1}{16}, 1)$

Kendall’s tau, upper and lower tail dependence coefficients obtained by the Bernstein-Bézier type Archimedean copula with control points for degree $(m = 3, 4, 5)$ are summarized in Table 2. Also, different distributions having the same dependence level at the control points α_2 and α_3 for polynomial degree 5 are given. All the Bernstein-Bézier control points and dependence coefficients are obtained by applying the simulation procedure (1–4). All cases in Table 2 are examined in the Subsections 3.1–3.3.

Table 2: Parameters and dependence coefficients.

Degree	$K(t)$	α_0	α_1	α_2	α_3	α_4	α_5	τ	λ_U	λ_L
$m = 3$	K_1	0	0.7173	0.7928	1	—	—	0.4899	0.7	0.45
$m = 4$	K_2	0	0.3537	0.5828	0.9815	1	—	0.68	0.1	0.75
$m = 5$	K_3	0	0.4	0.43	0.8531	0.9169	1	0.6	0.5	0.5
	K_4	0	0.4	0.63	0.6531	0.9169	1	0.6	0.5	0.5

3.1. Bernstein-Bézier type Archimedean copula with degree three

A Bernstein-Bézier type Archimedean copula with degree 3 has the following distribution function,

$$K(m = 3, \alpha; t) = \sum_{k=0}^3 \alpha_k \binom{3}{k} t^k (1 - t)^{3-k}, t \in [0, 1].$$

From Lemma 2.1, $\alpha_0 = 0, \alpha_3 = 1$, $\alpha_0 < \alpha_1 < \alpha_2 < \alpha_3$ and $\alpha_1 > \frac{1}{3}, \alpha_2 > \frac{2}{3}$. Kendall's tau of the distribution is given as

$$\tau = 3 - 4 \sum_{k=0}^3 \alpha_k \binom{3}{k} \beta(k+1, 3-k+1) = 2 - \alpha_1 - \alpha_2$$

and lower and upper tail dependence coefficients are

$$\lambda_L = 2^{1-3\alpha_1}, \lambda_U = 2 - 2^{3\alpha_2-2}.$$

(1–4) procedure is applied to determine the Kendall's tau and the tail dependence coefficients of the distribution. The arbitrary value of the upper tail dependence λ_U is determined primarily in the range $\lambda_U \in (0,1)$. We select λ_U as 0.7, so α_2 is equal to 0.7928. From Lemma 2.4, $1 > \lambda_L > 0.3846$. Then, λ_L is determined arbitrarily as 0.45. So, α_1 is equal to 0.7173. The stage conditions for control points given Lemma 2.1 are satisfied. Finally, Kendall's tau is 0.4899. $K(3, \alpha; t)$ with control points $\alpha_0 = 0, \alpha_1 = 0.7173, \alpha_2 = 0.7928$ and $\alpha_3 = 1$ has the Kendall's tau value as $\tau = 0.4899$ and the value tail dependence coefficients as $\lambda_L = 0.45$ and $\lambda_U = 0.7$. Simulated data and $K(m = 3, \alpha; t)$ with the sample of size 150 are visualized in Figure 1.

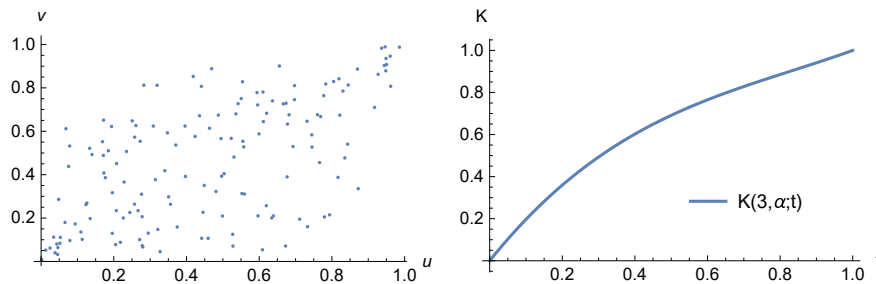


Figure 1: Simulated data from $K(3, \alpha; t)$ with $\tau = 0.4899, \lambda_L = 0.45, \lambda_U = 0.7$.

3.2. Bernstein-Bézier type Archimedean copula with degree four

Bernstein-Bézier type Archimedean copula with degree 4 has the following distribution function with the dependence characteristics, Kendall's tau, lower and upper tail dependence:

$$K(4, \alpha; t) = \sum_{k=0}^4 \alpha_k \binom{4}{k} t^k (1-t)^{4-k}, t \in [0, 1],$$

$$\tau = \frac{1}{5} \left(11 - 4(\alpha_1 + \alpha_2 + \alpha_3) \right),$$

$$\lambda_L = 2^{1-4\alpha_1}, \lambda_U = 2 - 2^{4\alpha_3-3}.$$

(1–4) procedure is applied to determine the Kendall’s tau and the tail dependence values of the distribution. The arbitrary value of the upper tail dependence λ_U is determined primarily in range $\lambda_U \in (0, 1)$. We select λ_U as 0.1 and so α_3 is equal to 0.9815. From Lemma 2.4, $1 > \lambda_L > 0.1315$. Then, λ_L is determined arbitrarily as 0.75. So, α_1 is equal to 0.3537. Finally from Lemma 2.1, Kendall’s tau should be selected in the range $\tau \in (0.3610, 0.7462)$. We determine Kendall’s tau arbitrarily as 0.68. So, α_2 is 0.5828. $K(4, \alpha; t)$ with control points $\alpha_0 = 0, \alpha_1 = 0.3537, \alpha_2 = 0.5828, \alpha_3 = 0.9815$ and $\alpha_4 = 1$ has the value of Kendall’s tau $\tau = 0.68$ and the values of tail dependences as $\lambda_L = 0.75$ and $\lambda_U = 0.1$. Simulated data and $K(m = 4, \alpha; t)$ with the sample of size 150 is visualized in Figure 2.

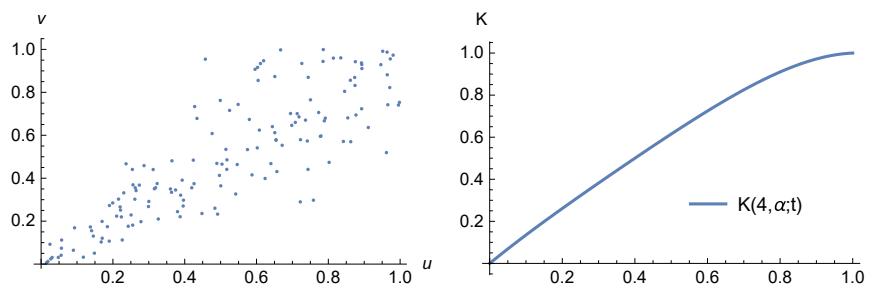


Figure 2: Simulated data from $K(4, \alpha; t)$ with $\tau = 0.68, \lambda_L = 0.75, \lambda_U = 0.1$.

3.3. Bernstein-Bézier type Archimedean copula with degree five

Bernstein-Bézier type Archimedean copula with degree 5 has the following distribution function with the dependence characteristics Kendall’s tau, lower and upper tail dependence,

$$K(5, \alpha; t) = \sum_{k=0}^5 \alpha_k \binom{5}{k} t^k (1-t)^{5-k}, t \in [0, 1],$$

$$\tau = \frac{1}{3} \left(7 - 2(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) \right),$$

$$\lambda_L = 2^{1-5\alpha_1}, \lambda_U = 2 - 2^{5\alpha_4-4}.$$

(1–4) procedure is again applied to determine the Kendall’s tau and the tail dependence values of the distribution. The arbitrary value of the upper tail dependence λ_U is determined primarily in range $\lambda_U \in (0, 1)$. We select λ_U as 0.5 and so α_4 is equal to 0.9169. From Lemma 2.4, $1 > \lambda_L > 0.0833$. Then, λ_L is determined arbitrarily as 0.5. So, α_1 is equal to 0.4. Finally from Lemma 2.1, Kendall’s tau should be selected in the range $\tau \in (0.2328, 0.6220)$. We determine Kendall’s tau arbitrarily as 0.6. α_2 and α_3 can be derived from solving equations $\alpha_2 + \alpha_3 = 1.2831$. From the last equation and Lemma 2.1, α_2 and α_3 should be selected in the range $\alpha_2 \in (0.4, 0.6415)$ and $\alpha_3 \in (0.6415, 0.8831)$, respectively. Different α_2 and α_3 values can be selected in order to provide $\alpha_2 + \alpha_3 = 1.2831$ in the range of α_2 and α_3 . This case is important, because we can create different distributions with the same dependence level by selecting different α_2 and α_3 values. One possible selection is $\alpha_2 = 0.43$ and $\alpha_3 = 0.8531$.

Another possible selection is $\alpha_2 = 0.63$ and $\alpha_3 = 0.6531$. $K_1(5, \alpha; t)$ with control points $\alpha_0 = 0, \alpha_1 = 0.4, \alpha_2 = 0.43, \alpha_3 = 0.8531, \alpha_4 = 0.9169, \alpha_5 = 1$ and $K_2(5, \alpha; t)$ with control points $\alpha_0 = 0, \alpha_1 = 0.4, \alpha_2 = 0.63, \alpha_3 = 0.6531, \alpha_4 = 0.9169, \alpha_5 = 1$ with the same dependence level are visualized in Figure 3.

For the higher order polynomial degree, for example $m = 6$, the range of τ, λ_L and λ_U are determined as the same as for degree $m < 6$. But the range of α_2, α_3 and α_4 for the solutions of $\alpha_2 + \alpha_3 + \alpha_4 = a$ cannot be determined easily.

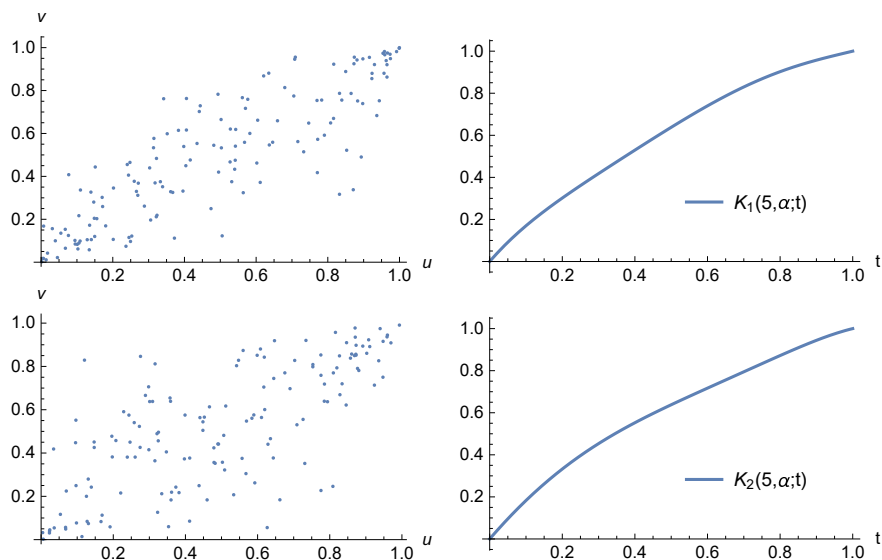


Figure 3: Simulated data from $K_1(5, \alpha; t)$ and $K_2(5, \alpha; t)$ with the same $\tau = 0.6, \lambda_L = 0.5, \lambda_U = 0.5$.

4. PARAMETER ESTIMATION BASED ON CRAMÉR-VON-MISES MEASURE

Genest and Rivest (1993) [8] proposed a nonparametric procedure using empirical estimate K_n of K . The psuedo observations of \hat{T}_i were obtained by

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

Then, $K(t)$ was estimated by the empirical distribution function as

$$(4.1) \quad \hat{K}_n(t) = \sum_{i=1}^n (\hat{T}_i \leq t) / n.$$

Barbe *et al.* (1996) [1] investigated consistency of $\hat{K}_n(t)$. Alternatively, Susam and Ucer (2018) [14] defined the empirical Bernstein estimator of order ($m_1 > 0$) for the Kendall distribution function as

$$(4.2) \quad \hat{K}_{m_1,n}(t) = \sum_{k=0}^{m_1} \hat{K}_n(k/m_1)P_{k,m_1}(t),$$

where $P_{k,m_1}(t) = \binom{m_1}{k}t^k(1-t)^{m_1-k}$ is the binomial probability. Also, they showed that the Bernstein Kendall distribution function outperforms the empirical Kendall distribution function according to its performance by Monte Carlo simulation study.

In this study, through the parameter estimation process, we first estimate the Bernstein-Bézier type Archimedean copula parameters by using empirical estimate of \hat{K}_n . Then, Cramér-von-Mises (CvM) distance between the empirical Kendall distribution function and the Bernstein-Bézier type Kendall distribution function is obtained as

$$\begin{aligned} CvM_{\hat{K}_n} &= \int_0^1 n(\hat{K}_n(t) - K(\alpha, m_2; t))^2 d\hat{K}_n(t) \\ &= \frac{1}{n} \sum_{i=1}^n (\hat{K}_n(\hat{T}_i) - K(\alpha, m_2; \hat{T}_i))^2. \end{aligned}$$

Then the parameters are estimated by

$$\hat{\alpha}_{\hat{K}_n} = \operatorname{argmin}_{\alpha \in \Theta} \{CvM_{\hat{K}_n}\}$$

where $\Theta = \{\alpha_k > \frac{k}{m_2}, \alpha_{k+1} > \alpha_k ; k = 1, \dots, m_2 - 1\}$ and $\alpha_0 = 0, \alpha_{m_2} = 1$.

Secondly, the Bernstein-Bézier type Archimedean copula parameters are estimated by using empirical Bernstein estimator $\hat{K}_{m_1,n}(t)$. Since the empirical Bernstein Kendall distribution function is a continuous approximation of the empirical Kendall distribution function \hat{K}_n , we use empirical Bernstein Kendall distribution function which is upgraded version of \hat{K}_n to obtain Cramér-von-Mises (CvM) distance as

$$(4.3) \quad CvM_{\hat{K}_{n,m}} = \int_0^1 n(\hat{K}_{n,m_1}(t) - K(\alpha, m_2; t))^2 dt.$$

The estimation of the dependence parameter α_i for $i = 0, \dots, m_2$ can be selected as the value that minimizes the CvM distance.

Lemma 4.1. *Let $K(\alpha, m_2; t)$ be the Bernstein-Bézier type Kendall distribution function with order ($m_2 > 0$) and let $\hat{K}_{m,n}(t)$ be the empirical Bernstein estimator of Kendall*

distribution function with order ($m_1 > 0$). Then the Cramér-von-Mises distance is defined as

$$\begin{aligned}
CvM &= n \sum_{k=0}^{m_1} \binom{m_1}{k}^2 \hat{K}_n^2\left(\frac{k}{m_1}\right) \beta(2k+1, 2m_1-2k+1) \\
&+ 2n \sum_{k=0}^{m_1-1} \sum_{s=k+1}^{m_1} \binom{m_1}{k} \binom{m_1}{s} \hat{K}_n\left(\frac{k}{m_1}\right) \hat{K}_n\left(\frac{s}{m_1}\right) \beta(k+s+1, 2m_1-k-s+1) \\
&+ n \sum_{k=0}^{m_2} \binom{m_2}{k}^2 \alpha_k^2 \beta(2k+1, 2m_2-2k+1) \\
&+ 2n \sum_{k=0}^{m_2-1} \sum_{s=k+1}^{m_2} \binom{m_2}{k} \binom{m_2}{s} \alpha_k \alpha_s \beta(k+s+1, 2m_2-k-s+1) \\
&- 2n \sum_{k=0}^{m_1} \sum_{s=0}^{m_2} \hat{K}_n\left(\frac{k}{m_1}\right) \alpha_s \binom{m_1}{k} \binom{m_2}{s} \beta(k+s+1, m_1+m_2-k-s+1)
\end{aligned}$$

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

Proof:

$$\begin{aligned}
CvM &= \int_0^1 (\hat{K}_{n,m_1}(t) - K(\alpha, m_2; t))^2 dt \\
&= n \int_0^1 \hat{K}_{n,m_1}^2(t) dt + n \int_0^1 (K(\alpha, m_2; t))^2 dt - 2n \int_0^1 \hat{K}_{n,m_1}(t) K(\alpha, m_2; t) dt \\
&= n \int_0^1 \left(\sum_{k=0}^{m_1} \binom{m_1}{k} t^k (1-t)^{m_1-k} \hat{K}_n\left(\frac{k}{m_1}\right) \right)^2 dt \\
&+ n \int_0^1 \left(\sum_{k=0}^{m_2} \alpha_k t^k \binom{m_2}{k} t^k (1-t)^{m_2-k} \right)^2 dt \\
&- 2n \sum_{k=0}^{m_1} \sum_{s=0}^{m_2} \hat{K}_n\left(\frac{k}{m_1}\right) \alpha_s \binom{m_1}{k} \binom{m_2}{s} \int_0^1 t^{k+s} (1-t)^{m_1+m_2-k-s} dt \\
&= I_1 + I_2 - I_3.
\end{aligned}$$

Now we calculate part of I_1 . We know that $(a_1 + a_2 + \dots + a_n)^2 = \sum_{i=1}^n a_i^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n a_i a_j$, then we can write

$$\begin{aligned}
I_1 &= n \sum_{k=0}^{m_1} \binom{m_1}{k}^2 \hat{K}_n^2\left(\frac{k}{m_1}\right) \int_0^1 t^{2k} (1-t)^{2m_1-2k} dt \\
&+ 2 \sum_{k=0}^{m_1-1} \sum_{s=k+1}^{m_1} \binom{m_1}{k} \hat{K}_n\left(\frac{k}{m_1}\right) \binom{m_1}{s} \hat{K}_n\left(\frac{s}{m_1}\right) \int_0^1 t^{k+s} (1-t)^{2m_1-k-s} dt \\
&= n \sum_{k=0}^{m_1} \binom{m_1}{k}^2 \hat{K}_n^2\left(\frac{k}{m_1}\right) \beta(2k+1, 2m_1-2k+1) \\
&+ 2n \sum_{k=0}^{m_1-1} \sum_{s=k+1}^{m_1} \binom{m_1}{k} \hat{K}_n\left(\frac{k}{m_1}\right) \binom{m_1}{s} \hat{K}_n\left(\frac{s}{m_1}\right) \beta(k+s+1, 2m_1-k-s+1).
\end{aligned}$$

Proof of the parts of I_2 and I_3 are the same as proof of part I_1 . □

Then, the parameter estimate which gives the minimum value of Cramér-von-Mises distance based on Bernstein empirical distribution is defined for Bernstein-Bézier type Archimedean copula by

$$\hat{\alpha}_{\hat{K}_{n,m}} = \operatorname{argmin}_{\alpha \in \Theta} \left\{ CvM_{\hat{K}_{n,m}} \right\}$$

where $\Theta = \left\{ \alpha_k > \frac{k}{m_2}, \alpha_{k+1} > \alpha_k ; k = 1, \dots, m_2 - 1 \right\}$ and $\alpha_0 = 0, \alpha_{m_2} = 1$.

Genest *et al.* (1993) [8] introduced a method-of-moment estimator for bivariate Archimedean copula based on empirical Kendall distribution function $\hat{K}_n(t)$. For one-parameter families, the parameter can be estimated by only using the first moment. However, for more than one parameters, we need the moments as much as the number of parameters.

We note that the estimation procedure explained in this section are not only available for Archimedean copulas but also available for all continuous copula classes. The empirical Kendall distribution function can also be used for all continuous copula classes. See Genest *et al.* (1993) [8].

A Monte Carlo simulation study is conducted to measure the performance of the estimation method with several values of Kendall’s tau, lower and upper tail dependence coefficients.

1,000 Monte Carlo samples of sizes $n = 50, 150$ are generated from each type of Bernstein-Bézier type Archimedean copulas given in Table 2 and investigated the performances of two parameter estimation methods as $\alpha_{\hat{K}_n}$ and $\alpha_{\hat{K}_{n,m}}$. For the empirical Bernstein estimator, we select the polynomial degree as $m_1 = 15$ for sample size $n = 50$ and $m_1 = 30$ for sample size $n = 150$.

Simulation results are shown in Table 3 and Table 4. When the results are examined, the minimum Cramér-von-Mises method based on Kendall distribution using Bernstein polynomials outperforms the method based on empirical Kendall distribution in almost all cases for all sample sizes.

Table 3: MSE of the parameter estimations for four Bernstein-Bézier type copula with sample size $n = 50$.

Dist.	Est. Mth.	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$
K_1	$\hat{\alpha}_{\hat{K}_n}$	0.00684	0.00431	—	—
	$\hat{\alpha}_{\hat{K}_{n,15}}$	0.00575	0.00313	—	—
K_2	$\hat{\alpha}_{\hat{K}_n}$	0.00903	0.01116	0.00221	—
	$\hat{\alpha}_{\hat{K}_{n,15}}$	0.00324	0.00688	0.00585	—
K_3	$\hat{\alpha}_{\hat{K}_n}$	0.00633	0.01580	0.01428	0.00349
	$\hat{\alpha}_{\hat{K}_{n,15}}$	0.00342	0.00925	0.01192	0.00193
K_4	$\hat{\alpha}_{\hat{K}_n}$	0.01544	0.00957	0.00992	0.00266
	$\hat{\alpha}_{\hat{K}_{n,15}}$	0.00534	0.01422	0.00923	0.00356

Table 4: MSE of the parameter estimations for four Bernstein-Bézier type copula with sample size $n = 150$.

Dist.	Est. Mth.	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$
K_1	$\hat{\alpha}_{\hat{K}_n}$	0.00261	0.00151	—	—
	$\hat{\alpha}_{\hat{K}_{n,30}}$	0.00303	0.00141	—	—
K_2	$\hat{\alpha}_{\hat{K}_n}$	0.00209	0.00437	0.00096	—
	$\hat{\alpha}_{\hat{K}_{n,30}}$	0.00123	0.00384	0.00177	—
K_3	$\hat{\alpha}_{\hat{K}_n}$	0.00177	0.00661	0.00827	0.00242
	$\hat{\alpha}_{\hat{K}_{n,30}}$	0.00229	0.00589	0.00614	0.00091
K_4	$\hat{\alpha}_{\hat{K}_n}$	0.00516	0.00775	0.00650	0.00144
	$\hat{\alpha}_{\hat{K}_{n,30}}$	0.00224	0.00753	0.00670	0.00165

5. CONCLUSION

In this study, we propose a new family of Archimedean copulas based on Kendall distribution function $K(t)$. We use Bernstein-Bézier polynomials to construct this new multi-parameter distribution. The method is illustrated for polynomial degree $m = 3, 4, 5$. There are several advantages of this new Archimedean copula class. It is shown that while working with the Bernstein-Bézier polynomial structures, a multi-parameter copula family can be constructed in an organized way. It is possible to create a new distribution function which has desirable dependence characteristics using Kendall's tau, lower and upper tail dependence. The parameters of the new model can be interpreted in terms of these dependence characteristics. And also, it is possible that we can create different distributions with the same dependence structures. Also, we obtain the parameter estimates minimizing the Cramér-von-Mises distance which is based on Bernstein-Bézier type Archimedean copulas. We measure the performance of the estimation method with several values of Kendall's tau, lower and upper tail dependence coefficients by a Monte Carlo simulation study. We can conclude that the minimum Cramér-von-Mises method based on Kendall distribution using Bernstein polynomials outperforms the method based on empirical Kendall distribution function.


ACKNOWLEDGMENTS



We thank the anonymous referees and the editor for their helpful suggestions which improved the presentation of the paper.

REFERENCES

- [1] BARBE, P.; GENEST, C.; GHOUDI, K. and RÉMILLARD, B. (1996). On Kendall's Process, *Journal of Multivariate Analysis*, **58**, 197–229.
- [2] CHARPENTIER, A. and SEGERS, J. (2009). Tails of multivariate Archimedean copulas. *Journal of Multivariate Analysis*, **100**, 1521–1537.
- [3] COORAY, K. (2018). Strictly Archimedean copulas with complete association for multivariate dependence based on the Clayton family, *Dependence Modelling*, **6**, 1–18.
- [4] DIMITROVA, D.; KAISHEV, K. and PENEV, I. (2008). GeD spline estimation of multivariate Archimedean copulas, *Computational Statistics and Data Analysis*, **52**, 3570–3582.
- [5] DUNCAN, M. (2005). *Applied Geometry for Computer Graphics and CAD*, Springer Verlag, London.
- [6] FARIN, G. (2001). *Curves and Surfaces for CAGD: A Practical Guide*, Morgan Kaufmann Publishers, San Francisco.
- [7] GENEST, C. and MACKAY, R.J. (1986). Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données, *The Canadian journal of Statistics*, **14**, 145–149.
- [8] GENEST, C. and RIVEST, L. (1993). Statistical inference procedures for bivariate Archimedean copulas, *Journal of the American Statistical Association*, **88**(423), 1034–1043.
- [9] GENEST, C.; GHOUDI, K. and RIVEST, L. (1998). Discussion to “Understanding Relationships Using Copulas” by E. Frees and E. Valdez, *North American Actuarial Journal*, **2**(1), 143–149.
- [10] MICHIELS, F.; KOCH, I. and DE SCHEPPER, A. (2011). A new method for the construction of bivariate Archimedean copulas based on the λ function, *Communications in Statistics – Theory and Methods*, **40**(15), 2670–2679.
- [11] NAJJARI, V.; BACIGAL, T. and BAL, H. (2014). An Archimedean copula family with hyperbolic cotangent generator, *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, **20**(5), 761–768.
- [12] NELSEN, R.B. (2006). *An Introduction to Copulas*, Springer Verlag, New York (NY).
- [13] SKLAR, A. (1959). Fonctions de répartition à n dimensions et leurs marges, *Publications de l'Institut de Statistique de l'Université de Paris*, **8**, 229–231.
- [14] SUSAM, S.O. and UCER HUDAVERDI, B. (2018). Testing independence for Archimedean copula based on Bernstein estimate of Kendall distribution function, *Journal of Statistical Computation and Simulation*, **88**(13), 2589–2599.

Wavelet Estimation of Regression Derivatives for Biased and Negatively Associated Data

Authors: JUNKE KOU 
– School of Mathematics and Computational Science,
Guilin University of Electronic Technology,
Guilin, P.R. China
kjkou@guet.edu.cn

CHRISTOPHE CHESNEAU  
– Laboratoire de Mathématiques Nicolas Oresme,
Université de Caen Normandie,
Caen, France
christophe.chesneau@gmail.com

Received: August 2019

Revised: April 2020

Accepted: April 2020

Abstract:

- This paper considers the estimation of the derivatives of a regression function based on biased data. The main feature of the study is to explore the case where the data comes from a negatively associated process. In this context, two different wavelet estimators are introduced: a linear wavelet estimator and a nonlinear wavelet estimator using the hard thresholding rule. Their theoretical performance is evaluated by determining sharp rates of convergence under L^p risk, assuming that the unknown function of interest belongs to a ball of Besov spaces $B_{p,q}^s(\mathbb{R})$. The obtained results extend some existing works on biased data in the independent case to the negatively associated case.

Keywords:

- *regression derivatives estimation; negatively associated; L^p risk; wavelets.*

AMS Subject Classification:

- 62G07, 62G20, 42C40.

1. INTRODUCTION

In this paper, the biased nonparametric regression model is considered. It is formulated as follows. Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ be identically distributed random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with the common density function

$$(1.1) \quad f(x, y) = \frac{\omega(x, y) g(x, y)}{\mu}, \quad (x, y) \in [0, 1] \times \mathbb{R},$$

where ω stands for a known positive function, g denotes the density function of the unobserved random variables (U, V) and $\mu := \mathbb{E}(\omega(X, Y)) < \infty$. In this setup g and f mean the target density and weighted density, respectively, and the resulting data are biased data. We want to estimate the d th derivative $r^{(d)}(x)$ of regression function

$$(1.2) \quad r(x) := \mathbb{E}(\rho(V) | U = x) = \int_{\mathbb{R}} \frac{\rho(y) g(x, y)}{h(x)} dy, \quad x \in [0, 1].$$

This above model arises in many applications. For example, in order to estimate the change rate of agricultural output V when the input U increase (decrease) in a country. We obtain data (X_i, Y_i) ($i = 1, 2, \dots, n$) from those regions where spend more in agriculture, then X_i and Y_i stands for the agricultural input and output. Because it is more likely to sample those special regions, the density f of (X_i, Y_i) satisfies $f(x, y) = \frac{\omega(x, y) g(x, y)}{\mu}$ with some weight function ω and the real density g of (U, V) . Then we can estimate the change rate $r^{(d)}$ of the country by the given data (X_i, Y_i) . Hence, the work about this regression estimation model is very important.

The former works have developed kernel or modified local polynomials estimators for the problem of estimating $r(x)$, i.e., $r^{(d)}(x)$ with $d = 0$. See, for instance, [1], [20], [10], [21], [11], [12] and [5]. In order to obtain theoretical results, as optimal rates of convergence, in a general statistical setting or to reach the goal of adaptivity, wavelet methods have been developed by [9], [4] and [6]. Always focusing on wavelet methods, the estimation of $r(x)$ for (strongly mixing) dependent $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ has been explored by [7], [8] and [17]. Also, for the prime goal, the estimation of the derivative $r^{(d)}(x)$ has been considered by [3] and [14], but only for independent $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$. More precisely, [3] provide an upper bound estimation over $L^p(\mathbb{R})$ ($1 \leq p < \infty$) risk for the derivative $r^{(d)}(x)$ of regression function with a linear wavelet estimator. Because this linear wavelet estimator is not adaptive, [14] construct a nonlinear wavelet estimator and study its convergence rate over $L^p(\mathbb{R})$ ($1 \leq p < \infty$) risk.

In this paper, we investigate a generalization of these works by considering the estimation of $r^{(d)}(x)$ from dependent $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$; the negatively associated case is considered. This kind of dependence naturally appear in many well-known multivariate distributions involved in a wide variety of applications. We refer to [2] and [16]. In this setting, a linear nonadaptive and nonlinear adaptive wavelet estimators are introduced. We determine their rates of convergence under the L^p risk with $1 \leq p < \infty$, assuming that $r^{(d)}(x)$ belongs to Besov spaces $B_{p,q}^s(\mathbb{R})$. We prove that, with mathematical efforts, the established results in the independent case can be transposed to the negatively associated case, showing the consistency of the wavelet methodology for this problem.

The rest of this paper is the following. The mathematical assumptions on the model are presented in Section 2. The necessary on the wavelets and Besov spaces are described in Section 3. The linear wavelet estimation is performed in Section 4. The nonlinear wavelet estimation is developed in Section 5. Some concluding remarks are postponed in Section 6.

2. ASUMPTIONS ON THE MODEL

In this section, we will introduce the definition and properties of negatively associated sample. In addition, some other assumptions for the model (1.1)–(1.2) are proposed.

Definition 2.1 ([2]). A sequence of random variable X_1, X_2, \dots, X_n is said to be negatively associated, if for each pair of disjoint nonempty subsets A and B of $\{i = 1, 2, \dots, n\}$,

$$\text{Cov}(f(X_i, i \in A), g(X_j, j \in B)) \leq 0,$$

where f and g are real-valued coordinate-wise nondecreasing functions and the corresponding covariances exist.

This definition can be extended to random vectors (see [16]). It is well known that $\text{Cov}(X_i, X_j) \equiv 0$ when the random variable X_1, X_2, \dots, X_n is independent. Hence, the independence case is a special case of negatively associated case. Also, let X_1, X_2, \dots, X_n be independent random variables with log concave densities. Then, if $\sum_{i=1}^n X_i = c$ (c is a constant), X_1, X_2, \dots, X_n are negatively associated.

For examples of negatively associated case, [16] showed that many well-known multivariate distributions possess the negatively associated property. Some examples include: the multinomial distribution, the multivariate hypergeometric distribution, the Dirichlet compound multinomial distribution, the permutation distribution and so on. Because of its wide application in multivariate statistical analysis and system reliability, many researches on negatively associated have already been considered, see, e.g., [19], [24], [18], [23]. In addition, an important property of negative association is given in the following lemma. It will be at the center of one of our main results.

Lemma 2.1 ([16]). Let X_1, X_2, \dots, X_n be a sequence of negatively associated random variables and B_1, B_2, \dots, B_m be some pairwise disjoint nonempty subsets of $\{i = 1, 2, \dots, n\}$. If f_i ($i = 1, 2, \dots, m$) are m coordinate-wise nondecreasing (nonincreasing) functions, then $f_1(X_i, i \in B_1), f_2(X_i, i \in B_2), \dots, f_m(X_i, i \in B_m)$ are also negatively associated.

In this paper, $A \lesssim B$ denotes $A \leq cB$ with a positive constant c which is independent of A and B ; $A \gtrsim B$ means $B \lesssim A$; $A \sim B$ stands for both $A \lesssim B$ and $B \lesssim A$.

For the problem (1.1)–(1.2), in addition to assume that $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ are negatively associated, we make the following other assumptions:

A1. The density function h of the random variable U is nonincreasing, and has a positive lower bound,

$$0 < c_1 \leq h(x), \quad x \in [0, 1].$$

A2. The weight function ω is coordinate-wise nonincreasing, and has both positive upper and lower bounds, i.e., for $(x, y) \in [0, 1] \times \mathbb{R}$,

$$\omega(x, y) \sim 1.$$

A3. The function ρ is known, nondecreasing and $\rho \in L^\infty(\mathbb{R})$.

A4. We have $r^{(u)}(0) = r^{(u)}(1) = 0$ for any $u \in \{0, \dots, d\}$.

A5. There exists a constant $c_2 > 0$ such that

$$\sup_{x \in [0,1]} |r^{(d)}(x)| \leq c_2.$$

These assumptions are quite standard for the considered problem (see [3] and [14]). Only those involving the non monotonicity of some functions are deeply link with the negatively associated dependence assumption. They will be used for technical purpose in the proofs.

3. WAVELETS AND BESOV SPACES

Throughout this paper, we work with the wavelet basis described below. A wavelet function ψ can be constructed from the scaling function ϕ in a simple way such that $\{2^{j/2}\psi(2^j x - k), j \in \mathbb{Z}, k \in \mathbb{Z}\}$ constitutes an orthonormal basis (wavelet basis) of $L^2(\mathbb{R})$. Then, each $f \in L^2(\mathbb{R})$,

$$f = \sum_{k \in \mathbb{Z}} \alpha_{j_0,k} \phi_{j_0,k} + \sum_{j=j_0}^{\infty} \sum_{k \in \mathbb{Z}} \beta_{j,k} \psi_{j,k}$$

holds in $L^2(\mathbb{R})$ sense, where $\alpha_{j_0,k} = \langle f, \phi_{j_0,k} \rangle$, $\beta_{j,k} = \langle f, \psi_{j,k} \rangle$ and

$$\phi_{j_0,k}(x) = 2^{\frac{j_0}{2}} \phi(2^{j_0} x - k), \quad \psi_{j,k}(x) = 2^{\frac{j}{2}} \psi(2^j x - k).$$

Let P_j be the orthogonal projection operator from $L^2(\mathbb{R})$ onto the space V_j with the orthonormal basis $\{\phi_{j,k}(\cdot) = 2^{j/2} \phi(2^j \cdot - k), k \in \mathbb{Z}\}$. Then, for $f \in L^2(\mathbb{R})$,

$$P_j f = \sum_{k \in \mathbb{Z}} \alpha_{j,k} \phi_{j,k}.$$

A scaling function ϕ is called m regular, if $\phi \in C^m(\mathbb{R})$ and $|D^\alpha \phi(x)| \leq c(1 + x^2)^{-l}$ for each $l \in \mathbb{Z}$ ($\alpha = 0, 1, \dots, m$). In this paper, we choose Daubechies scaling function D_{2N} . Then, ϕ is m regular when N gets large enough. Furthermore, it can be shown that, for $f \in L^p(\mathbb{R})$ ($1 \leq p < \infty$),

$$(3.1) \quad P_j f(x) = \sum_{k \in \mathbb{Z}} \alpha_{j,k} \phi_{j,k}(x)$$

holds almost everywhere on \mathbb{R} ([15]).

Lemma 3.1. *Let a scaling function $\phi \in L^2(\mathbb{R})$ satisfy m regular and $\{\alpha_k\} \in l_p$ ($1 \leq p \leq \infty$). Then*

$$\left\| \sum_{k \in \mathbb{Z}} \alpha_k 2^{\frac{j}{2}} \phi(2^j x - k) \right\|_p \sim 2^{j(\frac{1}{2} - \frac{1}{p})} \|(\alpha_k)\|_p.$$

The proof of lemma can be found in [15]. In addition, Lemma 3.1 holds if the scaling function ϕ is replaced by the corresponding wavelet ψ .

One advantage of wavelets is that it can characterize Besov spaces. Besov spaces are important in theory and applications, which contain Hölder and L^2 Sobolev spaces as special examples. The next lemma provides equivalent definition for Besov space.

Lemma 3.2. *Let ϕ be m regular, ψ be the corresponding wavelets and $f \in L^p(\mathbb{R})$. If $\alpha_{j,k} = \langle f, \phi_{j,k} \rangle$, $\beta_{j,k} = \langle f, \psi_{j,k} \rangle$, $p, q \in [1, \infty]$ and $0 < s < m$, then the following assertions are equivalent:*

- (1) $f \in B_{p,q}^s(\mathbb{R})$;
- (2) $\{2^{js} \|P_j f - f\|_p\} \in l_q$;
- (3) $\{2^{j(s - \frac{1}{p} + \frac{1}{2})} \|\beta_j\|_p\} \in l_q$.

The Besov norm of f can be defined by

$$(3.2) \quad \|f\|_{B_{p,q}^s} := \|(\alpha_{j_0})\|_p + \left\| \left(2^{j(s - \frac{1}{p} + \frac{1}{2})} \|\beta_j\|_p \right)_{j \geq j_0} \right\|_q,$$

where $\|\beta_j\|_p^p = \sum_{k \in \mathbb{Z}} |\beta_{j,k}|^p$.

In this paper, we will suppose the unknown function $r^{(d)}(x)$ belong to Besov balls $B_{p,q}^s(H)$ with $H > 0$, which means $f \in B_{p,q}^s(H) := \{f \in B_{p,q}^s(\mathbb{R}^d), \|f\|_{B_{p,q}^s} \leq H\}$.

4. LINEAR WAVELET ESTIMATION

This section will introduce a linear wavelet estimator and discuss its convergence rate over L^p ($1 \leq p < \infty$) risk. Now our linear wavelet estimator is defined by

$$(4.1) \quad \hat{r}_n^{(d)}(x) := \sum_{k \in \Omega} \hat{\alpha}_{j_0,k} \phi_{j_0,k}(x).$$

In this definition, we have set

$$(4.2) \quad \hat{\alpha}_{j_0,k} = (-1)^d \frac{\hat{\mu}_n}{n} \sum_{i=1}^n \frac{\rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \phi_{j_0,k}^{(d)}(X_i),$$

$$(4.3) \quad \hat{\mu}_n = \left[\frac{1}{n} \sum_{i=1}^n \frac{1}{\omega(X_i, Y_i)} \right]^{-1}$$

and $\Omega = \{k \in \mathbb{Z}, \text{supp } r^{(d)} \cap \text{supp } \phi_{j_0,k} \neq \emptyset\}$. Then, it follows from the compactly supported properties of the function $r^{(d)}$ and $\phi_{j_0,k}$ that the cardinality of Ω satisfies $|\Omega| \sim 2^{j_0}$.

On the other hand, some existing results on these estimators in the independent case remain true. Indeed, according to the [14, Lemma 2.1], under Condition A4, we know that

$$(4.4) \quad \mathbb{E} \left(\frac{1}{\widehat{\mu}_n} \right) = \frac{1}{\mu}$$

and

$$(4.5) \quad \mathbb{E} \left[(-1)^d \frac{\mu \rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \phi_{j_0, k}^{(d)}(X_i) \right] = \alpha_{j_0, k}.$$

These two equations mean that $\widehat{\mu}_n$ and $\widehat{\alpha}_{j_0, k}$ are unbiased estimators of μ and $\alpha_{j_0, k}$, respectively. Furthermore, the linear estimator $\widehat{r}_n^{(d)}(x)$ can also be as an unbiased estimator of $r^{(d)}(x)$. In the following, we present an important lemma, which will be used to prove our theorems.

Lemma 4.1. *For the problem (1.1)–(1.2) with Conditions A1–A5 hold. If $2^{j_0} \leq n$, then, for $1 \leq p < \infty$, we have*

$$\mathbb{E} \left| \widehat{\alpha}_{j_0, k} - \alpha_{j_0, k} \right|^p \lesssim 2^{j_0 d p} n^{-\frac{p}{2}}.$$

Proof of Lemma 4.1: According to the definition of $\widehat{\alpha}_{j_0, k}$, the following decomposition holds:

$$\widehat{\alpha}_{j_0, k} - \alpha_{j_0, k} = \frac{\widehat{\mu}_n}{\mu} \left[(-1)^d \frac{\mu}{n} \sum_{i=1}^n \frac{\rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \phi_{j_0, k}^{(d)}(X_i) - \alpha_{j_0, k} \right] + \alpha_{j_0, k} \cdot \widehat{\mu}_n \left(\frac{1}{\mu} - \frac{1}{\widehat{\mu}_n} \right).$$

Furthermore, one has

$$(4.6) \quad \begin{aligned} \mathbb{E} \left| \widehat{\alpha}_{j_0, k} - \alpha_{j_0, k} \right|^p &\lesssim \mathbb{E} \left| \frac{\widehat{\mu}_n}{\mu} \left[(-1)^d \frac{\mu}{n} \sum_{i=1}^n \frac{\rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \phi_{j_0, k}^{(d)}(X_i) - \alpha_{j_0, k} \right] \right|^p \\ &+ \mathbb{E} \left| \alpha_{j_0, k} \cdot \widehat{\mu}_n \left(\frac{1}{\mu} - \frac{1}{\widehat{\mu}_n} \right) \right|^p. \end{aligned}$$

Then, it follows from Condition A5, Hölder’s inequality and the orthonormality of $\{\phi_{j_0, k}\}$ that $|\alpha_{j_0, k}| = \left| \int_{[0,1]} r^{(d)}(x) \phi_{j_0, k}(x) dx \right| \lesssim 1$. Moreover, Condition A2 and the definition of $\widehat{\mu}_n$ imply that $|\widehat{\mu}_n| \lesssim 1$. Hence, the inequality (4.6) reduces to

$$(4.7) \quad \begin{aligned} \mathbb{E} \left| \widehat{\alpha}_{j_0, k} - \alpha_{j_0, k} \right|^p &\lesssim \mathbb{E} \left| \frac{\mu}{n} \sum_{i=1}^n (-1)^d \frac{\rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \phi_{j_0, k}^{(d)}(X_i) - \alpha_{j_0, k} \right|^p + \mathbb{E} \left| \frac{1}{\mu} - \frac{1}{\widehat{\mu}_n} \right|^p \\ &:= Q_1 + Q_2. \end{aligned}$$

Let us now bound Q_1 and Q_2 as sharp as possible.

- Upper bound of Q_1 .

Define $\xi_i := \frac{(-1)^d \mu \rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \phi_{j_0, k}^{(d)}(X_i) - \alpha_{j_0, k}$. Then, one gets

$$Q_1 := \mathbb{E} \left| \frac{1}{n} \sum_{i=1}^n \xi_i \right|^p = \left(\frac{1}{n} \right)^p \mathbb{E} \left| \sum_{i=1}^n \xi_i \right|^p.$$

Because $\phi^{(d)}$ is a bounded variation function, one can assume

$$\phi^{(d)} := \bar{\phi} - \tilde{\phi},$$

where $\bar{\phi}$ and $\tilde{\phi}$ are bounded, nonnegative and nondecreasing functions ([22]). Then, we can write

$$\phi_{j_0,k}^{(d)} := 2^{j_0d}(\bar{\phi}_{j_0,k} - \tilde{\phi}_{j_0,k}).$$

Moreover, one defines

$$\bar{\alpha}_{j_0,k} := \int (-1)^d 2^{j_0d} \bar{\phi}_{j_0,k}(x) r(x) dx, \quad \tilde{\alpha}_{j_0,k} := \int (-1)^d 2^{j_0d} \tilde{\phi}_{j_0,k}(x) r(x) dx$$

and

$$\bar{\xi}_i := \frac{(-1)^d 2^{j_0d} \mu \rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \bar{\phi}_{j_0,k}(X_i) - \bar{\alpha}_{j_0,k}, \quad \tilde{\xi}_i := \frac{(-1)^d 2^{j_0d} \mu \rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \tilde{\phi}_{j_0,k}(X_i) - \tilde{\alpha}_{j_0,k}.$$

Then, we have $\alpha_{j_0,k} = \bar{\alpha}_{j_0,k} - \tilde{\alpha}_{j_0,k}$, $\xi_i = \bar{\xi}_i - \tilde{\xi}_i$ and, by an elementary inequality of convexity, one gets

$$(4.8) \quad Q_1 = \left(\frac{1}{n}\right)^p \mathbb{E} \left| \sum_{i=1}^n (\bar{\xi}_i - \tilde{\xi}_i) \right|^p \lesssim \left(\frac{1}{n}\right)^p \left[\mathbb{E} \left| \sum_{i=1}^n \bar{\xi}_i \right|^p + \mathbb{E} \left| \sum_{i=1}^n \tilde{\xi}_i \right|^p \right].$$

Using (1.1), (1.2) and Condition A4, one knows that $\mathbb{E}\bar{\xi}_i = 0$. Note that $\frac{\rho(y)\bar{\phi}_{j_0,k}(x)}{\omega(x,y)h(x)}$ is a nondecreasing function by the monotonicity of $\bar{\phi}_{j_0,k}(x)$ and Conditions A1–A3. Furthermore, we get that $\{\bar{\xi}_i, i = 1, 2, \dots, n\}$ is negatively associated by Lemma 2.1. On the other hand, $|\bar{\xi}_i|^p \lesssim \left| \frac{(-1)^d 2^{j_0d} \mu \rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \bar{\phi}_{j_0,k}(X_i) \right|^p + |\bar{\alpha}_{j_0,k}|^p$ and $|\bar{\alpha}_{j_0,k}|^p = \left| \mathbb{E} \left[\frac{(-1)^d 2^{j_0d} \mu \rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \bar{\phi}_{j_0,k}(X_i) \right] \right|^p \leq \mathbb{E} \left| \frac{(-1)^d 2^{j_0d} \mu \rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \bar{\phi}_{j_0,k}(X_i) \right|^p$ thanks to Jensen’s inequality. Then, one has

$$\begin{aligned} \mathbb{E}|\bar{\xi}_i|^p &\lesssim \mathbb{E} \left| \frac{(-1)^d 2^{j_0d} \mu \rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \bar{\phi}_{j_0,k}(X_i) \right|^p \\ &= \int_{\mathbb{R}} \int_{[0,1]} \left| \frac{(-1)^d 2^{j_0d} \mu \rho(y)}{\omega(x, y) h(x)} \bar{\phi}_{j_0,k}(x) \right|^p f(x, y) dx dy. \end{aligned}$$

Using Conditions A1–A3 and (1.1), one finds that

$$(4.9) \quad \mathbb{E}|\bar{\xi}_i|^p \lesssim 2^{j_0dp} \int_{[0,1]} |\bar{\phi}_{j_0,k}(x)|^p dx \lesssim 2^{j_0} [(d+\frac{1}{2})^{p-1}].$$

In particular, $\mathbb{E}|\bar{\xi}_i|^2 \lesssim 2^{2j_0d}$. Recall Rosenthal’s inequality ([18]): If X_1, X_2, \dots, X_n are negatively associated random variables such that $\mathbb{E}X_i = 0$ and $\mathbb{E}|X_i|^p < \infty$, then

$$\mathbb{E} \left| \sum_{i=1}^n X_i \right|^p \lesssim \begin{cases} \sum_{i=1}^n \mathbb{E}|X_i|^p + \left(\sum_{i=1}^n \mathbb{E}X_i^2 \right)^{\frac{p}{2}}, & p > 2; \\ \left(\sum_{i=1}^n \mathbb{E}X_i^2 \right)^{\frac{p}{2}}, & 1 \leq p \leq 2. \end{cases}$$

According to this inequality and (4.9), one gets

$$\mathbb{E} \left| \sum_{i=1}^n \bar{\xi}_i \right|^p \lesssim \begin{cases} \left[2^{j_0} [(d+\frac{1}{2})^{p-1}] \cdot n + (n \cdot 2^{2j_0d})^{\frac{p}{2}} \right], & p \geq 2; \\ 2^{j_0dp} n^{p/2}, & 1 \leq p < 2. \end{cases}$$

This with $2^{j_0} < n$ shows that $\mathbb{E} \left| \sum_{i=1}^n \bar{\xi}_i \right|^p \lesssim 2^{j_0 d p} n^{p/2}$. Similarly, $\mathbb{E} \left| \sum_{i=1}^n \tilde{\xi}_i \right|^p \lesssim 2^{j_0 d p} n^{p/2}$. Combining those with (4.8), one knows that

$$(4.10) \quad Q_1 \lesssim 2^{j_0 d p} n^{-p/2}.$$

- Upper bound of Q_2 .

Using the definition of $\widehat{\mu}_n$, one has

$$(4.11) \quad \begin{aligned} \mathbb{E} \left| \frac{1}{\mu} - \frac{1}{\widehat{\mu}_n} \right|^p &= \mathbb{E} \left| \frac{1}{n} \sum_{i=1}^n \frac{1}{\omega(X_i, Y_i)} - \frac{1}{\mu} \right|^p \\ &= \frac{1}{n^p} \mathbb{E} \left| \sum_{i=1}^n \left[\frac{1}{\omega(X_i, Y_i)} - \frac{1}{\mu} \right] \right|^p. \end{aligned}$$

Define $\eta_i := \frac{1}{\omega(X_i, Y_i)} - \frac{1}{\mu}$. Then, $\mathbb{E}(\eta_i) = 0$ by (4.4). The monotonicity of $\omega(x, y)$ in Condition A2 and Lemma 2.1 imply that η_1, \dots, η_n are negatively associated. In addition, $\mathbb{E}|\eta_i|^p \lesssim 1$ thanks to Condition A2. According to Rosenthal's inequality, one has

$$(4.12) \quad \mathbb{E} \left| \frac{1}{\mu} - \frac{1}{\widehat{\mu}_n} \right|^p \lesssim n^{-\frac{p}{2}}.$$

Now it is easy to see from (4.7), (4.10) and (4.12) that

$$\mathbb{E} \left| \widehat{\alpha}_{j_0, k} - \alpha_{j_0, k} \right|^p \lesssim 2^{j_0 d p} n^{-\frac{p}{2}}.$$

This completes the proof of Lemma 4.1. □

In this position, we will state our first theorem.

Theorem 4.1. *For the problem (1.1)–(1.2) with Conditions A1–A5. Let $r^{(d)} \in B_{\tilde{p}, q}^s(H)$ ($\tilde{p}, q \in [1, \infty)$, $s > 0$), and $\tilde{p} \geq p \geq 1$, or $\tilde{p} \leq p < \infty$ and $s > \frac{1}{\tilde{p}}$. The linear wavelet estimator $\widehat{r}_n^{(d)}$ be defined in (4.1) with $2^{j_0} \sim n^{\frac{1}{2s'+2d+1}}$ and $s' = s - \left(\frac{1}{\tilde{p}} - \frac{1}{p}\right)_+$. Then, for $1 \leq p < \infty$, we have*

$$\mathbb{E} \int_{[0,1]} \left| \widehat{r}_n^{(d)}(x) - r^{(d)}(x) \right|^p dx \lesssim n^{-\frac{s' p}{2s'+2d+1}}.$$

Proof of Theorem 4.1: Note that

$$(4.13) \quad \mathbb{E} \int_{[0,1]} \left| \widehat{r}_n^{(d)}(x) - r^{(d)}(x) \right|^p dx \lesssim \mathbb{E} \left\| \sum_{k \in \Omega} (\widehat{\alpha}_{j_0, k} - \alpha_{j_0, k}) \phi_{j_0, k} \right\|_p^p + \left\| P_{j_0} r^{(d)} - r^{(d)} \right\|_p^p.$$

It follows from Lemma 3.1 that

$$\mathbb{E} \left\| \sum_{k \in \Omega} (\widehat{\alpha}_{j_0, k} - \alpha_{j_0, k}) \phi_{j_0, k} \right\|_p^p \lesssim 2^{p \left(\frac{j_0}{2} - \frac{j_0}{p}\right)} \sum_{k \in \Omega} \mathbb{E} \left| \widehat{\alpha}_{j_0, k} - \alpha_{j_0, k} \right|^p.$$

Using Lemma 4.1, $|\Omega| \sim 2^{j_0}$ and $2^{j_0} \sim n^{\frac{1}{2s'+2d+1}}$, one knows

$$(4.14) \quad \mathbb{E} \left\| \sum_{k \in \Omega} (\hat{\alpha}_{j_0,k} - \alpha_{j_0,k}) \phi_{j_0,k} \right\|_p^p \lesssim \left(\frac{2^{j_0(1+2d)}}{n} \right)^{\frac{p}{2}} \sim n^{-\frac{s'p}{2s'+2d+1}}.$$

Next, one estimates $\|P_{j_0} r^{(d)} - r^{(d)}\|_p^p$. When $\tilde{p} \leq p$ and $s > \frac{1}{\tilde{p}}$, $B_{\tilde{p},q}^s(\mathbb{R}) \subseteq B_{p,q}^{s'}(\mathbb{R})$. Then, $r^{(d)} \in B_{p,q}^{s'}(\mathbb{R})$ and

$$(4.15) \quad \left\| P_{j_0} r^{(d)} - r^{(d)} \right\|_p^p \lesssim 2^{-j_0 s' p}$$

thanks to Lemma 3.2. When $\tilde{p} > p$, $s' = s$. Using Hölder's inequality and the compact support of $r^{(d)}$ and ϕ , one gets

$$\left\| P_{j_0} r^{(d)} - r^{(d)} \right\|_p^p \lesssim \left\| P_{j_0} r^{(d)} - r^{(d)} \right\|_{\tilde{p}}^p.$$

Then, it is easy to see from Lemma 3.2 and $r^{(d)} \in B_{\tilde{p},q}^s(H)$ that $\|P_{j_0} r^{(d)} - r^{(d)}\|_p^p \lesssim 2^{-j_0 s' p}$. This result with (4.15) shows that, for $1 \leq p < \infty$,

$$(4.16) \quad \left\| P_{j_0} r^{(d)} - r^{(d)} \right\|_p^p \lesssim 2^{-j_0 s' p}.$$

Furthermore, by $2^{j_0} \sim n^{\frac{1}{2s'+2d+1}}$, one gets

$$(4.17) \quad \left\| P_{j_0} r^{(d)} - r^{(d)} \right\|_p^p \lesssim n^{-\frac{s'p}{2s'+2d+1}}.$$

Combining this with (4.13) and (4.14),

$$\mathbb{E} \int_{[0,1]} \left| \hat{r}_n^{(d)}(x) - r^{(d)}(x) \right|^p dx \lesssim n^{-\frac{s'p}{2s'+2d+1}}.$$

This ends the proof of Theorem 4.1. □

Since j_0 depends on s' which remains unknown, $\hat{r}_n^{(d)}(x)$ is not adaptive. Theorem 4.1 is however of interest to determine in a simple manner sharp rates of convergence in our statistical setting. We do not however claim that they are optimal in the minimax sense; the lower bounds in this case are not proved in this study. Also, Theorem 4.1 can be viewed as generalization to the [3, Theorem 3.3] to the negatively associated case.

5. NONLINEAR WAVELET ESTIMATION

In this section, we will construct a adaptive nonlinear wavelet estimator and consider its upper bound over L^p ($1 \leq p < +\infty$) risk. Now, we define our nonlinear wavelet estimator

$$(5.1) \quad \tilde{r}_n^{(d)}(x) := \sum_{k \in \Omega} \hat{\alpha}_{j_0,k} \phi_{j_0,k}(x) + \sum_{j=j_0}^{j_1} \sum_{k \in \Lambda_j} \hat{\beta}_{j,k} I_{\{|\hat{\beta}_{j,k}| \geq \kappa t_n\}} \psi_{j,k}(x),$$

where $t_n := 2^{jd} \sqrt{\frac{\ln n}{n}}$,

$$(5.2) \quad \widehat{\beta}_{j,k} = (-1)^d \frac{\widehat{\mu}_n}{n} \sum_{i=1}^n \frac{\rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \psi_{j,k}^{(d)}(X_i)$$

and I_A denotes the indicator function over a set A , i.e., $I_A = 1$ if A is satisfied and 0 otherwise. The positive integers j_0, j_1 (depend on n) and the positive number κ will be given later on. The main difference between $\widetilde{\tau}^{(d)}$ and the linear wavelet estimator is the individual selection of the $\widehat{\beta}_{j,k}$'s done by the hard thresholding rule (formalized by the indicator function over $\{|\widehat{\beta}_{j,k}| \geq \kappa t_n\}$). We refer to [13] and [15] for the deep link between this selection technique and the intrinsic properties of the wavelets.

It should be pointed out that $\mathbb{E} \left[(-1)^d \frac{\mu \rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \psi_{j,k}^{(d)}(X_i) \right] = \beta_{j,k}$ thanks to [14, Lemma 2.1] (which uses Condition A4).

Note that Lemma 4.1 is still true if $\widehat{\alpha}_{j_0,k}$ is replaced by $\widehat{\beta}_{j,k}$, which leads to the following lemma.

Lemma 5.1. *For the problem (1.1)–(1.2) with Conditions A1–A5 hold. If $2^j \leq n$, then for $1 \leq p < \infty$, we have*

$$\mathbb{E} \left| \widehat{\beta}_{j,k} - \beta_{j,k} \right|^p \lesssim 2^{jdp} n^{-\frac{p}{2}}.$$

Lemma 5.2. *For the problem (1.1)–(1.2) with Conditions A1–A5. Then, for $j2^j \leq n$ and each $w > 0$, there exists a constant $\kappa > 1$ such that*

$$\mathbb{P} \left(\left| \widehat{\beta}_{j,k} - \beta_{j,k} \right| \geq \kappa t_n \right) \lesssim 2^{-wj}.$$

Proof of Lemma 5.2: Via similar arguments to those used in (4.7), we obtain

$$\left| \widehat{\beta}_{j,k} - \beta_{j,k} \right| \lesssim \left| \frac{\mu}{n} \sum_{i=1}^n (-1)^d \frac{\rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \psi_{j,k}^{(d)}(X_i) - \beta_{j,k} \right| + \left| \frac{1}{\mu} - \frac{1}{\widehat{\mu}_n} \right|.$$

Hence, it suffices to prove

$$(5.3) \quad \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \left[\frac{(-1)^d \mu \rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \psi_{j,k}^{(d)}(X_i) - \beta_{j,k} \right] \right| \geq \frac{\kappa}{2} t_n \right) \lesssim 2^{-wj}$$

and

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \left[\frac{1}{\omega(X_i, Y_i)} - \frac{1}{\mu} \right] \right| \geq \frac{\kappa}{2} t_n \right) \lesssim 2^{-wj}.$$

One shows the first inequality (5.3) only, the second one is similar and even simpler.

Define $\gamma_i := \frac{(-1)^d \mu \rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \psi_{j,k}^{(d)}(X_i) - \beta_{j,k}$. Then, one has

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \left[\frac{(-1)^d \mu \rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \psi_{j,k}^{(d)}(X_i) - \beta_{j,k} \right] \right| \geq \frac{\kappa}{2} t_n \right) = \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \gamma_i \right| \geq \frac{\kappa}{2} t_n \right).$$

Because $\psi^{(d)}$ is a bounded variation function, one can assume

$$\psi^{(d)} := \bar{\psi} - \tilde{\psi},$$

where $\bar{\psi}$ and $\tilde{\psi}$ are bounded, nonnegative and nondecreasing functions ([22]). Then,

$$\psi_{j,k}^{(d)} := 2^{jd}(\bar{\psi}_{j,k} - \tilde{\psi}_{j,k}).$$

Moreover, one defines

$$\bar{\beta}_{j,k} := \int (-1)^d 2^{jd} \bar{\psi}_{j,k}(x) r(x) dx, \quad \tilde{\beta}_{j,k} := \int (-1)^d 2^{jd} \tilde{\psi}_{j,k}(x) r(x) dx,$$

and

$$\bar{\gamma}_i := \frac{(-1)^d 2^{jd} \mu \rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \bar{\psi}_{j,k}(X_i) - \bar{\beta}_{j,k}, \quad \tilde{\gamma}_i := \frac{(-1)^d 2^{jd} \mu \rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \tilde{\psi}_{j,k}(X_i) - \tilde{\beta}_{j,k}.$$

Then, $\beta_{j,k} = \bar{\beta}_{j,k} - \tilde{\beta}_{j,k}$, $\gamma_i = \bar{\gamma}_i - \tilde{\gamma}_i$ and

$$(5.4) \quad \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \gamma_i \right| \geq \frac{\kappa}{2} t_n \right) \lesssim \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \bar{\gamma}_i \right| \geq \frac{\kappa}{4} t_n \right) + \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \tilde{\gamma}_i \right| \geq \frac{\kappa}{4} t_n \right).$$

According to (1.1), (1.2) and Condition A4, one gets $\mathbb{E}\bar{\gamma}_i = \bar{\beta}_{j,k}$. Moreover, $\bar{\gamma}_1, \bar{\gamma}_2, \dots, \bar{\gamma}_n$ are negatively associated by Conditions A1–A3, Lemma 2.1 and the nondecreasing property of $\bar{\psi}_{j,k}$. On the other hand, by the bounded properties of functions in Conditions A1–A3, $\left| \frac{(-1)^d 2^{jd} \mu \rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \bar{\psi}_{j,k}(X_i) \right| \lesssim 2^{j(d+\frac{1}{2})}$ and

$$|\bar{\gamma}_i| \lesssim \left| \frac{(-1)^d 2^{jd} \mu \rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \bar{\psi}_{j,k}(X_i) \right| + \mathbb{E} \left| \frac{(-1)^d 2^{jd} \mu \rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \bar{\psi}_{j,k}(X_i) \right| \lesssim 2^{j(d+\frac{1}{2})}.$$

Similar to the arguments of (4.9) with $p = 2$, $\mathbb{E}(\bar{\gamma}_i)^2 \lesssim 2^{2jd}$. Recall Bernstein’s inequality: Let X_1, \dots, X_n be negatively associated random variables such that $\mathbb{E}X_i = 0$, $|X_i| \leq M$ and $\mathbb{E}X_i^2 = \sigma^2$. Then, for each $v \geq 0$,

$$\mathbb{P} \left(\frac{1}{n} \left| \sum_{i=1}^n X_i \right| \geq v \right) \leq 2 \cdot \exp \left\{ -\frac{nv^2}{2(\sigma^2 + \frac{vM}{3})} \right\}.$$

It follows from Bernstein’s inequality, $t_n = 2^{jd} \sqrt{\frac{\ln n}{n}}$ and $j2^j \leq n$ that

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \bar{\gamma}_i \right| \geq \frac{\kappa}{4} t_n \right) \lesssim \exp \left\{ -\frac{n \left(\frac{\kappa t_n}{4} \right)^2}{2 \left(2^{2jd} + \frac{\kappa t_n}{12} 2^{j(d+\frac{1}{2})} \right)} \right\} \lesssim \exp \left\{ -\frac{\ln n \kappa^2}{32 \left(1 + \frac{\kappa}{12} \right)} \right\}.$$

Obviously, there exists sufficiently large $\kappa > 1$ such that $\exp \left\{ -\frac{\ln n \kappa^2}{32 \left(1 + \frac{\kappa}{12} \right)} \right\} \lesssim 2^{-wj}$. Hence,

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \bar{\gamma}_i \right| \geq \frac{\kappa}{4} t_n \right) \lesssim 2^{-wj}.$$

Similarly, $\mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \tilde{\gamma}_i \right| \geq \frac{\kappa}{4} t_n \right) \lesssim 2^{-wj}$. Those results with (5.4) show that

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \left[\frac{(-1)^d \mu \rho(Y_i)}{\omega(X_i, Y_i) h(X_i)} \psi_{j,k}^{(d)}(X_i) - \beta_{j,k} \right] \right| \geq \frac{\kappa}{2} t_n \right) \lesssim 2^{-wj}.$$

This ends the proof of Lemma 5.2. □

Now we will give our last theorem in this position.

Theorem 5.1. For the problem (1.1)–(1.2) with Conditions A1–A5. Let $r^{(d)} \in B_{p,q}^s(H)$ ($\tilde{p}, q \in [1, \infty)$, $s > 0$), and $\tilde{p} \geq p \geq 1$, or $\tilde{p} \leq p < \infty$ and $s > \frac{1}{\tilde{p}}$. Then, the nonlinear wavelet estimator $\tilde{r}_n^{(d)}$ defined in (5.1) with $2^{j_0} \sim n^{\frac{1}{2m+2d+1}}$ ($m > s$) and $2^{j_1} \sim (\frac{n}{\ln n})^{\frac{1}{2d+1}}$ satisfies

$$(5.5) \quad \mathbb{E} \int_{[0,1]} \left| \tilde{r}_n^{(d)}(x) - r^{(d)}(x) \right|^p dx \lesssim (\ln n)^{\frac{3p}{2}} n^{-\alpha p},$$

where

$$(5.6) \quad \alpha = \begin{cases} \frac{s}{2s + 2d + 1}, & \tilde{p} \geq \frac{p(2d + 1)}{2s + 2d + 1}, \\ \frac{s - 1/\tilde{p} + 1/p}{2(s - 1/\tilde{p}) + 2d + 1}, & \tilde{p} < \frac{p(2d + 1)}{2s + 2d + 1}. \end{cases}$$

Proof of Theorem 5.1: For the proof of Theorem 5.1, we will prove it under two cases respectively.

(i) Upper bound estimation under $\tilde{p} \leq p < \infty$ and $s > \frac{1}{\tilde{p}}$.

In this case, (5.6) can be rewritten as

$$\alpha = \min \left\{ \frac{s}{2s + 2d + 1}, \frac{s - 1/\tilde{p} + 1/p}{2(s - 1/\tilde{p}) + 2d + 1} \right\}.$$

By the definition of $\tilde{r}_n^{(d)}(x)$,

$$(5.7) \quad \begin{aligned} \mathbb{E} \int_{[0,1]} \left| \tilde{r}_n^{(d)}(x) - r^{(d)}(x) \right|^p dx &\lesssim \mathbb{E} \left\| \sum_{k \in \Omega} (\hat{\alpha}_{j_0,k} - \alpha_{j_0,k}) \phi_{j_0,k} \right\|_p^p + \left\| r^{(d)} - P_{j_1+1} r^{(d)} \right\|_p^p \\ &+ \mathbb{E} \left\| \sum_{j=j_0}^{j_1} \sum_{k \in \Lambda_j} [\hat{\beta}_{j,k} I_{\{|\hat{\beta}_{j,k}| \geq \kappa t_n\}} - \beta_{j,k}] \psi_{j,k} \right\|_p^p. \end{aligned}$$

It follows from Lemma 3.1 that

$$\mathbb{E} \left\| \sum_{k \in \Omega} (\hat{\alpha}_{j_0,k} - \alpha_{j_0,k}) \phi_{j_0,k} \right\|_p^p \lesssim 2^{p(\frac{j_0}{2} - \frac{j_0}{p})} \sum_{k \in \Omega} \mathbb{E} \left| \hat{\alpha}_{j_0,k} - \alpha_{j_0,k} \right|^p.$$

Using Lemma 4.1, $|\Omega| \sim 2^{j_0}$ and $2^{j_0} \sim n^{\frac{1}{2m+2d+1}}$ ($m > s$), one knows

$$(5.8) \quad \mathbb{E} \left\| \sum_{k \in \Omega} (\hat{\alpha}_{j_0,k} - \alpha_{j_0,k}) \phi_{j_0,k} \right\|_p^p \lesssim n^{-\frac{mp}{2m+2d+1}} < n^{-\frac{sp}{2s+2d+1}} \leq n^{-\alpha p}.$$

Similar to the arguments of (4.15), when $\tilde{p} \leq p$ and $s > \frac{1}{\tilde{p}}$, one gets that

$$(5.9) \quad \left\| P_{j_1+1} r^{(d)} - r^{(d)} \right\|_p \lesssim 2^{-j_1(s-1/\tilde{p}+1/p)}.$$

On the other hand, $s - \frac{1}{p} + \frac{1}{p} \geq \alpha$ thanks to $\tilde{p} \leq p$ and $s > \frac{1}{p}$. Then, it follows from $2^{j_1} \sim \left(\frac{n}{\ln n}\right)^{\frac{1}{2d+1}}$ that

$$\left\| P_{j_1+1} r^{(d)} - r^{(d)} \right\|_p^p \lesssim \left(\frac{\ln n}{n}\right)^{\frac{(s-1/\tilde{p}+1/p)p}{2d+1}} \lesssim \left(\frac{\ln n}{n}\right)^{\alpha p}.$$

The main work for the proof of Theorem 5.1 is to show

$$(5.10) \quad Z := \mathbb{E} \left\| \sum_{j=j_0}^{j_1} \sum_{k \in \Lambda_j} \left[\widehat{\beta}_{j,k} I_{\{|\widehat{\beta}_{j,k}| \geq \kappa t_n\}} - \beta_{j,k} \right] \psi_{j,k} \right\|_p^p \lesssim (\ln n)^{\frac{3p}{2}} n^{-\alpha p}.$$

It is easy to see from Lemma 3.1 that

$$Z \lesssim (j_1 - j_0 + 1)^{p-1} \sum_{j=j_0}^{j_1} 2^{p(\frac{j}{2} - \frac{j}{p})} \sum_{k \in \Lambda_j} \mathbb{E} \left| \widehat{\beta}_{j,k} I_{\{|\widehat{\beta}_{j,k}| \geq \kappa t_n\}} - \beta_{j,k} \right|^p.$$

Then, the classical technique ([13]) gives

$$(5.11) \quad Z \lesssim (j_1 - j_0 + 1)^{p-1} (Z_1 + Z_2 + Z_3),$$

where

$$\begin{aligned} Z_1 &= \sum_{j=j_0}^{j_1} 2^{p(\frac{j}{2} - \frac{j}{p})} \sum_{k \in \Lambda_j} \mathbb{E} \left[\left| \widehat{\beta}_{j,k} - \beta_{j,k} \right|^p I_{\{|\widehat{\beta}_{j,k} - \beta_{j,k}| > \frac{\kappa t_n}{2}\}} \right], \\ Z_2 &= \sum_{j=j_0}^{j_1} 2^{p(\frac{j}{2} - \frac{j}{p})} \sum_{k \in \Lambda_j} \mathbb{E} \left[\left| \widehat{\beta}_{j,k} - \beta_{j,k} \right|^p I_{\{|\widehat{\beta}_{j,k}| \geq \kappa t_n, |\beta_{j,k}| \geq \frac{\kappa t_n}{2}\}} \right], \\ Z_3 &= \sum_{j=j_0}^{j_1} 2^{p(\frac{j}{2} - \frac{j}{p})} \sum_{k \in \Lambda_j} |\beta_{j,k}|^p I_{\{|\widehat{\beta}_{j,k}| < \kappa t_n, |\beta_{j,k}| \leq 2\kappa t_n\}}. \end{aligned}$$

- Upper bound of Z_1 .

It follows from Hölder’s inequality that

$$\mathbb{E} \left[\left| \widehat{\beta}_{j,k} - \beta_{j,k} \right|^p I_{\{|\widehat{\beta}_{j,k} - \beta_{j,k}| > \frac{\kappa t_n}{2}\}} \right] \leq \left[\mathbb{E} \left| \widehat{\beta}_{j,k} - \beta_{j,k} \right|^{2p} \right]^{\frac{1}{2}} \left[\mathbb{P} \left(\left| \widehat{\beta}_{j,k} - \beta_{j,k} \right| > \frac{\kappa t_n}{2} \right) \right]^{\frac{1}{2}}.$$

Furthermore, Lemmas 5.1 and 5.2 imply that

$$\mathbb{E} \left[\left| \widehat{\beta}_{j,k} - \beta_{j,k} \right|^p I_{\{|\widehat{\beta}_{j,k} - \beta_{j,k}| > \frac{\kappa t_n}{2}\}} \right] \lesssim 2^{jdp} n^{-\frac{p}{2}} 2^{-\frac{wj}{2}},$$

where $\kappa > 1$ is chosen for $w > p + 2dp$ in Lemma 5.2. This with the choice $2^{j_0} \sim n^{\frac{1}{2m+2d+1}}$ ($m > s$) shows that

$$(5.12) \quad \begin{aligned} Z_1 &\lesssim n^{-\frac{p}{2}} \sum_{j=j_0}^{j_1} 2^{j(\frac{p}{2} + dp - \frac{w}{2})} \lesssim n^{-\frac{p}{2}} 2^{j_0(\frac{p}{2} + dp)} \lesssim n^{-\frac{mp}{2m+2d+1}} \\ &\leq n^{-\frac{sp}{2s+2d+1}} \leq n^{-\alpha p}. \end{aligned}$$

- Upper bound of Z_2 .

Taking

$$2^{j_0^*} \sim \left(\frac{n}{\ln n}\right)^{\frac{1-2\alpha}{2d+1}}.$$

Because $0 < \alpha \leq \frac{s}{2s+2d+1}$ and $2^{j_0} \sim n^{\frac{1}{2m+2d+1}}$ ($m > s$), $2^{j_0^*} \leq 2^{j_1} \sim \left(\frac{n}{\ln n}\right)^{\frac{1}{2d+1}}$ and $2^{j_0^*} \geq \left(\frac{n}{\ln n}\right)^{\frac{1-\frac{2s}{2s+2d+1}}{2d+1}} = \left(\frac{n}{\ln n}\right)^{\frac{1}{2s+2d+1}} \gtrsim n^{\frac{1}{2m+2d+1}} \sim 2^{j_0}$. Furthermore, it follows from Lemma 5.1 that

$$\begin{aligned} Z_{21} &:= \sum_{j=j_0}^{j_0^*} 2^{p(\frac{j}{2}-\frac{j}{p})} \sum_{k \in \Lambda_j} \mathbb{E} \left[\left| \widehat{\beta}_{j,k} - \beta_{j,k} \right|^p I_{\{|\widehat{\beta}_{j,k}| \geq \kappa t_n, |\beta_{j,k}| \geq \frac{\kappa t_n}{2}\}} \right] \\ (5.13) \quad &\lesssim \sum_{j=j_0}^{j_0^*} 2^{p(\frac{j}{2}-\frac{j}{p})} \sum_{k \in \Lambda_j} 2^{jdp} n^{-\frac{p}{2}} \lesssim 2^{j_0^*(\frac{p}{2}+dp)} n^{-\frac{p}{2}} \lesssim n^{-\alpha p}. \end{aligned}$$

On the other hand, by Lemmas 5.1 and 3.2, and $t_n = 2^{jd} \sqrt{\frac{\ln n}{n}}$, one has

$$\begin{aligned} Z_{22} &:= \sum_{j=j_0^*+1}^{j_1} 2^{p(\frac{j}{2}-\frac{j}{p})} \sum_{k \in \Lambda_j} \mathbb{E} \left[\left| \widehat{\beta}_{j,k} - \beta_{j,k} \right|^p I_{\{|\widehat{\beta}_{j,k}| \geq \kappa t_n, |\beta_{j,k}| \geq \frac{\kappa t_n}{2}\}} \right] \\ &\lesssim \sum_{j=j_0^*+1}^{j_1} 2^{p(\frac{j}{2}-\frac{j}{p})} \sum_{k \in \Lambda_j} \mathbb{E} \left| \widehat{\beta}_{j,k} - \beta_{j,k} \right|^p \left(\frac{|\beta_{j,k}|}{\kappa t_n/2} \right)^{\widetilde{p}} \\ (5.14) \quad &\lesssim \sum_{j=j_0^*+1}^{j_1} (\ln n)^{-\widetilde{p}/2} n^{-\frac{p-\widetilde{p}}{2}} 2^{-j \left[s\widetilde{p} - \frac{(p-\widetilde{p})(2d+1)}{2} \right]}. \end{aligned}$$

Define

$$\varepsilon := s\widetilde{p} - \frac{(p-\widetilde{p})(2d+1)}{2}.$$

Then, $\varepsilon > 0$ holds if and only if $\widetilde{p} > \frac{p(2d+1)}{2s+2d+1}$, and (5.14) can be rewritten as

$$(5.15) \quad Z_{22} \lesssim (\ln n)^{-\widetilde{p}/2} n^{-\frac{p-\widetilde{p}}{2}} \sum_{j=j_0^*+1}^{j_1} 2^{-j\varepsilon}.$$

When $\varepsilon > 0$, $\widetilde{p} > \frac{p(2d+1)}{2s+2d+1}$ and $\alpha = \frac{s}{2s+2d+1}$ thanks to (5.6). Moreover, it can be easily checked that $\frac{p-\widetilde{p}}{2} + \frac{1-2\alpha}{2d+1} \left[s\widetilde{p} - \frac{(p-\widetilde{p})(2d+1)}{2} \right] = \alpha p$. This with the choice of $2^{j_0^*}$ leads to

$$\begin{aligned} Z_{22} &\lesssim (\ln n)^{-\widetilde{p}/2} n^{-\frac{p-\widetilde{p}}{2}} 2^{-j_0^* \varepsilon} \leq (\ln n) \left(\frac{1}{n}\right)^{\frac{p-\widetilde{p}}{2} + \frac{1-2\alpha}{2d+1} \left[s\widetilde{p} - \frac{(p-\widetilde{p})(2d+1)}{2} \right]} \\ (5.16) \quad &= (\ln n) n^{-\alpha p}. \end{aligned}$$

For the case $\varepsilon \leq 0$, $\widetilde{p} \leq \frac{p(2d+1)}{2s+2d+1}$ and $\alpha = \frac{s-\frac{1}{\widetilde{p}}+\frac{1}{p}}{2\left(s-\frac{d}{\widetilde{p}}\right)+2d+1}$. Define $p_1 := (1-2\alpha)p$. Then,

$\alpha \leq \frac{s}{2s+2d+1}$ and $\tilde{p} \leq \frac{p(2d+1)}{2s+2d+1} < (1 - 2\alpha)p = p_1$. Similarly to (5.14), one has

$$\begin{aligned} Z_{22} &\lesssim \sum_{j=j_0^*+1}^{j_1} 2^{p(\frac{j}{2}-\frac{j}{p})} \sum_{k \in \Lambda_j} \mathbb{E} \left| \widehat{\beta}_{j,k} - \beta_{j,k} \right|^p \left(\frac{|\beta_{j,k}|}{\kappa t_n/2} \right)^{p_1} \\ &\lesssim \sum_{j=j_0^*+1}^{j_1} 2^{p(\frac{j}{2}-\frac{j}{p})} 2^{jdp} n^{-\frac{p}{2}} t_n^{-p_1} \|\beta_j\|_{p_1}^{p_1}. \end{aligned}$$

Because $\tilde{p} \leq p_1$ and $r^{(d)} \in B_{\tilde{p},q}^s(H)$, we get $\|\beta_j\|_{p_1}^{p_1} \leq \|\beta_j\|_{\tilde{p}}^{p_1} \lesssim 2^{-j(s-\frac{1}{\tilde{p}}+\frac{1}{2})p_1}$ and

$$\begin{aligned} Z_{22} &\lesssim \sum_{j=j_0^*+1}^{j_1} 2^{p(\frac{j}{2}-\frac{j}{p})} 2^{jdp} n^{-\frac{p}{2}} t_n^{-p_1} 2^{-j(s-\frac{1}{\tilde{p}}+\frac{1}{2})p_1} \\ &\leq \left(\frac{1}{n}\right)^{\frac{p-p_1}{2}} \sum_{j=j_0^*+1}^{j_1} 2^{-j(sp_1-\frac{p_1}{\tilde{p}}+\frac{p_1}{2}+dp_1-dp-\frac{p}{2}+1)}. \end{aligned}$$

By the definitions of p_1 and α , $sp_1 - \frac{p_1}{\tilde{p}} + \frac{p_1}{2} + dp_1 - dp - \frac{p}{2} + 1 = 0$ and $Z_{22} \lesssim \left(\frac{1}{n}\right)^{\frac{p-p_1}{2}} (\ln n) = (\ln n) \left(\frac{1}{n}\right)^{\alpha p}$. This with (5.13) and (5.16) shows in both cases,

$$(5.17) \quad Z_2 = Z_{21} + Z_{22} \lesssim (\ln n) n^{-\alpha p}.$$

- Upper bound of Z_3 .

It is easy to see that

$$\begin{aligned} Z_{31} &:= \sum_{j=j_0}^{j_0^*} 2^{p(\frac{j}{2}-\frac{j}{p})} \sum_{k \in \Lambda_j} |\beta_{j,k}|^p I_{\{|\widehat{\beta}_{j,k}| < \kappa t_n, |\beta_{j,k}| \leq 2\kappa t_n\}} \\ &\leq \sum_{j=j_0}^{j_0^*} 2^{p(\frac{j}{2}-\frac{j}{p})} \sum_{k \in \Lambda_j} |2\kappa t_n|^p \lesssim \sum_{j=j_0}^{j_0^*} 2^{j(\frac{p}{2}+dp)} \left(\frac{\ln n}{n}\right)^{\frac{p}{2}} \\ (5.18) \quad &\lesssim \left(\frac{\ln n}{n}\right)^{\frac{p}{2}} 2^{j_0^*(\frac{p}{2}+dp)} \lesssim \left(\frac{\ln n}{n}\right)^{\alpha p}. \end{aligned}$$

On the other hand, one has

$$\begin{aligned} Z_{32} &:= \sum_{j=j_0^*+1}^{j_1} 2^{p(\frac{j}{2}-\frac{j}{p})} \sum_{k \in \Lambda_j} |\beta_{j,k}|^p I_{\{|\widehat{\beta}_{j,k}| < \kappa t_n, |\beta_{j,k}| \leq 2\kappa t_n\}} \\ &\leq \sum_{j=j_0^*+1}^{j_1} 2^{p(\frac{j}{2}-\frac{j}{p})} \sum_{k \in \Lambda_j} |\beta_{j,k}|^p \left| \frac{2\kappa t_n}{\beta_{j,k}} \right|^{p-\tilde{p}} \\ (5.19) \quad &\lesssim \sum_{j=j_0^*+1}^{j_1} 2^{p(\frac{j}{2}-\frac{j}{p})} t_n^{p-\tilde{p}} \|\beta_j\|_{\tilde{p}}^{\tilde{p}} \lesssim \left(\frac{\ln n}{n}\right)^{\frac{p-\tilde{p}}{2}} \sum_{j=j_0^*+1}^{j_1} 2^{-j\varepsilon}. \end{aligned}$$

The same arguments as (5.15) shows that, for $\varepsilon > 0$,

$$(5.20) \quad Z_{32} \lesssim \left(\frac{\ln n}{n}\right)^{\alpha p}.$$

For the case of $\varepsilon \leq 0$, one defines

$$2^{j_1^*} \sim \left(\frac{n}{\ln n}\right)^{\frac{\alpha}{s-1/\tilde{p}+1/p}}.$$

Note that $\varepsilon \leq 0$ and $s > \frac{1}{\tilde{p}}$. Then, $\tilde{p} \leq \frac{p(2d+1)}{2s+2d+1}$, $\alpha = \frac{s-\frac{1}{\tilde{p}}+\frac{1}{p}}{2\left(s-\frac{1}{\tilde{p}}\right)+2d+1}$ and $\alpha \leq s - \frac{1}{\tilde{p}} + \frac{1}{p}$. Hence, $n^{\frac{1-2\alpha}{2d+1}} \lesssim 2^{j_0^*} \leq 2^{j_1^*} \leq 2^{j_1} \sim \left(\frac{n}{\ln n}\right)^{\frac{1}{2d+1}}$ and $Z_{32} = Z_{321} + Z_{322}$, where

$$\begin{aligned} Z_{321} &:= \sum_{j=j_0^*+1}^{j_1^*} 2^{p\left(\frac{j}{2}-\frac{j}{p}\right)} \sum_{k \in \Lambda_j} |\beta_{j,k}|^p I_{\{|\hat{\beta}_{j,k}| < \kappa t_n, |\beta_{j,k}| \leq 2\kappa t_n\}}, \\ Z_{322} &:= \sum_{j=j_1^*+1}^{j_1} 2^{p\left(\frac{j}{2}-\frac{j}{p}\right)} \sum_{k \in \Lambda_j} |\beta_{j,k}|^p I_{\{|\hat{\beta}_{j,k}| < \kappa t_n, |\beta_{j,k}| \leq 2\kappa t_n\}}. \end{aligned}$$

By the arguments of (5.15) and the choice of $2^{j_1^*}$, one has

$$Z_{321} \lesssim \left(\frac{\ln n}{n}\right)^{\frac{p-\tilde{p}}{2}} 2^{-j_1^* \varepsilon} = \left(\frac{\ln n}{n}\right)^{\frac{p-\tilde{p}}{2} + \frac{\alpha \varepsilon}{s-1/\tilde{p}+1/p}}.$$

It is easy to check that $\frac{p-\tilde{p}}{2} + \frac{\alpha \varepsilon}{s-1/\tilde{p}+1/p} = \alpha p$. Then,

$$Z_{321} \lesssim \left(\frac{\ln n}{n}\right)^{\alpha p}.$$

On the other hand, using $\|\beta_j\|_{\tilde{p}} \lesssim 2^{-j\left(s-\frac{1}{\tilde{p}}+\frac{1}{2}\right)}$, $s > \frac{1}{\tilde{p}}$ and $2^{j_1^*} \sim \left(\frac{n}{\ln n}\right)^{\frac{\alpha}{s-1/\tilde{p}+1/p}}$.

$$\begin{aligned} Z_{322} &\leq \sum_{j=j_1^*+1}^{j_1} 2^{p\left(\frac{j}{2}-\frac{j}{p}\right)} \sum_{k \in \Lambda_j} |\beta_{j,k}|^p \leq \sum_{j=j_1^*+1}^{j_1} 2^{p\left(\frac{j}{2}-\frac{j}{p}\right)} \|\beta_j\|_{\tilde{p}}^p \\ &\lesssim \sum_{j=j_1^*+1}^{j_1} 2^{-j(1+sp-p/\tilde{p})} \lesssim 2^{-j_1^*(1+sp-p/\tilde{p})} \sim \left(\frac{\ln n}{n}\right)^{\alpha p}. \end{aligned}$$

Now, it follows that for $\varepsilon \leq 0$,

$$Z_{32} = Z_{321} + Z_{322} \lesssim \left(\frac{\ln n}{n}\right)^{\alpha p}.$$

Combining this with (5.18) and (5.20), one knows

$$(5.21) \quad Z_3 \lesssim \left(\frac{\ln n}{n}\right)^{\alpha p}.$$

Then, it follows from (5.11), (5.12), (5.17) and (5.21) that

$$Z \lesssim (\ln n)^{\frac{3p}{2}} n^{-\alpha p}.$$

Hence,

$$(5.22) \quad \mathbb{E} \int_{[0,1]} \left| \tilde{r}_n^{(d)}(x) - r^{(d)}(x) \right|^p dx \lesssim (\ln n)^{\frac{3p}{2}} n^{-\alpha p}$$

in the case of $\tilde{p} \leq p < \infty$ and $s > \frac{1}{\tilde{p}}$.

(ii) Upper bound estimation under $\tilde{p} > p$.

From the above arguments, one finds that when $\tilde{p} = p$, the inequality (5.22) still holds without the assumption $s > \frac{1}{p}$. It remains to conclude (5.22) for $\tilde{p} > p \geq 1$. By Hölder's inequality,

$$\int_{[0,1]} \left| \tilde{r}_n^{(d)}(x) - r^{(d)}(x) \right|^p dx \lesssim \left[\int_{[0,1]} \left| \tilde{r}_n^{(d)}(x) - r^{(d)}(x) \right|^{\tilde{p}} dx \right]^{\frac{p}{\tilde{p}}}.$$

Using Jensen's inequality and (5.22) with $\tilde{p} = p$, one gets

$$\mathbb{E} \int_{[0,1]} \left| \tilde{r}_n^{(d)}(x) - r^{(d)}(x) \right|^p dx \lesssim \left[\mathbb{E} \int_{[0,1]} \left| \tilde{r}_n^{(d)}(x) - r^{(d)}(x) \right|^{\tilde{p}} dx \right]^{\frac{p}{\tilde{p}}} \lesssim (\ln n)^{\frac{3p}{2}} n^{-\alpha p}.$$

This completes the proof of Theorem 5.1. \square

Contrary to the linear wavelet estimator given by (4.1), $\tilde{r}_n^{(d)}(x)$ is fully adaptive; its construction does not depend on s . The convergence rate of the nonlinear estimator keeps the same as that of the linear one up to a logarithmic factor when $\tilde{p} > p$. However, it gets better in the case of $\tilde{p} \leq p$. This aspect remains standard in nonlinear wavelet estimation in the standard regression (or density) estimation framework (see [15]). Also, Theorem 5.1 can be viewed as generalization to the [14, Theorem 1] to the negatively associated case.

6. CONCLUDING REMARKS

In this paper, the estimation of the derivatives of a regression function for biased data is considered. The feature of the study is to investigate the negatively dependent assumption on the data, beyond the independent assumption, opening new perspective of applications. Two wavelet estimators are introduced. The first estimator is based on wavelet projection of wavelet coefficient estimators only, the second estimator is nonlinear; a selection of the wavelet coefficient estimators are applied according to their magnitude via a hard thresholding rule. Sharp rates of convergence are obtained under the L^p risk with $1 \leq p < \infty$, assuming that the function of interest belongs to a ball of Besov spaces $B_{p,q}^s(\mathbb{R})$. These rates correspond to those obtained in the independent setting, showing that the wavelet methodology is consistent for this problem. Perspectives of this work are to prove the optimal lower bounds in the minimax sense, to relax some assumptions on the model, mainly the compact support of $r^{(d)}$ and explore the practical aspects of the proposed estimators. These points needs further investigations that we leave for a future work.

ACKNOWLEDGMENTS


The authors would like to thank the referees for their important comments and suggestions. This paper is supported by the National Natural Science Foundation of China (No. 12001133), Guangxi Natural Science Foundation (Nos. 2018GXNSFBA281076, 2019GXNSFFA245012).

REFERENCES

- [1] AHMAD, I.A. (1995). On multivariate kernel estimation for samples from weighted distributions, *Statistics and Probability Letters*, **22**, 121–129.
- [2] ALAM, K. and SAXENA, K.M.L. (1981). Positive dependence in multivariate distributions, *Communications in Statistics – Theory and Methods*, **10**, 1183–1196.
- [3] CHAUBEY, Y.P.; CHESNEAU, C. and NAVARRO, F. (2017). Linear wavelet estimation of the derivatives of a regression function based on biased data, *Communications in Statistics – Theory and Methods*, **46**(19), 9541–9556.
- [4] CHAUBEY, Y.P.; CHESNEAU, C. and SHIRAZI, E. (2013). Wavelet-based estimation of regression function for dependent biased data under a given random design, *Journal of Nonparametric Statistics*, **25**(1), 53–71.
- [5] CHAUBEY, Y.P.; LAÏB, N. and LI, J. (2012). Generalized kernel regression estimator for dependent size-biased data, *Journal of Statistical and Planning Inference*, **142**, 708–727.
- [6] CHAUBEY, Y.P. and SHIRAZI, E. (2015). On MISE of a non linear wavelet estimator of the regression function based on biased data under strong mixing, *Communications in Statistics – Theory and Methods*, **44**(5), 885–899.
- [7] CHAUBEY, Y.P.; CHESNEAU, C. and SHIRAZI, E. (2013). Wavelet-based estimation of regression function for dependent biased data under a given random design, *Journal of Nonparametric Statistics*, **25**(1), 53–71.
- [8] CHAUBEY, Y.P. and SHIRAZI, E. (2015). On MISE of a non linear wavelet estimator of the regression function based on biased data under strong mixing, *Communications in Statistics – Theory and Methods*, **44**(5), 885–899.
- [9] CHESNEAU, C. and SHIRAZI, E. (2014). Nonparametric wavelet regression based on biased data, *Communications in Statistics – Theory and Methods*, **43**(13), 2642–2658.
- [10] CRISTÓBAL, J.A. and ALCALÁ, J.T. (2000). Nonparametric regression estimators for length biased data, *Journal of Statistical and Planning Inference*, **89**, 145–168.
- [11] CRISTÓBAL, J.A. and ALCALÁ, J.T. (2001). An overview of nonparametric contributions to the problem of functional estimation from biased data, *Test*, **10**(2), 309–332.
- [12] CRISTÓBAL, J.A.; OJEDA, J.L. and ALCALÀ, J.T. (2004). Confidence bands in nonparametric regression with length biased data, *Annals of the Institute of Statistical Mathematics*, **56**(3), 475–496.
- [13] DONOHO, D.L.; JOHNSTONE, M.I.; KERKYACHARIAN, G. and PICARD, D. (1996). Density estimation by wavelet thresholding, *The Annals of Statistics*, **24**(2), 508–539.
- [14] GUO, H.J. and KOU, J.K. (2019). Non linear wavelet estimation of regression derivatives based on biased data, *Communications in Statistics – Theory and Methods*, **48**(13), 3219–3235.
- [15] HÄRDLE, W.; KERKYACHARIAN, G.; PICARD, D. and TSYBAKOV, A. (1997). *Wavelets, Approximation and Statistical Application*, New York: Springer-Verlag.
- [16] JOAG-DEV, K. and PROSCHAN, F. (1983). Negative association of random variables with applications, *Annals of Statistics*, **11**, 286–295.
- [17] KOU, J. and LIU, Y. (2018). Wavelet regression estimations with strong mixing data, *Statistical Methods and Applications*, **27**(4), 667–688.
- [18] LIU, Y.M. and XU, J.L. (2014). Wavelet density estimation for negatively associated stratified size-biased sample, *Journal of Nonparametric Statistics*, **26**(3), 537–554.
- [19] ROUSSAS, G.G. (1999). *Positive and negative dependence with some statistical application*. In “Asymptotics, Nonparametrics and Time Series” (S. Ghosh, Ed.), New York: Marcel Dekker.

- [20] SKÖLD, M. (1999). Kernel regression in the presence of size-bias, *Journal of Nonparametric Statistics*, **12**, 41–51.
- [21] WU, C.O. (2000). Local polynomial regression with selection biased data, *Statistica Sinica*, **10**(3), 789–817.
- [22] XU, J.L. (2014). Wavelet linear estimations of density derivatives from a negatively associated stratified size-biased sample, *Frontiers of Mathematics in China*, **9**(3), 623–640.
- [23] XU, F.; WANG, B.H. and HOU, Y.W. (2019). Almost sure local central limit theorem for the product of some partial sums of negatively associated sequences, *Journal of Inequalities and Applications*, **1**, 309.
- [24] ZHANG, Y.; YANG, X.Y.; DONG, Z.S. and WANG, D.H. (2011). The limit theorem for dependent random variables with applications to autoregression models, *Journal of Systems Science and Complexity*, **24**, 565–579.

Approximation Results for the Sums of Independent Random Variables

Author: PRATIMA EKNATH KADU 
– Department of Maths & Stats, K.J. Somaiya College of Arts and Commerce,
Vidyavihar, Mumbai-400077, India
pratima.kadu@somaiya.edu

Received: November 2019

Revised: June 2020

Accepted: July 2020

Abstract:

- In this article, we consider Poisson and Poisson convoluted geometric approximation to the sums of n independent random variables under moment conditions. We use Stein's method to derive the approximation results in total variation distance. The error bounds obtained are either comparable to or improvement over the existing bounds available in the literature. Also, we give an application to the waiting time distribution of 2-runs.

Keywords:

- *Poisson and geometric distribution; perturbations; probability generating function; Stein operator; Stein's method.*

AMS Subject Classification:

- Primary: 62E17, 62E20; Secondary: 60E05, 60F05.

1. INTRODUCTION

Let $\xi_1, \xi_2, \dots, \xi_n$ be n independent random variables concentrated on $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$ and

$$(1.1) \quad W_n := \sum_{i=1}^n \xi_i,$$

their convolution of n independent random variables. The distribution of W_n has received special attention in the literature due to its applicability in many settings such as rare events, the waiting time distributions, wireless communications, counts in nuclear decay, and business situations, among many others. For large values of n , it is in practice hard to obtain the exact distribution of W_n in general, in fact, it becomes intractable if the underlying distribution is complicated such as hyper-geometric and logarithmic series distribution, among many others. It is therefore of interest to approximate the distribution of such W_n with some well-known and easy to use distributions. Approximations to W_n have been studied by several authors such as, saddle point approximation (Lugannani and Rice [22] and Murakami [24]), compound Poisson approximation (Barbour *et al.* [4], Serfoso [28], and Roos [25]), Poisson approximation (Barbour *et al.* [7]), the centred Poisson approximation (Čekanavičius and Vaitkus [8]), compound negative binomial approximation (Vellaisamy and Upadhye [33]), and negative binomial approximation (Vellaisamy *et al.* [32] and Kumar and Upadhye [17]). In this article, we consider Poisson and Poisson convoluted geometric approximation to W_n . Let X and Y follow Poisson and geometric distribution with parameter λ and $p = 1 - q$ with probability mass function (PMF)

$$(1.2) \quad P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!} \quad \text{and} \quad P(Y = k) = q^k p, \quad k = 0, 1, 2, \dots,$$

respectively. Also, assume X and Y are independent. We use Stein's method to obtain bounds for the approximation of the law of W_n with that of X and $X + Y$. Stein's method (Stein [29]) requires identification of a Stein operator and there are several approaches to obtain Stein operators (see Reinert [26]) such as density approach (Stein [29], Stein *et al.* [30], Ley and Swan [19, 20]), generator approach (Barbour [2] and Götze [12]), orthogonal polynomial approach (Diaconis and Zabell [10]), and probability generating function (PGF) approach (Upadhye *et al.* [31]). We use the PGF approach to obtain Stein operators.

This article is organized as follows. In Section 2, we introduce some notations to simplify the presentation of the article. Also, we discuss some known results of Stein's method. In Section 3, Stein operators for W_n and $X + Y$ are obtained as a perturbation of the Poisson operator. In Section 4, the error bounds for X and $X + Y$ approximation to W_n are derived in total variation distance. In Section 5, we demonstrate the relevance of our results through an application to the waiting time distribution of 2-runs. In Section 6, we point out some relevant remarks.

2. NOTATIONS AND PRELIMINARIES

Recall that $W_n = \sum_{i=1}^n \xi_i$, where $\xi_1, \xi_2, \dots, \xi_n$ are n independent random variables concentrated on \mathbb{Z}_+ . Throughout, we assume that ψ_{ξ_i} , the PGF of ξ_i , satisfies

$$(2.1) \quad \frac{\psi'_{\xi_i}(w)}{\psi_{\xi_i}(w)} = \sum_{j=0}^{\infty} g_{i,j+1} w^j =: \phi_{\xi_i}(w),$$

at all $w \in \mathbb{Z}_+$. Note that this assumption is satisfied for the series (2.1) converges absolutely. Also, one can show that the hyper-geometric and logarithmic series distribution do not satisfy (2.1). See Yakshyavichus [34], and Kumar and Upadhye [17] for more details. Note that

1. If $\xi_i \sim Po(\lambda_i) \implies g_{i,j+1} = \begin{cases} \lambda_i, & \text{for } j = 0, \\ 0, & \text{for } j \geq 1. \end{cases}$
2. If $\xi_i \sim Ge(p_i) \implies g_{i,j+1} = q_i^{j+1}$.
3. If $\xi_i \sim Bi(n, p_i) \implies g_{i,j+1} = n(-1)^j (p_i/(1 - p_i))^{j+1}$.

Next, let μ and σ^2 be the mean and variance of W_n , respectively. Also, let μ_2 and μ_3 denote the second and third factorial cumulant moments of W_n , respectively. Then, it can be easily verified that

$$(2.2) \quad \begin{aligned} \mu &= \sum_{i=1}^n \phi_{\xi_i}(1) = \sum_{i=1}^n \sum_{j=0}^{\infty} g_{i,j+1}, \sigma^2 = \sum_{i=1}^n [\phi_{\xi_i}(1) + \phi'_{\xi_i}(1)] = \sum_{i=1}^n \sum_{j=0}^{\infty} (j+1)g_{i,j+1}, \\ \mu_2 &= \sum_{i=1}^n \phi'_{\xi_i}(1) = \sum_{i=1}^n \sum_{j=0}^{\infty} jg_{i,j+1}, \text{ and } \mu_3 = \sum_{i=1}^n \phi''_{\xi_i}(1) = \sum_{i=1}^n \sum_{j=0}^{\infty} j(j-1)g_{i,j+1}. \end{aligned}$$

For more details, see Vellaisamy *et al.* [32], and Kumar and Upadhye [17].

Next, let $H := \{f|f : \mathbb{Z}_+ \rightarrow \mathbb{R} \text{ is bounded}\}$ and

$$(2.3) \quad H_{\bar{X}} := \{h \in H | h(0) = 0, \text{ and } h(j) = 0 \text{ for } j \notin \text{Supp}(\bar{X})\}$$

for a random variable \bar{X} and $\text{Supp}(\bar{X})$ denotes the support of random variable \bar{X} .

Now, we discuss Stein's method which can be carried out in the following three steps.

We first identify a suitable operator $\mathcal{A}_{\bar{X}}$ for a random variable \bar{X} (known as Stein operator) such that

$$\mathbb{E}(\mathcal{A}_{\bar{X}}h(\bar{X})) = 0, \quad \text{for } h \in H.$$

In the second step, we find a solution to the Stein equation

$$(2.4) \quad \mathcal{A}_{\bar{X}}h(j) = f(j) - \mathbb{E}f(\bar{X}), \quad j \in \mathbb{Z}_+ \text{ and } f \in H_{\bar{X}}$$

and obtain the bound for $\|\Delta h\|$, where $\|\Delta h\| = \sup_{j \in \mathbb{Z}_+} |\Delta h(j)|$ and $\Delta h(j) = h(j+1) - h(j)$ denotes the first forward difference operator.

Finally, substitute a random variable \bar{Y} for j in (2.4) and taking expectation and supremum, the expression leads to

$$(2.5) \quad d_{TV}(\bar{X}, \bar{Y}) := \sup_{f \in \mathcal{H}} |\mathbb{E}f(\bar{X}) - \mathbb{E}f(\bar{Y})| = \sup_{f \in \mathcal{H}} |\mathbb{E}[\mathcal{A}_{\bar{X}}h(\bar{Y})]|,$$

where $\mathcal{H} = \{\mathbf{1}_A \mid A \subseteq \mathbb{Z}_+\}$ and $\mathbf{1}_A$ is the indicator function of A . Equivalently, (2.5) can be represented as

$$d_{TV}(\bar{X}, \bar{Y}) = \frac{1}{2} \sum_{j=0}^{\infty} |P(\bar{X} = j) - P(\bar{Y} = j)|.$$

For more details, we refer the reader to Barbour *et al.* [7], Chen *et al.* [9], Goldstein and Reinert [11], and Ross [27]. For recent developments, see Barbour and Chen [3], Ley *et al.* [21], Upadhye *et al.* [31], and references therein.

Next, it is known that a Stein operator for $X \sim Po(\lambda)$, the Poisson random variable with parameter λ , is given by

$$(2.6) \quad \mathcal{A}_X h(j) = \lambda h(j + 1) - j h(j), \quad \text{for } j \in \mathbb{Z}_+ \text{ and } h \in H.$$

Also, from Section 5 of Barbour and Eagleson [6], the bound for the solution to the stein equation (say h_f) is given by

$$(2.7) \quad \|\Delta h_f\| \leq \frac{1}{\max(1, \lambda)}, \quad \text{for } f \in \mathcal{H}, h \in H.$$

In terms of $\|f\|$, we have the following bound

$$(2.8) \quad \|\Delta h_f\| \leq \frac{2\|f\|}{\max(1, \lambda)}, \quad \text{for } f \in \mathcal{H}, h \in H.$$

See Section 3 of Upadhye *et al.* [31] for more details. Note that the condition $h(0) = 0$ in (2.3) is used while obtaining the bound (2.7), see Barbour and Eagleson [6] for more details. Next, suppose we have three random variables X_1, X_2 , and X_3 defined on some common probability space. Define $\mathcal{U} = \mathcal{A}_{X_2} - \mathcal{A}_{X_1}$ then the upper bound for $d_{TV}(X_2, X_3)$ can be obtained by the following lemma which is given by Upadhye *et al.* [31].

Lemma 2.1 (Lemma 3.1, Upadhye *et al.* [31]). *Let X_1 be a random variable with support \mathcal{S} , Stein operator \mathcal{A}_{X_1} , and h_0 be the solution to Stein equation (2.4) satisfying*

$$\|\Delta h_0\| \leq w_1 \|f\| \min(1, \alpha^{-1}),$$

where $w_1, \alpha > 0$. Also, let X_2 be a random variable whose Stein operator can be written as $\mathcal{A}_{X_2} = \mathcal{A}_{X_1} + \mathcal{U}_1$ and X_3 be a random variable such that, for $h \in H_{X_1} \cap H_{X_2}$,

$$\|\mathcal{U}_1 h\| \leq w_2 \|\Delta h\| \quad \text{and} \quad |\mathbb{E} \mathcal{A}_{X_2} h(X_3)| \leq \varepsilon \|\Delta h\|,$$

where $w_1 w_2 < \alpha$. Then

$$d_{TV}(X_2, X_3) \leq \frac{\alpha}{2(\alpha - w_1 w_2)} (\varepsilon w_1 \min(1, \alpha^{-1}) + 2P(X_2 \in \mathcal{S}^c) + 2P(X_3 \in \mathcal{S}^c)),$$

where \mathcal{S}^c denote the complement of set \mathcal{S} .

Finally, from Corollary 1.6 of Mattner and Roos [23], we have

$$(2.9) \quad d_{TV}(W_n, W_n + 1) \leq \sqrt{\frac{2}{\pi}} \left(\frac{1}{4} + \sum_{i=1}^n (1 - d_{TV}(\xi_i, \xi_i + 1)) \right)^{-1/2}.$$

For more details about these results, we refer the reader to Barbour *et al.* [5], Upadhye *et al.* [31], Vellaisamy *et al.* [32], Kumar and Upadhye [17], and references therein.

3. STEIN OPERATOR FOR THE CONVOLUTION OF RANDOM VARIABLES

In this section, we derive Stein operators for W_n and $X + Y$ as a perturbation of Poisson operator which are used to obtain the main results in Section 4.

Proposition 3.1. *Let $\xi_1, \xi_2, \dots, \xi_n$ be independent random variables satisfying (2.1) and $W_n = \sum_{i=1}^n \xi_i$. Then, a Stein operator for W_n is*

$$\mathcal{A}_{W_n} h(j) = \mu h(j + 1) - j h(j) + \sum_{i=1}^n \sum_{k=0}^{\infty} \sum_{l=1}^k g_{i,k+1} \Delta h(j + l),$$

where μ is defined in (2.2).

Proof: It can be easily verified that the PGF of W_n , denoted by ψ_{W_n} , is

$$\psi_{W_n}(w) = \prod_{i=1}^n \psi_{\xi_i}(w)$$

as $\xi_1, \xi_2, \dots, \xi_n$ are independent random variables. Differentiating with respect to w , we have

$$\begin{aligned} \psi'_{W_n}(w) &= \psi_{W_n}(w) \sum_{i=1}^n \phi_{\xi_i}(w) \\ &= \sum_{i=1}^n \psi_{W_n}(w) \sum_{j=0}^{\infty} g_{i,j+1} w^j, \end{aligned}$$

where $\phi_{\xi_i}(\cdot)$ is defined in (2.1). Using definition of the PGF, the above expression can be expressed as

$$\sum_{j=0}^{\infty} (j + 1) \gamma_{j+1} w^j = \sum_{i=1}^n \sum_{k=0}^{\infty} \gamma_k w^k \sum_{j=0}^{\infty} g_{i,j+1} w^j = \sum_{j=0}^{\infty} \left(\sum_{i=1}^n \sum_{k=0}^j \gamma_k g_{i,j-k+1} \right) w^j,$$

where $\gamma_j = P(W_n = j)$. Comparing the coefficients of w^j , we get

$$\sum_{i=1}^n \sum_{k=0}^j \gamma_k g_{i,j-k+1} - (j + 1) \gamma_{j+1} = 0.$$

Let $h \in H_{W_n}$ as defined in (2.3), then

$$\sum_{j=0}^{\infty} h(j + 1) \left[\sum_{i=1}^n \sum_{k=0}^j \gamma_k g_{i,j-k+1} - (j + 1) \gamma_{j+1} \right] = 0.$$

Therefore,

$$\sum_{j=0}^{\infty} \left[\sum_{i=1}^n \sum_{k=0}^{\infty} g_{i,k+1} h(j + k + 1) - j h(j) \right] \gamma_j = 0.$$

Hence, a Stein operator for W_n is given by

$$(3.1) \quad \mathcal{A}_{W_n}h(j) = \sum_{i=1}^n \sum_{k=0}^{\infty} g_{i,k+1}h(j+k+1) - jh(j).$$

It is well known that

$$(3.2) \quad h(j+k+1) = \sum_{l=1}^k \Delta h(j+l) + h(j+1).$$

Using (3.2) in (3.1), the proof follows. □

Proposition 3.2. *Let $X \sim Po(\lambda)$ and $Y \sim Ge(p)$ as defined in (1.2). Also, assume X and Y are independent random variables. Then a Stein operator for $X + Y$ is given by*

$$\bar{\mathcal{A}}_{X+Y}h(j) = \left(\lambda + \frac{q}{p}\right)h(j+1) - jh(j) + \sum_{k=0}^{\infty} \sum_{l=1}^k q^{k+1} \Delta h(j+l).$$

Proof: It is known that the PGF of X and Y are

$$\psi_X(w) = e^{-\lambda(1-w)} \quad \text{and} \quad \psi_Y(w) = \frac{p}{1-qw},$$

respectively. Then, the PGF of $Z = X + Y$ is given by

$$\psi_Z(w) = \psi_X(w) \cdot \psi_Y(w).$$

Differentiating with respect to w , we get

$$\psi'_Z(w) = \left(\lambda + \frac{q}{1-qw}\right)\psi_Z(w) = \left(\lambda + q \sum_{j=0}^{\infty} q^j w^j\right)\psi_Z(w), \quad |w| < q^{-1}.$$

Let $\bar{\gamma}_j = P(Z = j)$ be the PMF of Z . Then, using definition of the PGF, we have

$$\sum_{j=0}^{\infty} (j+1)\bar{\gamma}_{j+1}w^j = \lambda \sum_{j=0}^{\infty} \bar{\gamma}_j w^j + \sum_{j=0}^{\infty} q^{j+1}w^j \sum_{k=0}^{\infty} \bar{\gamma}_k w^k.$$

This implies

$$\sum_{j=0}^{\infty} (j+1)\bar{\gamma}_{j+1}w^j - \lambda \sum_{j=0}^{\infty} \bar{\gamma}_j w^j - \sum_{j=0}^{\infty} \left(\sum_{k=0}^j \bar{\gamma}_k q^{j-k+1}\right) w^j = 0.$$

Collecting the coefficients of w^j , we get

$$(j+1)\bar{\gamma}_{j+1} - \lambda\bar{\gamma}_j - \sum_{k=0}^j \bar{\gamma}_k q^{j-k+1} = 0.$$

Let $h \in H_Z$ as defined in (2.3), then

$$\sum_{j=0}^{\infty} h(j+1) \left[\lambda\bar{\gamma}_j - (j+1)\bar{\gamma}_{j+1} + \sum_{k=0}^j \bar{\gamma}_k q^{j-k+1}\right] = 0.$$

Further simplification leads to

$$\sum_{j=0}^{\infty} \left[\lambda h(j+1) - jh(j) + \sum_{k=0}^{\infty} q^{k+1} h(j+k+1) \right] \bar{\gamma}_j = 0.$$

Therefore,

$$\bar{A}_{X+Y} h(j) = \lambda h(j+1) - jh(j) + \sum_{k=0}^{\infty} q^{k+1} h(j+k+1).$$

Using (3.2), the proof follows. □

4. APPROXIMATION RESULTS

In this section, we derive an error bound for the Poisson and Poisson convoluted geometric approximation to W_n . The following theorem gives the bound for Poisson, with parameter μ , approximation.

Theorem 4.1. *Let $\xi_1, \xi_2, \dots, \xi_n$ be independent random variables satisfying (2.1) and $W_n = \sum_{i=1}^n \xi_i$. Then*

$$d_{TV}(W_n, X) \leq \frac{|\mu_2|}{\max(1, \mu)},$$

where $X \sim Po(\mu)$.

Proof: From Proposition 3.1, a Stein operator for W_n is given by

$$\begin{aligned} \mathcal{A}_{W_n} h(j) &= \mu h(j+1) - jh(j) + \sum_{i=1}^n \sum_{k=0}^{\infty} \sum_{l=1}^k g_{i,k+1} \Delta h(j+l) \\ &= \mathcal{A}_X h(j) + \mathcal{U}_{W_n} h(j), \end{aligned}$$

where \mathcal{A}_X is a Stein operator for X as discussed in (2.6). Observe that \mathcal{A}_{W_n} is a Stein operator for W_n which can be seen as a perturbation of Poisson operator. Now, for $h \in H_X \cap H_{W_n}$, taking expectation of perturbed operator \mathcal{U}_{W_n} with respect to W_n and using (2.7), the result follows. □

Next, we derive $Z = X + Y$ approximation to W_n , where $X \sim Po(\lambda)$ and $Y \sim Ge(p)$, by matching first two moments, that is, $\mathbb{E}(Z) = \mathbb{E}(W_n)$ and $\text{Var}(Z) = \text{Var}(W_n)$ which give the following choice of parameters

$$(4.1) \quad \lambda = \mu - \sqrt{\sigma^2 - \mu} \quad \text{and} \quad p = \frac{1}{1 + \sqrt{\sigma^2 - \mu}}.$$

Theorem 4.2. *Let $\xi_1, \xi_2, \dots, \xi_n$ be independent random variables satisfying (2.1) and the mean and variance of $W_n = \sum_{i=1}^n \xi_i$ satisfying (4.1). Also, assume that $\sigma^2 > \mu$ and $\lambda > 2(q/p)^2$. Then*

$$d_{TV}(W_n, Z) \leq \frac{\lambda \sqrt{\frac{2}{\pi}} \left| \mu_3 - 2(q/p)^3 \right| \left(\frac{1}{4} + \sum_{i=1}^n (1 - d_{TV}(\xi_i, \xi_i + 1)) \right)^{-1/2}}{\left(\lambda - 2(q/p)^2 \right) \max(1, \lambda)},$$

where $Z = X + Y$, $X \sim Po(\lambda)$ and $Y \sim Ge(p)$.

Remark 4.1. Note that, in Theorem 4.2, the choice of parameters are valid as

$$\mu = \lambda + \frac{q}{p} > \frac{q}{p} = \sqrt{\sigma^2 - \mu} \quad \text{and} \quad p = \frac{1}{1 + \sqrt{\sigma^2 - \mu}} \leq 1,$$

since $\sigma^2 > \mu$.

Proof of Theorem 4.2: From (3.1), the Stein operator for W_n is given by

$$A_{W_n} h(j) = \sum_{i=1}^n \sum_{k=0}^{\infty} g_{i,k+1} h(j+k+1) - jh(j).$$

Using (3.2), with $\sum_{i=1}^n \sum_{k=0}^{\infty} g_{i,k+1} = \mathbb{E}(W_n) = \mathbb{E}(Z) = \lambda + q/p$, we get

$$\begin{aligned} A_{W_n} h(j) &= \left(\lambda + \frac{q}{p}\right) h(j+1) - jh(j) + \sum_{k=0}^{\infty} \sum_{l=1}^k q^{k+1} \Delta h(j+l) \\ &\quad + \sum_{i=1}^n \sum_{k=0}^{\infty} \sum_{l=1}^k g_{i,k+1} \Delta h(j+l) - \sum_{k=0}^{\infty} \sum_{l=1}^k q^{k+1} \Delta h(j+l) \\ &= A_Z h(j) + \bar{U}_{W_n} h(j). \end{aligned}$$

This is a Stein operator for W_n which can be seen as perturbation of $Z = X + Y$ operator, obtained in Proposition 3.2. Now, consider

$$(4.2) \quad \bar{U}_{W_n} h(j) = \sum_{i=1}^n \sum_{k=0}^{\infty} \sum_{l=1}^k g_{i,k+1} \Delta h(j+l) - \sum_{k=0}^{\infty} \sum_{l=1}^k q^{k+1} \Delta h(j+l).$$

We know that

$$\Delta h(j+l) = \sum_{m=1}^{l-1} \Delta^2 h(j+m) + \Delta h(j+1).$$

Substituting in (4.2) and using $\text{Var}(Z) = \text{Var}(W_n)$ with $\sum_{i=1}^n \sum_{k=0}^{\infty} g_{i,k+1} = \mathbb{E}(W_n) = \mathbb{E}(Z) = \lambda + q/p$, we have

$$\bar{U}_{W_n} h(j) = \sum_{i=1}^n \sum_{k=0}^{\infty} \sum_{l=1}^k \sum_{m=1}^{l-1} g_{i,k+1} \Delta^2 h(j+m) - \sum_{k=0}^{\infty} \sum_{l=1}^k \sum_{m=1}^{l-1} q^{k+1} \Delta^2 h(j+m).$$

Now, taking expectation with respect to W_n , we get

$$\begin{aligned} \mathbb{E}[\bar{U}_{W_n} h(W_n)] &= \sum_{j=0}^{\infty} \left[\sum_{i=1}^n \sum_{k=0}^{\infty} \sum_{l=1}^k \sum_{m=1}^{l-1} g_{i,k+1} \Delta^2 h(j+m) \right. \\ &\quad \left. - \sum_{k=0}^{\infty} \sum_{l=1}^k \sum_{m=1}^{l-1} q^{k+1} \Delta^2 h(j+m) \right] P[W_n = j]. \end{aligned}$$

Therefore,

$$\begin{aligned} |\mathbb{E}[\bar{U}_{W_n} h(W_n)]| &\leq 2d_{TV}(W_n, W_n + 1) \|\Delta h\| \left| \sum_{i=1}^n \sum_{k=0}^{\infty} \frac{k(k-1)}{2} g_{i,k+1} - \sum_{k=0}^{\infty} \frac{k(k-1)}{2} q^{k+1} \right| \\ &\leq d_{TV}(W_n, W_n + 1) \|\Delta h\| \left| \mu_3 - 2 \frac{q^3}{p^3} \right|. \end{aligned}$$

Using (2.9), we have

$$(4.3) \quad \left| \mathbb{E}[\mathcal{U}_{W_n} h(W_n)] \right| \leq \|\Delta h\| \sqrt{\frac{2}{\pi}} \left(\frac{1}{4} + \sum_{i=1}^n (1 - d_{TV}(\xi_i, \xi_i + 1)) \right)^{-1/2} \left| \mu_3 - 2 \frac{q^3}{p^3} \right|.$$

From Proposition 3.2, we have

$$(4.4) \quad \|\mathcal{U}_{X+Y} h\| \leq \frac{q^2}{p^2} \|\Delta h\|.$$

Using (2.8), (4.3), and (4.4) with Lemma 2.1, the proof follows. \square

5. AN APPLICATION TO THE WAITING TIME DISTRIBUTION OF 2-RUNS

The concept of runs and patterns is well-known in the literature due to its applicability in many real-life applications such as reliability theory, machine maintenance, statistical testing, and quality control, among many others. In this section, we consider the set up discussed by Hirano [13] and generalized by Huang and Tsai [15] as follows:

Let N denote the number of two consecutive successes in n Bernoulli trials with success probability p . Then, Huang and Tsai [15] (with $k_1 = 0$ and $k_2 = 2$ in their notation) have shown that the waiting time for n -th occurrence of 2-runs can be written as the sum of n independent and identical distributed (iid) random variables, say U_1, U_2, \dots, U_n , concentrated on $\{2, 3, \dots\}$. Here U_i is 2 plus the number of trials between the $(j - 1)$ -th and j -th occurrence of 2-runs. The PGF of U_i is given by

$$\psi_U(t) = \frac{p^2 t^2}{1 - t + p^2 t^2},$$

where U is the iid copy of $U_i, i = 1, 2, \dots, n$ (see Hung and Tsai [15] for more details).

Now, let $V_i = U_i - 2$ concentrated on \mathbb{Z}_+ . Then, Kumar and Upadhye [17] have given the PGF of V_i and which is given by

$$\psi_{V_i}(t) = \frac{p^2}{1 - t + p^2 t^2} = \sum_{j=0}^{\infty} \left(\sum_{\ell=0}^{\lfloor j/2 \rfloor} \binom{j-\ell}{\ell} (-1)^\ell p^{2(\ell+1)} \right) t^j = \sum_{j=0}^{\infty} g_{i,j+1} t^j,$$

where $g_{i,j+1} = \sum_{\ell=0}^{\lfloor j/2 \rfloor} \binom{j-\ell}{\ell} (-1)^\ell p^{2(\ell+1)}$, for each $i = 1, 2, \dots, n$. For more details, we refer the reader to Huang and Tsai [15], Kumar and Upadhye [17], and Balakrishnan and Koutras [1], and references therein.

Now, let $W_{\bar{n}} = \sum_{i=1}^{\bar{n}} V_i$ then $W_{\bar{n}}$ denotes the number of failures before \bar{n} th occurrence of 2-runs. Therefore, from Theorem 4.1, we have

$$d_{TV}(W_{\bar{n}}, Po(\mu)) \leq \frac{|\mu_2|}{\max(1, \mu)},$$

where $\mu = \bar{n} \sum_{j=0}^{\infty} g_{i,j+1}$ and $\mu_2 = \bar{n} \sum_{j=0}^{\infty} j g_{i,j+1}$. In a similar manner, from Theorem 4.2, we can also obtain the bound for the Poisson convoluted geometric approximation. For more details, we refer the reader to Section 4 of Kumar and Upadhye [17].

6. CONCLUDING REMARKS

1. Note that, if $\xi_i \sim Po(\lambda_i)$, $i = 1, 2, \dots, n$ then $d_{TV}(W_n, X) = 0$ in Theorem 4.1, as expected.
2. If $\xi_1 \sim Po(\lambda)$ and $\xi_2 \sim Ge(p)$, for $i = 1, 2$, and $W_2 = \xi_1 + \xi_2$ then $d_{TV}(W_2, Z) = 0$ in Theorem 4.2, as expected.
3. The bounds obtained in Theorems 4.1 and 4.2 are either comparable to or improvement over the existing bounds available in the literature. In particular, some comparison can be seen as follows:
 - (a) If $\xi_i \sim Ber(p_i)$, for $i = 1, 2, \dots, n$ then, from Theorem 4.1, we have

$$d_{TV}(W_n, Po(\mu)) \leq \frac{1}{\max(1, \mu)} \sum_{i=1}^n p_i^2,$$

where $\mu = \sum_{i=1}^n p_i$. The above bound is same as given by Barbour *et al.* [7] and is an improvement over the bound $d_{TV}(W_n, Po(\mu)) \leq \sum_{i=1}^n p_i^2$ given by Khintchine [16] and Le Cam [18].

- (b) If $\xi_i \sim Ge(p_i)$, $i = 1, 2, \dots, n$ then, from Theorem 4.1, we have

$$d_{TV}(W_n, X) \leq \frac{1}{\max(1, \mu)} \sum_{i=1}^n \left(\frac{q_i}{p_i}\right)^2.$$

This bound is an improvement over negative binomial approximation given by Kumar and Upadhye [17] in Corollary 3.1.

- (c) If $\xi_i \sim NB(\alpha_i, p_i)$, $i = 1, 2, \dots, n$ then, from Theorems 4.1, we have

$$(6.1) \quad d_{TV}(W_n, Po(\mu)) \leq \frac{1}{\max(1, \mu)} \sum_{i=1}^n \alpha_i \left(\frac{q_i}{p_i}\right)^2,$$

where $\mu = \sum_{i=1}^n \frac{\alpha_i q_i}{p_i}$. Vellaisamy and Upadhye [33] obtained bound for $S_n = \sum_{i=1}^n \xi_i$ and is given by

$$(6.2) \quad d_{TV}(S_n, Po(\lambda)) \leq \min\left(1, \frac{1}{\sqrt{2\lambda e}}\right) \sum_{i=1}^n \frac{\alpha_i q_i^2}{p_i},$$

where $\lambda = \sum_{i=1}^n \alpha_i q_i = \alpha q$. Under identical set up with $\alpha = 5$ and various values of n and q , the numerical comparison of (6.1) and (6.2) as follows:

Table 1: Comparison of bounds.

n	q	From (6.1)	From (6.2)
10	0.1	0.1111	0.3370
30		0.1111	1.0109
50		0.1111	1.6848
10	0.2	0.2500	1.0722
30		0.2500	3.2166
50		0.2500	5.3610

Note that our bound (from (6.1)) is better than the bound given in (6.2). In particular, graphically, the closeness of these two distributions can be seen as follows:

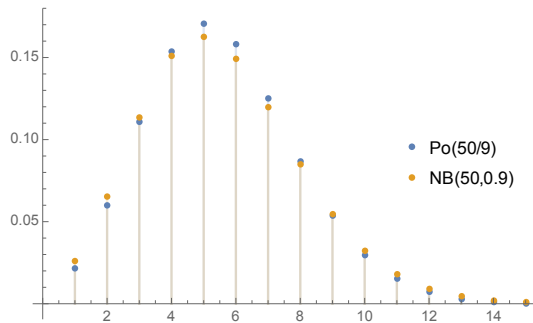


Figure 1: $n = 10, q = 0.1$.

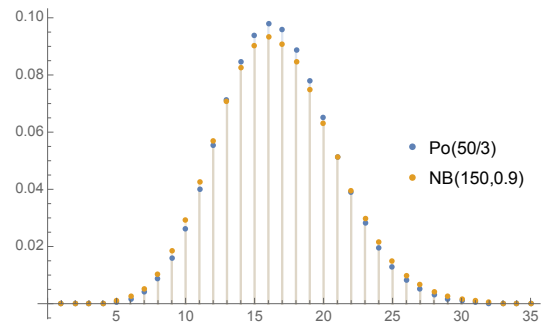


Figure 2: $n = 30, q = 0.1$.

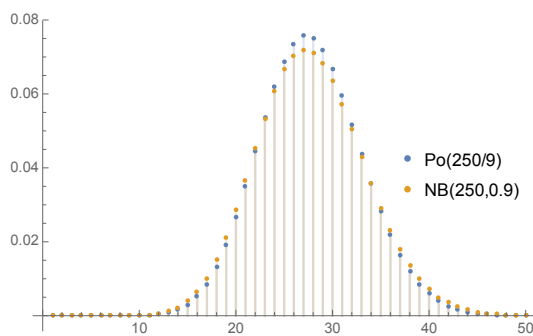


Figure 3: $n = 50, q = 0.1$.

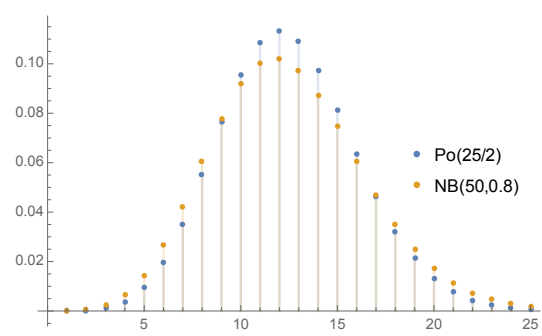


Figure 4: $n = 10, q = 0.2$.

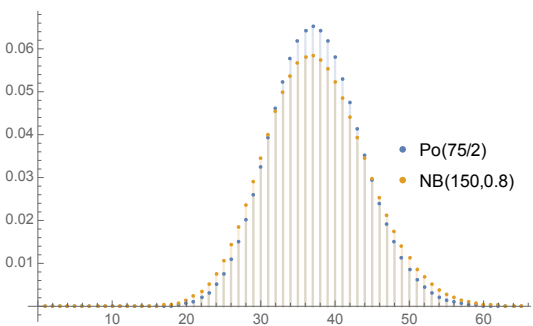


Figure 5: $n = 30, q = 0.2$.

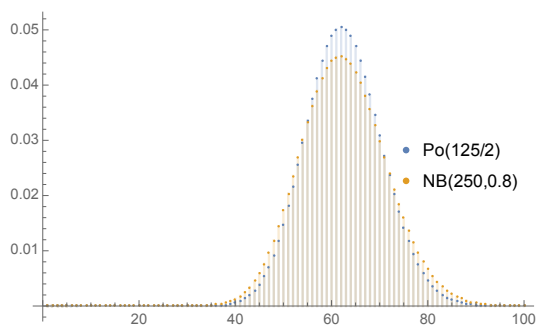


Figure 6: $n = 50, q = 0.2$.

The above graphs are obtained by using the moment matching conditions. Also, from the numerical table and graphs, observe that the distributions are closer for sufficiently small values of q and large values of n , as expected.

(d) From Theorem 1 of Hung and Giang [14], it is given that, for $A \subset \mathbb{Z}_+$,

$$(6.3) \quad \sup_A \left| P(W_n \in A) - \sum_{k \in A} \frac{\lambda_n^k e^{-\lambda_n}}{k!} \right| \leq \sum_{i=1}^n \min \left\{ \lambda_n^{-1} (1 - e^{-\lambda_n}) r_{n,i} (1 - p_{n,i}), 1 - p_{n,i} \right\} (1 - p_{n,i}) p_{n,i}^{-1},$$

where $W_n = \sum_{i=1}^n X_{n,i}$, $X_{n,i} \sim NB(r_{n,i}, p_{n,i})$ with $\lambda_n = \mathbb{E}(W_n)$. Note that if $\min \left\{ \lambda_n^{-1} (1 - e^{-\lambda_n}) r_{n,i} (1 - p_{n,i}), 1 - p_{n,i} \right\} = 1 - p_{n,i}$, for all $i = 1, 2, \dots, n$, then

$$(6.4) \quad \sup_A \left| P(W_n \in A) - \sum_{k \in A} \frac{\lambda_n^k e^{-\lambda_n}}{k!} \right| \leq \sum_{i=1}^n (1 - p_{n,i})^2 p_{n,i}^{-1},$$


which is of order $O(n)$. Clearly, for large values of n , Theorem 4.1 is an improvement over (6.4).


REFERENCES



- [1] BALAKRISHNAN, N. and KOUTRAS, M.V. (2002). *Runs and Scans with Applications*, John Wiley, New York.
- [2] BARBOUR, A.D. (1990). Stein's method for diffusion approximations, *Probab. Theory and Related Fields*, **84**(3), 297–322.
- [3] BARBOUR, A.D. and CHEN, L.H.Y. (2014). Stein's (magic) method, *Preprint: arXiv:1411.1179*.
- [4] BARBOUR, A.D.; CHEN, L.H.Y. and LOH, W.L. (1992). Compound Poisson approximation for nonnegative random variables via Stein's method, *Ann. Prob.*, **20**(4), 1843–1866.
- [5] BARBOUR, A.D.; ČEKANAČIUS, V. and XIA, A. (2007). On Stein's method and Perturbations, *ALEA*, **3**, 31–53.
- [6] BARBOUR, A.D. and EAGLESON, G.K. (1983). Poisson approximation for some statistics based on exchangeable trials, *Adv. Appl. Prob.*, **15**(3), 585–600.
- [7] BARBOUR, A.D.; HOLST, L. and JANSON, S. (1992). *Poisson Approximation*, Oxford: Clarendon Press.
- [8] ČEKANAČIUS, V. and VAITKUS, P. (2001). Centered Poisson approximation via Stein's method, *Lith. Math. J.*, **41**, 319–329.
- [9] CHEN, L.H.Y.; GOLDSTEIN, L. and SHAO, Q.-M. (2011). *Normal Approximation by Stein's Method*, Springer, Heidelberg.
- [10] DIACONIS, P. and ZABELL, S. (1991). Closed form summation for classical distributions. Variations on a theme of de Moivre, *Statist. Sci.*, **6**(43), 284–302.
- [11] GOLDSTEIN, L. and REINERT, G. (2005). Distributional transformations, orthogonal polynomials and Stein characterizations, *J. Theoret. Probab.*, **18**(1), 237–260.
- [12] GÖTZE, F. (1991). On the rate of convergence in the multivariate CLT, *Ann. Prob.*, **19**(2), 724–739.


- [13] HIRANO, K. (1984). *Some properties of the distributions of order k* . In “Proceedings of the First International Conference on Fibonacci Numbers and Their Applications (Gutenberg, Athens)” (A.N. Philippou and A.F. Horadam, Eds.), 43–53.
- [14] HUNG, T.L. and GIANG, L.T. (2016). On bounds in Poisson approximation for distributions of independent negative-binomial distributed random variables, *Springerplus*, **5**, 79.
- [15] HUANG, W.T. and TSAI, C.S. (1991). On a modified binomial distribution of order k , *Statist. Probab. Lett.*, **11**(2), 125–131.
- [16] KHINTCHINE, A.YA. (1933). *Asymptotische Gesetze der Wahrscheinlichkeitsrechnung*, Springer-Verlag, Berlin.
- [17] KUMAR, A.N. and UPADHYE, N.S. (2017). On perturbations of Stein operator, *Comm. Statist. Theory Methods*, **46**(18), 9284–9302.
- [18] LE CAM, L. (1960). An approximation theorem for the Poisson binomial distribution, *Pacific J. Math.*, **10**, 1181–1197.
- [19] LEY, C. and SWAN, Y. (2013). Stein’s density approach and information inequalities, *Electron. Commun. Probab.*, **18**(7), 1–14.
- [20] LEY, C. and SWAN, Y. (2013). Local Pinsker inequalities via Stein’s discrete density approach, *IEEE Trans. Inform. Theory*, **59**(9), 5584–5591.
- [21] LEY, C.; REINERT, G. and SWAN, Y. (2014). Stein’s method for comparison of univariate distributions, *Probab. Surv.*, **14**, 1–52.
- [22] LUGANNANI, R. and RICE, S.O. (1980). Saddle point approximation for the distribution of the sum of independent random variables, *Adv. Appl. Probab.*, **12**(2), 475–490.
- [23] MATTNER, L. and ROOS, B. (2007). A shorter proof of Kanter’s Bessel function concentration bound, *Probab. Theory and Related Fields*, **139**(1–2), 407–421.
- [24] MURAKAMI, H. (2015). Approximations to the distribution of sum of independent non-identically gamma random variables, *Math. Sci.*, **9**, 205–213.
- [25] ROOS, B. (2003). Kerstan’s method for compound Poisson approximation, *Ann. Probab.*, **9**(41), 1754–1771.
- [26] REINERT, G. (2005). *Three general approaches to Stein’s method*. In “An Introduction to Stein’s Method”, **4**, Singapore University Press, Singapore, 183–221.
- [27] ROSS, N. (2011). Fundamentals of Stein’s method, *Probab. Surv.*, **8**, 210–293.
- [28] SERFOZO, R.F. (1986). Compound Poisson approximation for sum of random variables, *Ann. Probab.*, **14**(4), 1391–1398.
- [29] STEIN, C. (1986). Approximate computation of expectations, *Institute of Mathematical Statistics Lecture Notes – Monograph Series*, **7**, Institute of Mathematical Statistics, Hayward, CA.
- [30] STEIN, C.; DIACONIS, P.; HOLMES, S. and REINERT, G. (2004). *Use of exchangeable pairs in the analysis of simulations*. In “Stein’s Method: Expository Lectures and Applications” (P. Diaconis and S. Holmes, Eds.), IMS Lecture Notes Monogr. Ser 46, 1–26, Beachwood, Ohio, USA: Institute of Mathematical Statistics.
- [31] UPADHYE, N.S.; ČEKANAČIUS, V. and VELLAISAMY, P. (2014). On Stein operators for discrete approximations, *Bernoulli*, **23**(4A), 2828–2859.
- [32] VELLAISAMY, P.; UPADHYE, N.S. and ČEKANAČIUS, V. (2013). On negative binomial approximation, *Theory Probab. Appl.*, **57**(1), 97–109.
- [33] VELLAISAMY, P. and UPADHYE, N.S. (2009). Compound negative binomial approximations for sums of random variables, *Probab. Math. Statist.*, **29**(2), 205–226.
- [34] YAKSHYAVICHUS, SH. (1998). On a method of expansion of the probabilities of lattice random variables, *Theory Probab. Appl.*, **42**(2), 271–282.

Modeling Heavy-Tailed Bounded Data by the Trapezoidal Beta Distribution with Applications

Authors: JORGE I. FIGUEROA-ZÚÑIGA 
– Department of Statistics, Universidad de Concepción,
Concepción, Chile
jifiguer@gmail.com

SEBASTIÁN NIKLITSCHK-SOTO 
– Department of Statistics, Universidad de Concepción,
Concepción, Chile
sniklitschek@udec.cl

VÍCTOR LEIVA  
– School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso,
Valparaíso, Chile
victorleivasanchez@gmail.com

SHUANGZHE LIU 
– Faculty of Science and Technology, University of Canberra,
Canberra, Australia
shuangzhe.liu@canberra.edu.au

Received: June 2020

Revised: August 2020

Accepted: August 2020

Abstract:


- In this paper, by using a new method, we derive the trapezoidal beta (TB) distribution and its properties. The TB distribution is a mixture model, generalizes both the beta and rectangular beta distributions, and allows one to describe bounded data with heavy right and/or left tails. In relation to the two-parameter beta distribution, we add two additional parameters which have an intuitive interpretation. The four TB parameters are estimated with the expectation-maximization algorithm. We conduct a simulation study to evaluate performance of the TB distribution. An application with real data is carried out, which includes a comparison among the beta, rectangular beta and TB distributions indicating that the TB one describes these data better.

Keywords:

- *bounded-support distributions; EM algorithm; mixture distributions; R software; trapezoidal distributions.*

AMS Subject Classification:

- 60E05, 62E15.

 Corresponding author.

1. INTRODUCTION

Distribution theory is an emerging field of statistics which has received an increasing attention recently, with different methods that have been proposed to generate new distributions; see [4, 21, 23, 24, 45]. To the best of our knowledge, the method used in the present work has not been previously considered.

When modeling continuous data restricted to a bounded interval, the beta distribution is a natural choice providing a wide variety of shapes; see [12]. Some of its extensions, derived by using general classes of distributions, are the beta-Gumbel [37], beta-Fréchet [36], beta-exponential [38], beta-Pareto [1], beta-generalized-exponential [3], beta-normal [11], beta-power [8], beta-Marshall-Olkin [2], and beta-Marshall-Olkin-Lomax [45] distributions. These extensions of the beta distribution have provided good fits to different types of data. However, all of such extensions lose the essence of the beta distribution of having its support in the unit interval, that is, to model data between zero and one.

An alternative to the beta distribution is a double-bounded distribution first defined in [25] and after named the Kumaraswamy distribution in [22]. The cumulative distribution function (CDF) of the Kumaraswamy distribution has a closed analytical form. Some of its extensions are the Kumaraswamy-G [6], Kumaraswamy-Gumbel [7], Kumaraswamy-Weibull [9], Kumaraswamy-generalized-gamma [10], and trapezoidal-Kumaraswamy [42] distributions. The extensions of the Kumaraswamy distribution include additional parameters, are able to model bathtub-shaped hazard rates, and are widely applied in engineering.

In general, as mentioned, the beta distribution is very flexible and often employed in practice. However, it is common in many cases to have bounded data which follow heavy left-and-right tailed distributions. Therefore, as noted in [14, 18], the beta and Kumaraswamy distributions, as well as their extensions above mentioned, are not suitable to model heavy tails. In order to add flexibility into the beta distribution, the rectangular beta (RB) distribution was proposed in [18]. In practice, the beta and RB distributions have been powerful tools for modeling bounded data, but the RB distribution permits the modeling of heavy-tailed bounded data in equal proportions in both tails. An approach to solve the above mentioned limitations was presented in [19], but the parameters of such an approach do not have a clear interpretation and there is no an efficient method for estimating these parameters. Another attempt for obtaining alternative beta distributions is provided in [24]. To the best of our knowledge, there is no distributions that allow the modeling of heavy left-and-right tailed bounded data in different proportions.

The objective of this paper is to propose a bounded-support distribution based on a novel method to circumvent the above-mentioned limitations. This new distribution is the trapezoidal beta (TB) model, which has high flexibility to describe the tails in different proportions for its probability density function (PDF). The TB distribution is a mixture model, extends both the beta and rectangular beta distributions, and permits one to describe bounded data with heavy right and/or left tails in different proportions. We estimate the TB distribution parameters by using the maximum likelihood method. We take advantage of the finite mixture representation of the TB distribution to implement the expectation-maximization (EM) algorithm. This algorithm has two main steps: the expectation (E) step and the maximization (M) step.

The EM algorithm is a widely applicable approach to the iterative computation of maximum likelihood estimates, which is useful in a variety of incomplete data settings. The idea behind the EM algorithm applied to mixture models is to assume that the mixture is generated by missing observations. For more details of this algorithm, see [34].

The rest of the paper is organized as follows. In Section 2, we provide background of the beta and RB distributions and propose the new TB distribution specifying its mathematical properties. In addition, in this section, a shape analysis is performed to show the flexibility of the TB distribution graphically. Section 3 describes a methodology to estimate the TB distribution parameters based on the EM algorithm. In Section 4, the proposed distribution is evaluated throughout Monte Carlo simulation studies. A comparison of the proposed distribution and the beta and RB distributions is also conducted in this section. Furthermore, we include an empirical illustration with education data corresponding to a university selection score of 1295 institutions in the Metropolitan region of Chile. Finally, some concluding remarks and possible directions for future research are given in Section 5.

2. THE NEW DISTRIBUTION

In this section, background with respect to the beta and RB distribution is provided and then the proposed TB distribution is derived specifying its mathematical properties and a shape analysis to graphically show the flexibility of the TB distribution.

2.1. Background

Let Y follow a beta distribution of parameters $\alpha > 0$ and $\beta > 0$, which we denote by $Y \sim \text{Beta}(\alpha, \beta)$. The PDF of Y is given by

$$(2.1) \quad f_Y(y; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1 - y)^{\beta-1}, \quad 0 < y < 1,$$

where Γ is the gamma function. The mean and variance of Y are established respectively as

$$(2.2) \quad \begin{aligned} E(Y) &= \frac{\alpha}{\alpha + \beta}, \\ \text{Var}(Y) &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \end{aligned}$$

In order to add flexibility into the beta distribution, the RB distribution was proposed. If a random variable Z follows an RB distribution of parameters $0 \leq \theta \leq 1$, $\alpha > 0$ and $\beta > 0$, the notation $Z \sim \text{RB}(\theta, \alpha, \beta)$ is adopted. The PDF of Z is stated as

$$(2.3) \quad f_Z(z; \theta, \alpha, \beta) = \theta + (1 - \theta)f_Y(z; \alpha, \beta), \quad 0 < z < 1,$$

where θ is a mixture parameter. From (2.2) and (2.3), we obtain that

$$(2.4) \quad \begin{aligned} E(Z) &= \frac{\theta}{2} + (1 - \theta)\frac{\alpha}{\alpha + \beta}, \\ \text{Var}(Z) &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}(1 - \theta)(1 - \theta(1 + (\alpha + \beta))) + \frac{\theta}{12}(4 - 3\theta). \end{aligned}$$

By taking $\theta = 1$ and $\theta = 0$ in the RB distribution, we get the uniform and beta distributions, so that its mean and variance are given in (2.4). The RB distribution permits one to model heavy-tailed bounded data in equal proportions on both tails as noted in the shape analysis; see Figure 1(a).

2.2. The trapezoidal beta distribution

Consider a non-negative polynomial P such that $0 \leq \int_0^1 P(t; a, b) dt \leq 1$. By choosing $P(t; a, b) = a + (b - a)t$, the PDF of the TB distribution is obtained as

$$\begin{aligned}
 f_T(t; a, b, \alpha, \beta) &= a + (b - a)t + \left(1 - \int_0^1 (a + (b - a)t) dt\right) f_Y(t; \alpha, \beta) \\
 (2.5) \qquad \qquad \qquad &= a + (b - a)t + \left(1 - \frac{a + b}{2}\right) f_Y(t; \alpha, \beta), \quad 0 < t < 1,
 \end{aligned}$$

with $0 \leq a, b \leq 2$, $0 \leq a + b \leq 2$, and f_Y being the beta PDF of parameters α and β as defined in (2.1). In this case, the notation $T \sim \text{TB}(a, b, \alpha, \beta)$ is used. Note that the TB PDF defined by (2.5) can be rewritten as a mixture of three beta distributions by considering

$$\begin{aligned}
 (2.6) \qquad f_T(t; a, b, \alpha, \beta) &= \omega_1 f_1(t) + \omega_2 f_2(t) + \omega_3 f_3(t), \\
 &= \frac{a}{2}(2 - 2t) + \frac{b}{2}(2t) + \left(1 - \frac{a + b}{2}\right) f_Y(t; \alpha, \beta),
 \end{aligned}$$

where $f_1(t) = f_Y(t; 1, 2) = 2 - 2t$, $f_2(t) = f_Y(t; 2, 1) = 2t$ and $f_3(t) = f_Y(t; \alpha, \beta)$ correspond to particular cases of the beta PDF described in (2.1). In addition,

$$(2.7) \qquad \qquad \qquad \omega_1 = \frac{a}{2}, \quad \omega_2 = \frac{b}{2}, \quad \omega_3 = \left(1 - \frac{a + b}{2}\right)$$

are the weights such that $\omega_1 + \omega_2 + \omega_3 = 1$ and $0 \leq \omega_1, \omega_2, \omega_3 \leq 1$.

We now present some properties of the TB distribution. Let $T \sim \text{TB}(a, b, \alpha, \beta)$. Then, the k -th moment of T is given by

$$(2.8) \qquad \qquad \qquad m_k = E(T^k) = \frac{a}{k + 1} + \frac{b - a}{k + 2} + \left(1 - \frac{a + b}{2}\right) m_k^*,$$

where m_k^* is the k -th moment of the $\text{Beta}(\alpha, \beta)$ distribution. Thus, from (2.8), we have

$$(2.9) \qquad \qquad \qquad m_k = \frac{a}{k + 1} + \frac{b - a}{k + 2} + \left(1 - \frac{a + b}{2}\right) \left(\prod_{r=0}^{k-1} \frac{\alpha + r}{\alpha + \beta + r}\right).$$

In addition, the moment generating and characteristic functions of $T \sim \text{TB}(a, b, \alpha, \beta)$ are stated respectively as

$$\begin{aligned}
 (2.10) \qquad M_T(v) &= E(e^{vT}) = 1 + \sum_{k=1}^{\infty} m_k \frac{v^k}{k!}, \quad v \in \mathbb{R}, \\
 \varphi_T(v) &= E(e^{ivT}) = 1 + \sum_{k=1}^{\infty} m_k \frac{(iv)^k}{k!}, \quad v \in \mathbb{R}.
 \end{aligned}$$

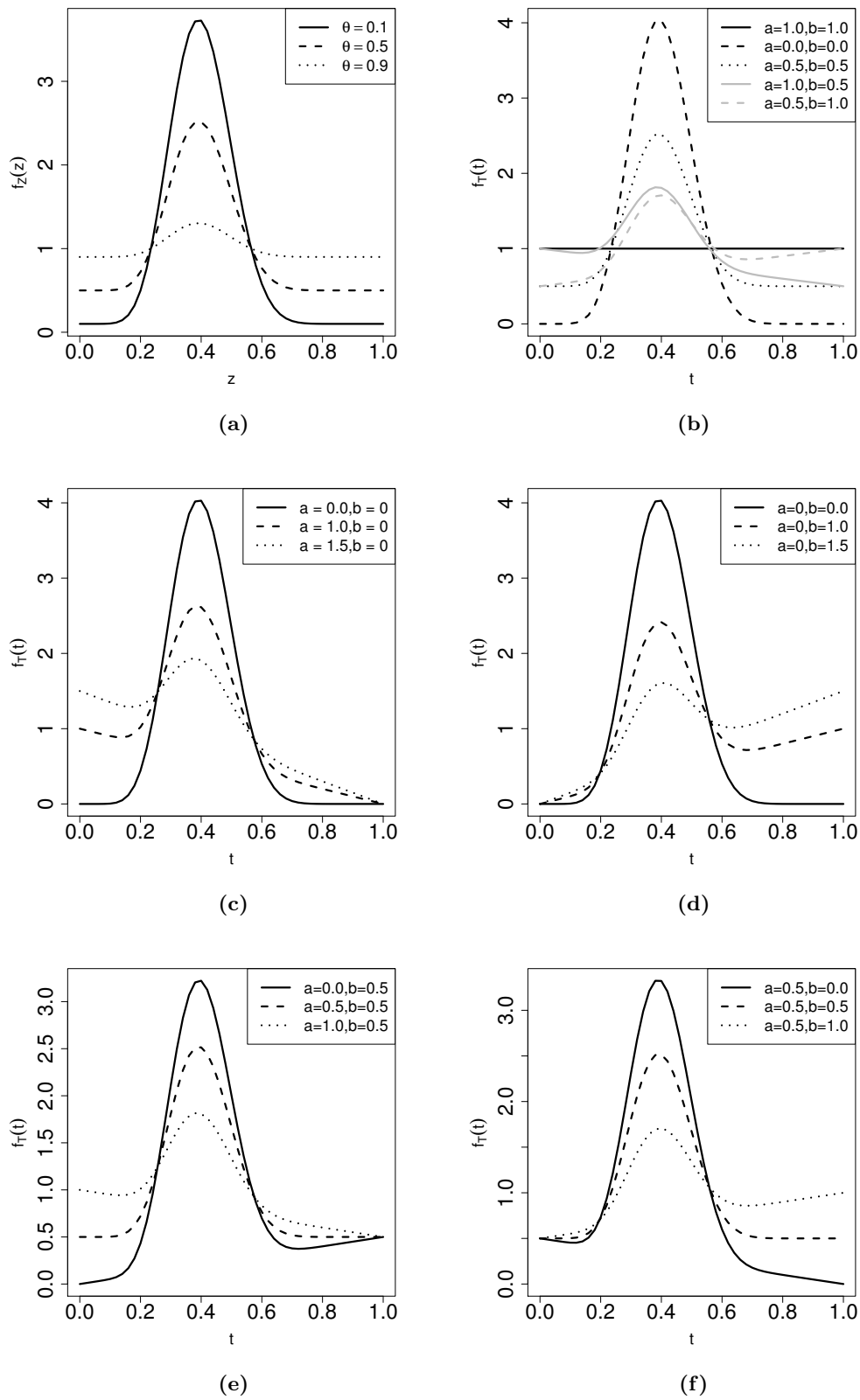


Figure 1: Plots of the (a) RB($\theta, \alpha = 10, \beta = 15$) PDF with θ as indicated, and (b)–(f) TB($a, b, \alpha = 10, \beta = 15$) PDF with a, b as listed.

Based on (2.9) or (2.10), we deduce that the mean and variance of T are given respectively as

$$(2.11) \quad E(T) = \frac{a + 2b}{6} + \left(1 - \frac{a + b}{2}\right) \frac{\alpha}{\alpha + \beta},$$

$$\text{Var}(T) = \left(\frac{3a + 9b - (a + 2b)^2}{36}\right) + \left(\frac{\alpha}{\alpha + \beta}\right) \left(1 - \frac{a + b}{2}\right) \left(\frac{\alpha + 1}{\alpha + \beta + 1} - \frac{\alpha(2 - a - b)}{2(\alpha + \beta)} - \frac{a + 2b}{3}\right).$$

Note that taking $a = b = 0$ (beta distribution), and $a = b = \theta$ (RB distribution) in (2.11), the mean and variance established in (2.2) and (2.4) are obtained, respectively.

Figure 1(a) shows how the RB distribution allow us to model heavy tails in equal proportions in both tails, but not in different proportions, such as the TB distribution does. Figure 1(b) reflects a global vision of the TB distribution with its diverse particular cases, which are the uniform (solid line in black), beta (segmented line in black), RB (dotted line in black) and two different types of TB (in gray) distributions. Observe that the parameters a and b presented in the PDF of the TB distribution defined in (2.5) can be intuitively interpreted as the lift at the left and right tails, respectively; see Figure 1(b)–(e). For example, Figure 1(c) lifts the left tails but not the right tails, whereas Figure 1(d) does the opposite. Similarly, Figure 1(e) lifts the left tails and also the right tails, whereas Figure 1(f) does the opposite. In summary, particular cases of the TB distribution, plotted in Figure 1(a)–(f), are: (i) $a = b = 1$ (uniform distribution); (ii) $a = b = 0$ (beta distribution); and (iii) $a = b = \theta$ (RB distribution), with PDFs defined in (2.1) and (2.3), respectively. Special and interesting situations occur when $a = 0, b \neq 0$ and when $a \neq 0, b = 0$, in whose case extreme-tail events are concentrated close to zero or to one, respectively, as noted in Figure 1(c)–(d).

3. ESTIMATION AND EM ALGORITHM

In this section, a methodology to estimate the parameters of the TB distribution is provided. We implement the EM algorithm to efficiently obtain the corresponding estimates.

3.1. Estimation of TB distribution parameters

Note that the parameters of the TB distribution can be estimated by the maximum likelihood method. Then, by taking advantage of the finite mixture representation of the TB distribution stated in (2.6), the EM algorithm may be implemented to efficiently estimate the TB distribution parameters.

First, based on a sample $\mathbf{T} = (T_1, \dots, T_n)^\top$ of size n from the TB distribution of PDF as given in (2.5), with observations $\mathbf{t} = (t_1, \dots, t_n)^\top$, the likelihood function for $\Theta = (a, b, \alpha, \beta)^\top$ is written as

$$(3.1) \quad \mathcal{L}(\Theta; \mathbf{t}) = \prod_{i=1}^n \left(a + (b - a)t_i + \left(1 - \frac{a + b}{2}\right) f_Y(t_i; \alpha, \beta) \right).$$

Then, in order to build estimators for the parameter Θ of the TB distribution, we can maximize the log-likelihood function defined as

$$(3.2) \quad \ell(\Theta; \mathbf{t}) = \sum_{i=1}^n \log \left(a + (b - a)t_i + \left(1 - \frac{a + b}{2} \right) f_Y(t_i; \alpha, \beta) \right).$$

The maximum likelihood estimates of a, b, α and β are obtained by differentiating the function (3.2) with respect to the mentioned parameters, generating the corresponding score vector. This vector must be equated to zero and the associated solution are the maximum likelihood estimates. However, such equations do not have closed-form and then they need to be solved numerically to maximize the log-likelihood function defined in (3.2). Subsequently, a non-linear optimization method is needed. For instance, the Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton method can be used; see [26]. We use the EM algorithm to obtain the parameter estimates.

3.2. EM algorithm

An efficient computationally strategy for estimating the parameter $\Theta = (a, b, \alpha, \beta)^\top$ of the TB distribution is to optimize the function (3.2) as a missing data framework.

The optimization problem can be solved with the EM algorithm and the finite mixture structure of the TB distribution. Consider a discrete random variable U for the missing (unobserved) data, where $u_i = j$, with $j \in \{1, 2, 3\}$, indicates which mixture component generates t_i . Thus, the complete data \mathbf{v} are formed by $\mathbf{v} = (\mathbf{t}^\top, \mathbf{u}^\top)^\top$, where the unobserved data are $\mathbf{u} = (u_1, \dots, u_n)^\top$ and the observed data are $\mathbf{t} = (t_1, \dots, t_n)^\top$. Thus, the likelihood function for Θ , considering the finite mixture representation of the TB distribution given in (2.6), under a complete data setting with n observations is now given by

$$(3.3) \quad \mathcal{L}^{(c)}(\Theta; \mathbf{v}) = \prod_{i=1}^n (\omega_1 f_1(t))^{1_1} (\omega_2 f_2(t))^{1_2} (\omega_3 f_3(t))^{1_3},$$

where 1 is the indicator function, such that $1_j = 1$ if $u_i = j$, with $j \in \{1, 2, 3\}$, and $1_j = 0$ otherwise. Hence, the log-likelihood function based on (3.3) for complete data is defined as

$$(3.4) \quad \ell^{(c)}(\Theta; \mathbf{v}) = \sum_{i=1}^n 1_1 \log (\omega_1 f_1(t)) + \sum_{i=1}^n 1_2 \log (\omega_2 f_2(t)) + \sum_{i=1}^n 1_3 \log (\omega_3 f_3(t)).$$

Note that the complete data log-likelihood function defined in (3.4) contains missing data, so that parameter estimates obtained directly from it cannot be calculated. Thus, in order to compute the estimates of a, b, α and β , we use the EM algorithm, recalling it has the E-step and M-step.

In order to implement its E-step, we need to find the expected value of the log-likelihood function stated in (3.4) and consequently of 1_j , for $j \in \{1, 2, 3\}$, given T_i . Therefore, it is necessary to specify an auxiliary function Q , which is the mentioned conditional expectation, using the random vector $\mathbf{V} = (\mathbf{T}^\top, \mathbf{U}^\top)^\top$, associated with the complete data \mathbf{v} , given the

observed data $\mathbf{T} = \mathbf{t}$, established as

$$\begin{aligned}
 (3.5) \quad Q(\Theta) &= E(\ell^{(c)}(\Theta; \mathbf{V}) | \mathbf{T} = \mathbf{t}) \\
 &= \sum_{i=1}^n E(\ell^{(c)}(\Theta; V_i) | T_i = t_i) \\
 &= \sum_{i=1}^n \sum_{j=1}^3 p_{ij} \ell^{(c)}(\Theta; v_i, t_i) \\
 &= \sum_{i=1}^n \sum_{j=1}^3 p_{ij} \log(\omega_j f_j(t_i; \Theta)),
 \end{aligned}$$

where

$$(3.6) \quad p_{ij} = P(U_i = j | T_i = t_i; \Theta) = \frac{\omega_j f_j(t_i; \Theta)}{\sum_{l=1}^3 \omega_l f_l(t_i; \Theta)}, \quad i \in \{1, \dots, n\}, \quad j \in \{1, 2, 3\}.$$

In order to initiate the EM algorithm, in its E-step, we need a starting value $\widehat{\Theta}^{(0)}$; see details about how to establish this starting value in Subsection 4.2. Thus, from (3.5), we have

$$(3.7) \quad Q(\Theta) |_{\Theta = \widehat{\Theta}^{(r-1)}} = \sum_{i=1}^n \sum_{j=1}^3 \widehat{p}_{ij}^{(r-1)} \log(\widehat{\omega}_j^{(r-1)} f_j(t_i; \widehat{\Theta}^{(r-1)})),$$

where $\widehat{\Theta}^{(r-1)}$ is the value of Θ for the $(r - 1)$ -th iteration at which the function $Q(\Theta)$ must be evaluated in order to iterate the EM algorithm. In addition, for $j \in \{1, 2, 3\}$, ω_j and f_j are defined in (2.6), with $\widehat{\omega}_j^{(r-1)}$ being the value of ω_j given in (2.7) for the $(r - 1)$ -th iteration and $\widehat{\omega}_j^{(0)}$ as established in Subsection 4.2. Furthermore, we have

$$(3.8) \quad \widehat{p}_{ij}^{(r-1)} = \frac{\widehat{\omega}_j^{(r-1)} f_j(t_i; \widehat{\Theta}^{(r-1)})}{\sum_{l=1}^3 \widehat{\omega}_l^{(r-1)} f_l(t_i; \widehat{\Theta}^{(r-1)})}, \quad i \in \{1, \dots, n\}, \quad j \in \{1, 2, 3\}.$$

Note that the expression given in (3.8) is obtained from $E(\mathbb{1}_j | T_i = t_i) |_{\Theta = \widehat{\Theta}^{(r-1)}}$.

In the M-step, we must find $\widehat{\Theta}^{(r)}$, which maximizes $Q(\Theta) |_{\Theta = \widehat{\Theta}^{(r-1)}}$ defined in (3.7). By taking the derivatives of Q with respect to ω_1, ω_2 , and ω_3 , under the restriction $\omega_1 + \omega_2 + \omega_3 = 1$, it is possible obtain the estimates

$$(3.9) \quad \widehat{\omega}_j^{(r)} = \frac{\sum_{i=1}^n \widehat{p}_{ij}^{(r-1)}}{\sum_{i=1}^n \sum_{j=1}^3 \widehat{p}_{ij}^{(r-1)}} = \frac{\widehat{n}_j^{(r-1)}}{n}, \quad j \in \{1, 2, 3\}.$$

In addition, the derivatives with respect to α and β lead to the usual maximum likelihood estimates of the beta distribution, which solves the equations

$$\begin{aligned}
 (3.10) \quad \psi(\widehat{\alpha}^{(r)}) - \psi(\widehat{\alpha}^{(r)} + \widehat{\beta}^{(r)}) &= \frac{\sum_{i=1}^n \widehat{p}_{i3}^{(r-1)} \log(t_i)}{\widehat{n}_3^{(r-1)}}, \\
 \psi(\widehat{\beta}^{(r)}) - \psi(\widehat{\alpha}^{(r)} + \widehat{\beta}^{(r)}) &= \frac{\sum_{i=1}^n \widehat{p}_{i3}^{(r-1)} \log(1 - t_i)}{\widehat{n}_3^{(r-1)}},
 \end{aligned}$$

where ψ is the digamma function that is defined as the logarithmic derivative of the gamma function Γ stated in (2.1) and given by

$$\psi(x) = \frac{d}{dx} \log(\Gamma(x)) = \frac{1}{\Gamma(x)} \frac{d}{dx} \Gamma(x).$$

The estimating equations presented in (3.10) can be solved using a quasi-Newton algorithm and the estimates of ω_1 , ω_2 , and ω_3 , subject to $\omega_1 + \omega_2 + \omega_3 = 1$, are obtained from (3.9). Once the parameters are updated in each iteration, repeat both the E and M steps iteratively until a certain criterion of convergence is obtained. The algorithm EM must be iterated until reaching convergence, for example, when $|\ell^{(c)}(\widehat{\Theta}^{(r)}) - \ell^{(c)}(\widehat{\Theta}^{(r-1)})| < 10^{-5}$, where $\widehat{\Theta}^{(r)}$ is the current ML estimate of Θ and $\widehat{\Theta}^{(r-1)}$ its previous estimate, with $\ell^{(c)}$ being given in (3.4); see McLachlan and Krishnan [35, pp.21–23]. Note that, in some cases, the EM algorithm does not admit an analytical solution in its E-step or M-step. Then, it becomes necessary to use iterative methods for the computation of the expectation or for the maximization. For variants of the EM algorithm based on approximations of its E-step or M-step, which preserve its convergence properties, see [32]. In our case, in the M-step of the algorithm, we use the BFGS method to iteratively solve the corresponding non-linear maximization problem. The BFGS method is implemented in the R software by the functions `optim` and `optimx`; see www.R-project.org and R Core Team [39].

4. NUMERICAL STUDIES

In this section, the TB distribution is evaluated throughout Monte Carlo simulations, comparing it with the beta and RB distributions. Here, we also include an empirical illustration with education data to show potential applications of the results obtained in the present investigation.

4.1. Simulation study

We start this section with an important remark about the data generation from the TB distribution. As noted in (2.6), this distribution can be seen as the mixture of three beta distributions. Except in some extreme cases such as the L-J-U-shaped beta distribution, the weights of the first two distributions on the mixture precisely capture the behavior of their tails. From Figure 2, note that, if we generate a small sample of data from the TB distribution with parameter $\Theta = (0.2, 0.5, 10, 15)$, we might not have data in any of its tails. Therefore, the corresponding histogram may not represent the true shape of the TB distribution. This small sample behavior is improved as the sample size increases and noted in Figure 2 for different values of the sample size n . For this reason, in our simulation study, we consider a sample size $n = 1000$.

We carry out a Monte Carlo simulation study to compare the performance of the beta, RB and TB distributions with samples generated from each of them. In order to capture the particular tail behavior of each one of these distributions, we use a sample size of $n = 1000$ and generate 100 samples for calculating the mean of the log-likelihood and Akaike information criterion (AIC). The AIC is given by $AIC = -2\ell(\widehat{\Theta}) + 2d$, where $\ell(\widehat{\Theta})$ is the log-likelihood function for Θ , associated with the underlying distribution, evaluated at $\Theta = \widehat{\Theta}$, d is the dimension of the parameter space, and n is the size of the data set. Note that this criterion is based on the log-likelihood function and penalize the distribution with more parameters. A distribution whose information criterion has a smaller value is better [13, 47].

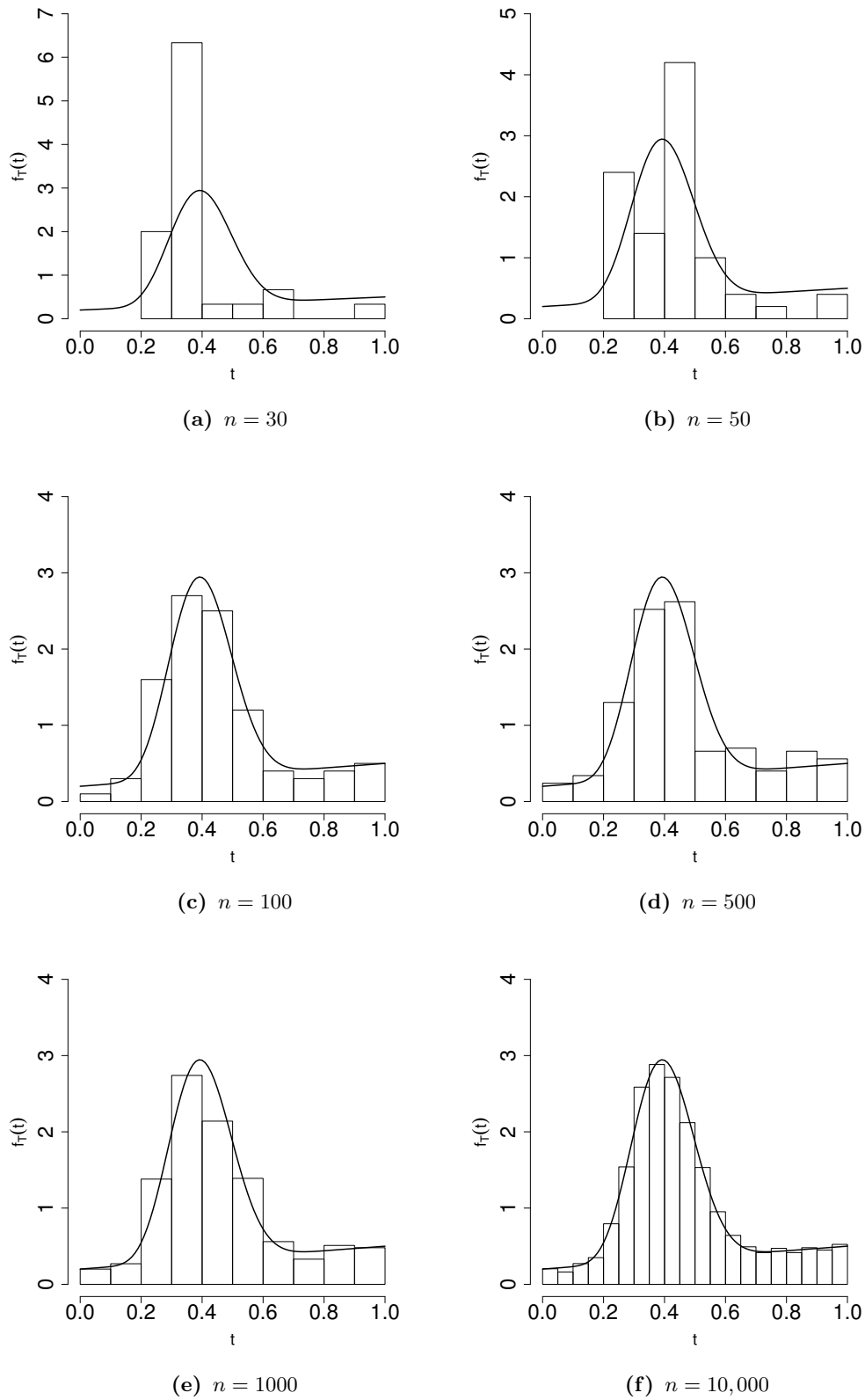


Figure 2: Histograms for the indicated sample size n from the $TB(0.2,0.5,10,15)$ distribution with simulated data, where the true TB PDF is drawn in solid line.

Firstly, we simulate data from the TB distribution with parameter $\Theta = (0.3, 0.7, 10, 15)$. In Table 1, we observe that the TB distribution achieves a better fit than the RB and beta distributions. Table 2 reports that the RB distribution fits the data by finding a value for θ between a and b . The beta distribution fits the data by increasing the variance, that is, by finding smaller values for α and β compensating the inability of this distribution to lift the tails. Secondly, we simulate data from the RB(0.4,10,15) distribution. In Table 3, note that the TB distribution fits the data with the same good level than the RB distribution. Table 4 reports that the TB distribution gives similar parameter estimates compared to the RB distribution. As in the first scenario, the beta distribution fits the data by increasing the variance. We collect a sample from the Beta(10,15) distribution. In Table 5, notice that the TB and RB distributions fit the data with the same good level in comparison to the beta distribution. Table 6 reports that the TB and RB distributions give similar parameter estimates in comparison to the beta distribution.

Table 1: Mean log-likelihood and AIC of the listed distributions for samples drawn from a TB(0.3,0.7,10,15) distribution with simulated data.

Distribution	Log-likelihood	AIC
TB	193.3288	-378.6576
RB	181.0892	-356.1783
Beta	64.0552	-124.1103

Table 2: Mean estimated parameter of the indicated distribution for samples drawn from a TB distribution with simulated data.

Distribution	$a = 0.3$ \hat{a}	$b = \theta = 0.7$ $\hat{b}, \hat{\theta}$	$\alpha = 10$ $\hat{\alpha}$	$\beta = 15$ $\hat{\beta}$
TB(a, b, α, β)	0.3023	0.7187	10.0799	15.1376
RB(θ, α, β)	—	0.5435	11.0549	16.2195
Beta(α, β)	—	—	1.6037	1.6018

Table 3: Mean log-likelihood and AIC of the listed distributions for samples drawn from an RB(0.4,10,15) distribution with simulated data.

Distribution	Log-likelihood	AIC
TB	278.6866	-549.3732
RB	278.1757	-550.3514
Beta	132.9706	-261.9412

Table 4: Mean estimated parameter of the indicated distribution for samples drawn from a TB distribution with simulated data.

Distribution	$a = 0.4$ \hat{a}	$b = \theta = 0.4$ $\hat{b}, \hat{\theta}$	$\alpha = 10$ $\hat{\alpha}$	$\beta = 15$ $\hat{\beta}$
TB(a, b, α, β)	0.4188	0.4141	9.7293	14.7257
RB(θ, α, β)	—	0.4161	9.7188	14.7168
Beta(α, β)	—	—	1.7944	2.1850

Table 5: Mean log-likelihood and AIC of the listed distributions for samples drawn from a Beta(10, 15) distribution with simulated data.

Distribution	Log-likelihood	AIC
TB	942.6532	-1877.306
RB	942.6532	-1879.306
Beta	942.6532	-1881.306

Table 6: Mean estimated parameter of the indicated distribution for samples drawn from a Beta(10, 15) distribution with simulated data.

Distribution	$a = 0$ \hat{a}	$b = \theta = 0$ $\hat{b}, \hat{\theta}$	$\alpha = 10$ $\hat{\alpha}$	$\beta = 15$ $\hat{\beta}$
TB(a, b, α, β)	9.88e-324	4.94e-324	10.3288	15.5109
RB(θ, α, β)	—	9.88e-324	10.3294	15.5120
Beta(α, β)	—	—	10.3274	15.5087

4.2. Empirical illustration

To illustrate the TB distribution in practice, we apply the proposed methods to a real-world data set and we compare the goodness of fit of the beta, RB and TB distributions. We analyze the data collected in the year 2016 of the average score of a university selection test for 1295 school establishments in the Metropolitan Region of Chile. This test is applied to students who have graduated from school in Chile at a national level and covers different areas of knowledge. In Chile, this test is named “Prueba de Selección Universitaria” (PSU) and the results obtained by the students in this test define the available possibilities to continue their studies in different universities in the country. The data set is publicly available on the “datachile” website (<https://es.datachile.io>).

We are interested in describing the distribution of the performance of the students who have applied to the PSU. To measure the performance, a total of 1295 average scores per establishment have been taken in the Metropolitan Region of Chile and scored in the interval (0, 1) throughout the transformation proposed by [44] defined as

$$t = \frac{(N - 1)(t^* - a_1)}{N(a_2 - a_1)} + \frac{1}{2N}, \quad t^* \in [a_1, a_2].$$

In our case, $a_1 = 293.5$, $a_2 = 715.5$ and $N = 1295$.

From the histogram presented in Figure 3, note that the distribution of the data has a lifted right tail and slightly lifted left tail. Thus, it is justifiable to propose the TB distribution to model these data, that is, we assume that $T \sim \text{TB}(a, b, \alpha, \beta)$. From Table 7, observe that the TB distribution achieves the best fit compared to the RB and beta distributions. In Table 8, we present the estimated parameters according to the method described in Section 3.

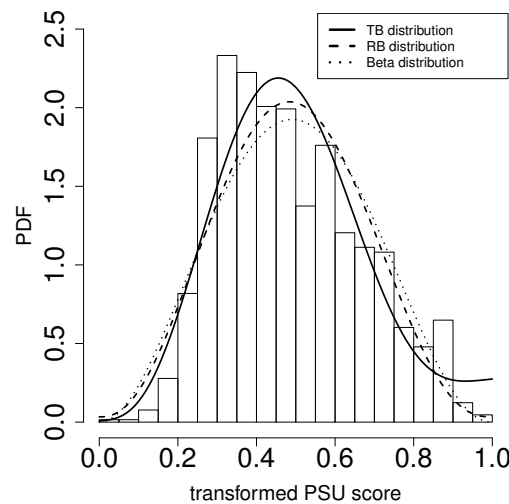


Figure 3: Histogram with estimated PDFs for the indicated distribution with the education data.

As starting values of $\Theta = (a, b, \alpha, \beta)^\top$ to initiate the EM algorithm, we consider the maximum likelihood estimates of α and β of the beta distribution. To obtain a and b , we consider the relation given in (2.7) with ω_1 and ω_2 , respectively, according to a visual conjecture detected at the tails of the histogram of the data such as mentioned above. This is corroborated by the estimates obtained, mainly at its right tail ($\hat{a} = 0.0066$ and $\hat{b} = 0.2742$). Observe that these estimates have a very intuitive interpretation, since the tails of the PDF are lifted in these quantities. The RB distribution attempts to compensate for this fact by assigning weight in both tails ($\hat{a} = \hat{b} = \hat{\theta} = 0.0334$), whereas the beta distribution tries to compensate it by increasing the variance (decreasing $\hat{\alpha}$ and $\hat{\beta}$). In Figure 3, we see the adjusted PDFs for the three different distributions, with the TB distribution being the model that captures the empirical behavior of the data better.

Table 7: Log-likelihood/AIC of the indicated distribution for education data.

Indicator	Distribution		
	TB	RB	Beta
Log-likelihood	413.896	401.647	371.711
AIC	-819.791	-797.293	-739.422

Table 8: Estimates of the indicated distribution parameter with education data.

Distribution	\hat{a}	$\hat{b} = \hat{\theta}$	$\hat{\alpha}$	$\hat{\beta}$
TB(a, b, α, β)	0.0066	0.2742	4.3566	5.0824
RB(θ, α, β)	—	0.0334	3.5307	3.6990
Beta(α, β)	—	—	3.1095	3.1901

5. CONCLUSIONS AND FUTURE RESEARCH

This paper reported the following findings:

- (i) By using a new method, we have proposed a new family of four-parameter distributions, called the trapezoidal beta distribution. The new model is widely flexible and generalizes the beta and rectangular beta distributions, being the new distribution an alternative to the beta distribution when both left and right tails are heavy.
- (ii) It was shown that the trapezoidal beta distribution can be rewritten conveniently as a mixture of three beta distributions, two with specific values in their parameters, and one traditional beta distribution with two arbitrary parameters.
- (iii) By taking advantage of the finite mixture representation of the new family of distributions, the expectation-maximization algorithm was implemented to efficiently estimate its parameters.
- (iv) Monte Carlo simulations based on the new family of distributions proposed in this research were provided to detect its performance.
- (v) An example with a real data set was conducted to illustrate the potential applications with the new family of distributions proposed in the paper. In addition, we compare the new distributions to their natural competitors, corresponding to the beta and rectangular beta distributions, showing the convenience of using the new distribution.

In summary, we have proposed a new family of distributions based on new method, which allows us to model data with support between zero and one as well as heavy left and/or right tails. We estimated the parameters of the new distribution with the expectation-maximization algorithm. Numerical studies with simulated and real data were performed to show the good empirical behavior of the estimators and to illustrate potential applications. In the simulation study, we observed that the trapezoidal beta distribution performed as well as the rectangular beta and beta distributions when the samples are generated from any of these two distributions. Moreover, we noted marked differences in favor of the trapezoidal beta distribution when the samples were generated from the trapezoidal beta distribution. In the empirical illustration, the trapezoidal beta distribution turned out to be the model that fits the data best, based on the Akaike information criterion. Furthermore, it is the only distribution that adequately addresses the essence of the data distribution when heavy left and/or right tails are present. We conclude that the trapezoidal beta distribution seems to be a new robust alternative for modeling bounded data. Therefore, this investigation may be a knowledge addition to the tool-kit of diverse practitioners, including educators, statisticians, and data scientists.

Some open problems that arose from the present investigation are the following:

- (i) It is possible to extend the benefits of the trapezoidal beta distribution to any bounded distribution.
- (ii) A re-parametrization of the trapezoidal beta model in terms of its mean is of interest. This will allow us to connect its mean to a regression structure in a similar manner to generalized linear models.

- (iii) Identifiability problems can be present in the case of the parameter estimation of the new distribution and they must be studied further.
- (iv) The use of covariates when modeling a response with support in $[0, 1]$ following the new family of distributions is of interest.
- (v) An extension of the present study to the multivariate case is also of practical relevance [27, 31, 41].
- (vi) Incorporation of temporal, spatial, functional, and quantile regression structures in the modeling, as well as errors-in-variables, and PLS regression, are also of interest [5, 16, 17, 20, 28, 29, 33, 40, 43].
- (vii) The derivation of diagnostic techniques to detect potential influential cases are needed, which are an important tool to be used in all statistical modeling [5, 15, 30].
- (viii) Robust estimation methods when outliers are present into the data set can be applied [46].
- (ix) Applications of the new methodology proposed here can be of interest in diverse areas [23].

Therefore, the proposed results in this study promotes new challenges and offers an open door to explore other theoretical and numerical issues. Research on these and other issues are in progress and their findings will be reported in future articles.

ACKNOWLEDGMENTS

The authors thank the editors and reviewers for their constructive comments on an earlier version of this manuscript. The research was partially supported by grant VRID Enlace N.º 217.014.027-1, from the Universidad de Concepción, Chile (J.I. Figueroa-Zúñiga), and by grant FONDECYT 1200525, from the National Agency for Research and Development (ANID) of the Chilean government (V. Leiva).

REFERENCES

- [1] AKINSETE, A. and FAMOYE, F. (2008). The beta-Pareto distribution, *Statistics*, **42**, 547–563.
- [2] ALIZADEH, M.; CORDEIRO, G.; BRITO, E. and DEMETRIO, C. (2015). The beta Marshall-Olkin family of distributions, *Journal of Statistical Distributions and Applications*, **2**, 1–18.
- [3] BARRETO-SOUZA, W.; SANTOS, A. and CORDEIRO, G. (2010). The beta generalized exponential distribution, *Journal of Statistical Computation and Simulation*, **80**, 159–172.
- [4] BOURGUIGNON, M.; LEO, J.; LEIVA, V. and SANTOS-NETO, M. (2017). The transmuted Birnbaum-Saunders distribution, *REVSTAT*, **15**, 601–628.

- [5] CARRASCO, J.M.F.; FIGUEROA-ZÚÑIGA, J.I.; LEIVA, V.; RIQUELME, M. and AYKROYD, R.G. (2020). An errors-in-variables model based on the Birnbaum-Saunders and its diagnostics with an application to earthquake data, *Stochastic Environmental Research and Risk Assessment*, **34**, 369–380.
- [6] CORDEIRO, G. and DE CASTRO, M. (2011). A new family of generalized distributions, *Journal of Statistical Computation and Simulation*, **81**, 883–898.
- [7] CORDEIRO, G.; NADARAJAH, S. and ORTEGA, E. (2012). The Kumaraswamy-Gumbel distribution, *Statistical Methods and Applications*, **21**, 139–168.
- [8] CORDEIRO, G.M. and DOS SANTOS BRITO, R. (2012). The beta power distribution, *Brazilian Journal of Probability and Statistics*, **26**, 88–112.
- [9] CORDEIRO, G.M.; ORTEGA, E.M. and NADARAJAH, S. (2010). The Kumaraswamy-Weibull distribution with application to failure data, *Journal of the Franklin Institute*, **347**, 1399–1429.
- [10] DE PASCOA, M.; ORTEGA, E. and CORDEIRO, G. (2011). The Kumaraswamy generalized gamma distribution with application in survival analysis, *Statistical Methodology*, **8**, 411–433.
- [11] EUGENE, N.; LEE, C. and FAMOYE, F. (2002). Beta-normal distribution and its applications, *Communications in Statistics: Theory and Methods*, **3**, 497–512.
- [12] FERRARI, S.L.P. and CRIBARI-NETO, F. (2004). Beta regression for modelling rates and proportions, *Journal of Applied Statistics*, **31**, 799–815.
- [13] FERREIRA, M.; GOMES, M.I. and LEIVA, V. (2012). On an extreme value version of the Birnbaum-Saunders distribution, *REVSTAT*, **10**, 181–210.
- [14] GARCÍA, C.B.; PÉREZ, J.G. and VAN DORP, J.R. (2011). Modeling heavy-tailed, skewed and peaked uncertainty phenomena with bounded support, *Statistical Methods and Applications*, **20**, 463–486.
- [15] GARCIA-PAPANI, F.; LEIVA, V.; URIBE-OPAZO, M.A. and AYKROYD, R.G. (2018). Birnbaum-Saunders spatial regression models: diagnostics and application to chemical data, *Chemometrics and Intelligent Laboratory Systems*, **177**, 114–128.
- [16] GARCIA-PAPANI, F.; URIBE-OPAZO, M.A.; LEIVA, V. and AYKROYD, R.G. (2017). Birnbaum-Saunders spatial modelling and diagnostics applied to agricultural engineering data, *Stochastic Environmental Research and Risk Assessment*, **31**, 105–124.
- [17] GIRALDO, R.; HERRERA, L. and LEIVA, V. (2020). Cokriging prediction using as secondary variable a functional random field with application in environmental pollution, *Mathematics*, **8**(8), 1305.
- [18] HAHN, E.D. (2008). Mixture densities for project management activity times: a robust approach to pert, *European Journal of Operational Research*, **188**, 450–459.
- [19] HAHN, E.D. and MARTIN, M.D. (2015). Robust project management with the tilted beta distribution, *SORT*, **39**, 253–272.
- [20] HUERTA, M.; LEIVA, V.; RODRIGUEZ, M.; LIU, S. and VILLEGAS, D. (2019). On a partial least squares regression model for asymmetric data with a chemical application in mining, *Chemometrics and Intelligent Laboratory Systems*, **190**, 55–68.
- [21] JOHNSON, N.L.; KOTZ, S. and BALAKRISHNAN, N. (1995). *Continuous Univariate Distributions*, New York: Wiley.
- [22] JONES, M.C. (2009). Kumaraswamy's distribution: A beta-type distribution with some tractability advantages, *Statistical Methodology*, **6**, 70–81.
- [23] KOTZ, S.; LEIVA, V. and SANHUEZA, A. (2010). Two new mixture models related to the inverse Gaussian distribution, *Methodology and Computing in Applied Probability*, **12**, 199–212.
- [24] KOTZ, S. and VAN DORP, J.R. (2004). *Beyond Beta: Other Continuous Families of Distributions with Bounded Support and Applications*, Singapore: World Scientific.

- [25] KUMARASWAMY, P. (1980). A generalized probability density function for double-bounded random processes, *Journal of Hydrology*, **46**, 79–88.
- [26] LANGE, K. (2000). *Numerical Analysis for Statisticians*, Springer: New York.
- [27] AYKROYD, R.G.; LEIVA, V. and MARCHANT, C. (2018). Multivariate Birnbaum-Saunders distributions: modelling and applications, *Risks*, **6**, 21.
- [28] LEIVA, V.; SÁNCHEZ, L.; GALEA, M. and SAULO, H. (2020). Global and local diagnostic analytics for a geostatistical model based on a new approach to quantile regression, *Stochastic Environmental Research and Risk Assessment*, **34**, 1457–1471.
- [29] LEIVA, V.; SAULO, H.; SOUZA, R.; AYKROYD, R.G. and VILA, R. (2021). A new BISARMA time series model for forecasting mortality using weather and particulate matter data, *Journal of Forecasting*, **40**, 346–364.
- [30] LIU, Y.; MAO, G.; LEIVA, V.; LIU, S. and TAPIA, A. (2020). Diagnostic analytics for an autoregressive model under the skew-normal distribution, *Mathematics*, **8**(5), 693.
- [31] MARCHANT, C.; LEIVA, V.; CHRISTAKOS, G. and CAVIERES, M.F. (2019). Monitoring urban environmental pollution by bivariate control charts: new methodology and case study in Santiago, Chile, *Environmetrics*, **30**, e2551.
- [32] MARCHANT, C.; LEIVA, V.; CYSNEIROS, F. and VIVANCO, J.F. (2016). Diagnostics in multivariate Birnbaum-Saunders regression models, *Journal of Applied Statistics*, **43**, 2829–2849.
- [33] MARTINEZ, S.; GIRALDO, R. and LEIVA, V. (2019). Birnbaum-Saunders functional regression models for spatial data, *Stochastic Environmental Research and Risk Assessment*, **33**, 1765–1780.
- [34] MCLACHLAN, G. and PEEL, D. (2004). *Finite Mixture Models*, New York: Wiley.
- [35] MCLACHLAN, G. and KRISHNAN, T. (1997). *The EM Algorithm and Extensions*, Wiley: New York.
- [36] NADARAJAH, S. and GUPTA, A.K. (2004). The beta-Fréchet distribution, *Far East Journal of Theoretical Statistics*, **14**, 15–24.
- [37] NADARAJAH, S. and KOTZ, S. (2004). The beta-Gumbel distribution, *Mathematical Problems in Engineering*, **10**, 323–332.
- [38] NADARAJAH, S. and KOTZ, S. (2006). The beta exponential distribution, *Reliability Engineering and System Safety*, **91**, 689–697.
- [39] R CORE TEAM (2021). R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing: Vienna*. Available at <http://www.r-project.org>.
- [40] SÁNCHEZ, L.; LEIVA, V.; GALEA, M. and SAULO, H. (2021). Birnbaum-Saunders quantile regression and its diagnostics with application to economic data, *Applied Stochastic Models in Business and Industry*, **37**, 53–73.
- [41] SÁNCHEZ, L.; LEIVA, V.; GALEA, M. and SAULO, H. (2020). Birnbaum-Saunders quantile regression models with application to spatial data, *Mathematics*, **8**(5), 1000.
- [42] SANHUEZA, R.A. and FIGUEROA-ZÚÑIGA, J. (2018). *Trapezoidal Kumaraswamy Distribution*, MSc Thesis in Statistics, Universidad de Concepción, Chile. Available at <http://repositorio.udec.cl/jspui/handle/11594/3535>.
- [43] SAULO, H.; LEO, J.; LEIVA, V. and AYKROYD, R.G. (2019). Birnbaum-Saunders autoregressive conditional duration models applied to high-frequency financial data, *Statistical Papers*, **60**, 1605–1629.
- [44] SMITHSON, M. and VERKUILEN, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables, *Psychological Methods*, **11**, 54–71.

- [45] TABLADA, C.J. and CORDEIRO, G.M. (2019). The beta Marshall-Olkin Lomax distribution, *REVSTAT*, **17**, 321–344.
- [46] VELASCO, H.; LANIADO, H.; TORO, M.; LEIVA, V. and LIO, Y. (2020). Robust three-step regression based on comedian and its performance in cell-wise and case-wise outliers, *Mathematics*, **8**(8), 1259.
- [47] VENTURA, M.; SAULO, H.; LEIVA, V. and MONSUETO, S. (2019). Log-symmetric regression models: information criteria, application to movie business and industry data with economic implications, *Applied Stochastic Models in Business and Industry*, **34**, 963–977.

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)*; *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

Julio SINGER, University of Sao Paulo, Brazil

Manuel SCOTTO, University of Lisbon, Portugal

Lisete SOUSA, University of Lisbon, Portugal

Milan STEHLÍK, University of Valparaíso, Chile and LIT-JK University Linz, Austria

María Dolores UGARTE, Public University of Navarre, Spain

Executive Editor

José A. PINTO MARTINS, Statistics Portugal

Assistant Editors

José CORDEIRO, Statistics Portugal

Olga BESSA MENDES, Statistics Portugal