




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: Month 0000 Revised: Month 0000 Accepted: Month 0000

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.

Key-Words:

- *Bounded-support distributions; EM algorithm; Mixture distributions; R software; Trapezoidal distributions.*

AMS Subject Classification:

- 60E05; 62E15.

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, 44]. 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 [36], beta-Fréchet [35], beta-exponential [37], beta-Pareto [1], beta-generalized-exponential [3], beta-normal [11], beta-power [8], beta-Marshall-Olkin [2], and beta-Marshall-Olkin-Lomax [44] 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 [41] 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 new method to circumvent the above-mentioned limitations. This new distribution is the trapezoidal beta (TB) model, which has high flexibility to model 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 model 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 [33]

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 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

$$(2.5) \quad \begin{aligned} f_T(t; a, b, \alpha, \beta) &= a + (b - a)t + \left(1 - \int_0^1 (a + (b - a)t) dt\right) f_Y(t; \alpha, \beta) \\ &= a + (b - a)t + \left(1 - \frac{a + b}{2}\right) f_Y(t; \alpha, \beta), \quad 0 < t < 1, \end{aligned}$$

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

$$(2.6) \quad \begin{aligned} f_T(t; a, b, \alpha, \beta) &= \omega_1 f_1(t) + \omega_2 f_2(t) + \omega_3 f_3(t), \\ &= \frac{a}{2}(2 - 2t) + \frac{b}{2}(2t) + \left(1 - \frac{a + b}{2}\right) f_Y(t; \alpha, \beta), \end{aligned}$$

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,

$$(2.7) \quad \omega_1 = \frac{a}{2}, \quad \omega_2 = \frac{b}{2}, \quad \omega_3 = \left(1 - \frac{a + b}{2}\right)$$

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

$$(2.8) \quad m_k = \mathbb{E}(T^k) = \frac{a}{k+1} + \frac{b-a}{k+2} + \left(1 - \frac{a+b}{2}\right) m_k^*,$$

where m_k^* is the k -th moment of the $\text{Beta}(\alpha, \beta)$ distribution. Thus, from (2.8), we have

$$(2.9) \quad m_k = \frac{a}{k+1} + \frac{b-a}{k+2} + \left(1 - \frac{a+b}{2}\right) \left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r}\right).$$

In addition, the moment generating and characteristic functions of $T \sim \text{TB}(a, b, \alpha, \beta)$ are stated respectively as

$$(2.10) \quad \begin{aligned} M_T(v) &= \mathbb{E}(e^{vT}) = 1 + \sum_{k=1}^{\infty} m_k \frac{v^k}{k!}, \quad v \in \mathbb{R}, \\ \varphi_T(v) &= \mathbb{E}(e^{ivT}) = 1 + \sum_{k=1}^{\infty} m_k \frac{(iv)^k}{k!}, \quad v \in \mathbb{R}. \end{aligned}$$

Based on (2.9) or (2.10), we deduce that the mean and variance of T are given respectively as

$$(2.11) \quad \begin{aligned} \mathbb{E}(T) &= \frac{a+2b}{6} + \left(1 - \frac{a+b}{2}\right) \frac{\alpha}{\alpha+\beta}, \\ \text{Var}(T) &= \left(\frac{3a+9b-(a+2b)^2}{36}\right) \\ &\quad + \left(\frac{\alpha}{\alpha+\beta}\right) \left(1 - \frac{a+b}{2}\right) \left(\frac{\alpha+1}{\alpha+\beta+1} - \frac{\alpha(2-a-b)}{2(\alpha+\beta)} - \frac{a+2b}{3}\right). \end{aligned}$$

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 (1)-(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 cases 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).

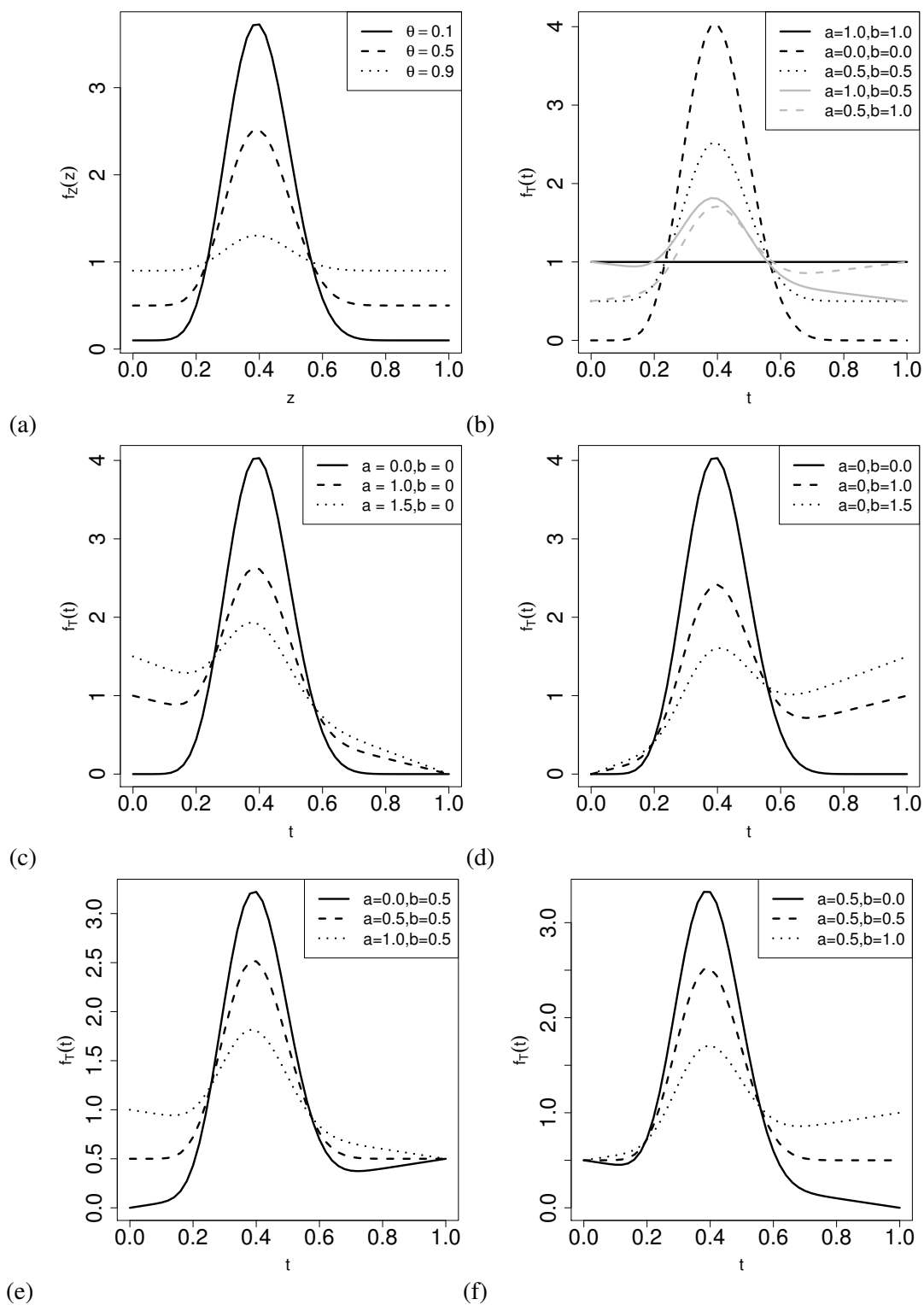


Figure 1: Plots of the (a) RB(θ , $\alpha = 10$, $\beta = 15$) PDF with θ as indicated, and (b)-(f) TB(a, b , $\alpha = 10$, $\beta = 15$) PDF with a, b as listed.

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. Hence, 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 $j = 1, 2, 3$, 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

$$(3.5) \quad \begin{aligned} Q(\Theta) &= \mathbb{E}(\ell^{(c)}(\Theta; \mathbf{V}) | \mathbf{T} = \mathbf{t}) \\ &= \sum_{i=1}^n \mathbb{E}(\ell^{(c)}(\Theta; V_i) | T_i = t_i) \\ &= \sum_{i=1}^n \sum_{j=1}^3 p_{ij} \ell^{(c)}(\Theta; v_i, t_i) \\ &= \sum_{i=1}^n \sum_{j=1}^3 p_{ij} \log(\omega_j f_j(t_i; \Theta)), \end{aligned}$$

where

$$(3.6) \quad p_{ij} = \mathbb{P}(U_i = j | T_i = t_i; \Theta) = \frac{\omega_j f_j(t_i; \Theta)}{\sum_{l=1}^3 \omega_l f_l(t_i; \Theta)}, \quad i = 1, \dots, n, j = 1, 2, 3.$$

In order to initiate the EM algorithm, in its E-step, we need a starting value $\hat{\Theta}^{(0)}$; see details about how to establish this starting value in Subsection 3. Thus, from (3.5), we have

$$(3.7) \quad Q(\Theta) |_{\Theta = \hat{\Theta}^{(r-1)}} = \sum_{i=1}^n \sum_{j=1}^3 \hat{p}_{ij}^{(r-1)} \log(\hat{\omega}_j^{(r-1)} f_j(t_i; \hat{\Theta}^{(r-1)})),$$

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 $j = 1, 2, 3$, 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 3. Furthermore, we have

$$(3.8) \quad \widehat{p}_{ij}^{(r-1)} = \frac{\widehat{\omega}_j^{(r-1)} f_j(t_i; \widehat{\Theta}^{(r-1)})}{\sum_{l=1}^3 \widehat{\omega}_l^{(r-1)} f_l(t_i; \widehat{\Theta}^{(r-1)})}, \quad i = 1, \dots, n, j = 1, 2, 3.$$

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

$$(3.9) \quad \widehat{\omega}_j^{(r)} = \frac{\sum_{i=1}^n \widehat{p}_{ij}^{(r-1)}}{\sum_{i=1}^n \sum_{j=1}^3 \widehat{p}_{ij}^{(r-1)}} = \frac{\widehat{n}_j^{(r-1)}}{n}, \quad j = 1, 2, 3.$$

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

$$(3.10) \quad \begin{aligned} \psi(\widehat{\alpha}^{(r)}) - \psi(\widehat{\alpha}^{(r)} + \widehat{\beta}^{(r)}) &= \frac{\sum_{i=1}^n \widehat{p}_{i3}^{(r-1)} \log(t_i)}{\widehat{n}_3^{(r-1)}}, \\ \psi(\widehat{\beta}^{(r)}) - \psi(\widehat{\alpha}^{(r)} + \widehat{\beta}^{(r)}) &= \frac{\sum_{i=1}^n \widehat{p}_{i3}^{(r-1)} \log(1 - t_i)}{\widehat{n}_3^{(r-1)}}, \end{aligned}$$

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 [34, 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. 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 [38].

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 $AIC = -2\ell(\hat{\Theta}) + 2d$, where $\ell(\hat{\Theta})$ is the log-likelihood function for Θ , associated with the underlying distribution, evaluated at $\Theta = \hat{\Theta}$, 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, 46].

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 fit 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.

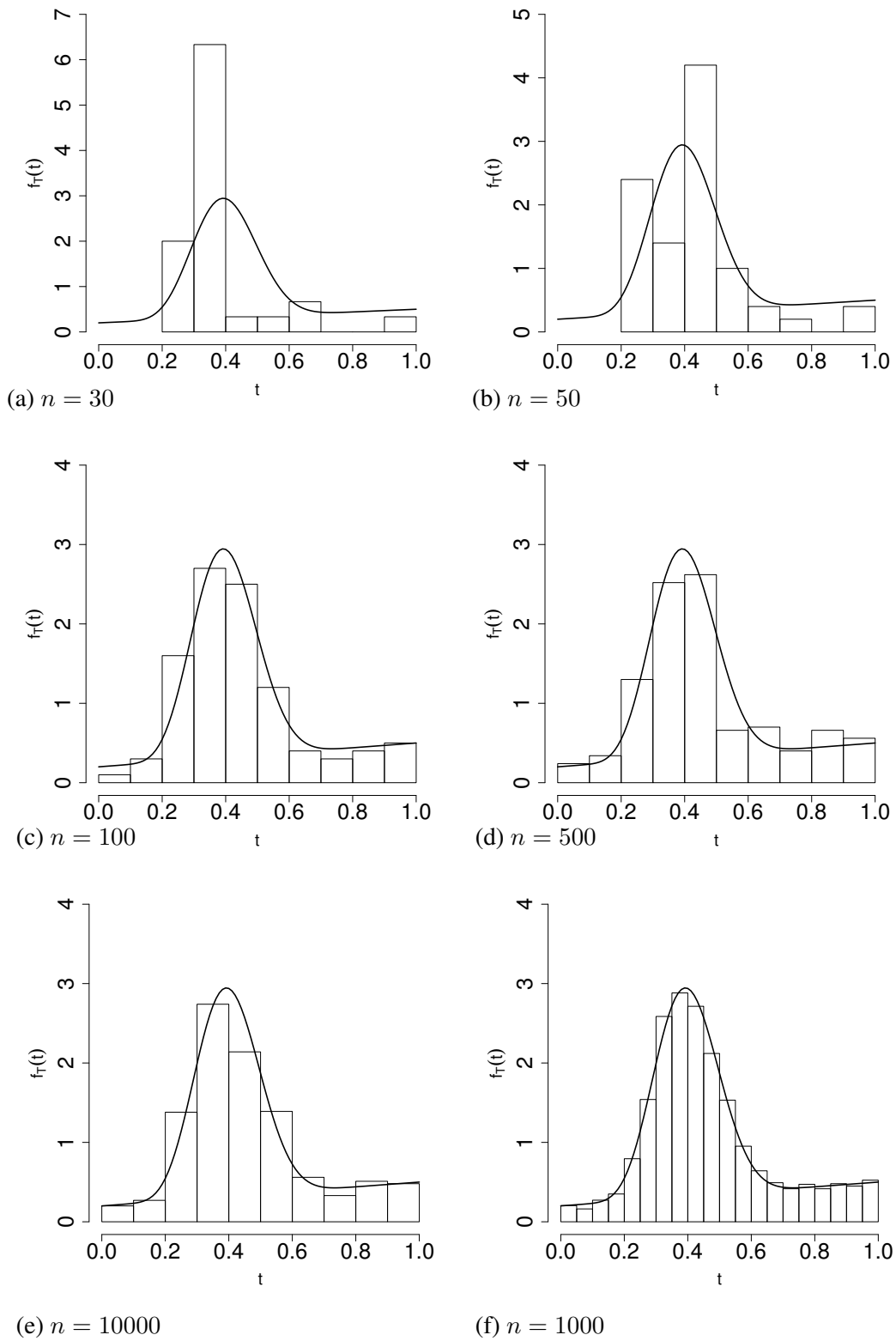


Figure 2:

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.

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 [43] 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. 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, whereas a and b are obtained from 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}	$\hat{b} = \hat{\theta}$	$\hat{\alpha}$	$\hat{\beta}$
$\text{TB}(a, b, \alpha, \beta)$	0.0066	0.2742	4.3566	5.0824
$\text{RB}(\theta, \alpha, \beta)$	-	0.0334	3.5307	3.6990
$\text{Beta}(\alpha, \beta)$	-	-	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, which 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.

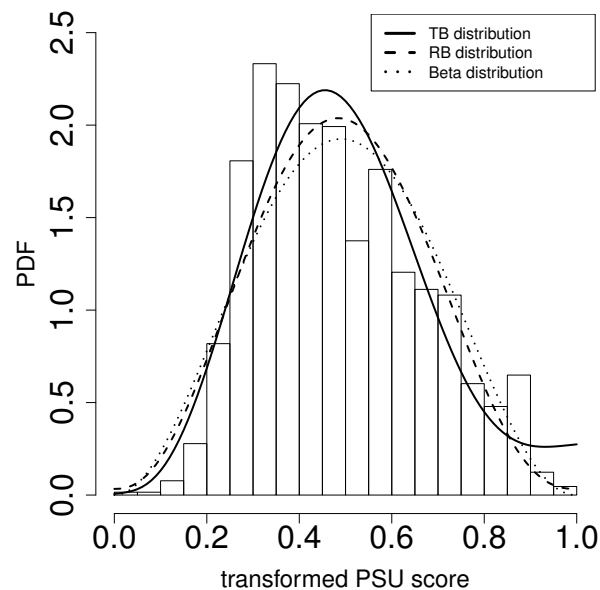


Figure 3: Histogram with estimated PDFs for the indicated distribution with the education data.

- (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 distributions.

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 and 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 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 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 that as in generalized linear models.
- (iii) Identifiability problem can be present in the case of the parameter estimation of the new distribution and it 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, 40].
- (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, 32, 39, 42]
- (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 [45].
- (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.

ACKNOWLEDGEMENTS

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° 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, Famoye F (2008) The beta-Pareto distribution. *Statistics* 42:547–563
- [2] Alizadeh M, Cordeiro G, Brito E, 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, Cordeiro G (2010) The beta generalized exponential distribution. *Journal of Statistical Computation and Simulation* 80:159–172
- [4] Bourguignon M, Leao J, Leiva V, Santos-Neto M (2017) The transmuted Birnbaum-Saunders distribution. *REVSTAT* 15:601–628
- [5] Carrasco JMF, Figueroa-Zúñiga JI, Leiva V, Riquelme M, Aykroyd RG (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, de Castro M (2011) A new family of generalized distributions. *Journal of Statistical Computation and Simulation* 81:883–898
- [7] Cordeiro G, Nadarajah S, Ortega E (2012) The Kumaraswamy-Gumbel distribution. *Statistical Methods and Applications* 21:139–168
- [8] Cordeiro GM, dos Santos Brito R (2012) The beta power distribution. *Brazilian Journal of Probability and Statistics* 26:88–112
- [9] Cordeiro GM, Ortega EM, 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, Cordeiro G (2011) The Kumaraswamy generalized gamma distribution with application in survival analysis. *Statistical Methodology* 8:411–433
- [11] Eugene N, Lee C, Famoye F (2002) Beta-normal distribution and its applications. *Communications in Statistics: Theory and methods* 3:497-512.
- [12] Ferrari SLP, Cribari-Neto F (2004) Beta regression for modelling rates and proportions. *Journal of Applied Statistics* 31:799–815
- [13] Ferreira M, Gomes MI, Leiva V (2012) On an extreme value version of the Birnbaum-Saunders distribution. *REVSTAT* 10:181–210

- [14] García CB, Pérez JG, van Dorp JR (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 MA, Aykroyd RG (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 MA, Leiva V, Aykroyd RG (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, Leiva, V (2020) Cokriging prediction using as secondary variable a functional random field with application in environmental pollution. *Mathematics* 8(8):1305.
- [18] Hahn ED (2008) Mixture densities for project management activity times: A robust approach to pert. *European Journal of Operational Research* 188:450–459
- [19] Hahn ED, Martin MD (2015) Robust project management with the tilted beta distribution. *SORT* 39:253–272
- [20] Huerta M, Leiva V, Rodriguez M, Liu S, 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 NL, Kotz S, Balakrishnan N (1995) *Continuous Univariate Distributions*. New York: Wiley
- [22] Jones MC (2009) Kumaraswamy’s distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology* 6:70–81
- [23] Kotz S, Leiva V, 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, van Dorp JR (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 RG, Leiva V, Marchant C (2018) Multivariate Birnbaum-Saunders distributions: Modelling and applications. *Risks* 6(1) article 21 pages:1-25.
- [28] Leiva V, Sánchez L, Galea M, Saulo H (2020) Global and local diagnostic analytics for a geo-statistical model based on a new approach to quantile regression. *Stochastic Environmental Research and Risk Assessment* doi:10.1007/s00477-020-01831-y
- [29] Leiva, V, Saulo H, Souza R, Aykroyd RG, Vila R (2020) A new BISARMA time series model for forecasting mortality using weather and particulate matter data. *Journal of Forecasting* doi:10.1002/for.2718
- [30] Liu Y, Mao G, Leiva V, Liu S, 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, Cavieres MF (2019) Monitoring urban environmental pollution by bivariate control charts: New methodology and case study in Santiago, Chile. *Environmetrics* 30:e2551
- [32] Martinez S, Giraldo R, Leiva V (2019) Birnbaum-Saunders functional regression models for spatial data. *Stochastic Environmental Research and Risk Assessment* 33:1765-1780
- [33] McLachlan G, Peel D (2004) *Finite Mixture Models*. New York: Wiley
- [34] McLachlan G, Krishnan T (1997) *The EM Algorithm and Extensions*. Wiley: New York.
- [35] Nadarajah S, Gupta AK (2004) The beta-Fréchet distribution. *Far East Journal of Theoretical Statistics* 14:15-24
- [36] Nadarajah S, Kotz S (2004) The beta-Gumbel distribution. *Mathematical Problems in Engineering* 10:323-332
- [37] Nadarajah S, Kotz S (2006) The beta exponential distribution. *Reliability Engineering and System Safety* 91:689-697
- [38] R Core Team (2018) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing:Vienna. Available at <http://www.r-project.org>
- [39] Sánchez L, Leiva V, Galea M, Saulo H (2020) Birnbaum-Saunders quantile regression and its diagnostics with application to economic data. *Applied Stochastic Models in Business and Industry* doi:10.1002/asmb.2556
- [40] Sánchez L, Leiva V, Galea M, Saulo H (2020) Birnbaum-Saunders quantile regression models with application to spatial data. *Mathematics* 8(5):1000
- [41] Sanhueza RA, 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>
- [42] Saulo H, Leao J, Leiva V, Aykroyd RG (2019) Birnbaum-Saunders autoregressive conditional duration models applied to high-frequency financial data. *Statistical Papers* 60:1605-1629
- [43] Smithson M, Verkuilen J (2006) A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods* 11:54-71
- [44] Tablada CJ, Cordeiro GM (2019) The beta Marshall-Olkin Lomax distribution. *REVSTAT* 17:321-344
- [45] Velasco H, Laniado H, Toro M, Leiva V, Lio Y(2020) Robust three-step regression based on comedian and its performance in cell-wise and case-wise outliers. *Mathematics* 8(8):1259.
- [46] Ventura M, Saulo H, Leiva V, 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