CHILEAN JOURNAL OF STATISTICS

Edited by Víctor Leiva and Carolina Marchant

A free open access journal indexed by

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

Published by the
Chilean Statistical Society
SOCHE
SOCIEDAD CHILENA DE ESTADÍSTICA

Aims

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

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

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

Contents

Víctor Leiva and Carolina Marchant
Confirming our international presence with publications and submissions from all continents in COVID-19 pandemic 69

Ibrahim M. Almanjahie, Mohammed Kadi Attouch, Omar Fetitah, and Hayat Louhab
Robust kernel regression estimator of the scale parameter for functional ergodic data with applications 73

Ricardo Puziol de Oliveira, Marcos Vinicius de Oliveira Peres, Jorge Alberto Achcar, and Nasser Davarzani
Inference for the trivariate Marshall-Olkin-Weibull distribution in presence of right-censored data 95

Henrique José de Paula Alves and Daniel Furtado Ferreira
On new robust tests for the multivariate normal mean vector with high-dimensional data and applications 117

Josmar Mazucheli, André F.B. Menezes, Sanku Dey, and Saralees Nadarajah
Improved parameter estimation of the Chaudhry and Ahmad distribution with climate applications 137

André Leite, Abel Borges, Geiza Silva, and Raydonal Ospina
A timetabling system for scheduling courses of statistics and data science: Methodology and case study 151

Jorge Figueroa-Zúñiga, Rodrigo Sanhueza-Parkes, Bernardo Lagos-Álvarez, and Germán Ibacache-Pulgar
Modeling bounded data with the trapezoidal Kumaraswamy distribution and applications to education and engineering 163

Chilean Journal of Statistics
Vol. 11, No. 2, December 2020, 163–176

Distribution Theory
Research Paper

Modeling bounded data with the trapezoidal Kumaraswamy distribution and applications to education and engineering

Jorge Figueroa-Zúñiga[1,*], Rodrigo Sanhueza-Parkes[1], Bernardo Lagos-Álvarez[1], and Germán Ibacache-Pulgar[2,3]

[1]Department of Statistics, Universidad de Concepción, Concepción, Chile,
[2]Department of Statistics, Universidad de Valparaíso, Valparaíso, Chile.
[3]Centro Interdisciplinario de Estudios Atmosféricos y Astroestadística
Universidad de Valparaíso, Valparaíso, Chile.

(Received: 13 October 2020 · Accepted in final form: 30 November 2020)

Abstract

The Kumaraswamy distribution has been a very studied tool in the analysis and modeling of limited-range continuous random variables. Several variants of this distribution have been studied, but they do not have the possibility of lifting the tails of this distribution. However, in many situations, scenarios where the data are bounded and tail-area events occur at one or both tails independently. In order to model these scenarios, we propose the trapezoidal Kumaraswamy distribution. This paper is centered on the trapezoidal Kumaraswamy distribution, which has two intuitive additional parameters with respect to the Kumaraswamy distribution and generalizes this. We study its probability density function and derive some fundamental properties, such as the moments, moment generating function, and characteristic function. Then, the trapezoidal Kumaraswamy distribution is rewritten conveniently as a finite mixture showing that its parameters can be easily estimated using the expectation-maximization algorithm. We report results of a simulation and an application to a real data set. Comparison with several competing distributions indicates that the trapezoidal Kumaraswamy distribution presents a better fit and so it can be quite useful in empirical applications.

Keywords: EM algorithm · Maximum likelihood · Mixture distributions.

Mathematics Subject Classification: 62E15 · 62F10.

1. Introduction

A good alternative for modeling continuous data restricted to a bounded interval is the double bounded distribution (Kumaraswamy, 1980), named after as the Kumaraswamy distribution (Jones, 2009). This distribution provides a wide variety of shapes for its probability density function (PDF) allowing different type of data to be accommodated.

The Kumaraswamy distribution is very flexible. However, it does not consider tail-area events nor high flexibility in the variance specification. In order to add flexibility into the model, other distributions derived from the Kumaraswamy distribution have been pro-

*Jorge Figueroa-Zúñiga. Email: jifiguer@gmail.com

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

http://www.soche.cl/chjs

posed. For example, the Kumaraswamy Weibull (Cordeiro et al., 2010) and Kumaraswamy-G (Cordeiro and de Castro, 2011) distributions have been derived including two additional positive parameters. The authors studied some of their mathematical properties by presenting special submodels such as: the Kumaraswamy generalized gamma distribution (de Pascoa et al., 2011), which is able to model bathtub-shaped hazard rate functions. The importance of Kumaraswamy generalized gamma distribution is in its capacity to model functions of monotonous failure frequency and non-monotone, which are fairly common in life-time data analysis and reliability. Another case is the Kumaraswamy Gumbel distribution (Cordeiro et al., 2012), which is probably the most widely applied statistical distribution to problems in engineering. Similarly, the Kumaraswamy-log-logistic (De Santana et al., 2012), Kumaraswamy-geometric (Akinsete et al., 2014), and Kumaraswamy Fréchet (Mead and Abd-Eltawab, 2014) distributions, among others of the same family have been proposed. Furthermore, in the same direction, in order to make some existing distributions flexible, other models have been proposed as in Liang et al. (2014), Nadarajah and Kotz (2004), Nadarajah and Kotz. (2006), Akinsete and Famoye (2008), Eugene et al (2002), Cordeiro and dos Santos Brito (2012), among others. Note that the Kumaraswamy distribution, and its extensions, are unable to fit data which are concentrated at both tails. The main objective of this work is to propose a new bounded distribution which is able to model data which are concentrated at both tails.

The reminder of this article is organized as follows. In Section 2, the trapezoidal Kumaraswamy (TK) distribution is proposed and its basic properties are discussed. In Section 3, we estimate parameters through a convenient reparametrization of the TK distribution given in Section 2. Section 4 conducts a Monte Carlo simulation study for both the TK and Kumaraswamy distributions, comparing them. In Section 5, two empirical illustrations are provided corresponding to (i) percent slacks for reduction in pollutant emissions/discharges for carbon dioxide (CO2) and water (H2O) in Angolan thermal power plants, and (ii) scores of a university admission test in 1295 school establishments in Metropolitan region of Chile. The results are compared with the classical Kumaraswamy distribution. Finally, discussions, conclusions and further research of the proposed distribution appear in Section 6.

2. The new distribution

In this section, we discuss some properties of the Kumaraswamy distribution and we present the TK distribution as well as its properties.

2.1 Background

The PDF of a random variable Y following a Kumaraswamy distribution is given by

$$f_{\mathrm{K}}(y;\alpha,\beta) = \alpha\beta y^{\alpha-1}(1-y^{\alpha})^{\beta-1}, \quad y \in (0,1), \tag{1}$$

where $\alpha > 0$ and $\beta > 0$. Then, note that

$$\mathrm{E}(Y) = m_1, \quad \mathrm{Var}(Y) = m_2 - m_1^2,$$

with m_k denoting the k-th moment of the Kumaraswamy distribution stated as

$$m_k = \frac{\beta\Gamma(1+\frac{k}{\alpha})\Gamma(\beta)}{\Gamma(1+\frac{k}{\alpha}+\beta)} = \beta B\left(1+\frac{k}{\alpha},\beta\right),$$

where B is the beta function.

In practice, the Kumaraswamy distribution has been a useful tool for modeling bounded data. However, it is common in many cases to have data concentrated at both tails independently. Hence, we propose the TK distribution as an extension which allows to model this situation and that it conserve the flexibility of the Kumaraswamy distribution.

2.2 The trapezoidal Kumaraswamy distribution

Let Y follow a TK distribution of parameters a, b, α, β which we denote by $Y \sim \mathrm{TK}(a, b, \alpha, \beta)$. Then, the PDF of Y is established as

$$f_{\mathrm{TK}}(y; a, b, \alpha, \beta) = a + (b-a)y + \left(1 - \frac{a+b}{2}\right) f_{\mathrm{K}}(y; \alpha, \beta), \tag{2}$$

with $0 < y < 1$, $0 \leq a, b \leq 2$, $0 \leq a+b \leq 2$ and $f_{\mathrm{K}}(y; \alpha, \beta)$ being the Kumaraswamy PDF of parameters α and β given in Equation (1). The parameters a and b can be intuitively interpreted as the lift at the left and right tails of the PDF respectively; see Figure 1. As a particular case, we have that, when $a = b = 0$, the standard Kumaraswamy distribution is recovered –see Equation (1)– and we propose the rectangular Kumaraswamy distribution when $a = b = \theta$.

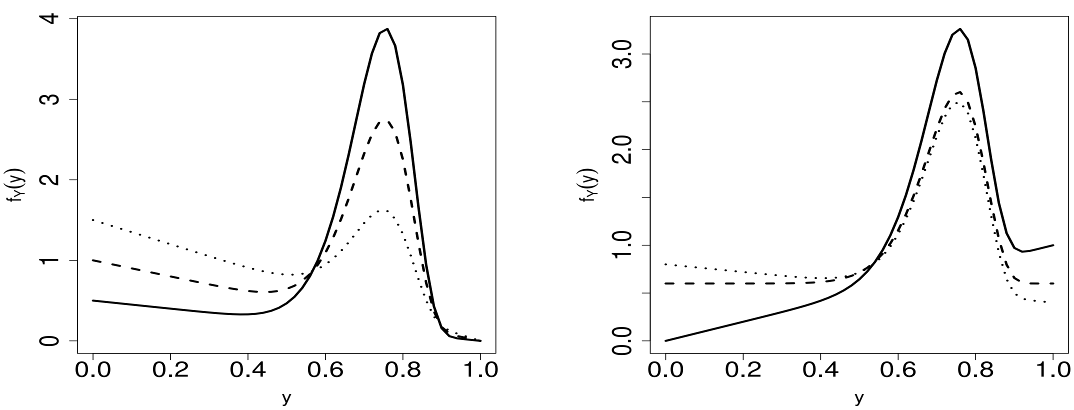

Figure 1. Examples of TK PDF with $\alpha = 10, \beta = 15$ and different values of the parameters (a, b). Left: $(a, b) = (0.5, 0)$ (solid line), $(a, b) = (1, 0)$ (dashed line) and $(a, b) = (1.5, 0)$ (dotted line); right: $(a, b) = (0, 1)$ (solid line), $(a, b) = (0.6, 0.6)$ (dashed line) and $(a, b) = (0.8, 0.4)$ (dotted line).

We now present some properties of the TK distribution. Let $Y \sim \mathrm{TK}(a, b, \alpha, \beta)$. Then, the k-th moment of Y is given by

$$m_k = \mathrm{E}(Y^k) = \frac{a}{k+1} + \frac{b-a}{k+2} + \left(1 - \frac{a+b}{2}\right) m_k^*, \tag{3}$$

where m_k^* is the k-th moment of the Kumaraswamy distribution of parameters α, β. Then, Equation (3) can be written as

$$\begin{aligned} m_k &= \frac{a}{k+1} + \frac{b-a}{k+2} + \left(1 - \frac{a+b}{2}\right) \frac{\beta\Gamma\left(1 + k/\alpha\right)\Gamma\left(\beta\right)}{\Gamma\left(1 + \beta + k/\alpha\right)} \\ &= \frac{a}{k+1} + \frac{b-a}{k+2} + \left(1 - \frac{a+b}{2}\right) \beta B\left(1 + k/\alpha, \beta\right). \end{aligned} \tag{4}$$

With the expression defined in Equation (4), it is easy to deduce that

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

The moment generating function of the random variable Y is given by

$$M_Y(t) = \mathrm{E}\left(\mathrm{e}^{tY}\right) = 1 + \sum_{k=1}^{\infty} m_k \frac{t^k}{k!}, \quad t \in \mathrm{R},$$

and its characteristic function is stated as

$$\varphi_Y(t) = \mathrm{E}\left(\mathrm{e}^{itY}\right) = 1 + \sum_{k=1}^{\infty} m_k \frac{(it)^k}{k!}, \quad t \in \mathrm{R}.$$

3. Estimation of trapezoidal Kumaraswamy distribution parameters

In this section, we discuss how to estimate the parameters of the TK distribution efficiently.

3.1 Log-likelihood function

The likelihood function for a sample of n observations from the TK distribution is specified as

$$\mathcal{L}(a, b, \alpha, \beta) = \prod_{i=1}^{n}\left(a + (b-a)y_i + \left(1 - \frac{a+b}{2}\right) f_{\mathrm{K}}(y_i; \alpha, \beta)\right). \tag{5}$$

Then, one strategy to build estimators for its parameters is to maximize the corresponding log-likelihood given by

$$\ell(a, b, \alpha, \beta) = \sum_{i=1}^{n} \log\left(a + (b-a)y_i + \left(1 - \frac{a+b}{2}\right) f_{\mathrm{K}}(y_i; \alpha, \beta)\right). \tag{6}$$

The maximum likelihood estimators of a, b, α and β are obtained from the differentiation of Equation (6) with respect to the mentioned parameters and equating to zero. However, in this case, the obtained equations do not have closed-form. Hence, they need to be obtained by numerically maximizing the log-likelihood function using a nonlinear optimization algorithm, such as the Newton algorithm or the quasi-Newton algorithm, such the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Nocedal and Wright, 1999).

An efficiently strategy to estimate the parameters of the TK distribution is solving this problem as a missing data problem, specifying the likelihood function defined in Equation (5) conveniently, as described in next subsection.

3.2 The EM algorithm

First, we can observe that Equation (2) can be rewrite as a mixture of beta distributions and a Kumaraswamy distribution, that is, by means of

$$f_{\mathrm{TK}}(y;a,b,\alpha,\beta)=\frac{a}{2}(2-2y)+\frac{b}{2}2y+\left(1-\frac{a+b}{2}\right)f_{\mathrm{K}}(y;\alpha,\beta), \tag{7}$$

where $f_1(y)=f_B(y;1,2)=2-2y$ and $f_2(y)=f_B(y;2,1)=2y$ are particular cases of the beta PDF defined as $f_B(y;\alpha^*,\beta^*)$, whereas $f_3(y)=f_{\mathrm{K}}(y;\alpha,\beta)$ corresponds to Kumaraswamy PDF described in Equation (1). In addition, here $w_1=a/2$, $w_2=b/2$ and $w_3=(1-(a+b)/2)$ are the weights such that $w_1+w_2+w_3=1$ and $0\le w_1,w_2,w_3\le 1$. Then, this problem can be solved as a finite mixture of distributions by using the expectation-maximization (EM) algorithm (McLachlan and Peel, 2004). The EM algorithm is a general method for finding maximum likelihood estimates when there are missing values or latent variables. The idea behind the EM algorithm applied to mixture models is to assume that the mixture is generated by missing observations of a discrete random variable Z, where $z_i\in\{1,2,3\}$ indicates which mixture component generated the observation y_i. The likelihood function of the complete data formed by the observed data ($\boldsymbol{y}$) and the unobserved data ($\boldsymbol{z}$), for a sample of n, is established by

$$\begin{aligned}p_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y},\boldsymbol{z};\Theta)=\prod_{i=1}^{n}p_{\boldsymbol{Y},\boldsymbol{Z}}(y_i,z_i;\Theta)&=\prod_{i=1}^{n}\left(\frac{a}{2}(2-2y_i)\right)^{\mathbb{1}_{z_i=1}}\left(\frac{b}{2}(2y_i)\right)^{\mathbb{1}_{z_i=2}}\\&\times\left(\left(1-\frac{a+b}{2}\right)f_{\mathrm{K}}(y_i;\alpha,\beta)\right)^{\mathbb{1}_{z_i=3}},\end{aligned}$$

where $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are the random vectors associated with ($\boldsymbol{y}$) and ($\boldsymbol{z}$), respectively. In addition, $\Theta=(a,b,\alpha,\beta)$ is the parameter vector and $\mathbb{1}$ is the indicator function, that is $\mathbb{1}_{z_i=j}=1$ if $z_i=j$ (with $j\in\{1,2,3\}$) holds, and $\mathbb{1}_{z_i=j}=0$, otherwise. Note that, in the EM algorithm, it is necessary to specify an auxiliary function Q, corresponding to the conditional expectation of the log-likelihood function with complete data $(\boldsymbol{y},\boldsymbol{z})$ given the observed data $Y=y$, and a parameterization $\Theta^{(p-1)}$, that is, we have that

$$\begin{aligned}Q\left(\Theta,\Theta^{(p-1)}\right)&=\mathrm{E}_{\boldsymbol{Y},\boldsymbol{Z},\Theta^{(p-1)}}(\log(p_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{Y},\boldsymbol{Z};\Theta)))\\&=\sum_{i=1}^{n}\mathrm{E}_{\boldsymbol{Y},\boldsymbol{Z},\Theta^{(p-1)}}(\log(p_{\boldsymbol{Y},\boldsymbol{Z}}(Y_i,Z_i;\Theta)))\\&=\sum_{i=1}^{n}\sum_{j=1}^{3}r_{ij}^{(p-1)}\log(p_{\boldsymbol{Y},\boldsymbol{Z}}(y_i,z_i;\Theta))\\&=\sum_{i=1}^{n}\sum_{j=1}^{3}r_{ij}^{(p-1)}(\log(w_jf_j(y_i;\Theta))),\end{aligned}$$

where $w_1=a/2$, $w_2=b/2$, $w_3=(1-(a+b)/2)$, $f_1(y_i;\Theta)=2-2y_i$, $f_2(y_i;\Theta)=2y_i$, $f_3(y_i;\Theta)=f_{\mathrm{K}}(y_i;\alpha,\beta)$ as in Equation (7), and

$$r_{ij}^{(p-1)}=\mathrm{P}(Z_i=j;Y_i=y_i,\Theta^{(p-1)})=\frac{w_j^{(p-1)}f_j(y_i;\Theta^{(p-1)})}{\sum_{l=1}^{3}w_l^{(p-1)}f_l(y_i;\Theta^{(p-1)})}.$$

In the E-Step, we need to find the expected value of $\mathbb{1}_{z_i=j}$ for $j = 1, 2, 3$ given y_i and the current parameterization $\Theta^{(p-1)}$, stated as

$$\mathrm{E}\left[\mathbb{1}_{z_i=j}; y_i, \Theta^{(p-1)}\right] = r_{ij}^{(p-1)}.$$

In the M-Step, we find $\Theta^{(p)}$ which maximizes $Q(\Theta, \Theta^{(p-1)})$. Calculating the derivates of Q with respect to w_1, w_2, w_3 under the restriction $w_1 + w_2 + w_3 = 1$, is possible obtain the estimators

$$w_j^{(p)} = \frac{\sum_{i=1}^{n} r_{ij}^{(p-1)}}{\sum_{i=1}^{n}\sum_{j=1}^{3} r_{ij}^{(p-1)}} = \frac{n_j^{(p-1)}}{n}.$$

Additionally, the derivates with respect to α and β lead to the usual maximum likelihood estimators of the Kumaraswamy distribution, which solve the equations expressed as

$$(\beta - 1)\,\frac{\sum_{i=1}^{n} r_{i3}^{(p-1)} y_i^{\alpha}\log(y_i)}{1 - y_i^{\alpha}} - \frac{n_3^{(p-1)}}{\alpha} - \sum_{i=1}^{n} r_{i3}^{(p-1)}\log(y_i) = 0 \tag{8}$$

$$\frac{n_3^{(p-1)}}{\beta} + \sum_{i=1}^{n} r_{i3}^{(p-1)}\log(1 - y_i^{\alpha}) = 0. \tag{9}$$

The corresponding estimates generated from Equations (8) and (9) can be obtained using the quasi-Newton algorithm. Once we update the parameters, we must repeat both the E and M steps, iteratively. In our case, in the M-step of the algorithm, we use the BFGS method to iteratively solve the non-linear maximization problem associated. The BFGS method is implemented in the `R` software by the functions `optim` and `optimx`; see `www.R-project.org` and R Core Team (2018).

4. Simulation study

In this section, we conduct a simulation study to compare the performance of the TK distribution with the Kumaraswamy distribution for samples generated from each of them.

4.1 Scenario of the simulations

In order to capture the particular tail behavior of each one, we use a sample size of 1000 and generate 100 sample sets to calculate the mean log-likelihood function and the Akaike information criterion (AIC). First, we simulate from the TK distribution with parameters given by $\Theta = (0.2, 0.5, 7, 10)$, that is, we simulate an asymmetric distribution with independent lifting in both tails to capture the essense of the proposed TK distribution. Second, we collect a sample from the Kumaraswamy distribution with parameters stated as $\Theta_B = (7, 10)$, that is, an asymmetric distribution but without lifted tails in its PDF.

4.2 Results of the simulations

In our first simulation from the TK distribution, we can observe in Table 1 that the TK distribution achieves a better fit than the Kumaraswamy distribution. In Table 2, we can appreciate that the Kumaraswamy distribution tries to fit the model by increasing the variance, that is, finding small values for α and β to overcome the inability of this distribution to raise the tails.

Table 1. Comparison between the mean log-likelihood and mean AIC of the TK and Kumaraswamy distributions for 100 samples of size 1000 drawn from a TK distribution with parameters $(0.2, 0.5, 7, 10)$

Distribution	Log-likelihood	AIC
TK	363.26	-718.53
Kumaraswamy	237.38	-470.75

Table 2. Comparison between the mean of the estimated parameters of the TK and Kumaraswamy distributions for 100 samples of size 1000 drawn from a TK distribution with parameters $(0.2, 0.5, 7, 10)$

	Estimated parameter			
Distribution	a	b	α	β
True	0.20	0.50	7.00	10.00
TK	0.20	0.49	7.03	10.28
Kumaraswamy	-	-	2.72	1.94

In Figure 2, we can see the histogram for simulated data from the TK distribution and the adjusted PDFs for the TK and Kumaraswamy distributions. The interpretation of the estimated parameters a, b is straightforward and corresponds exactly to the lifting of the tails of PDF in left and right tails respectively. In addition, note that the Kumaraswamy distribution is unable to capture this lifting.

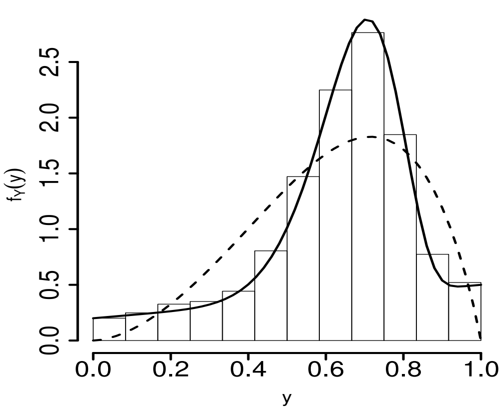

Figure 2. Histogram for simulate data set from TKD and adjusted PDFs for two different models: In solid line, the TK model; In dashed line the Kumaraswamy model.

Table 3 reports the relative bias (RB) and the root-mean-squared error (RMSE) for each parameter estimator over the 100 simulated samples under the TK distribution. They are defined as

$$\mathrm{RB}(\theta) = \frac{1}{100}\sum_{i=1}^{100}\left(\frac{\widehat{\theta}^{(i)} - \theta}{\theta}\right), \quad \mathrm{MSE}(\theta) = \frac{1}{100}\sum_{i=1}^{100}(\widehat{\theta}^{(i)} - \theta)^2,$$

where θ represents any particular parameter, and $\widehat{\theta}^{(i)}$ is the estimate of θ for the i-th sample. Table 3 reports that the estimate of each parameter in each data set is reasonable when fitting the TK distribution.

Table 3. RB and RMSE of each parameter under 100 samples of size 1000 drawn from a TK distribution with parameters $(0.2, 0.5, 7, 10)$.

Indicator	Parameter a	b	α	β
RB	0.00088	−0.00287	0.00038	0.00276
RMSE	0.00554	0.04537	0.08497	0.87242

In our second simulation from the Kumaraswamy distribution, we can observe in Table 4 that the TK distribution achieve an equally good fit than the Kumaraswamy distribution. In Table 5, note that the TK distribution gives similar estimates for the parameters, compared to the Kumaraswamy distribution.

Table 4. Log-likelihood and AIC for simulated data

Distribution	Log-likelihood	AIC
TK	843.52	−1679.03
Kumaraswamy	843.29	−1682.58

Table 5. Comparison between the mean of the estimated parameters of the TK and Kumaraswamy distributions for 100 samples of size 1000 drawn from a Kumaraswamy distribution with parameters $(7, 10)$

Distribution	Estimated parameter a	b	α	β
True	0.00	0.00	7.00	10.00
TK	2.85e-04	1.12e-03	7.07	10.29
Kumaraswamy	-	-	7.05	10.22

Unsurprisingly, when the sample is generated from the Kumaraswamy distribution, we do not see significant differences on the mean log-likelihood and AIC achieved by the two adjusted Kumaraswamy and TK distributions. When the sample is drawn from the TK distribution with a difference between the its two tails, $a = 0.2$ and $b = 0.5$, the best fit in terms of the mean log-likelihood and AIC is achieved by the TK distribution. This can be explained by the fact that the data generated from the tails of the distribution cannot be captured only by using a Kumaraswamy distribution.

5. Empirical illustrations with real data

In this section, in order to illustrate the TK distribution in practice, we apply the proposed results to two real data sets. We compare the goodness of fit between the TK and Kumaraswamy distributions.

5.1 Pollutant emissions in Angolan thermal power plants

Data on Angolan thermal power plants span the period 2010 to 2014 were obtained from a enterprise named ENE-EP. They are based on the plants balance sheets and income statements, which are gathered and organized by ENE-EP as part of regular reporting. The variables of interest for our study are the percent slacks for reduction in pollutant emissions/discharges for CO2 and H2O. This scalar measure deals directly with the input excesses and the output shortfalls of the decision making unit concerned and is typically

Table 7. Log-likelihood and AIC values or H2O data

	Distribution	
Indicator	TK	Kumaraswamy
Log-likelihood	82.21	24.40
AIC	−156.43	−44.81

used as efficiency measure for modeling environmental performance (Barros and Wanke, 2017).

Efficiency scores computed from the slacks based model with undesirable (bad) outputs (SBM-Undesirable) range between 0 and 1, where 1 denotes a maximum or 100 % of efficiency. This suggests that a given thermal plant is operating at the frontier of the productive technology. In fact, efficiency is a productivity ratio between two DMUs: in data envelopment analysis (DEA) based models, all plants are assessed against a convex frontier of best practices formed by the most productive DMUs that can deliver higher outputs consuming lower inputs or benchmarks. In DEA, each production unit is known as a decision making unit (DMU).

Before proceeding, it is worth noting that if the variable assumes the extreme values of zero and one ($Y^* \in [0,1]$), then a practical transformation must be applied (Smithson and Verkuilen, 2006) by

$$y = \frac{(n-1)}{n} y^* + \frac{1}{2n}, \quad y^* \in [0,1],$$

where n is the sample size.

In our study, we consider 160 efficiency scores ($n = 160$) for the 32 Angolan thermal power plants from 2010 to 2014. This efficiency scores has been measures for CO2 and H2O. From Figures 3 and **??**, note that the data distribution have a lifted left tail. Then, it is justified to fit the TK distribution to model these data. The model under consideration is defined by

$$Y_i \overset{\text{IND}}{\sim} \text{TK}(a, b, \alpha, \beta), \quad i = 1, \ldots, 160,$$

where IND stands for independent. Note in Tables 6 and 7 that the TK distribution achieves a best fit compared to the Kumaraswamy distribution. In Tables 8 and 9, we report the estimated parameters. It is clear that the distribution in this example is lifted in the left tail, since for CO2 data we have $\widehat{a} = 0.3806$ and $\widehat{b} = 0$, whereas for H2O data, $\widehat{a} = 0.3303$ and $\widehat{b} = 0$, and then we can see that these estimates have a very intuitive interpretation since the tails of the PDF are lifted visually in these quantities. This fact is attempted to be compensated in the Kumaraswamy distribution by increasing the variance (decreasing $\widehat{\alpha}$ and $\widehat{\beta}$).

Table 6. Log-likelihood and AIC values for CO2 data

	Distribution	
Indicator	TK	Kumaraswamy
Log-likelihood	66.86	14.00
AIC	−125.73	−23.99

In Figure 3, we can see the adjusted PDFs for the two different models, with the TK distribution being the model that better captures the distribution of the data.

Table 8. Estimated parameters for CO2 data

	Estimated parameter			
Distribution	a	b	α	β
TK	0.3806	2.50e-45	7.0541	5.1930
Kumaraswamy	-	-	1.7546	1.2278

Table 9. Estimated parameters for H2O data

	Estimated parameter			
Distribution	a	b	α	β
TK	0.3303	1.12e-43	8.2015	5.5768
Kumaraswamy	-	-	2.1070	1.2778

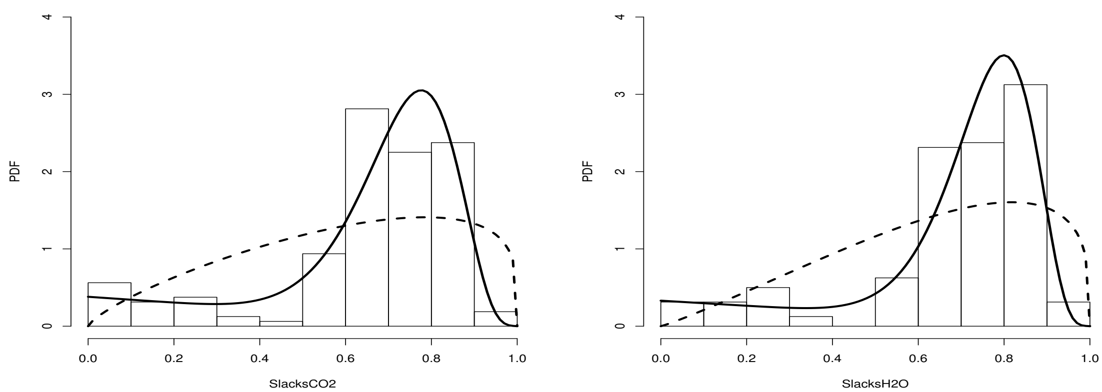

Figure 3. Adjusted PDFs for two different models: in solid line, the TK distribution; and in dotted line the Kumaraswamy distribution for CO2 (left) and H2O (right) data.

5.2 University admission score

We analyze the average score of university admission test in 1295 school establishments in Metropolitan region of Chile, 2016. This test is applied to students who have graduated from school in Chile, which is carried out at a national level and covers different areas of knowledge. In Chile, this test is named "prueba de selección universitaria (PSU)" and allows the student's admission to the different universities of the country, depending on the result obtained in this test. The data set is available in the website https://es.datachile.io.

We are interested in the performance of the students who have applied to the PSU. To measure performance, a total of 1295 average scores per establishment have been collected in the Metropolitan region of Chile and scored in the interval $(0, 1)$ through the transformation proposed by Smithson and Verkuilen (2006) formulated as

$$y = \frac{n-1}{n}\frac{y^* - a_1}{a_2 - a_1} + \frac{1}{2n}, \quad y^* \in [a_1, a_2].$$

Then, $y \in (0, 1)$ and in our case $a_1 = 293.5$, $a_2 = 715.5$ and $n = 1295$. We can see in Figure 4 that the data distribution have a lifted right tail and slightly lifted left tail. Thus, it is justified to fit the TK distribution to model these data. The model under consideration is

Table 10. Log-likelihood and AIC values for PSU data

	Distribution	
Indicator	TK	Kumaraswamy
Log-likelihood	393.68	352.95
AIC	−779.35	−701.90

Table 11. Estimated parameters for PSU data

	Estimated parameter			
Distribution	a	b	α	β
TK	0.0066	0.3072	2.9844	6.6608
Kumaraswamy	-	-	2.3976	3.3506

defined by

$$Y_i \overset{\text{IND}}{\sim} \text{TK}(a, b, \alpha, \beta), \quad i = 1, \ldots, 1295.$$

We can see in Table 10 that the TK distribution achieves a best fit compared to the Kumaraswamy distribution. In Table 11 we report the estimated parameters. It is clear that the distribution in this example is lifted in the tails ($\widehat{a} = 0.0066$ and $\widehat{b} = 0.3072$) and we can see that these estimates have a very intuitive interpretation since the tails of the PDF are lifted visually in these quantities. This fact is once again attempted to be compensated in the Kumaraswamy distribution by increasing the variance (decreasing $\widehat{\alpha}$ and $\widehat{\beta}$).

In Figure 4, we can see the adjusted PDFs for the two different models, with the TK distribution being the model that better captures the distribution of the data.

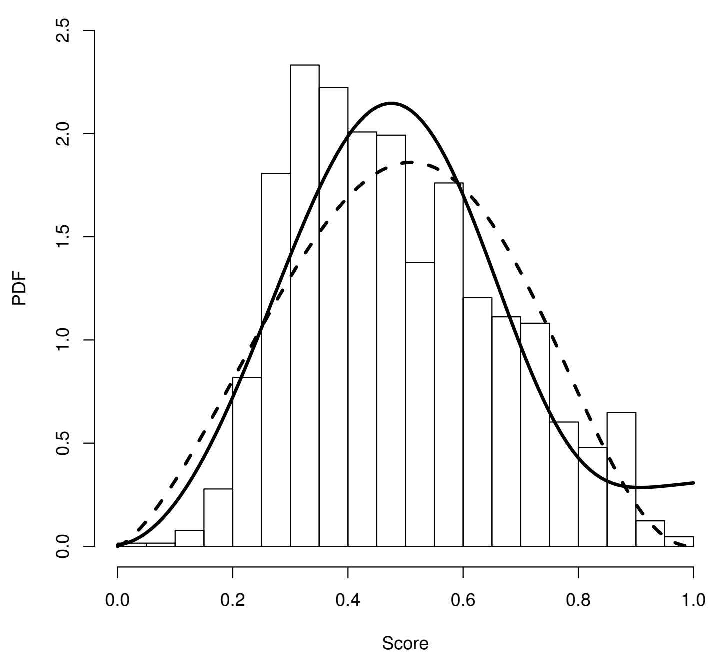

Figure 4. Adjusted PDFs for two different models: in solid line, the TK distribution; and in dotted line the Kumaraswamy distribution for PSU data.

6. Concluding remarks and future research

The Kumaraswamy distribution and other distributions derived from this have been very used in practice. However, until now, it has not been proposed a distribution that allows us to raise the tails of the probability density function in the case of having data accumulated in one or both ends. In this work, we introduced a new four-parameter model called the trapezoidal Kumaraswamy distribution, that is a generalization of the Kumaraswamy distribution which has the rectangular Kumaraswamy distribution as a particular case. The trapezoidal Kumaraswamy distribution comes to solve the problem of adjusting data with some concentration in the extremes. The trapezoidal Kumaraswamy distribution can be represented as a finite mixture model generated by two specific beta distributions and the Kumaraswamy distribution. The trapezoidal Kumaraswamy distribution presented two additional parameters with respect to the Kumaraswamy distribution and they have the advantage of being very intuitive, because they represent the lifting of the probability density function in the tails. The estimation procedure for their parameters is straightforward and in this paper was presented a methodology of estimation achieving good results both with the simulated and real data. In the simulation studies, we observed marked differences in favor of the trapezoidal Kumaraswamy distribution when the samples have some concentration in the tails. In the empirical illustration, the trapezoidal Kumaraswamy distribution turned out to be the model that best adjusted the data and that attended to the essence of the data distribution with some accumulation at the ends. Then, we can conclude that the trapezoidal Kumaraswamy distribution seems to be a new robust alternative for modeling data bounded on the unit interval.

Some open problems that arose from the present investigation are the following:

- An extension of this work that is under development is to propose the reparametrized trapezoidal Kumaraswamy distribution in terms of its mean and connect to it a regression structure, then we will propose a trapezoidal Kumaraswamy regression model.
- The development of a bayesian methodology can be of interest for an alternative implementation.
- The benefits of the distribution will be extended to any bounded distribution.
- A re-parametrization of the trapezoidal Kumaraswamy distribution in terms of its mode is of interest, as this will allow us to connect its mean to a regression structure in a similar manner to that as in generalized linear models.
- A quantile regression model with a trapezoidal Kumaraswamy distributed response will be studied.

Therefore, the proposed results in this study opens opportunities to explore other theoretical and numerical issues.

Acknowledgements

The authors thank Editors and Referees for their suggestions which allowed us to improve the presentation of this work. J.I. Figueroa-Zúñiga acknowledges funding support by grant VRID 217.014.027-1.0, from the Universidad de Concepción, Chile. G. Ibacache-Pulgar acknowledges funding support by grant FONDECYT 11130704, Chile. B. Lagos-Alvarez acknowledges funding support by grant VRID 216.014.026-1.0, from University of Concepción, Chile.

References

Akinsete, A., Famoye, F. and Lee, C., 2014. The Kumaraswamy-geometric distribution. Journal of Statistical Distributions and Applications, 1, 1–17.

Akinsete, A. and Famoye, F., 2008. The beta-Pareto distribution. Statistics, 42, 547–563

Barros, C. and Wanke, P., 2017. Efficiency in Angolan thermal power plants: Evidence from cost structure and pollutant emissions. Energy, 130, 129–143.

Cordeiro, G. and de Castro, M., 2011. A new family of generalized distributions. Journal of Statistical Computation and Simulation, 81, 883–898.

Cordeiro, G. and dos Santos, B., 2012. The beta power distribution. Brazilian Journal of Probability and Statistics, 26, 88–112

Cordeiro, G., Nadarajah, S., and Ortega, E., 2012. The Kumaraswamy Gumbel distribution. Statistical Methods and Applications, 21, 139–168.

Cordeiro, G., Ortega, M. and Nadarajah, S., 2010. The Kumaraswamy Weibull distribution with application to failure data. Journal of the Franklin Institute, 347, 1399–1429.

de Pascoa, M., Ortega, E. and Cordeiro G., 2011. The Kumaraswamy generalized gamma distribution with application in survival analysis. Statistical Methodology, 8, 411–433.

De Santana, T., Ortega, E., Cordeiro, G. and Silva, G., 2012. The Kumaraswamy-log-logistic distribution. Journal of Statistical Theory and Applications, 11, 265–291.

Eugene, N., Lee, C. and Famoye, F., 2002. Beta-normal distribution and its applications. Communications in Statistics: Theory and Methods, 3, 497–512.

Jones, M.C., 2009. Kumaraswamy distribution: A beta-type distribution with some tractability advantages. Statistical Methodology, 6, 70–81.

Kumaraswamy, P., 1980. A generalized probability density function for double-bounded random processes. Journal of Hydrology, 46, 79–88.

Liang, Y., Sun, D., He, C. and Schootman, M., 2014. Modeling bounded outcome scores using the binomial-logit-normal distribution. Chilean Journal of Statistics, 5, 3–14.

McLachlan, G. and Peel, D., 2004. Finite Mixture Models. Wiley, New York.

Mead, M. and Abd-Eltawab, A., 2014. A note on Kumaraswamy Fréchet distribution. Australian Journal of Basic and Applied Sciences, 8, 294–300.

Nadarajah, S. and Kotz, S., 2004. The beta-Gumbel distribution. Mathematical Problems in Engineering, 10, 323–332

Nadarajah, S. and Kotz, S., 2006. The beta exponential distribution. Reliability Engineering and System Safety, 91, 689–697

Nocedal, J. and Wright, S., 1999. Numerical Optimization. Springer, New York.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. Available at http://www.r-project.org

Smithson, M. and Verkuilen, J., 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychological Methods, 11, 54–71.

Appendix

This appendix presents one piece of R codes used for fitting the trapezoidal Kumaraswamy distribution.

```
library(extraDistr)
# For evaluation of Kumaraswamy probability density function (dkumar)
## Trapezoidal Kumaraswamy probability density function ##
dtrapkum<-function(data,w1,w2,alpha,beta){ # w1 and w2 are the weights
                                             # described in the paper
eval<-w1*dbeta(data,1,2)+w2*dbeta(data,2,1)+(1-w1-w2)*dkumar(data,alfa,beta)
return(eval)
}

# Function used in Algorithm to estimate the Kumaraswamy parameters
model<-function(x,data){
alfa0<-(sum(tau3)/x[1])+sum(tau3*log(data))
-sum(tau3*(x[2]-1)*data^x[1]*log(data)/(1-data^x[1]))
beta0<-(sum(tau3)/x[2])+sum(tau3*log(1-data^x[1]))
c(alfa0=alfa0,beta0=beta0)
}

# Initial values
a<-0.1
b<-0.2
alfa<-2
beta<-2
w1<-a/2
w2<-b/2
w3<-1-w1-w2

# EM algorithm #
for(k in 1:1000){
# E step
tau1<-w1*dbeta(data,1,2)/(dtrapkum(data,w1,w2,alpha,beta))
tau2<-w2*dbeta(data,2,1)/(dtrapkum(data,w1,w2,alpha,beta))
tau3<-(1-w1-w2)*dkumar(data,alfa,beta)/(dtrapkum(data,w1,w2,alpha,beta))
# M step
pi1<-sum(tau1)/length(data)
pi2<-sum(tau2)/length(data)
solution<-multiroot(f=model,start = c(alfa,beta),maxiter=5000,data=data)
solution
alfa<-solution$root[1]
beta<-solution$root[2]
}
```

Information for authors

The editorial board of the Chilean Journal of Statistics (ChJS) is seeking papers, which will be refereed. We encourage the authors to submit a PDF electronic version of the manuscript in a free format to the Editors-in-Chief of the ChJS (E-mail: `chilean.journal.of.statistics@gmail.com`). Submitted manuscripts must be written in English and contain the name and affiliation of each author followed by a leading abstract and keywords. The authors must include a"cover letter" presenting their manuscript and mentioning: "We confirm that this manuscript has been read and approved by all named authors. In addition, we declare that the manuscript is original and it is not being published or submitted for publication elsewhere".

Preparation of accepted manuscripts

Manuscripts accepted in the ChJS must be prepared in Latex using the ChJS format. The Latex template and ChJS class files for preparation of accepted manuscripts are available at `http://chjs.mat.utfsm.cl/files/ChJS.zip`. Such as its submitted version, manuscripts accepted in the ChJS must be written in English and contain the name and affiliation of each author, followed by a leading abstract and keywords, but now mathematics subject classification (primary and secondary) are required. AMS classification is available at `http://www.ams.org/mathscinet/msc/`. Sections must be numbered 1, 2, etc., where Section 1 is the introduction part. References must be collected at the end of the manuscript in alphabetical order as in the following examples:

Arellano-Valle, R., 1994. Elliptical Distributions: Properties, Inference and Applications in Regression Models. Unpublished Ph.D. Thesis. Department of Statistics, University of São Paulo, Brazil.

Cook, R.D., 1997. Local influence. In Kotz, S., Read, C.B., and Banks, D.L. (Eds.), Encyclopedia of Statistical Sciences, Vol. 1., Wiley, New York, pp. 380-385.

Rukhin, A.L., 2009. Identities for negative moments of quadratic forms in normal variables. Statistics and Probability Letters, 79, 1004-1007.

Stein, M.L., 1999. Statistical Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Tsay, R.S., Peña, D., and Pankratz, A.E., 2000. Outliers in multivariate time series. Biometrika, 87, 789-804.

References in the text must be given by the author's name and year of publication, e.g., Gelfand and Smith (1990). In the case of more than two authors, the citation must be written as Tsay et al. (2000).

Copyright