University of Concepción
Graduate address
Faculty of Physical Sciences and Mathematics
Master's Program in Statistics

Trapezoidal Kumaraswamy Distribution (Distribución Kumaraswamy Trapezoidal).

Thesis to qualify for the Master's Degree in Statistics

RODRIGO ALONSO SANHUEZA PARKES
CONCEPCIÓN-CHILE
2018

Guide professor: Jorge Isaac Figueroa Zuñiga
Department of Statistics
Faculty of Physical Sciences and Mathematics
University of Concepción

UNIVERSITY OF CONCEPCIÓN
FACULTY OF PHYSICAL SCIENCES AND MATHEMÁTICS
DEPARTAMENT OF STATÍSTICS

THESIS TO APPLY FOR THE MAGISTER DEGREE IN STATISTICS

Trapezoidal Kumaraswamy Distribution.

Guide professor	:	Jorge Isaac Figueroa Zuñiga	Firm	.....
Assistent professor	:	Guillermo Ferreira Cabezas	Firm	.....
Assistent professor	:	Bernardo Lagos Alvarez	Firm	.....
Assistent professor	:	German Ibacache Pulgar	Firm	.....
Memorant name	:	Rodrigo Alonso Sanhueza Parkes	Firm	.....
Phone	:	(9)71965477		
e-mail	:	rsanhueza@udec.cl		

Concepción, 2018

I dedicate my work and my life to God, because thanks to Him I have achieved everything I have, it has given me health, intelligence, strength and the necessary faith to achieve the goals I have set for myself.

To my family for being the reason to advance and fight against any adversity, being a fundamental part of my daily life, to those who are and to those who have left. Especially my beloved wife Natalia Riquelme and my children Dafne and Alexander, my great loves.

A Duchess who accompanied me for more than a decade showing me his great loyalty and love, being my loyal companion

Acknowledgements

Thanks to my beloved family for supporting me unconditionally and for giving me the energy necessary to achieve my goal; to my wife for giving me the necessary time to study and perfect myself in what I love and for being a fundamental pillar in my life, being the best woman that God could put in my way to go hand in hand and grow old together.

My daughter Daphne, my little princess, for her unconditional love and although many times I could not understand that her father could not play or share with her, she was always there, with a smile, with a hug or a kiss, supporting and motivating me.

To my son Alexander for being the last angel that came to my home and give me his love with smiles and opening my arms to receive a hug.

To my parents and my brothers for always being there, trusting and concerned about my well-being, always encouraging me to continue.

I also thank Dr. Jorge Figueroa for guiding me and supporting me in this work which I could not finish without his advice and the Magister teachers who gave me classes and gave me minutes of their time to clear up the doubts, they are: Mauricio Castro, Nora Serdiukova, Guillermo Ferreira and Bernardo Lagos. To the teacher Luisa Rivas who many times I lend her support and do not hesitate to clarify some doubts despite not being my teacher in the magister.

To my colleague Maria Jose, who always supported me and helped me to better understand some contents of different subjects.

To the authorities of my school, who gave me the time I needed to advance in my studies.

But first of all, I thank God for all that he has given me, because he gave me the best that I have had, my family and because he has always been and will be in my life.

Abstract

In the year 1980 Poondi Kumaraswamy proposed the Kuamaraswamy distribution, which is very similar to the beta distribution (it is also restricted to the interval (0,1)), but it has a great advantage over it, which is to have a distribution function accumulated in a closed form which is more beneficial for intensive calculation activities such as simulation modeling and estimation of models by methods based on simulation. The problem of this distribution and its extensions proposed in the following years is that they have not been able to adjust the data that sometimes are concentrated in each of the extremes or both ends independently.

This work has the purpose of showing the proposal of a new distribution, which has been called trapezoidal kumaraswamy distribution which has been originated by mixing the Kumaraswamy distribution and the Beta distribution, making the tails of the density function more flexible in one of the extremes or in both of them independently with which a greater adjustment of the data is achieved.

We can appreciate the properties of this new model and the estimation of parameters. Finally, a simulation study and an application of real data is presented, with the intention of showing the best adjustment obtained.

Table of Contents

List of Figures

List of Tables

Chapter 1

Trapezoidal Kumaraswamy Distribution

manuscript ID BJPS1810-020 to Brazilian Journal of Probability and Statistics

Authors: Rodrigo Sanhueza-Parkes, Jorge Figueroa-Zuñiga, German Ibacache-Pulgar, Bernardo Lagos-Alvarez

Abstract

Kumaraswamy Models have been a very studied tool in the analysis and modeling of limited-range continuous variables. This mainly because of the great flexibility of its density function, which can cover a wide range of different shapes and their cumulative distribution in closed form. As is discussed in this paper, many variants of the Kumaraswamy distribution have been studied but these do not have the possibility of lifting the tails of this distribution. However, in many situations, the data are bounded and tail-area events occur at either end or at both ends independently. To model this scenario, in this work we propose the "Trapezoidal Kumaraswamy Model".

This paper is centered on the development of the "Trapezoidal Kumaraswamy Model", which have two intuitive additional parameters respect to "Kumaraswamy Model" and is a generalization of this. We study its density function and we derive some fundamental properties such as the moments, moment generating function and characteristic function. Finally, the "Trapezoidal Kumaraswamy Model" is rewritten conveniently as a mixture model and we show that its parameters can be estimated by means of the EM algorithm. We report results of an application to a real data set. Model fitting comparisons with several alternative models indicates that the model proposed presents the best fit and so it can be quite useful in real applications.

keywords Maximum likelihood - Kumaraswamy distribution - Mixture Model - EM algorithm

Jorge Figueroa-Zuñiga
Departament of Statistics, University of Concepción, Concepción, Chile
E-mail:jifiguer@gmail.com

Rodrigo Sanhueza Parkes
Departament of Statistics, University of Concepción, Concepción, Chile

German Ibacache
Departament of Statistics, University of Concepción, Concepción, Chile

Bernardo Lagos
Departament of Statistics, University of Concepción, Concepción, Chile

1.1 Introduction

A good alternative for modeling continuous data restricted to a bounded interval, is the double bounded distribution Kumaraswamy (1980) (renamed as kumaraswamy distribution Jones (2009)). This given by the variety of density shapes that can be accommodated. The probability density function (pdf) of a random variable Y following a Kumaraswamy distribution of parameters $\alpha, \beta > 0$ is given by

$$f_K(y \mid \alpha, \beta) = \alpha\beta y^{\alpha-1}(1-y^{\alpha})^{\beta-1}, \quad y \in (0,1) \tag{1.1}$$

where $\alpha, \beta > 0$. Here,

$$\mathrm{E}(Y) = m_1 \quad \text{and} \quad \mathrm{Var}(Y) = m_2 - m_1^2, \tag{1.2}$$

where m_k is the k-th moment of the kumaraswamy distribution given by

$$m_k = \frac{\beta\Gamma(1+\frac{k}{\alpha})\Gamma(\beta)}{\Gamma(1+\frac{k}{\alpha}+\beta)} = \beta B\left(1+\frac{k}{\alpha}, \beta\right) \tag{1.3}$$

with B, the beta function.

The Kumaraswamy distribution is very flexible. However, do not consider tail-area events nor greater flexibility in the variance specification. In order to add flexibility into the model, other distributions derived from the Kumaraswamy distribution have been proposed. The Kumaraswamy Weibull distribution (Cordeiro et al, 2010) and the Kumaraswamy-G distribution (Cordeiro and de castro, 2011) includes two additional positive parameters (they studied some of their mathematical properties by presenting special submodels), the Kumaraswamy generalized gamma distribution (de Pascoa et al, 2011) which is able to model bathtub-shaped hazard rate functions (the importance of this distribution is in its capacity to model functions of monotonous failure frequency and not monotone, which are fairly common in life-time data analysis and reliability), the Kumaraswamy Gumbel distribution (Cordeiro et al, 2010) which is probably the most widely applied statistical distribution for problems in engineering, the Kumaraswamy-log-logistic distribution (de Santana et al, 2012). the Kumaraswamy-geometric distribution (Akinsete et al, 2014 and the kumaraswamy fréchet distribution (Mead and Abd-Eltawab, 2014), among other distributions of the same family.

However, the Kumaraswamy distribution, as their extensions, are unable to fit data in which some sample points are concentrated at either, one end or both

ends independently. In this work, we propose a new bounded distribution with is able to model this behavior.

The article is organized as follows. In Section 2, the trapezoidal Kumaraswamy distribution is proposed and their basic properties are discussed. In Section 3, the estimation of parameters is develop through a convenient reparametrization of the Trapezoidal Kumaraswamy distribution given in Section 2. In Section 4 we perform a simulation study, both the Trapezoidal Kumaraswamy distribution and the Kumaraswamy distribution, comparing the results obtained in both. In Section 5 an application of the proposed model is presented using the Australian Institute of Sport data set. The results are compared with the classical Kumaraswamy distribution. Finally, discussions and observations appear in Section 6 of the proposed model and the specific numerical results.

1.2 Trapezoidal Kumaraswamy Distribution

In practice, the Kumaraswamy distribution has been a useful tool for modelling bounded data. However, is common in many cases, to have data concentrated at either, one end or both ends independently, and hence, it misses an extension which allows to model this situation and that it conserve the great flexibility of the Kumaraswamy distribution. Hence, to this issues we propose the Trapezoidal Kumaraswamy distribution with the following pdf

$$f_{TK}(y \mid a, b, \alpha, \beta) = a + (b - a)y + \left(1 - \frac{a+b}{2}\right) f_K(y \mid \alpha, \beta). \tag{1.4}$$

with $0 < y < 1$, $0 \leq a, b \leq 2$, $0 \leq a + b \leq 2$ and $f_K(y \mid \alpha, \beta)$ is the Kumaraswamy density function of parameters α and β given in (1.1).

The parameters a and b can be intuitively interpreted as the lift at the left and right tails of the pdf respectively (see figure 1.1 and 1.2). The notation $Y \sim TK(a, b, \alpha, \beta)$ will be used through the paper.

As a particular case, we have that when $a = b = 0$, the standard Kumaraswamy distribution is recovered (1.1) and as particular case, the Rectangular Kumaraswamy distribution is proposed when $a = b = \theta$.

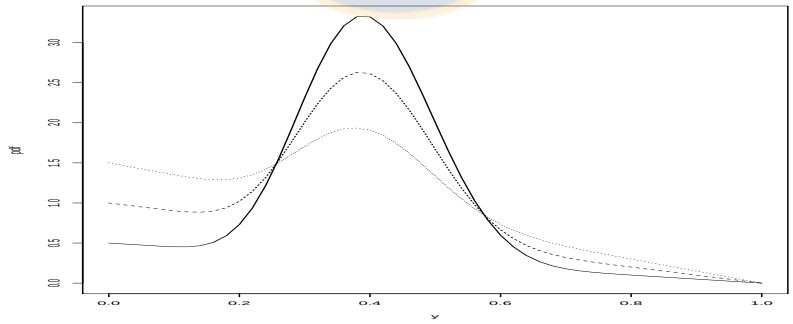

Figure 1.1: *Examples of Trapezoidal Kumaraswamy pdf with* $\alpha = 10, \beta = 15$ *and different values of the parameters* (a, b)*:* $(a, b) = (0.5, 0)$ *(solid line),* $(a, b) = (1, 0)$ *(dashed line) and* $(a, b) = (1.5, 0)$ *(dotted line).*

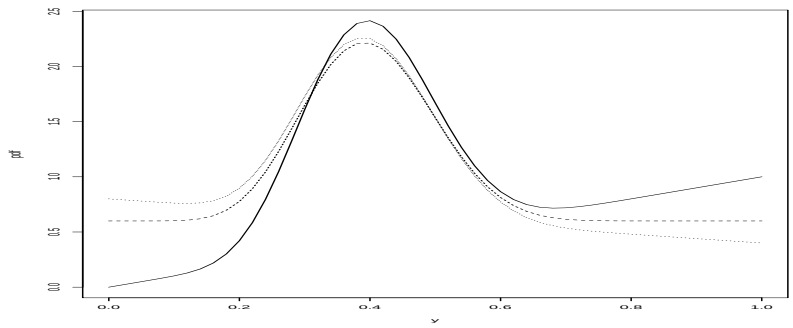

Figure 1.2: *Examples of Trapezoidal Kumaraswamy pdf with* $\alpha = 10, \beta = 15$ *and different values of the parameters* (a, b)*:* $(a, b) = (0, 1)$ *(solid line),* $(a, b) = (0.6, 0.6)$ *(dashed line) and* $(a, b) = (0.8, 0.4)$ *(dotted line).*

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

$$m_k = \mathrm{E}(Y^k) = \frac{a}{k+1} + \frac{b-a}{k+2} + \left(1 - \frac{a+b}{2}\right) m_k^*, \tag{1.5}$$

where m_k^* is the k-th moment of the kumaraswamy distribution of paramenters α, β. Then (2.1) can be casted as

$$\begin{aligned} m_k &= \frac{a}{k+1} + \frac{b-a}{k+2} + \left(1 - \frac{a+b}{2}\right) \frac{\beta\Gamma\left(1 + k/\alpha\right)\Gamma\left(\beta\right)}{\Gamma\left(1 + \beta + k/\alpha\right)} \\ &= \frac{a}{k+1} + \frac{b-a}{k+2} + \left(1 - \frac{a+b}{2}\right) \beta B\left(1 + k/\alpha, \beta\right). \end{aligned} \tag{1.6}$$

With the above expression it is easy to deduce that

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

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

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

and its characteristic function is given by

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

1.3 Estimation of parameters in Trapezoidal Kumaraswamy model

In this section we discuss how to estimate, efficiently, the parameters of the Trapezoidal Kumaraswamy distribution. Therefore, we first write the likelihood function for a sample of n observations as

$$\mathcal{L}(a,b,\alpha,\beta) = \prod_{i=1}^{n}\left(a + (b-a)y_i + \left(1 - \frac{a+b}{2}\right) f_K(y_i \mid \alpha,\beta)\right). \quad (1.7)$$

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

$$l(a,b,\alpha,\beta) = \sum_{i=1}^{n} ln\left(a + (b-a)y_i + \left(1 - \frac{a+b}{2}\right) f_K(y_i \mid \alpha,\beta)\right). \quad (1.8)$$

The maximum likelihood estimators of a, b, α and β are obtained from the differentiation of (1.8) with respect to the mentioned parameters and equating to zero, but 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; for details, see (Nocedal andWright (1999)

An efficiently strategy to obtain the parameters estimations is solving this problem as a missing data problem, specifing the likelihood function given in (1.7) conveniently, as described below.

First, we can observe that the equation given by (1.4) can be rewrite as a mixture of beta distributions and a Kumaraswamy distribution, i.e.,

$$f_{TK}(y \mid a,b,\alpha,\beta) = \frac{a}{2}(2-2y) + \frac{b}{2}2y + \left(1 - \frac{a+b}{2}\right) f_K(y \mid \alpha,\beta), \quad (1.9)$$

where, $f_1(y) = f_B(y \mid 1,2) = 2-2y$, $f_2(y) = f_B(y \mid 2,1) = 2y$ are particular cases of the beta density function $f_B(y \mid \alpha^*,\beta^*)$ and $f_3(y) = f_K(y \mid \alpha,\beta)$ correspond to Kumaraswamy density function described in (1.1). Besides $w_1 = \frac{a}{2}$, $w_2 = \frac{b}{2}$ and $w_3 = \left(1 - \frac{a+b}{2}\right)$ are the weights such that $w_1 + w_2 + w_3 = 1$ and $0 \leq w_1, w_2, w_3 \leq 1$. Then, this problem can be solved as a finite mixture of distributions by using the expectation-maximization (EM) algorithm; for details, see 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 of the complete data (Y, Z) for a sample of n observations is given by

$$p_{Y,Z}(\boldsymbol{y}, \boldsymbol{z} \mid \Theta) = \prod_{i=1}^{n} p_{Y,Z}(y_i, z_i \mid \Theta) = \prod_{i=1}^{n} \left(\frac{a}{2}(2-2y_i)\right)^{\mathbb{1}_{z_i=1}} \left(\frac{b}{2}(2y_i)\right)^{\mathbb{1}_{z_i=2}} \times \left(\left(1-\frac{a+b}{2}\right) f_K(y_i \mid \alpha, \beta)\right)^{\mathbb{1}_{z_i=3}},$$

where $\Theta = (a, b, \alpha, \beta)$ is the parameter vector and $\mathbb{1}$ is the indicator function, i.e. $\mathbb{1}_{z_i=j} = 1$ if $z_i = j$ (with $j \in \{1, 2, 3\}$) holds, and $\mathbb{1}_{z_i=j} = 0$ otherwise. Then, in the EM algorithm is necessary to specify an auxiliary function $\mathcal{Q}$, the conditional expectation of the complete data (Y, Z) given the observed data Y, and a parameterization $\Theta^{(p-1)}$, i.e.,

$$\begin{aligned}
\mathcal{Q}\left(\Theta, \Theta^{(p-1)}\right) &= \mathrm{E}_{Z|Y,\Theta^{(p-1)}}(\log p_{Y,Z}(\boldsymbol{y}, \boldsymbol{z} \mid \Theta)) \\
&= \sum_{i=1}^{n} \mathrm{E}_{Z|Y,\Theta^{(p-1)}}(\log p_{Y,Z}(y_i, z_i \mid \Theta)) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{3} r_{ij}^{(p-1)}(\log p_{Y,Z}(y_i, z_i \mid \Theta)) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{3} r_{ij}^{(p-1)}(\log(w_j f_j(y_i \mid \Theta))),
\end{aligned}$$

where $w_1 = \frac{a}{2}$, $w_2 = \frac{b}{2}$, $w_3 = \left(1 - \frac{a+b}{2}\right)$ and $f_1(y_i \mid \Theta) = 2 - 2y_i$, $f_2(y_i \mid \Theta) = 2y_i$, $f_3(y_i \mid \Theta) = f_K(y_i \mid \alpha, \beta)$as in (2.2), and

$$r_{ij}^{(p-1)} = P(Z_i = j \mid Y_i = y_i, \Theta^{(p-1)}) = \frac{w_j^{(p-1)} f_j(y_i \mid \Theta^{(p-1)})}{\sum_{l=1}^{3} w_l^{(p-1)} f_l(y_i \mid \Theta^{(p-1)})}. \quad (1.10)$$

For 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 $\Theta^{(p-1)}$, given by

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

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

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

On the other hand, the derivates with respect at α and β lead to the usual maximum likelihood estimators of the Kumaraswamy distribution, which solve the equations

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

and

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

They can be obtained using the quasi-Newton algorithm. Once updated the parameters, repeat both, the E and M steps, iteratively.

1.4 Simulation Study

We develop a simulation study to compare the performance of the Trapezoidal Kumaraswamy distribution (TKD) in comparison with the Kumaraswamy distribution for samples generated from each of them. In order to capture the particular tail behavior of each one, we use a sample size of 1000 and generate 100 sample sets in order to calculate the mean Log-Likelihood and the Akaike Information Criterion (AIC).

First, we simulate from the TKD with parameters given by $\Theta = (0.2, 0.5, 7, 10)$. From table 1.1 we can observe that the TKD achieves a better fit than the Kumaraswamy distribution. In table 1.2, we can appreciate that the Kumaraswamy distribution tries to fit the model by increasing the variance, i.e., finding small values for α and β to overcome the inability of this distribution to raise the tails.

Table 1.1: Comparison between the Mean Log-likelihood and Mean AIC of the Trapezoidal Kumaraswamy and Kumaraswamy distributions for 100 samples of size 1000 drawn from a Trapezoidal Kumaraswamy distribution with parameters $(0.2, 0.5, 7, 10)$

	Mean Log-Likelihood	Mean AIC
Trapezoidal Kumaraswamy	363.26	-718.53
Kumaraswamy	237.38	-470.75

Table 1.2: *Comparison between the mean of the estimated parameters of the Trapezoidal Kumaraswamy and Kumaraswamy distributions for 100 samples of size 1000 drawn from a Trapezoidal Kumaraswamy distribution with parameters* $(0.2, 0.5, 7, 10)$

	Mean Estimated Parameters			
	a	b	α	β
True	0.2	0.5	7	10
Trapezoidal Kumaraswamy	0.20	0.49	7.03	10.28
Kumaraswamy	-	-	2.72	1.94

In figure 1.3, we can see the histogram for simulated data from TKD and the adjusted densities for TKD amd Kumaraswamy distribution are represented.

The interpretation of the estimation in parameters a, b is straightforward and correspond exactly to the lifting of the tails of pdf in left and right hand respectively. In this figure we can appreciate that the Kumaraswamy distribution is not able of capture this lifting.

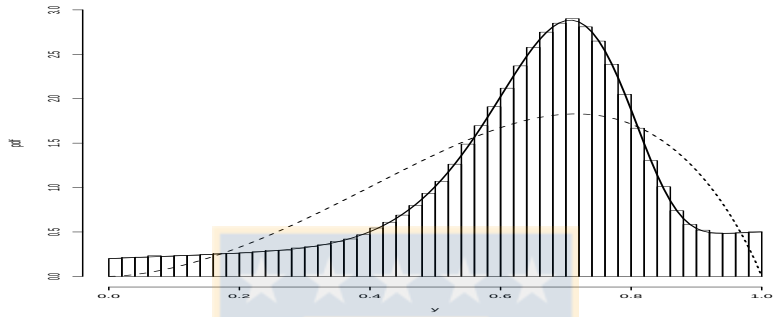

Figure 1.3: *Histogram for simulate data set from TKD and adjusted densities for two different models: In solid line, the Trapezoidal Kumaraswamy model; In dashed line the Kumaraswamy model.*

Table 1.3 present the relative bias (RelBias) and the root-mean-squared error ($\sqrt{\text{MSE}}$) for each parameter estimator over the 100 simulated samples under the TKD. They are defined as

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

where θ represents any particular parameter, and $\widehat{\theta}^{(i)}$ is the estimation of θ for the i-th sample. The Table 1.3 show that the estimation of each parameter in each data set is good when the TKD is adjusted.

Table 1.3: Relbias and root-squared error of each parameter under 100 samples of size 1000 drawn from a Trapezoidal Kumaraswamy distribution with parameters $(0.2, 0.5, 7, 10)$

	a	b	α	β
Relbias	0.00088	-0.00287	0.00038	0.00276
$\sqrt{MSE}$	0.00554	0.04537	0.08497	0.87242

Second, we take a sample from the Kumaraswamy distribution with parameters given by $\Theta_B = (7, 10)$. In table 1.4 we can observe that the TKD achieve an equally good fit than the Kumaraswamy distribution. In table 1.5 we can appreciate that the TKD give similar estimates for the parameters, compared to Kumaraswamy distribution.

Table 1.4: Log-Likelihood and AIC

	Mean Log-Likelihood	Mean AIC
Trapezoidal Kumaraswamy	843.52	-1679.03
Kumaraswamy	843.29	-1682.58

Table 1.5: *Comparison between the mean of the estimated parameters of the Trapezoidal Kumaraswamy and Kumaraswamy distributions for 100 samples of size 1000 drawn from a Kumaraswamy distribution with parameters* $(7, 10)$

	Mean Estimated Parameters			
	a	b	α	β
True	0	0	7	10
Trapezoidal Kumaraswamy	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 see non significant differences on the mean log-likelihood and mean AIC achieved by the two adjusted distributions (Kumaraswamy and TKD). When the sample is drawn from the TKD 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 mena AIC is achieved by the Trapezoidal beta model. This can be explained by the fact that the data generated from the tails of the distribution can not be capture only by using a Kumaraswamy distribution.

1.5 Real data application

To illustrate the Trapezoidal Kumaraswamy distribution in practice, we apply the proposed model to a real dataset and we compare the goodnes of fit of this flexible distribution with the godness of fit of the Kumaraswamy distribution. We analyse the Australian Institute of Sport (AIS) dataset available in the library sn in R (http://azzalini.stat.unipd.it/ SN/index.html). We consider only the data of the 102 male athletes in the AIS dataset. We are interested in the body fat percentage $(Bfat)$ of each athlete. Normal ranges for $Bfat$ in adult men are $5\% - 25\%$ approximately, see Jeukendrup and Gleeson (2010). Therefore, we consider the following transformation $Y = (Bfat-5)/(25-5)$. We can see, in figure 1.4, that the data distribution have a lifted left tail. Then, it is justified to fit the Trapezoidal Kumaraswamy distribution to model this data.

The model under consideration is defined by:

$$y_i \mid a, b, \alpha, \beta \overset{\text{ind}}{\sim} \text{TK}(a, b, \alpha, \beta) \quad , \quad i = 1, \ldots, 102.$$

We can see in table 1.6 that the TKD achieves the best fit compared to the Kumaraswamy distribution. In table 1.7 we present the estimated parameters. It is clear that the distribution in this example is lifted at the left ($\hat{a} = 0.5981$ and $\hat{b} = 0.00$), this fact is attempted to be compensated in the Kumaraswamy distribution by increasing the variance (decreasing $\hat{\alpha}$ and $\hat{\beta}$).

Table 1.6: Log-Likelihood and AIC

	Trapezoidal Kumaraswamy	Kumaraswamy
Log-Likelihood	70.7724	63.6881
AIC	-133.5447	-123.3762

Table 1.7: Estimations

	$\hat{a}$	$\hat{b}$	$\hat{\alpha}$	$\hat{\beta}$
Trapezoidal Kumaraswamy	0.5981	0.0000	2.0417	32.8129
Kumaraswamy	-	-	1.3055	5.6495

In figure 1.4, we can see the adjusted densities for the two different models, being the TKD the model that beter captures the distribution of the data.

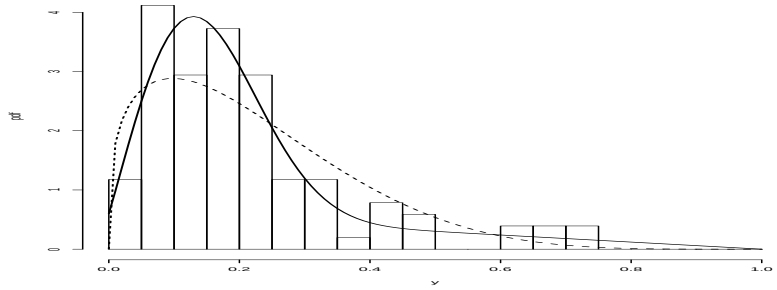

Figure 1.4: *Adjusted densities for two different models: In solid line, the Trapezoidal Kumaraswamy model; In dashed line the Kumaraswamy model.*

1.6 Concluding remarks

Kumaraswamy distribution and other distributions derived from this has been very used in the practice, but so far it has not been proposed a distribution that allows to raise the tails of the probability density function (pdf) in the case of having data accumulated in one or both ends. In this work, a new distribution called "Trapezoidal Kumaraswamy distribution" (TKD) has been proposed and that comes to solve the problem of adjusting data with some concentration in the extremes. The TKD is a mixture model generated by two specific beta distributions and the Kumaraswamy distribution, being the Kumaraswamy distribution a particular case of the TKD. The TKD present two additional parameter respect to Kumaraswamy distribution and these have the advantage of being very intuitive, because they represent the lifting of the pdf in the queues. The estimation procedure for their parameters is straightforward and is presented a clear methodology of estimation in this paper achieving good results both the simulation studies and the real data application.

1.7 References

Akinsete A, Famoye F, Lee C (2014). The Kumaraswamy-geometric distribution. Journal of statistical distributions and applications 1:1-17

Cordeiro G, de Castro M (2011). A new family of generalized distributions. Journal of statistical computation and simulation 81(7):883-898

Cordeiro G, Nadarajah S, Ortega E (2012). The kumaraswamy gumbel distribution. Statistical Methods and Applications 21(2):139-168

De Santana T, Ortega E, Cordeiro G, Silva G (2012). The kumaraswamy log-logistic distribution. Journal of Statistical Theory and Applications 11(3):265-291

Cordeiro GM, Ortega EM, Nadarajah S (2010). The kumaraswamy weibull distribution with application to failure data. Journal of the Franklin Institute 347(8):1399-1429

Jeukendrup A, Gleeson M (2010). Sport Nutrition-2nd Edition: An introduction to energy production and performance. Human Kinetics

Jones MC (2009). Kumaraswamys distribution: A beta-type distribution with some tractability advantages. Statistical Methodology 6(1):708-1

Kumaraswamy P (1980). A generalized probability density function for double-bounded random processes. Journal of Hydrology 46(1-2):79-88

McLachlan G, Peel D (2004). Finite mixture models. John Wiley and Sons

Mead ME, Abd-Eltawab AR (2014). A note on kumaraswamy frchet distribution. Australian Journal of Basic and Applied Sciences 8(15):294-300

Nocedal J, Wright SJ (1999). Numerical optimization. New York: Springer-Verlag

De Pascoa M, Ortega E, Cordeiro G (2011). The kumaraswamy generalized gamma distribution with application in survival analysis. Statistical methodology

8(5):411-433

Chapter 2

Apendice

Below you can see in more detail the steps that were applied to reach the results shown in this work

The moment generator function (Equation 1.5)

$$\begin{aligned} m_k = \mathrm{E}(Y^k) &= \int_0^1 y^k f(y) dy \\ &= \int_0^1 y^k (a + (b-a)y + \left(1 - \frac{a}{2} - \frac{b}{2}\right) f_k(y|\alpha, \beta) dy \\ &= a \int_0^1 y^k dy + (b-a) \int_0^1 y^{k+1} dy + \left(1 - \frac{a}{2} - \frac{b}{2}\right) \int_0^1 f_k(y|\alpha, \beta) dy \\ &= \frac{a}{k+1} + \frac{b-a}{k+2} + \left(1 - \frac{a}{2} - \frac{b}{2}\right) m_k^* \end{aligned}$$

With which, we conclude that:

$$m_k = \mathrm{E}(Y^k) = \frac{a}{k+1} + \frac{b-a}{k+2} + \left(1 - \frac{a+b}{2}\right) m_k^*, \tag{2.1}$$

The equation on page 4, regarding the variance

$$
\begin{aligned}
\text{Var}(y) &= \text{E}(y^2) - (\text{E}(y))^2 \\
&= \frac{a}{3} + \frac{b-a}{4} + \left(1 - \frac{a+b}{2}\right) m_k - \left(\frac{a+2b}{6} + \left(1 - \frac{a}{2} - \frac{b}{2}\right) m_k\right)^2 \\
&= \frac{a+3b}{12} + \left(1 - \frac{a+b}{2}\right) m_k - \frac{(a+2b)^2}{36} - \frac{a+2b}{3}\left(1 - \frac{a+b}{2}\right) m_k \\
&- \left(1 - \frac{a+b}{2}\right)^2 m_k^2 \\
&= \frac{3a+9b-(a+2b)^2}{36} \\
&+ \left(1 - \frac{a+b}{2}\right) \beta B\left(1 + \frac{k}{\alpha}, \beta\right)\left(1 - \frac{a+2b}{3} - \left(1 - \frac{a+b}{2}\right) \beta B\left(1 + \frac{k}{\alpha}, \beta\right)\right)
\end{aligned}
$$

In both cases, B is the beta function

Equation 1.9 on page 7

$$
\begin{aligned}
f_{TK}(y|a,b,\alpha,\beta) &= a + (b-a)y + \left(1 - \frac{a}{2} - \frac{b}{2}\right) f_k(y|\alpha,\beta) \\
&= a - ay + by + \left(1 - \frac{a}{2} - \frac{b}{2}\right) f_k(y|\alpha,\beta) \\
&= \frac{2a(1-y)}{2} + by + \left(1 - \frac{a}{2} - \frac{b}{2}\right) f_k(y|\alpha,\beta)
\end{aligned}
$$

And simplifying, we have to:

$$
f_{TK}(y \mid a,b,\alpha,\beta) = \frac{a}{2}(2-2y) + \frac{b}{2}2y + \left(1 - \frac{a+b}{2}\right) f_K(y \mid \alpha,\beta), \quad (2.2)
$$

Then, we will remember the density function of the beta distribution, which is:

$$
f_B(y \mid \alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \quad , \ 0 < y < 1,
$$

And for the particular cases in which the parameters α and β take the values 1 and 2, and vice versa, we would have the following

$$f_B(y \mid 1,2) = \frac{\Gamma(3)}{\Gamma(1)\Gamma(2)} y^0(1-y) = 2(1-y) \quad , \ 0 < y < 1,$$

$$f_B(y \mid 2,1) = \frac{\Gamma(3)}{\Gamma(2)\Gamma(1)} y(1-y)^0 = 2y \quad , \ 0 < y < 1,$$

That is to say, $f_\beta(y|1,2) = 2 - 2y \qquad f_\beta(y|2,1) = 2y$
And replacing the beta functions, we have to

$$f_{TK}(y|a,b,\alpha,\beta) \ = a + (b-a)y + \left(1 - \tfrac{a}{2} - \tfrac{b}{2}\right) f_k(y|\alpha,\beta)$$

it can be rewritten in the following way, with which we can see that the above can be written as a mixture of distributions (mix of two beta distributions and a kumaraswamy distribution)

$$f_{TK}(y|a,b,\alpha,\beta) = \frac{a}{2} f_\beta(y|1,2) + \frac{b}{2} f_\beta(y|2,1) + \left(1 - \frac{a+b}{2}\right) f_k(y|\alpha,\beta)$$