UNIVERSITY OF CONCEPCIÓN
FACULTY OF PHYSICAL SCIENCES AND MATHEMATICS
DEPARTAMENT OF STATISTICS

# TRAPEZOIDAL KUMARASWAMY DISTRIBUTION AND SEM ALGORITHM.

THESIS TO APPLY FOR THE MASTER DEGREE IN STATISTICS

By : Juan Guillermo Toledo Balboa Firm .....

Guide professor : Jorge Isaac Figueroa Zúñiga Firm .....

december, 2021
Concepción, Chile

# Table of Contents

**List of figures** iii

**List of tables** iv

**1 Trapezoidal Kumaraswamy Distribution** 1

1.1 Introduction . . . 3
1.2 Background . . . 4
1.2.1 Trapezoidal Kumaraswamy Distribution . . . 4
1.2.2 EM Algorithm . . . 5
1.3 Bayesian estimation for the Trapezoidal Kumaraswamy model . . . 5
1.3.1 The model . . . 5
1.3.2 SEM algorithm . . . 6
1.4 Simulation Study . . . 10
1.4.1 Scenario of the simulations . . . 10
1.4.2 Results of the simulations . . . 10
1.5 Empirical illustrations with real data . . . 14
1.5.1 Covid-19 cases . . . 14
1.6 Concluding remarks . . . 19

# List of Figures

1.1 Examples of TK (solid line) and K (dashed line) pdf with $(\alpha, \beta) = (5, 13)$ and different values of the parameters $(a, b)$ in TK: $(a, b) = (0.4; 0.4)$, $(a, b) = (0; 0.7)$, $(a, b) = (0.7; 0.0)$ respectively. . . . . 4
1.2 The real daily data set in the left side and the histogram with the applied transformation in the right. . . . . 15
1.3 Fitting for both models taking the parameter's mean. Solid line: TK distribution. Dashed line: Kumaraswamy distribution. . . . . 16
1.4 Fitting for TK and two-mixture TK models taking the parameter's mean. . 17

# List of Tables


1.1 Comparison between the Mean DIC, mean EAIC and Mean EBIC of the TK and Kumaraswamy distributions for 100 samples of size 1000 drawn from a TK distribution with parameters $(0.1, 0.3, 5, 10)$ . . . . . 11
1.2 Estimated posterior medians, means and credibility interval (CI) for 100 samples of size 1000 drawn from a TK distribution with parameters $(0.1, 0.3, 5, 10)$ 11
1.3 Relbias and root-squared error of each parameter under 100 samples of size 1000 drawn from a TK distribution with parameters $(0.1, 0.3, 5, 10)$ . 12
1.4 Comparison between the Mean DIC, mean EAIC and Mean EBIC of the TK and Kumaraswamy distributions for 100 samples of size 1000 drawn from a Kumaraswamy distribution with parameters $(5, 10)$ . . . . . 12
1.5 Estimated posterior medians and means for 100 samples of size 1000 drawn from a Kumaraswamy distribution with parameters $(5, 10)$ . . . . . 12
1.6 Comparison between the Mean DIC, mean EAIC and Mean EBIC of the TK and Kumaraswamy distributions and its mixture models for Daily confirmed Covid19 data set. . . . . 16
1.7 Estimated posterior means, medians and credibility interval (CI) for Daily confirmed Covid19 data set. . . . . 17
1.8 two-mixture TK distribution's Mean DIC, mean EAIC and Mean EBIC for Daily confirmed Covid19 data set. . . . . 18
1.9 Estimated posterior means, medians and credibility interval (CI) in two-mixture TK distribution for Daily confirmed Covid19 data set. . . . . 18

# Chapter 1

# Trapezoidal Kumaraswamy Distribution

Authors: Juan Toledo-Balboa, Jorge Figueroa-Zuñiga, Bernardo Lagos-Alvarez

## Abstract

Models involving the Kumaraswamy distribution have been a very studied in the past years in the analysis and modeling of bounded continuos variables. In this paper we focus on one in particular: the Trapezoidal Kumaraswamy model. We present an estimation method for its parameters based on bayesian approaches: the Stochastic EM algorithm (SEM), which avoids the most common issues of the classical EM. Then, we apply this method to the daily covid-19 cases in Chile using this model.



**keywords** SEM algorithm; Kumaraswamy distribution; Bayesian Analysis; Metropolis-Hastings; Mixture Model.



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

Juan Toledo Balboa
Departament of Statistics, University of Concepción, Concepción, Chile

Bernardo Lagos Álvarez
Departament of Statistics, University of Concepción, Concepción, Chile
Grupo de Matemática Aplicada (GMA), Facultad de Ciencias, Universidad del Bío-Bío, Concepción-Chillán, Chile.

## 1.1 Introduction

The Kumaraswamy distribution was seen for first time when prof. Kumaraswamy proposed it under the name "double-bounded probability density function" (Kumaraswamy, 1980) for modeling hydrological processes. Later, it was renamed as "Kumaraswamy distribution" by Jones (2009).

The Trapezoidal Kumaraswamy (TK) model (Figueroa et al., 2020) is a new proposal, even more flexible than it, being based on it and keeping its properties. The TK's distinctive feature unlike others proposed previously, like Kumaraswamy Weibull distribution (Cordeiro et al., 2010) or Kumaraswamy generalized Gamma distribution (de Pascoa et al., 2011) among others, while they have been developed to be more flexible than their base models, do not allow to fit scenarios where tail-area events occur.

Rewriting properly the TK model as a mixture, it will allow us to leverage the existing tools to work with, just like the EM algorithm by Dempster et al. (1977). However, this algorithm may present some disadvantages at the step maximization. For example, the dependence on initial values for multimodal likelihoods (Celeux and Govaert, 1992). To solve this and other issues presented in the following sections, Celeux and Diebolt (1985) proposed an alternative based on a bayesian approach: the Stochastic Expectation Maximization (SEM) algorithm, by including a new step called S-step, which consists in simulation methods like Gibbs Sampling and Metropolis-Hastings algorithms from the posterior distribution of the parameters, i. e., it takes reasonable prior information and use it in order to deal with these problems.

This article is organized as follows: Section 1.2 presents in a general way the TK model and the EM algorithm applied to mixture models. In section 1.3 we describe how SEM algorithm is implemented for the model and its main difference with EM: first, the proposal of parameters as random variables and leveraging the prior information known about them. Second, proposing joint prior distributions and third, through Gibbs Sampling and Metropolis-Hastings we simulate them from their posterior distribution. Section 1.4 shows how this algorithm works for 100 data sets at different scenarios. Finally, in section 1.5 we applied SEM algorithm to a real data set about new daily cases of Covid-19 disease in Chile.

## 1.2 Background

### 1.2.1 Trapezoidal Kumaraswamy Distribution

A random variable $Y$ follows a Trapezoidal Kumaraswamy distribution with parameters $a, b, \alpha, \beta$ if its probability density function (pdf) is given by

$$f_{TK}(y; a, b, \alpha, \beta) = a + (b - a)y + \left(1 - \frac{a+b}{2}\right) f_K(y; \alpha, \beta), \quad y \in (0, 1) \tag{1.1}$$

where $0 \leq a, b \leq 2,\ a + b \leq 2$ and $f_K(\alpha, \beta)$ is the Kumaraswamy's pdf (Kumaraswamy, 1980) with parameters $\alpha, \beta > 0$. Here, the expectation and variance of TK distribution are

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

where $m_k$ denotes the k-th moment of the TK distribution, that is

$$m_k = \frac{a}{k+1} + \frac{b-a}{k+2} + \left(1 - \frac{a+b}{2}\right) \beta B(1 + k/\alpha, \beta),$$

where $B(\alpha^*, \beta^*)$ is the Beta function of $\alpha^*$ and $\beta^*$.

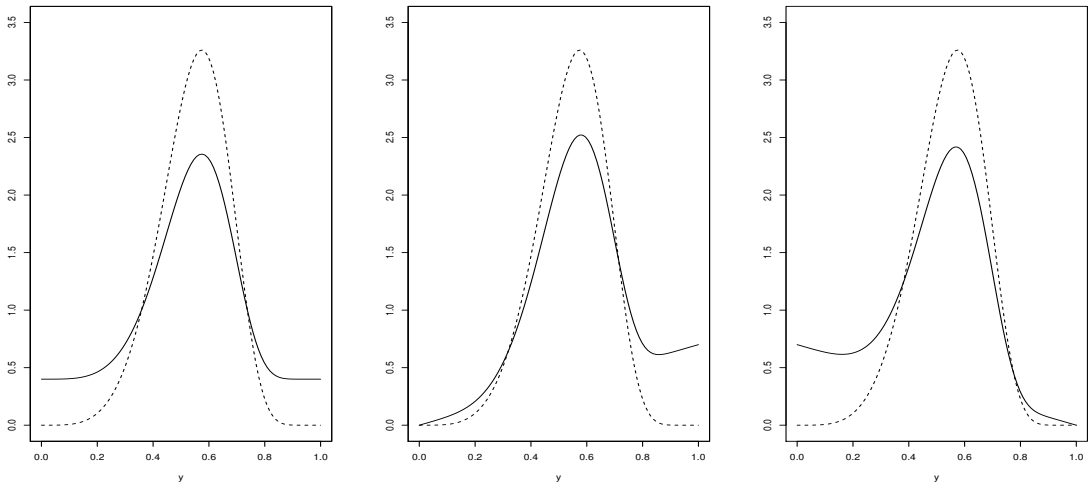


Figure 1.1: Examples of TK (solid line) and K (dashed line) pdf with $(\alpha, \beta) = (5, 13)$ and different values of the parameters $(a, b)$ in TK: $(a, b) = (0.4; 0.4)$, $(a, b) = (0; 0.7)$, $(a, b) = (0.7; 0.0)$ respectively.

### 1.2.2 EM Algorithm

EM algorithm is a general method for finding maximum likelihood estimates when there are missing values or latent variables. It can be used for mixture density models, beginning from an observed data set of random variable $Y$, that will be classified to a mixture component according to a probability. The basic idea is assuming that the data set comes from a non observable discrete random variable $Z$, which indicates what mixture component generated the observation $y_i$ and fit these probabilities in each iteration until some convergence criterion. The following describes in general the steps of the algorithm:

For each variable $Y_i,\ 1 \leq i \leq n$, where $n$ denotes the sample size; $Z_i = (Z_{i1}, Z_{i2}, Z_{i3})$ is the vector indicating to which component $Y_i$ belongs, such that $Z_{ij} = \{1, 0\}$ if $Y_i$ belongs to the component $j$ or not, respectively. Thus, the EM algorithm consists in repeating the steps below until convergence:

1. E-step: Compute $\hat{Z}_{ij}^{(t)}$ parameter estimates from initial values at $t = 0$.
2. M-step: Update the parameters estimates according to:
   $\hat{\Theta}^{(t)} = argmax_{\Theta} L(\Theta^{(t-1)}, Z, Y)$

where $t$ indexes the $t$-th iteration, $\hat{Z}_{ij}^{(t)}$ is the probability that the $i$-th observation comes from the $j$-th component of the mixture and $L$ is the loglikelihood function. Unfortunately, EM algorithm has some disadvantages, like the dependence on initial values for the case of multimodal likelihoods that may carry out to saddle points (Bouguila et al, 2006) and slow convergence in several others (Celeux and Govaert, 1992). In order to avoid these problems, a lot of extensions of the EM algorithm have been proposed, very many based on bayesian approaches; SEM algorithm is one of them. For details regarding to classic EM, see McLachlan and Peel (2004). In section 1.3 we give some necessary details to deal with SEM, a very popular extension of the EM algorithm to estimate mixture model's parameters.

## 1.3 Bayesian estimation for the Trapezoidal Kumaraswamy model

In this section we talk about how to estimate efficiently the parameters of the TK distribution through SEM algorithm.

### 1.3.1 The model

As mentioned in section 1.2.2 EM algorithm and its variants require to represent the model as a mixture of densities. In order to satisfy this requirement, we rewrite the density

function of the equation (1.1) as follows

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

where $f_1(y) = f_B(y;1,2) = 2-2y$ and $f_2(y) = f_B(y;2,1) = 2y$ are the densities of the Beta distribution $f_B(y;\alpha^*,\beta^*)$ for random variable $Y$. Then, the model expressed as a mixture is

$$f_{TK}(y;a,b,\alpha,\beta) = \frac{a}{2} f_B(y;1,2) + \frac{b}{2} f_B(y;2,1) + \left(1 - \frac{a+b}{2}\right) f_K(y;\alpha,\beta) \tag{1.2}$$

and $w_1 = \frac{a}{2}$ $w_2 = \frac{b}{2}$, $w_3 = 1 - \frac{a+b}{2}$ represent the weights of each density they are next to, respectively. Therefore, the parameters to estimate in this model are $\Theta = (w,\theta)$ where $w = (w_1, w_2, w_3)$ and $\theta = (\alpha,\beta)$.

Such as the EM algorithm, SEM requires the model be a mixture and expressed in terms of missing data. If, for each variable $Y_i$, $1 \leq i \leq n$; $Z_i = (Z_{i1}, Z_{i2}, Z_{i3})$ is a tridimensional vector indicating to which component $j$, where $j = \{1,2,3\}$ of the model the i-th observed data from $Y$ belongs, such that:

$$Z_{ij} = \begin{cases} 1 & \text{if } Y_i \text{ belongs to component } j \\ 0 & \text{otherwise.} \end{cases}$$

Then, the likelihood function for the complete data $(Y,Z)$ es given by

$$f_{YZ}(y,z;\Theta) = \prod_{i=1}^{n}\prod_{j=1}^{3}(w_j f_j(y_i;\theta_j))^{Z_{ij}} \tag{1.3}$$

and the loglikelihood is

$$L(\Theta,y,z) = \sum_{j=1}^{3}\sum_{i=1}^{n} Z_{ij} log(w_j f_j(y_i|\theta_j)), \tag{1.4}$$

where $f_j$ indicates the $j$-th density function from the TK model.

### 1.3.2 SEM algorithm

In this subsection we describe step by step the application of the algorithm to the TK model from initial values for $\Theta$, i.e., for $t = 0$, $w^{(0)} = (w_1^{(0)}, w_2^{(0)}, w_3^{(0)})$, $\theta^{(0)} = (\alpha^{(0)}, \beta^{(0)})$.

### E-step

As $Z_{ij}$ represent the belonging of each observed data to some of the components of the model, we can estimate the real value by its expectation (McLachlan and Peel, 2004), that is, compute

$$\hat{Z}_{ij}^{(t)} = P\left(Z_{ij} = 1 | Y_i = y_i, \Theta^{(t-1)}\right) = \frac{w_j^{(t-1)} f_j(y_i|\alpha^{(t-1)}, \beta^{(t-1)})}{\sum_{l=1}^{3} w_l^{(t-1)} f_l(y_i|\alpha^{(t-1)}, \beta^{(t-1)})} \tag{1.5}$$

In practical terms, $\hat{Z}_{ij}^{(t)}$ is the posterior probability that the $i$-th observation arises from the $j$-th component of the model given the observations and the parameters in a previous iteration.

### S-step

The EM algorithm could present some problems after this point in the case of multimodal likelihoods, what implies dependence on initial conditions and without guarantee of obtain a local maximum, but to saddle points as we mentioned in subsection 1.2.2. As an alternative specially for this inconvenient, it has been proposed a different approach based on bayesian approach. In particular, the information of the complete data $(Y, Z)$ is combined with prior information about the parameters $\Theta$, that is, assigning a prior probability distribution $\pi(\Theta)$ and according to Bayes theorem:

$$\pi(\Theta|y, z) = \frac{f(y, z|\Theta)\pi(\Theta)}{\int f(y, z|\Theta)\pi(\Theta)d\Theta} \propto f(y, z|\Theta)\pi(\Theta), \tag{1.6}$$

which implies that we could know the posterior distribution for each unknown parameter. The inclusion of the S-step of SEM algorithm may be considered a bayesian extension of the classical EM, since it consists in simulating $Z$, $W$, $\theta$ from their posterior distribution as indicated in equation (1.6). As the reader will observe in subsection 1.3.2, the simulation techniques we are going to use are known as Gibbs Sampling and Metropolis-Hastings algorithms (Casella and Robert, 2010). In the S-step, we simulate $Z$ from it's posterior distribution, which we can assume intuitively as Multinomial of $\hat{Z}$ of size one with probabilities $(\hat{Z}_{i1}, \hat{Z}_{i2}, \hat{Z}_{i3})$. Therefore, $Z_i^{(t)} \sim \mathcal{M}(1; \hat{Z}_{i1}^{(t)}, \hat{Z}_{i2}^{(t)}, \hat{Z}_{i3}^{(t)})$. This gives a vector with 1 as one of its components and two others as 0, i. e., the data has been assignated randomly to a component indicating with the value of 1.

### M-step

The M-step depends on the form of the density and often its solution does not exist in closed form. In addition, it is possible that the parameters can become too high during the process, causing numeric problems (Bouguila et al, 2006). In order to avoid this issue, a

viable option could be simulating the parameters instead computing them.
According to Diebolt and Robert (1994) this method is one of the most used in bayesian estimation for mixture models. The association of each observation from $Y_i$ with a non observable variable $Z_i$ allows us to simulate $W$ from its posterior distribution; $\pi(w|y,z)$. The reader can observe that $W$ is independent of $Y$, that means $\pi(w|y,z) = \pi(w|z)$. The Gibbs sampler standard for mixture models is based on successive simulations of $Z$, $W$ and $\theta$.

We know that

$$\pi(w|z) \propto \pi(z|w)\pi(w). \tag{1.7}$$

Then, given the prior known information we have about $w = (w_1, w_2, w_3)$, $0 < w_j < 1$, $\sum_{j=1}^{3} w_j = 1$, it is intuitive to assume that $W$ comes from a Dirichlet distribution, i. e.,

$$\pi(w) = \frac{\Gamma(\sum_{j=1}^{3} \eta_j)}{\prod_{j=1}^{3} \Gamma(\eta_j)} \prod_{j=1}^{3} w_j^{\eta_j - 1} \tag{1.8}$$

where $\eta = (\eta_1, \eta_2, \eta_3)$ is the parameter vector of the Dirichlet distribution. Besides, we have that

$$\pi(z|w) = \prod_{i=1}^{n} \pi(z_i|w) = \prod_{i=1}^{n} w_1^{z_{i1}} w_2^{z_{i2}} w_3^{z_{i3}} = \prod_{i=1}^{n} \prod_{j=1}^{3} w_j^{z_{ij}} = \prod_{j=1}^{3} w_j^{n_j} \tag{1.9}$$

where $n_j = \sum_{i=1}^{n} \mathbb{1}_{z_{ij}=j}$ are the number of observations assignated to component $j$ in S-step. This means that every $n_j$ is the sum of each $zij = 1$ simulated from $Z_i$ in every iteration. Then, Assembling both of equations (1.8) and (1.9), we can prove that

$$\pi(w|z) \propto \mathcal{D}(\eta_1 + n_1, \eta_2 + n_2, \eta_3 + n_3) \tag{1.10}$$

where $\mathcal{D}$ is the Dirichlet distribution with parameters $(\eta_1 + n_1, \eta_2 + n_2, \eta_3 + n_3)$. Note that $\eta_1, \eta_2, \eta_3$ are hyperparameters. Therefore, $w^{(t)} \sim \mathcal{D}(\eta_1 + n_1^{(t)}, \eta_2 + n_2^{(t)}, \eta_3 + n_3^{(t)})$.

For the estimation of parameters from Kumaraswamy distribution $\theta = (\alpha, \beta)$, it is effective to simulate them from the joint posterior distribution $\pi(\theta^{(t)}|z^{(t)}, y) = \pi(\alpha^{(t)}, \beta^{(t)}|z^{(t)}, y)$. For this case, we use Metropolis-Hastings algorithm as follows:

Given $\epsilon \sim \mathcal{N}_2\left(0, \Sigma_{2x2}\right)$, $\Sigma = \sigma^2 \mathbb{I}_2$ and propose

$$\begin{pmatrix} log(\alpha) \\ log(\beta) \end{pmatrix} = \begin{pmatrix} log(\alpha^{(t-1)}) \\ log(\beta^{(t-1)}) \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \tag{1.11}$$

what means that $\alpha, \beta$ emulate each of them a Log-Normal distribution as indicated below:

1. Generate the candidates $\alpha^*, \beta^*$ from:

$$\alpha^* \sim \mathcal{LN}\left(log(\alpha^{(t-1)}), \sigma^2\right) \tag{1.12}$$
$$\beta^* \sim \mathcal{LN}\left(log(\beta^{(t-1)}), \sigma^2\right) \tag{1.13}$$

where $\sigma$ is a hyperparameter. Also, we know that

$$\pi(\alpha^{(t)}, \beta^{(t)}|z^{(t)}, y) = \pi(y, z^{(t)}|\alpha^{(t)}, \beta^{(t)})\pi(\alpha^{(t)}, \beta^{(t)}) \tag{1.14}$$
$$= \left(\prod_{i=1}^{n} f(y_i|\alpha, \beta)\right)^{\mathbb{1}_{z_i=3}} \pi(\alpha^{(t)})\pi(\beta^{(t)}). \tag{1.15}$$

Then, with probability $\delta$ given by

$$\delta\left(\begin{pmatrix}\alpha^* \\ \beta^*\end{pmatrix}\middle|\begin{pmatrix}\alpha^{(t-1)} \\ \beta^{(t-1)}\end{pmatrix}\right) = min\left\{1, \frac{\pi(\alpha^*, \beta^*|z^{(t)}, y)}{\pi(\alpha^{(t-1)}, \beta^{(t-1)}|z^{(t)}, y)}\right\} \tag{1.16}$$

we accept or reject the new values of the parameters (both of them).

2. Accept/reject:

For a value $u$ generated from $U \sim \mathcal{U}_{[0,1]}$

$$\begin{pmatrix}\alpha^{(t)} \\ \beta^{(t)}\end{pmatrix} = \begin{cases}\begin{pmatrix}\alpha^* \\ \beta^*\end{pmatrix} & \text{if} \quad \delta \geq u, \\ \begin{pmatrix}\alpha^{(t-1)} \\ \beta^{(t-1)}\end{pmatrix} & \text{if} \quad \delta < u.\end{cases} \tag{1.17}$$

Summarizing the three basic steps, the application of the SEM algorithm to TK model must consider:

1. Initialization: set arbitrary initial values for the parameters vector $\Theta = \Theta^{(0)}$
2. E-step: Compute $\hat{Z}_{ij}^{(t)}$ from equation (1.5).
3. S-step: Simulate a sample from $Z_i^{(t)} \sim \mathcal{M}(1; \hat{Z}_{i1}^{(t)}, \hat{Z}_{i2}^{(t)}, \hat{Z}_{i3}^{(t)})$ for each $Y_i$.
4. M-step:
   (a) Simulate the weights from $w \sim \mathcal{D}(\eta_1 + n_1^{(t)}, \eta_2 + n_2^{(t)}, \eta_3 + n_3^{(t)})$.
   (b) Generate the candidates $(\alpha^*, \beta^*)$ from equations (1.12) and (1.13).
   (c) Compute $\delta$ from equation (1.16), generate $u$ from $U \sim \mathcal{U}_{[0,1]}$ and accept or reject $(\alpha^*, \beta^*)$ according to equation (1.17).
5. Repeat E, S and M steps until some converge criterion is achieved.

## 1.4 Simulation Study

We perform a simulation study to compare the performance of the TK distribution under different scenarios, conducting a brief discussion on the value of the $\sigma^2$ and the hyperparameters $\eta_i,\ (i = 1, 2, 3)$. Then, we compare the TK distribution performance with the Kumaraswamy distribution for samples generated from each of them.

### 1.4.1 Scenario of the simulations

We can observe that the $\alpha^* = log(\alpha)$ and $\beta^* = log(\beta)$ parameters has been generated from the Log-Normal distribution with variance $\sigma^2$ and mean $\log(\alpha^{(t-1)})$ and $\log(\beta^{(t-1)})$ respectively. Hence, the variance to $\alpha^*$ is equal to $(exp(\sigma^2) - 1)exp(2\alpha^{(t-1)} + \sigma^2)$ and the assigned value to $\sigma^2$ hyperparameter should be done with care. For example, if we consider $log(\alpha^{(t-1)}) = 5$ and standard deviation equal to $\sigma = 0.5,\ 1$, then $V(\alpha^*) = 8032.961,\ 102880.6$ respectively, so a value to hyperparameter $\sigma = 1$ may seem extremely high to explore a new $\alpha^*$ update as we can see in section 1.4.2 (analogous to $\beta^*$). In addition, we have from equation (1.10) that $w^{(t)} \propto \mathcal{D}(\eta_1 + n_1^{(t)}, \eta_2 + n_2^{(t)}, \eta_3 + n_3^{(t)})$, where $n_j$ is the total number of data assignated to component $j$, then we suppose that $\eta_i$ $(i = 1, 2, 3)$ hyperparameter can take, for example, values $\eta_i = 0.1$ or $\eta_i = 1$ indistinctly and in fact if we have lift tails, both values achieve good results, but if the tails are not lifted, then with $\eta_i = 0.1$ the estimations of $a$ and $b$ achieve be more close to zero as we can see in Kumaraswamy simulation, section 1.4.2. All the numerical calculations are obtained considering 100000 Monte Carlo replications and discarding the first 40000 as burn-in.

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 average of the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002), the expected Akaike information criterion (EAIC) introduced by Brooks (2002), and the expected Bayesian information criterion (EBIC) given in Carlin and Louis (2001). First, we simulate from the TK distribution with parameters given by $\Theta = (0.1, 0.3, 5, 10)$, that is, we simulate an asymmetric distribution with independent lifting in both tails to capture the essence of the proposed TK distribution. Second, we take a sample from the Kumaraswamy distribution with parameters given by $\Theta_{\mathrm{K}} = (5, 10)$, that is, an asymmetric distribution but without lifted tails in its density.

### 1.4.2 Results of the simulations

In our first simulation from the TK distribution, we can observe in Table 1.1 that the TK distribution achieves a better fit than the Kumaraswamy distribution and we can appreciate that the value of $\sigma$ hyperparameter must be chosen carefully, delivering better results the choice of $\sigma = 0.5$ as commented in section 1.4.1. In Table 1.2, we consider results from the TK model with $(\sigma = 0.5, \eta_i = 0.1)$ hyperparameters and from the Kumaraswamy

model, and we can appreciate that the Kumaraswamy distribution tries to fit the model by increasing the variance, that is, finding small values for $\alpha$ and $\beta$ to overcome the inability of this distribution to raise the tails.

Table 1.3 present the empirical relative bias (RelBias) and the root-mean-squared error ($\sqrt{\text{MSE}}$) for each parameter estimator over the $100$ simulated samples under the TK distribution. 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 $\theta$ represents any particular parameter, and $\widehat{\theta}^{(i)}$ is the posterior estimate of $\theta$ for the $i$-th sample. Table 1.3 shows that the estimation of each parameter in each data set is good when the TK distribution is adjusted.

Table 1.1: Comparison between the Mean DIC, mean EAIC and Mean EBIC of the TK and Kumaraswamy distributions for 100 samples of size 1000 drawn from a TK distribution with parameters $(0.1, 0.3, 5, 10)$

| | $(\sigma,\eta_i)$ | Mean DIC | Mean EAIC | Mean EBIC |
|---|---|---|---|---|
| Trapezoidal Kumaraswamy | (0.5, 0.1) | **-802.301** | -794.466 | -774.835 |
| | (0.5, 1) | -801.519 | **-794.616** | **-774.985** |
| | (1, 0.1) | -780.712 | -782.756 | -763.125 |
| | (1, 1) | -783.543 | -783.138 | -763.507 |
| Kumaraswamy | $\sigma = 0.5$ | -526.796 | -604.993 | -595.177 |

In Table 1.2 we show the results of the estimation process for each parameter for both models, where $a$ and $b$ come from $w$, with $w_1 = \frac{a}{2}$ and $w_2 = \frac{b}{2}$ according to the definition of the model in section 1.3.1.

Table 1.2: Estimated posterior medians, means and credibility interval (CI) for 100 samples of size 1000 drawn from a TK distribution with parameters $(0.1, 0.3, 5, 10)$

| | Parameter | True | Mean | Standard deviation | Median | 95% CI |
|---|---|---|---|---|---|---|
| Trapezoidal Kumaraswamy | $a$ | 0.1 | 0.098 | 0.018 | 0.095 | (0.077,0.117) |
| | $b$ | 0.3 | 0.3 | 0.027 | 0.298 | (0.263,0.335) |
| | $\alpha$ | 5 | 4.944 | 0.271 | 4.933 | (4.637,5.237) |
| | $\beta$ | 10 | 9.932 | 1.234 | 9.646 | (8.303, 11.233) |
| Kumaraswamy | $\alpha$ | - | 3.006 | 0.154 | 2.996 | (2.828,3.170) |
| | $\beta$ | - | 3.139 | 0.327 | 3.097 | (2.825,3.398) |

Table 1.3: Relbias and root-squared error of each parameter under 100 samples of size 1000 drawn from a TK distribution with parameters $(0.1, 0.3, 5, 10)$

| | $a$ | $b$ | $\alpha$ | $\beta$ |
|---|---|---|---|---|
| RelBias | -0.0041 | -0.0283 | -0.0174 | -0.0231 |
| $\sqrt{\text{MSE}}$ | 0.0139 | 0.0212 | 0.2170 | 0.9884 |

In our second simulation from the Kumaraswamy distribution, we can observe in Table 1.4 that the TK distribution achieve an equally good fit than the Kumaraswamy distribution. In Table 1.5 we can appreciate that the TK distribution give similar estimates for the parameters, compared to Kumaraswamy distribution.

Table 1.4: Comparison between the Mean DIC, mean EAIC and Mean EBIC of the TK and Kumaraswamy distributions for 100 samples of size 1000 drawn from a Kumaraswamy distribution with parameters $(5, 10)$

| | $(\sigma,\eta_i)$ | Mean DIC | Mean EAIC | Mean EBIC |
|---|---|---|---|---|
| Trapezoidal Kumaraswamy | (0.5, 0.1) | -1257.515 | -1321.922 | -1302.291 |
| | (0.5, 1) | -1232.631 | -1311.324 | -1291.693 |
| Kumaraswamy | $\sigma = 0.5$ | -1257.887 | -1326.381 | -1316.566 |

When the true value to $a$ and $b$ parameters is zero, the hyperparameter $\eta_i = 0.1$ may be a better option than taking $\eta_i = 1$. Then in conclusion, the hyperparameter combination for $(\sigma, \eta_i)$ equal to $(0.5, 0.1)$ delivers good results in different scenarios.

Table 1.5: Estimated posterior medians and means for 100 samples of size 1000 drawn from a Kumaraswamy distribution with parameters $(5, 10)$

| | Parameter | True | Mean | Standard deviation | Median | 95% CI |
|---|---|---|---|---|---|---|
| Trapezoidal Kumaraswamy | $a$ | 0 | 5.58e-04 | 4.44e-04 | 2.65e-05 | (3.52e-08, 6.32e-04) |
| | $b$ | 0 | 7.98e-04 | 1.24e-03 | 2.25e-04 | (3.04e-05,1.12e-03) |
| | $\alpha$ | 5 | 4.980 | 0.149 | 4.970 | (4.715,5.249) |
| | $\beta$ | 10 | 10.054 | 0.715 | 9.839 | (8.666,11.272) |
| Kumaraswamy | $\alpha$ | 5 | 4.971 | 0.152 | 4.963 | (4.709,5.239) |
| | $\beta$ | 10 | 10.009 | 0.728 | 9.804 | (8.641,11.214) |

In summary and unsurprisingly, when the sample is generated from the Kumaraswamy distribution, we see similar values on the mean DIC, mean EAIC and mean EBIC achieved by the two adjusted distributions (Kumaraswamy and TK distribution). When the sample is drawn from the TK distribution with a difference between the its two tails, $a = 0.1$ and

$b = 0.3$, the best fit in terms of the mean DIC, mean EAIC and mean EBIC is achieved by the TK 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 Empirical illustrations with real data

To illustrate the TK model in practice, we apply the proposed method to real dataset and we compare the well of the fitting between the TK distribution and the Kumaraswamy distribution. Codes used for fitting TK models in the simulated data example are presented in the appendix. And if you need more detail contact with the author.

### 1.5.1 Covid-19 cases

The Coronavirus disease (COVID-19) has spread worldwide leading to a pandemic, which remains to the last day of this paper's writing. Note in figure 1.2 that there are two peaks: the first and second massive outbreaks during June, 2020 and January, 2021.

The data that we analyze, contains the daily confirmed and probable cases in Chile from the first one to be detected on march 2, 2020 to march 3, 2021 (366 records), where every of them corresponds to the cases of the previous day. The data is available in Ministerio de Ciencia, Tecnología, Conocimiento e Innovación of Chile (www.minciencia.gob.cl/covid19).

We are interested in fitting the TK model to the new daily cases behavior in this country and compare this with the fitting of the Kumaraswamy model. Due to this curve corresponds a daily measures, we first must transform it in a probability model of a random variable that takes values in the (0,1) interval, both 0 and 1 not included, i.e., taking from day 1 to 366 to (0,1). Then, we consider all new cases at the day in which have been observed and transform them in the value of their proper day, this means giving all days a proportional weight according to the number of new cases recorded, achieving for our distribution the shape and growing of the original data, resulting in the histogram shown below in Figure 1.2.
Therefore, we use the equation (1.18) proposed by Smithson and Verkuilen (2006)

$$y = \frac{N-1}{N}\frac{y^* - a_1}{a_2 - a_1} + \frac{1}{2N}, \quad y^* \in [a_1, a_2]. \tag{1.18}$$

Here $N = 800569$ represents the sum of the all new daily cases, $a_1 = 1,\ a_2 = 366$ are the first and last day of the observations and $y \in (0, 1)$. Now, for each $y_i,\ i = 1, ..., 366$ we repeat its value as many times as indicates the total daily cases. For example, if the last day (day 366, $y_{366} = 0.9999$) there were 2747 new cases, in our sample the value $y_{366}$ is repeated 2747 times. Then, because of the required time in the processing this much information and the shape of distribution is not affected nor our purposes, we take for all approaches given below a proportion of the total sample (805 data).

To go back and know what day is represented by the value of any $y$, return from transforming and solve the equation (1.18) for $y^*$.

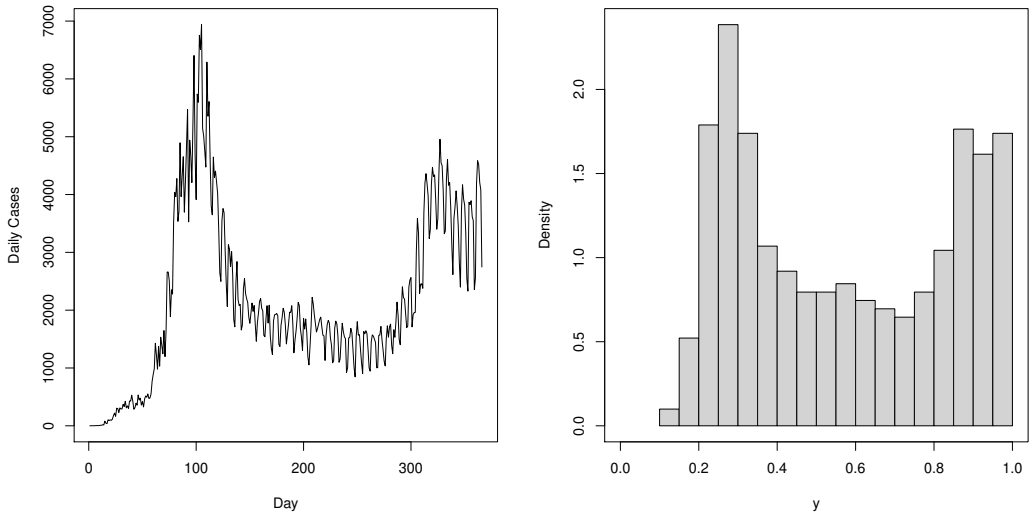


Figure 1.2: The real daily data set in the left side and the histogram with the applied transformation in the right.

The purpose is to show that a fit performed by the TK distribution gives us a better one than the Kumaraswamy distribution, because sometimes we might have enough data in the tails to be in the need of having a model that perform this behavior.
We can see in Figure 1.3 and Table 1.6 that the TK model achieves a better fit compared to the Kumaraswamy distribution. It is clear that the distribution in this data set is lifted in the right tail, captured by $\hat{b}$ (from $w$) which value is $\hat{b} = 1.3264$. Moreover, in accordance with Mean EAIC and Mean EBIC we conclude without a doubt that the TK distribution is a better choice for this data.

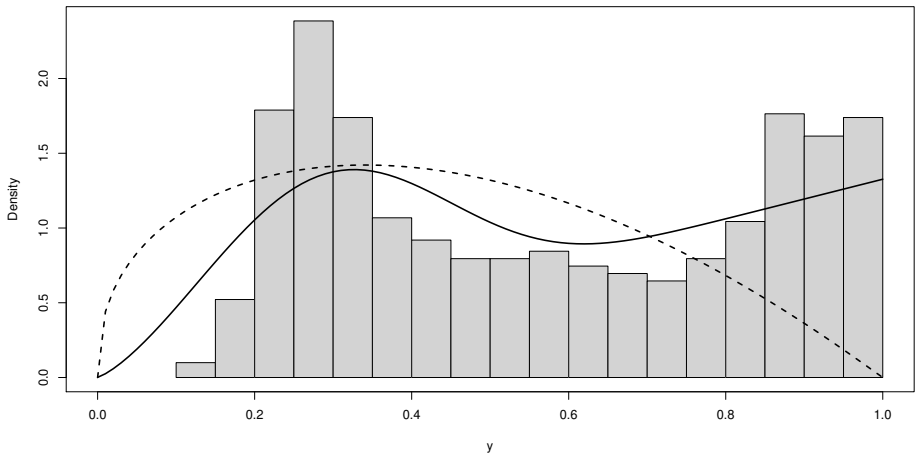


Figure 1.3: Fitting for both models taking the parameter's mean. Solid line: TK distribution. Dashed line: Kumaraswamy distribution.

Table 1.6: Comparison between the Mean DIC, mean EAIC and Mean EBIC of the TK and Kumaraswamy distributions and its mixture models for Daily confirmed Covid19 data set.

| | Mean DIC | Mean EAIC | Mean EBIC |
|---|---|---|---|
| Trapezoidal Kumaraswamy | 876.3951 | -207.6895 | -188.9261 |
| Kumaraswamy | 424.0051 | 428.0051 | 437.3868 |

In Table 1.7 are presented the mean, standard deviation, median and credibility intervals for each model. We can see that $\hat{\beta}$ has high standard deviation compared with the mean in both models (less in TK). This might be because there are too many data in the right tail and causing the estimation have some trouble to assign the $\beta$ value in every iteration, assuming that the peak is in this side. Furthermore, the $\hat{\alpha}$ standard deviation is considered greater compared in the same way that $\hat{\beta}$ in the Kumaraswamy model.

Table 1.7: Estimated posterior means, medians and credibility interval (CI) for Daily confirmed Covid19 data set.

| | Parameter | Mean | Standard deviation | Median | 95% CI |
|---|---|---|---|---|---|
| Trapezoidal Kumaraswamy | $a$ | 0.0009 | 0.0030 | 0 | (0,0.0050) |
| | $b$ | 1.3264 | 0.0487 | 1.3267 | (1.2373, 1.4211) |
| | $\alpha$ | 2.4929 | 0.5118 | 2.4966 | (1.4474, 3.4459) |
| | $\beta$ | 12.9142 | 8.4907 | 10.7872 | (1.5407, 29.9875) |
| Kumaraswamy | $\alpha$ | 1.4027 | 0.5590 | 1.2789 | (0.6033, 2.4522) |
| | $\beta$ | 2.0127 | 7.5916 | 0.8599 | (0.4080, 2.1823) |

In conclusion, the TK model performs a better fitting for this data set.
The described situation above about the standard deviation made us think if we would can get a better fitting than we already had. This would be possible if we consider a model based on a mixture of two TK distributions, which can be justified because this data set is actually bimodal. Therefore, we can have a model of two-mixture TK, which is defined as

$$y_i; a_j, b_j, \alpha_j, \beta_j, p_j \sim \sum_{j=1}^{2} p_j \cdot \mathrm{TK}(a_j, b_j, \alpha_j, \beta_j) \quad ; \quad i = 1, \ldots, 805;\ j = 1, 2.$$

where $0 < p_j \leq 1,\ \sum p_j = 1$ are the weights for each individual TK distribution.

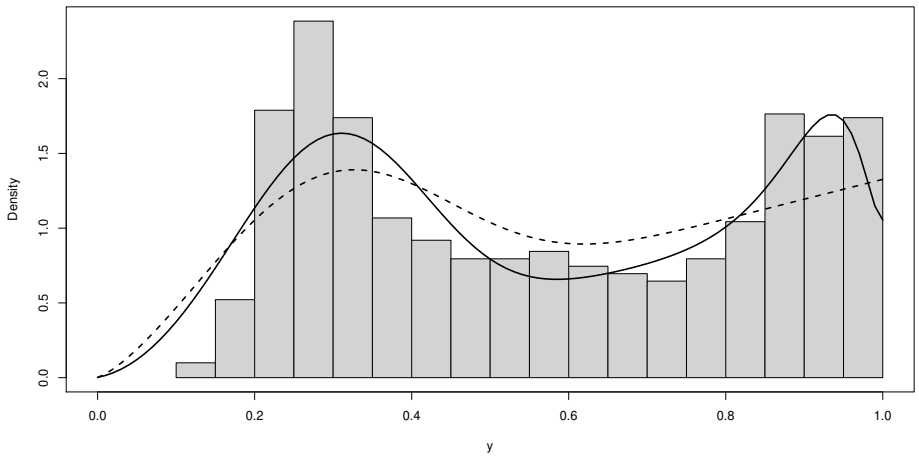


Figure 1.4: Fitting for TK and two-mixture TK models taking the parameter's mean.

Table 1.8: two-mixture TK distribution's Mean DIC, mean EAIC and Mean EBIC for Daily confirmed Covid19 data set.

| | Mean DIC | Mean EAIC | Mean EBIC |
|---|---|---|---|
| Trapezoidal Kumaraswamy Mixture | -603.6924 | -585.6924 | -543.4748 |
| Trapezoidal Kumaraswamy | 876.3951 | -207.6895 | -188.9261 |

Table 1.9: Estimated posterior means, medians and credibility interval (CI) in two-mixture TK distribution for Daily confirmed Covid19 data set.

| Parameter | | Mean | Standard deviation | Median | 95% CI |
|---|---|---|---|---|---|
| TK Mix | $a_1$ | 0.0013 | 0.0043 | 0 | (0, 0.0081) |
| | $b_1$ | 1.6641 | 0.1651 | 1.6680 | (1.3678, 2) |
| | $\alpha_1$ | 15.8136 | 6.8317 | 15.4616 | (2.2682, 27.8903) |
| | $\beta_1$ | 2.9145 | 2.5286 | 2.2340 | (0.0589, 8.1871) |
| | $a_2$ | 0.0026 | 0.0097 | 0 | (0, 0.0162) |
| | $b_2$ | 0.0938 | 0.01952 | 0 | (0, 0.6082) |
| | $\alpha_2$ | 3.0370 | 0.4966 | 3.065 | (2.0029, 3.9354) |
| | $\beta_2$ | 26.4600 | 15.5967 | 22.7700 | (4.3897, 59.6894) |
| | $p$ | 0.6120 | 0.0601 | 0.6248 | (0.4671, 0.6994) |

## 1.6 Concluding remarks

TK distribution is a new proposal derived from Kumaraswamy distribution with the particularity of raising its tails, this means a generalization that allows to fit the accumulated data in the extremes of the distribution by adding two very intuitive new parameters.
In particular, this paper's effort was about a fitting comparison between these two models under an alternative estimation procedure than proposed before: a bayesian approach based on simulation from the posterior distrubutions, justified in the intuitive nature of the parameters, that achieved very good results in both simulation and real data application as we saw above. Also, of equal importance, avoiding the dependence on initial values of the classic EM algorithm, accomplishing more reliable results.
We can conclude the TK distribution is the model that performs the best adjustment for data with some accumulation at the ends by far. Even more, by observing the real data set a new propose came up: to consider two TK distributions, that is, taking a 9 parameter mixture model, which could give the potential benefit of performing a bimodal distribution, and it did. The two TK mixture model achieved a better fit than the TK model. The importance of this is the possibility of extension of the model for finite mixture of TK distributions in future work.

# Bibliography


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

Bouguila, Z., Monga, E. and Ziou, D. (2006). Practical Bayesian estimation of a finite beta mixture through gibbs sampling and its applications. Stat Comput, 16:215–225.

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.

Brooks, S. P. (2002). Discussion on the paper by Spiegelhalter, Best, Carlin, and van der Linde. Journal of the Royal Statistical Society. Series B, 64(3):616–618.

Carlin, B., Louis, T., (2001). Bayes and Empirical Bayes Methods for Data Analysis. Second Edition. Chapman & Hall/CRC, Boca Raton.

Casella, G. and Robert, C. (2010). Introducing Monte Carlo Methods with R. Springer, New York.

Celeux, G. and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Computational Statistics Quarterly, 2:73–82.

Celeux, G. and Govaert, G. (1992). A Classification EM algorithm for clustering and two stochastic versions. Computational Statistics & Data Analysis, 14:315–332.

Cordeiro, G. and de Castro, M. (2011). A new family of generalized distributions. Journal of statistical computation and simulation, 81(7):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(2):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(8):1399–1429.

de Pascoa, M., Ortega, E. and Cordeiro, G. (2011). The kumaraswamy generalized gamma distribution with application in survival analysis. Statistical methodology, 8(5):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(3):265–291.

Dempster, A., Laird, N. and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, 39(1):1–38.

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

Diebolt, J. and Robert, C. (1994). Estimation of Finite Mixture Distributions Through Bayesian Sampling. Journal of the Royal Statistical Society, 56(2):363–375.

Figueroa, J., Sanhueza, R., Lagos, B. and Ibacache, G. (2020). Modeling bounded data with the trapezoidal Kumaraswamy distribution and applications to education and engineering. Chilean Journal of Statistics, 11(2):163–176.

Jones, M. C. (2009). Kumaraswamy distribution: A beta-type distribution with some tractability advantages. Statistical Methodology, 6(1):70–81.

Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes. Journal of Hydrology, 46(1-2):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(2):3–14.

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

Mead, M. and Abd-Eltawab, A. (2014). A note on kumaraswamy fréchet distribution. Australian Journal of Basic and Applied Sciences, 8(15):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. New York: Springer-Verlag.

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.

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

Spiegelhalter, D., Best, N., Carlin, B. and Van Der Linde, A. (2002), Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B (Statistical Methodology) 64:583–639.