ISSN 0104-0499

PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA UFSCar-USP

DEs-UFSCar e SME-ICMC-USP

BAYESIAN ESTIMATION FOR MIXTURE OF SIMPLEX DISTRIBUTION WITH UNKNOWN NUMBER OF COMPONENTS: HDI ANALYSIS IN BRAZIL

Rosineide F. da Paz
Jorge Luis Bazán
Luis Aparecido Milan

RELATÓRIO TÉCNICO

TEORIA E MÉTODO – SÉRIE A

Junho/2015
nº 261

Bayesian Estimation for Mixture of Simplex Distribution with Unknown Number of Components: HDI Analysis in Brazil

Rosineide F. da Paz[1,2,3], Jorge Luis Bazán [1], Luis Aparecido Milan[2]

Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, São Carlos-SP, Brazil.[1]

Universidade Federal de São Carlos, Departamento de Estatística, São Carlos-SP, Brazil[2]

E-mail: rfpaz@icmc.usp.br[3]

Abstract

Variable taking value on (0, 1), such as rates or proportions, are frequently analysed by researchers, for instance, political and social data as well as Human Development Index. However, sometime this type of data cannot be modelled adequately using a unique distribution. In this case, we can use a mixture of distribution that is a powerful and flexible probabilistic tool. This manuscript deals with a mixture of simplex distribution for model proportional data. A Full Bayesian approach is considered in the inference process with Reversible-jump Markov Chain Monte Carlo method. The usefulness of the proposed approach is confirmed by use of the simulated mixture data from several different scenarios and through an application of the methodology to analyses municipal Human Development Index data of the cities (or towns) of the Northeast region and São Paulo state in Brazil.

Key words: Bayesian Analysis, Markov chain Monte Carlo, Mixture model, Simplex distribution, Human development index.

1 Introduction

Variable taking values on (0, 1), such as index and proportions, are frequently analysed by researchers, for instance, Impartial Anonymous Culture (Stensholt, 1999) and the Human Development

Index (HDI) (McDonald & Ransom, 2008; Cifuentes *et al.*, 2008). Sometimes, the data cannot be modelled adequately using a unique distribution as is the case of the proportion of votes obtained by a political party in Presidential Elections in each cities of an country analysed in (Paz *et al.*, 2015). In addition, in the data of HDI of several regions in Brazil different components can be identified, see index in Fundação Instituto Brasileiro de Geografia e Estatística (2014).

The models with mixture of distributions can be a powerful and flexible probabilistic tool for modelling many kind of the data, see for example McLachlan & Peel (2004). In financial data we can cite Faria & Gonalves (2013). In addition, mixture of distributions have been widely analysed for normal data, see for example Tanner & Wong (1987); Gelfand & Smith (1990); Diebolt & Robert (1994); Richardson & Green (1997). For data in (0,1) there are some works which consider a finite mixture of Beta distributions (Bouguila *et al.*, 2006; Bouguila & Elguebaly, 2012). However there are in the literature other distributions which take values on (0,1), such as simplex distribution, for instance. The simplex distribution was proposed by Barndorff-Nielsen & Jorgensen (1991) and recently has been considered as a complementary and alternative regression model to the beta regression model (López, 2013; Song & Tan, 2000).

This manuscript deals with a new framework for modelling the bounded variables with multimodality as a complementary model to the correspondent beta model. The model proposed considerer a mixture of simplex distribution with the number of components unknown (simplex mixture model). This work is motivated by the municipal HDI data in Brazil. Thus, at present we focus on the identification of the numbers of components and the characteristics of each population identified by the model considering the HDI of the cities of São Paulo state and Northeast region of Brazil. For the inference process a fully Bayesian analysis is assumed where the unknown number of components and the parameters should be regarded as drawn from appropriate prior distribution. For dealing with the problem of estimating the number of components of the mixture model we adopt a reversible-jump Markov chain Monte Carlo (RJMCMC) approach proposed, in the case of mixture of normal distributions, by Green (1995) and Richardson & Green (1997). The results obtained are promising since

that the performance of the method is tested by applying it to simulated data sets from mixture of simplex distributions, considering several different scenarios.

In future developments we can consider that the phenomenon can be explained by sociological and economical factors which should be included. In addition the response variable might be associated considering geospatial information as potential covariates.

The remainder of the manuscript is organized as follows: The Section 2 is dedicated to give a description of general mixture model. In Section 3 we present the mixture of simplex distributions. Section 4 addresses the Bayesian Inference approach considering estimation RJMCMC. The Section 5 is dedicate to investigate if our algorithm is able to estimate the mixture parameters and select the number of components considering several scenarios of generated data. In the Section 6 we present an analysis of the municipal HDI data set considered show that are strong evidence for two component of cities we found that some cities in the Northeast region of Brazil show HDI similar to cities of the Paulo state. Finally, some conclusions are drawn in Section 7 .

2 The general mixture model

Finite mixture of distributions is a flexible method of modelling. Its more direct role in data analysis and inference is to provide a convenient and flexible family of distributions to estimate or approximate distributions which are not well modelled by any standard parametric family. This type of model is useful in the modelling of data from a heterogeneous population, that is, a population which can be divided in clusters or components. In this sense, the components in the data can be modelled for uni-modal distributions. For more details about modelling and applications of finite mixture models, see for example McLachlan & Peel (2000).

Consider initially a sequence of n continuous random variables $\boldsymbol{Y} = (Y_1, ..., Y_n)$ each following a distribution with probability density function (pdf) $f_i(.|\theta_i)$ for $i = 1, \ldots, n$. Now suppose that $f_i(.|\theta_i) = f_j(.|\theta_j)$ for some i's where $j = 1, .., k$ with $1 \leqslant k \leqslant n$. Thus, a random variable $Y \in \boldsymbol{Y}$ is

said to follow a mixture of distributions with k components and its probability density function (pdf) can be write as

$$f(y|\boldsymbol{\theta}, \boldsymbol{\omega}, k) = \sum_{j=1}^{k} \omega_j f_j(y|\boldsymbol{\theta}_j) \tag{1}$$

where each $f_j(y|\boldsymbol{\theta}_j)$ is a pdf called component density of the mixture, indexed by a parameter vector $\boldsymbol{\theta}_j$ (here we write $f(y|\boldsymbol{\theta}_j)$ without the index j because the component density belong to the same parametric family), $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$ is a vector containing all the parameters of the components in the mixture and the components of the vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_k)$ are called weights of the mixture where $0 < \omega_j < 1$ with $\sum_{j=1}^{k} \omega_j = 1$. In the equation (1) k is the number of components in the mixture. We call the model defined by the pdf in (1) mixture model, whose distribution is called mixtures of distributions. For a review on exwhoseisting techniques for Bayesian modelling and inference on mixtures of distributions, see for example Marin *et al.* (2005).

In order to make inference about the parameters of the mixture model, suppose $\boldsymbol{Y} = (Y_1, ..., Y_n)$ a random sample from the distribution defined by equation (1). The likelihood function related to a sample $\boldsymbol{y} = (y_1, ..., y_n)$, where each y_i is a observation of Y_i for $i = 1, ..., n$, is given by

$$L(\boldsymbol{\theta}, \boldsymbol{\omega}, k|\boldsymbol{y}) = \prod_{i=1}^{n} \sum_{j=1}^{k} \omega_j f(y_i|\theta_j).$$

A way to simplify the inference process of mixture model is to consider a unobserved random vector $Z_i = (Z_{i1}, ..., Z_{ik})$ such that $Z_{ij} = 1$ if the ith observation is from the jth mixture component and $Z_{ij} = 0$ otherwise, $i = 1, \ldots, n$. Note that $\sum_{j=1}^{k} Z_{ij} = 1$ then we suppose each random vector $Z_1, .., Z_n$ is distributed according to the multinomial distribution with parameters 1 and $\boldsymbol{\omega} = (\omega_1, ..., \omega_k) = (P(Z_{i1} = 1|\omega, k), ..., P(Z_{ik} = 1|\omega, k))$, for $i = 1, ..., n$. Then

$$P(Z_{ij} = 1|y_i, \theta_j, \omega, k) \propto P(Z_{ij} = 1|\boldsymbol{\omega}, k) f(y_i|Z_{ij} = 1, \theta_j, \boldsymbol{\omega}, k),$$

$j = 1, ..., k,\ i = 1, \ldots, n$. To simplify the notation we consider $\mathbf{Z} = (Z_1, ..., Z_n)$ an vector nk containing all the unobserved indicator vectors Z_i. Note that the distribution of each Y_i given Z_i has pdf given

by

$$f(y_i|Z_i,\boldsymbol{\theta},k)=\prod_{j=1}^{k}\big[f(y_i|\theta_j)\big]^{Z_{ij}} \tag{2}$$

then the joint distribution of (Y_i,Z_i) can be written as

$$f(y_i,Z_i|\boldsymbol{\theta},\boldsymbol{\omega},k)=P(Z_i|\boldsymbol{\omega},k)f(y_i|Z_i,\boldsymbol{\theta},k)=\prod_{j=1}^{k}\big[\omega_j f(y_i|\theta_j)\big]^{Z_{ij}}. \tag{3}$$

Note that, the vector Z_i have just one component equal to 1 and the others equal to zero then

$$\prod_{j=1}^{k}\big[\omega_j f(y_i|\theta_j)\big]^{Z_{ij}}=\begin{cases}\omega_1 f(y_i|\theta_1) & \text{if } Z_i=(1,0,...,0)\\ \omega_2 f(y_i|\theta_2) & \text{if } Z_i=(0,1,...,0)\\ \quad\vdots & \quad\vdots\\ \omega_k f(y_i|\theta_k) & \text{if } Z_i=(0,0,...,1)\end{cases}$$

thus,

$$f(y_i|\boldsymbol{\theta},\boldsymbol{\omega},k)=\sum_{Z_i}f(y_i,Z_i|\boldsymbol{\theta},\boldsymbol{\omega},k)=\sum_{j=1}^{k}\omega_j f(y_i|\theta_j). \tag{4}$$

After the inclusion of the indicator vectors in the model, the augmented data likelihood to $(\boldsymbol{y},\mathbf{Z})$ can be written as

$$L(\boldsymbol{\theta},\boldsymbol{\omega},k|\boldsymbol{y},\boldsymbol{Z})=\prod_{i=1}^{n}\prod_{j=1}^{k}\big[\omega_j f(y_i|\theta_j)\big]^{Z_{ij}}. \tag{5}$$

Finally, the joint distribution of all variables of the model including the augmented version and the prior specifications is

$$P(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{Z},\boldsymbol{\omega},k)=f(\boldsymbol{y}|\boldsymbol{\theta},\mathbf{Z},\boldsymbol{\omega},k)P(\boldsymbol{\theta}|\mathbf{Z},\boldsymbol{\omega},k)P(\mathbf{Z}|\boldsymbol{\omega},k)P(\boldsymbol{\omega}|k)P(k).$$

Here, we assume conditional independence such that

$$P(\boldsymbol{\theta}|\mathbf{Z},\boldsymbol{\omega},k)=P(\boldsymbol{\theta}|\mathbf{Z},k)\text{ and }P(\mathbf{y}|\boldsymbol{\theta},\mathbf{Z},\boldsymbol{\omega},k)=P(\mathbf{y}|\boldsymbol{\theta},\mathbf{Z})$$

to obtain

$$P(\mathbf{y},\boldsymbol{\theta},\mathbf{Z},\boldsymbol{\omega},k)=P(\mathbf{y}|\boldsymbol{\theta},\mathbf{Z})P(\boldsymbol{\theta}|\mathbf{Z},k)P(\mathbf{Z}|\boldsymbol{\omega},k)P(\boldsymbol{\omega}|k)P(k), \tag{6}$$

where $P(\mathbf{Z}|\boldsymbol{\omega},k)=\prod_{i=1}^{n}P(Z_i|\boldsymbol{\omega},k)=\prod_{i=1}^{n}\left(\prod_{j=1}^{k}\omega_j^{Z_{ij}}\right)$.

3 Simplex Mixture Distribution

Now consider a sequence of n continuous random variables $\boldsymbol{Y}$ where each $Y \in \boldsymbol{Y}$ assume values in (0,1) and follow the distribution whose pdf is given by (1). Let consider that the component densities, $f_j(.|\theta_j)$ for $j = 1, ..., k$, are taken to belong the simplex distribution (Jørgensen, 1997), whose pdf is given by

$$S(y|\mu, \sigma^2) = \left(2\pi\sigma^2 \left(y(1-y)\right)^3\right)^{-1/2} \exp\left\{-\left(\frac{1}{2\sigma^2}\right)\left(\frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}\right)\right\} I_{(0,1)}(y), \tag{7}$$

where $0 < \mu < 1$ is the location parameter and $\sigma^2 > 0$ is the dispersion parameter. The mean of simplex distribution is given by $E(Y) = \mu$. Since the components density $f_j(.|\theta_j)$ are taken to belong to the simplex family, we shall refer the component density in the mixture as simplex component, the model given by (1) as Simplex Mixture (SM) and to rewrite its pdf as

$$P(y|\boldsymbol{\theta}, \boldsymbol{\omega}, k) = \sum_{j=1}^{k} \omega_j S(y|\mu_j, \sigma_j^2) \tag{8}$$

where $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$ with each $\theta_j = (\mu_j, \sigma_j^2)$

4 Inference

Consider $\mathbf{y} = (y_1, .., y_n)$ a realization of $\mathbf{Y}$ where y_i is the observed value of the Y_i, for $i = 1, ..., n$, then the likelihood corresponding to a SM model with k-component is given by

$$L(\boldsymbol{\theta}, \boldsymbol{\omega}, k|\mathbf{y}) = \prod_{i=1}^{n} \sum_{j=1}^{k} w_j S(y_i|\mu_j, \sigma_j^2). \tag{9}$$

Thus, the augmented data likelihood to $(\mathbf{y}, \mathbf{Z})$ can be written as

$$L(\boldsymbol{\theta}, \boldsymbol{\omega}, k|\mathbf{y}, \mathbf{Z}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \left[\omega_j S(y_i|\mu_j, \sigma_j^2)\right]^{Z_{ij}}. \tag{10}$$

The representation of a mixture model, presented in this thesis, precludes the use of improper prior. This is because improper prior lead to improper posterior when some the component became empty. We define the prior, which we suppose that are all drawn independently since that is a common assumption taken generally when defining Bayesian models. For the distribution of $P(\boldsymbol{\theta}|k)$ consider

$\phi_j = \sigma_j^{-2}$, then for the parameters $\theta_j = (\mu_j, \phi_j)$ given k we assume the following independent prior

$$\mu_j|k \sim Uniform(0,1) \text{ and } \phi_j|k \sim Gamma(a,b), \;\; j = 1, \ldots, k, \tag{11}$$

where the hyperparameters a and b are fixed.

Also, for $P(\boldsymbol{\omega}|k)$ since the vector of weights $\boldsymbol{\omega}$ is defined on the simplex $\{\boldsymbol{\omega} \in \mathbb{R}^k : 0 < \omega_j < 1, j = 1, ..., k, \sum_{j=1}^k \omega_j = 1\}$ we consider a Dirichlet prior distribution for $\boldsymbol{\omega}$ given k, that is $\boldsymbol{\omega}|k \sim Dirichlet(\nu_1, ..., \nu_k)$. Finally, for $P(k)$, that is to the parameter k we adopt a uniform distribution between 1 and k_{max}.

Therefore, the full conditional posterior distributions can be obtained and consequently a Markov chain Monte Carlo method (MCMC) (Ross, 2006, pages, 245 - 271) can be used to sample from the joint probability distribution of the parameters $(\boldsymbol{\theta}, \boldsymbol{\omega}, k)$, given the observed data **y, Z**. Then the sample of the joint posterior distribution produced by MCMC is used for Bayesian inference.

The full conditional distributions of the parameters for jth components as given by

$$P(\phi_j|\mathbf{y}, \mathbf{Z}, \mu_j) \propto \phi_j^{n_j/2+a-1} \exp\left\{-\phi_j \left(\sum_{i \in \{i:Z_{ij}=1\}} \frac{(y_i - \mu_j)^2}{2y_i(1-y_i)\mu_j^2(1-\mu_j)^2} + b \right)\right\} \tag{12}$$

$$P(\mu_j|\mathbf{y}, \mathbf{Z}, \phi_j) \propto \exp\left\{-\frac{\phi_j}{2\mu_j^2(1-\mu_j)^2} \sum_{i \in \{i:Z_{ij}=1\}} \left(\frac{(y_i - \mu_j)^2}{y_i(1-y_i)} \right)\right\}, \tag{13}$$

where $n_j = \sum_{i=1}^n Z_{ij}$ denotes the number of observations drawn from a jth component of the mixture. Note that $(\phi_j|\mathbf{y}, Z, \mu_j) \sim Gamma(n_j/2 + a, \sum_{i \in \{i:Z_{ij}=1\}} \frac{(y_i - \mu_j)^2}{2y_i(1-y_i)\mu_j^2(1-\mu_j)^2} + b)$. In addition, the full conditional density of $\boldsymbol{\omega}$ is

$$P(\boldsymbol{\omega}|\mathbf{y}, \mathbf{Z}) \propto \prod_{j=1}^k \omega_j^{\nu_j + n_j - 1}, \tag{14}$$

that is the pdf of a Dirichlet distribution, that is, $(\boldsymbol{\omega}|\mathbf{y}, Z) \sim Dirichlet\,(\nu_1 + n_1, ..., \nu_k + n_k)$.

The parameter k is estimated by use the *reversible-jump* step, which is described in details in the Subsection 4.1. A step by step description of the whole algorithm specific to simulate this distributions is given in Appendix A.1.

4.1 Reversible-jump for the number of component in the mixture

Reversible-jump MCMC was introduced by Green (1995) as an extension to MCMC in which the dimension or number of components of the model is uncertain and need to be estimated. The move in the RJ step, called split-combine moves, allow to increases or reduces the number of components by one in each step. In each move, the reversible-jump compare two models with different number of simplex components.

The split-combine move form a reversible pair. For these pair, we choosing the proposal distribution $T_{k\to k^*}$ according to informal considerations in order to obtain a reasonable probability of acceptance. The notation $T_{k\to k^*}$ means proposal transition function for the move of model with k simplex component to model with k^* simplex component. This move is chosen with probability $p_{k^*|k}$. Since the parametric space of parameters $(\boldsymbol{\theta}, \boldsymbol{\omega}, k)$ is different from $(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*, k^*)$, the smaller parameter space should be increased. We generate a three-dimensional random vector $\mathbf{u}$ from a $g(\mathbf{u})$ to complete the parameters space. Green (1995) show that the balance condition is determined by the acceptance probability to this move given by $\alpha((\boldsymbol{\theta}^*, \boldsymbol{\omega}^*, k^*)|(\boldsymbol{\theta}, \boldsymbol{\omega}, k)) = \min\{1, A\}$ where

$$A = \frac{L((\boldsymbol{\theta}^*, \boldsymbol{\omega}^*, k^*)|\mathbf{y}, Z)P((\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)|k^*)P(k^*)p_{k|k^*}}{L((\boldsymbol{\theta}, \boldsymbol{\omega}, k)|\mathbf{y}, Z)P((\boldsymbol{\theta}, \boldsymbol{\omega})|k)P(k)p_{k^*|k}g(\mathbf{u})}\,|J|\,, \tag{15}$$

where J is the Jacobian of transformation. The probability of the inverse of move is given by $\alpha((\boldsymbol{\theta}, \boldsymbol{\omega}, k)|(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*, k^*)) = \min\{1, A^{-1}\}$.

The choice between split or combine move is made randomly with probability b_k and $d_k = 1 - b_k$ respectively, depending on k. Note that $d_1 = 0$ and $b_{k_{max}} = 0$ with k_{max} being a constant representing the maximum value allowed for k, as seen in the previous subsection. If $2 < k < k_{max}$ we adopt $b_k = d_k = 0.5$.

If the split move is chosen, we select randomly one component j_* to break into two new components (j_1, j_2) and create a new state with $k^* = k + 1$ component. In order to specify the new values of parameters for the two components, generate the vector $\mathbf{u} = (u_1, u_2, u_3)$ from beta distributions, i.e., $u_1 \sim Beta(2, 2)$, $u_2 \sim Beta(1, 1)$ and $u_3 \sim Beta(2, 2)$. Than the new parameters

are set as

$$\begin{aligned}
\omega_{j_1} &= \omega_{j_*} u_1, & \omega_{j_2} &= \omega_{j_*}(1-u_1), \\
\mu_{j_1} &= \mu_{j_*} - u_2 u_1(\mu_{j_*} - \mu_{j_*}^2), & \mu_{j_2} &= \mu_{j_*} + u_2(1-u_1)(\mu_{j_*} - \mu_{j_*}^2), \\
\sigma_{j_1}^2 &= \sigma_{j_*}^2 u_3(1-u_2^2)/u_1, & \sigma_{j_2}^2 &= \sigma_{j_*}^2(1-u_3)(1-u_2^2)/(1-u_1).
\end{aligned} \tag{16}$$

All the observation previously allocated to j_* is reallocated doing $z_i = j_1$ or $z_i = j_2$ follow the same criteria used on the Gibbs sampling algorithm (2a).

The combine proposal begins by choosing a pair of component (j_1, j_2), which the first is chosen randomly, in a uniform way, and the second is chosen by making $j_2 = j_1 + 1$, the kth component can not chosen in the first place. This two components are merged, reducing k by 1. The new component is labelled j_* and contain all the observation previously allocated to j_1 and j_2 doing $z_i = j_*$. The parameters for the component j_* are set as $\omega_{j_*} = \omega_{j_1} + \omega_{j_2}$, $\mu_{j_*} = \frac{\mu_{j_1}\omega_{j_2} + \mu_{j_2}\omega_{j_1}}{\omega_{j_*}}$ and $\sigma_{j_*}^2 = \frac{\sigma_{j_2}^2\left(\frac{\omega_{j_2}}{\omega_{j_*}}\right)}{\left(1-\left(\frac{\mu_{j_2}-\mu_{j_1}}{\mu_{j_*}-\mu_{j_*}^2}\right)^2\right)\left(\frac{\sigma_{j_2}^2\omega_{j_2}}{\sigma_{j_1}^2\omega_{j_1}+\sigma_{j_2}^2\omega_{j_2}}\right)}$. This process is reversible, i.e., if we first split one component in two and then combine the components j_1 and j_2 we can recover the previous state. Also we can compute corresponding value of u_i's in the merge move as $u_1 = \frac{\omega_{j_1}}{\omega_{j_*}}$, $u_2 = \frac{\mu_{j_2}-\mu_{j_1}}{\mu_{j_*}-\mu_{j_*}^2}$ and $u_3 = \frac{\omega_{j_1}\sigma_{j_1}^2}{\omega_{j_1}\sigma_{j_1}^2+\omega_{j_2}\sigma_{j_2}^2}$. The acceptance probability for split and combine are $\min\{1, A\}$ and $\min\{1, A^{-1}\}$ respectively, according to (15), with

$$\begin{aligned}
A &= \frac{(k+1)\left(\prod_{i\in\{i:Z_{ij_1}=1\}} S(y_i|\mu_{j_1}\sigma_{j_1}^2)\right)\left(\prod_{i\in\{i:Z_{ij_2}=1\}} S(y_i|\mu_{j_2}\sigma_{j_2}^2)\right)}{\left(\prod_{i\in\{i:Z_{ij_*}=1\}} S(y_i|\mu_{j_*}\sigma_{j_*}^2)\right)} \\
&\times \frac{P(k+1)}{P(k)} \frac{\omega_{j_1}^{\nu-1+n_1}\omega_{j_2}^{\nu-1+n_2}}{\omega_{j_*}^{\nu-1+n_1+n_2}} \frac{P(\sigma_{j_2}^2)P(\sigma_{j_2}^2)}{P(\sigma_{j_*}^2)} \frac{P(\mu_{j_2})P(\mu_{j_2})}{P(\mu_{j_*})} \\
&\times \frac{d_{k+1}}{b_k P_{alloc} g(\mathbf{u})} \frac{1}{2}\left(\sigma_{j_1}^2 + \sigma_{j_2}^2\right)(\omega_{j_1} + \omega_{j_2})\left[2\left(\mu_{j_1} + \mu_{j_2}\right) - (\mu_{j_1} + \mu_{j_2})^2\right],
\end{aligned}$$

where d_{k+1} is the probability of choosing the merge movement between the components j_1 and j_2, b_k is the probability of choosing the split movement of the component j_*, P_{alloc} is the probability of a specific allocation defined as the product of conditional posterior probabilities used to allocate the observations, $g(\mathbf{u})$ is the joint distribution of $\mathbf{u}$ given by product of density of beta distributions, (k+1) is the ratio $\frac{(k+1)!}{k!}$ from the order statistics densities for the parameters (μ, σ^2) and the last term of equation is the Jacobian of the transformations used to complete the dimension. The second term

in (17) is the rate of the density of the prior distribution.

5 Analysis of simulated data sets in several scenarios

This subsection is dedicated to investigate if our algorithm is able to estimate the mixture parameters and select the number of clusters effectively considering several scenarios of generated data. For this purpose, we implemented the algorithm described in Appendix A.1 by using the R program (R Development Core Team, 2015). The analysis was conduced to simulated data set considering six scenarios to simplex mixture models We simulated independent values $Y \sim SM(\mu, \sigma^{\mathbf{2}}, \boldsymbol{\omega}, k)$ with $k \in \{2, 3\}$. The parameters of the six models are shown in the first column of Table 1 and are noted as $\mathcal{M}_1, ..., \mathcal{M}_6$. For each model we simulated tree data sets, being the first with size $n = 1000$ and the others two with size $n < 1000$, as seen in the second column of Table 1. The value of k_{max} was fixed in 5 and the hyper-parameters for gamma prior were fixed in $a = 2$ and $b = 1/2$. After discarding the first 100000 iterations, we used 100000 iterations with thinning equal to 10 to the inference process.

The posterior relative frequency of k, shown in Table 1, gives evidence that the reversible-jump estimated correctly the number of components to these simulated data sets. In addition, Tables 2 and 3 shows the estimated values of parameters to the six models. Posterior mean and empirical standard deviation (SD) are shown in this table. We can observe that the SD decrease as n increase and the estimated values of the parameters are always close the true values. Finally, we show the real histogram and estimated density in Figure 1 which confirm the adequate performance of the estimation method to the simulated data sets.

Table 1: Parameters used to simulate the data sets and the posterior relative frequency for the number of components obtained from the each simulated data set of the size n.

Model		n	Posterior relative frequency				
			k=1	k=2	k=3	k=4	k=5
$\mathcal{M}_1$	$\mu = (0.34, 0.72)$	1000	0	**0.9874**	0.0125	0.0001	0
	$\sigma^2 = (0.8, 1.5)$	500	0	**0.9804**	0.0194	0.0002	0
	$w = (0.5, 0.5)$	100	0.2790	**0.6866**	0.0333	0.0010	0.
$\mathcal{M}_2$	$\mu = (0.08, 0.40)$	1000	0	**0.9941**	0.0059	0	0
	$\sigma^2 = (2, 1)$	500	0	**0.9852**	0.0145	0.0003	0
	$w = (0.65, 0.35)$	100	0.1060	**0.8532**	0.0389	0.002	0.0001
$\mathcal{M}_3$	$\mu = (0.23, 0.58)$	1000	0	**0.9876**	0.0123	0.0001	0
	$\sigma^2 = (1.8, 0.8)$	500	0	**0.9843**	0.0154	0.0002	0.0001
	$w = (0.30, 0.70)$	100	0.1648	**0.7711**	0.0591	0.005	0
$\mathcal{M}_4$	$\mu = (0.30, 0.55, 0.80)$	1000	0.01	0.051	**0.9370**	0.0024	0
	$\sigma^2 = (0.20, 0.10, 0.20)$	500	0.0062	0.0523	**0.9387**	0.0028	0
	$w = (0.20, 0.30, 0.50)$	300	0.0199	0.3393	**0.6316**	0.009	0.0003
$\mathcal{M}_5$	$\mu = (0.15, 0.47, 0.75)$	1000	0.0001	0.1140	**0.8667**	0.0189	0.0003
	$\sigma^2 = (5.0, 0.2, 0.8)$	500	0.0005	0.1660	**0.8157**	0.0177	0.0001
	$w = (0.25, 0.45, 0.30)$	400	0.0023	0.3831	**0.5906**	0.0240	0
$\mathcal{M}_6$	$\mu = (0.10, 0.50, 0.90)$	1000	0.0047	0.0208	**0.9518**	0.0224	0.0003
	$\sigma^2 = (6.0, 0.5, 8.0)$	500	0.0126	0.0583	**0.8984**	0.0388	0.001
	$w = (0.3, 0.50, 0.20)$	400	0.0106	0.1409	**0.7929**	0.0536	0.0020

Table 2: Posterior mean of the parameters and empirical standard deviation (SD) for each simulated data set considering six models with $k = 2$ described in Table 1.

Model	n		Estimative to μ	Estimative to σ^2	Estimative to ω
$\mathcal{M}_1$	1000	Mean	(0.33, 0.71)	(0.69, 1.57)	(0.46, 0.54)
		SD's	(0.005, 0.006)	(0.061, 0.134)	(0.019, 0.016)
	500	Mean	(0.33, 0.72)	(0.77, 1.43)	(0.47, 0.53)
		SD's	(0.010, 0.011)	(0.1225, 0.1990)	(0.031, 0.031)
	100	Mean	(0.35, 0.69)	(0.61, 1.83)	(0.41, 0.59)
		SD's	(0.041,0.047)	(0.556, 0.741)	(0.123,0.123)
$\mathcal{M}_2$	1000	Mean	(0.082,0.40)	(1.99, 0.98)	(0.70, 0.3)
		SD's	(0.001, 0.006)	(0.1178, 0.0915)	(0.015,0.015)
	500	Mean	(0.079, 0.40)	(2.00, 0.89)	(0.68, 0.32)
		SD's	(0.002, 0.008)	(0.3417, 0.1115)	(0.021, 0.021)
	100	Mean	(0.091,0.40)	(2.96, 0.77)	(0.72,0.28)
		SD's	(0.008, 0.026)	(0.811, 0.661)	(0.058,0.058)
$\mathcal{M}_3$	1000	Mean	(0.21,0.61)	(1.86,0.91)	(0.29, 0.71)
		SD's	(0.009, 0.005)	(0.2386, 0.0659)	(0.018, 0.018)
	500	Mean	(0.19,0.60)	(2.03, 1.08)	(0.27, 0.73)
		SD's	(0.012, 0.003)	(0.970, 0.1327)	(0.028, 0.028)
	100	Mean	(0.24, 0.58)	(2.00, 1.11)	(0.23, 0.77)
		SD's	(0.078, 0.028)	(2.085, 0.706)	(0.111, 0.111)

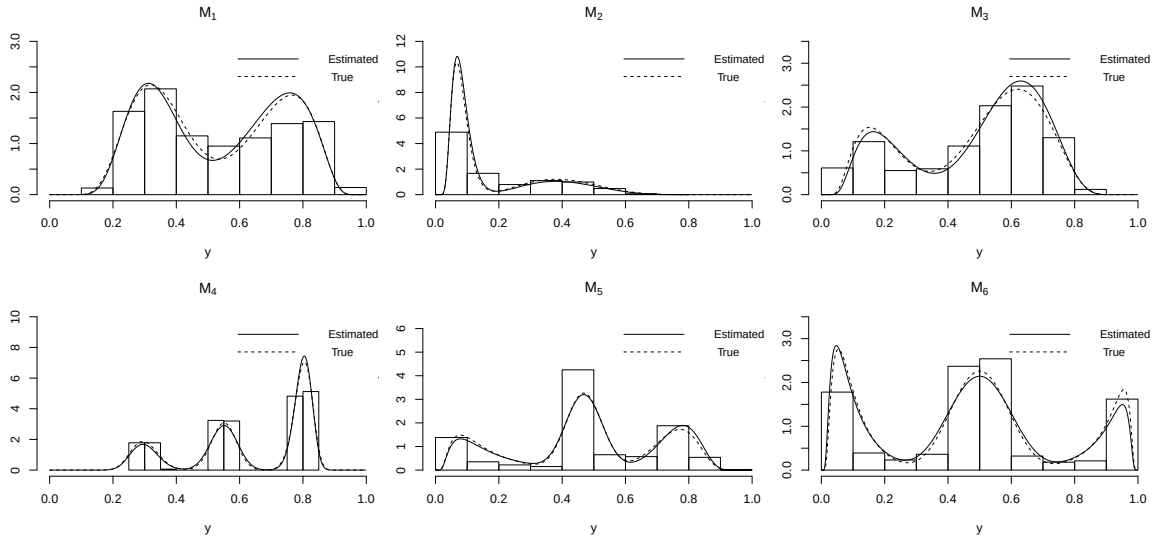

Figure 1: Histograms and estimated density function

Table 3: Posterior mean of the parameters and empirical standard deviation (SD) for simulated data set considering six models with $k = 3$ described in Table 1.

Model	n		Estimative to μ	Estimative to σ^2	Estimative to ω
$\mathcal{M}_4$	1000	Mean	(0.30, 0.55, 0.80)	(0.20, 0.12, 0.18)	(0.18, 0.31, 0.51)
		SD's	(0.007, 0.004, 0.03)	(0.07, 0.08, 0.05)	(0.01, 0.02, 0.02)
	500	Mean	(0.30 0.55 0.80)	(0.23,0.10,0.20)	(0.17 0.31 0.52)
		SD's	(0.009 0.007 0.007)	(0.10, 0.10 0.06)	(0.02 0.02 0.02)
	300	Mean	(0.30, 0.55, 0.80)	(0.27, 0.13, 0.22)	(0.17, 0.32, 0.52)
		SD's	(0.025, 0.016, 0.014)	(0.32, 0.25, 0.13)	(0.034, 0.042, 0.041)
$\mathcal{M}_5$	1000	Mean	(0.16, 0.47, 0.76)	(5.25, 0.21, 0.77)	(0.24, 0.45, 0.31)
		SD's	(0.01, 0.01, 0.01)	(0.7, 0.1, 0.1)	(0.02, 0.02, 0.02)
	500	Mean	(0.14, 0.46, 0.75)	(3.19, 0.26, 0.76)	(0.20 0.47 0.33)
		SD's	(0.02, 0.01, 0.01)	(0.65, 0.26, 0.19)	(0.02, 0.03, 0.03)
	400	Mean	(0.15, 0.48, 0.76)	(4.41, 0.37, 0.83)	(0.20, 0.45, 0.35)
		SD's	(0.031, 0.038, 0.028)	(2.9, 1.8, 0.35)	(0.04, 0.06, 0.07)
$\mathcal{M}_6$	1000	Mean	(0.10, 0.50, 0.89)	(7.34, 0.58, 7.50)	(0.31, 0.51, 0.18)
		SD's	(0.012, 0.009, 0.026)	(0.9, 0.9, 1.9)	(0.02, 0.02, 0.02)
	500	Mean	(0.098, 0.49, 0.88)	(6.05, 0.73, 9.25)	(0.31, 0.51, 0.18)
		SD's	(0.02, 0.01, 0.06)	(6,3,4)	(0.03, 0.04, 0.06)
	400	Mean	(0.097, 0.49, 0.87)	(4.843, 1.00, 10.45)	(0.31, 0.49, 0.20)
		SD's	(0.02, 0.03, 0.08)	(1, 5, 6)	(0.05, 0.05, 0.06)

We observed in the simulation process that the convergence speed is improved if initial value of number of component is set as $k^{(0)} = k_{max}$. The convergence is also affected by the acceptance of MR step in the *Gibbs sampling* algorithm then a strategy to avoid low rate of acceptance in the MR step is presented in Appendix A.1.

6 Analysis of municipal HDI data set in Brazil

The Human Development Index (HDI) is a summary measure of long-term progress in three basic dimensions of human development that takes into account education, income and longevity indexes. The HDI is the geometric mean of normalized indexes for each of the three dimensions of human development. In this work we analyse the municipal HDI data set, that is, the HDI of the cities

(or towns) of São Paulo state and Northeast region of Brazil. The Northeast was chosen because it is the third largest region of Brazil and the largest in number of states and considered a region with poor distribution of resources, see index in Fundação Instituto Brasileiro de Geografia e Estatística (2014). In this region are the states of Alagoas, Bahia, Ceará, Maranhão, Paraíba, Pernambuco, Piauí, Rio Grande do Norte and Sergipe. There are 1794 cities in the Northeastern region and 645 in São Paulo state leading to a sample of size $n = 2439$. The histogram of the data is showed in Figure 2 where we can see the multimodality phenomenon. This phenomenon is already expected because the HDI depend of characteristics that can be similar to some cities or towns.

Table 4: Relative frequency of k to the municipal HDI data set considering alternative SM models.

k	1	2	3	4	5
	0.073	0.923	0.005	0.0004	0

The municipal HDI data set was analysed with the SD model where we set $a = 2$ and $b = 1/2$ in the gamma prior distribution to σ_j^{-2}, for $j = 1, ..., k$. In order to reduce prior information we set $\nu_1 = \nu_2 = ... = \nu_k = 1$ to Dirichlet prior distribution and $k_{max} = 5$ to prior distribution of k. Table 4 show the posterior distribution of parameter k with high posterior probability to $k = 2$. Then, there are evidence for two components in the data. The mean and empirical SD of the parameter estimate are showed in Table 5.

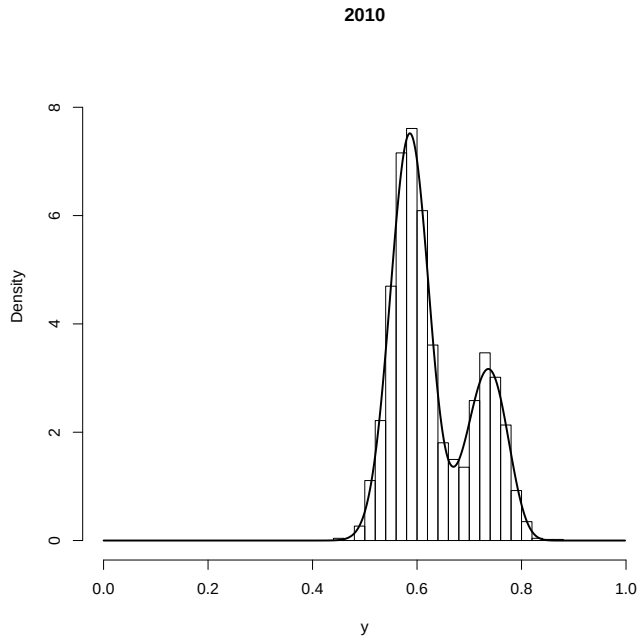

Figure 2: Real histogram and Estimated density function to the HDI data set.

Table 5: Posterior estimates of the parameters and the empirical standard deviation to the municipal HDI data set.

$\hat{\mu}$		$\hat{\sigma^2}$		$\hat{w}$	
means	SD's	means	SD's	means	SD's
(0.59, 0.73)	(0.004, 0.010)	(0.09, 0.21)	(0.03, 0.076)	(0.69, 0.31)	(0.03, 0.031)

The results of analysis show that there are strong evidence for two component of cities with similar characteristics. The first component have, in mean, smaller municipal HDI than second component. The component with less municipal HDI have larger mixing proportions than second component, as expected. The Figure 3(A) shows in red the cities classified as belonging to the first component. We can observe that there are some cities in the Northeastern region classified in the first component. This cities has better municipal HDI than those which are classified in the second component, shown in (3)(B) cities detached with colour blue.

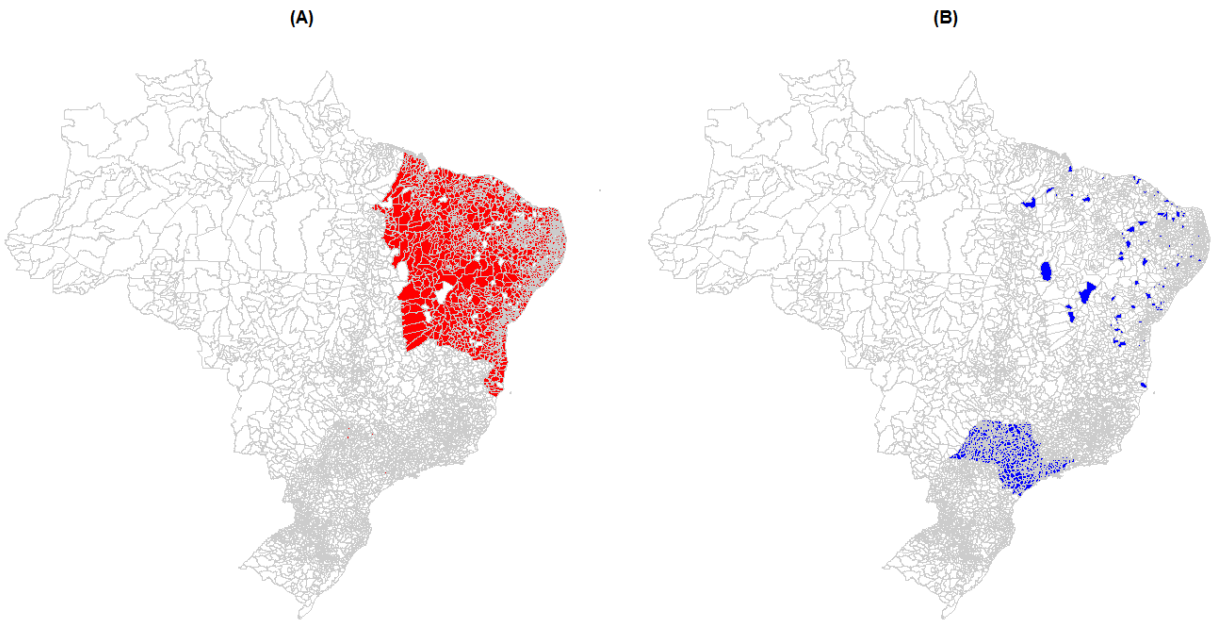

Figure 3: Classification of HDI of cities of the states São Paulo and Northeastern region of Brazil where the cities classified in the first component is in red (A) and cities classified in the second component is in blue (B).

7 Final comments

The main advantage of mixture of simplex models is its flexibility for working with bounded data with multimodality identified the components or populations in the data. A Full Bayesian approach considering an MCMC with reversible-jump algorithm similar to the methodology proposed by Richardson & Green (1997) and Green (1995) was developed.

An application to generated data sets from a mixture of simplex distributions with 2 and 3 components were conducted. For these applications, we found that the method provides a good estimate to the number of component as well as the other parameters of the model since the estimated values lie close to the real values of the parameters. In addition, the results from the simulated data sets with different size show that the empirical standard deviation decrease as the size of sample increase, as expected when the method works well. Another application was conduced with real data set and we fond small empirical standard deviations to the sample of the estimates of the parameters

of the components and mixing proportions, as we can see in Table 5.

The proposed model can be extended to the another problems, for instance, where the random variable Y (response) can be modelled as a function of another variable x (predictor variable). In this case, the mixing proportions can or can not be modelled as functions of a vector of predictor variable, not necessarily having some elements in common with the vector of covariates x, a example of the link function that can be used to mixing proportions is presented in McLachlan & Peel (2004, p. 145). In addition, since we observed that the acceptance of RJ decrease as size of the sample increases, strategy to avoid persistent rejection of proposed moves in a RJ algorithm can be added to improve the *Gibbs sampling* algorithm, strategy are discussed in Green & Mira (2001); Al-Awadhi *et al.* (2004).

References

Al-Awadhi, F., Hurn, M. & Jennison, C. (2004). Improving the acceptance rate of reversible jump MCMC proposals. *Statistics and Probability Letters*, **69**(2), 189–198.

Barndorff-Nielsen, O. & Jorgensen, B. (1991). Some parametric models on the simplex. *Journal of Multivariate Analysis*, **39**(1), 106 – 116.

Bouguila, N. & Elguebaly, T. (2012). A fully Bayesian model based on reversible jump MCMC and finite beta mixtures for clustering. *Expert Syst. Appl.*, **39**(5), 5946–5959.

Bouguila, N., Ziou, D. & Monga, E. (2006). Practical bayesian estimation of a finite beta mixture through gibbs sampling and its applications. *Statistics and Computing*, **16**(2), 215–225.

Cifuentes, M., Sembajwe, G., Tak, S., Gore, R., Kriebel, D. & Punnett, L. (2008). The association of major depressive episodes with income inequality and the human development index. *Social Science and Medicine*, **67**(4), 529 – –539.

Diebolt, J. & Robert, C. P. (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B*, **56**(2), 363–375.

Faria, S. & Gonalves, F. (2013). Financial data modeling by poisson mixture regression. *Journal of Applied Statistics*, **40**(10), 2150–2162.

Fundação Instituto Brasileiro de Geografia e Estatística, D. d. E. e. R. (2014). *Pesquisa nacional por amostra de domicílios, PNAD.: Síntese de indicadores*. IBGE.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**(410), 398–409.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Green, P. J. & Mira, A. (2001). Delayed rejection in reversible jump Metropolis-Hastings. *Biometrika*, **88**, 1035–1053.

Jørgensen, B. (1997). *The Theory of Dispersion Models*. Monographs on Statistics and Applied Probability. Taylor & Francis.

López, F. O. (2013). A bayesian approach to parameter estimation in simplex regression model: A comparison with beta regression. *Revista Colombiana de Estatística*, **36**(1), 1–21.

Marin, J. M., Mengersen, K. & Robert, C. P. (2005). *Bayesian modelling and inference on mixtures of distributions*. Elsevier.

McDonald, J. & Ransom, M. (2008). The generalized beta distribution as a model for the distribution of income: Estimation of related measures of inequality. In *Modeling Income Distributions and Lorenz Curves*, volume 5 of *Economic Studies in Equality, Social Exclusion and Well-Being*, pages 147–166. Springer New York.

McLachlan, G. & Peel, D. (2000). *Finite Mixture Models (Wiley Series in Probability and Statistics)*. Wiley-Interscience, first edition.

McLachlan, G. & Peel, D. (2004). *Finite Mixture Models*. Wiley series in probability and statistics: Applied probability and statistics. Wiley.

Paz, R. F., Brazan, J. L. & Elher, R. l. (2015). A Weibull mixture model for the votes of a Brazilian political party. In *EBEB: Interdisciplinary Bayesian Statistics*, volume 118 of *Springer Proceedings in Mathematics and Statistics*, pages 229–241. Springer.

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Richardson, S. & Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society*, **59**(4), 731–792.

Ross, S. M. (2006). *Simulation, Fourth Edition*. Academic Press, Inc., Orlando, FL, USA.

Song, P. X.-K. & Tan, M. (2000). A marginal models for longitudinal continuous proportional data. *Biometrics*, **56**(2), 496–502.

Stensholt, E. (1999). Beta distributions in a simplex and impartial anonymous cultures. *Mathematical Social Sciences*, **37**(1), 45–57.

Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**(398), 528–540.

A Appendix

Let's now summarize the MCMC technique mentioned above by giving a description of the Gibbs sampling algorithm used to sample from the joint probability distribution. The *Gibbs sampling* algorithm is used combined with *Metropolis-Hastings* (MR and reversible-jump) algorithm for obtain the sample of the posterior distribution of parameters $(\boldsymbol{\theta}, \boldsymbol{\omega}, k)$ and Z_i, for $i = 1, ..., N$. A scheme of the algorithm is shown below.

A.1 Algorithm

1. Initialize choosing $k^{(t)} = k^{(0)}$, $\boldsymbol{\omega}^{(t)} = \boldsymbol{\omega}^{(0)}$, $\mu_j^{(t)} = \mu_j^{(0)}$, $(\sigma_j^2)^{(t)} = (\sigma_j^2)^{(0)}$ and $Z_{ij}^{(t)} = Z_{ij}^{(0)}$, for $i = 1, ..., n$ $j = 1, ..., k^{(t)}$.

2. For $t = 0, 1, 2, \ldots$ repeat

   (a) For $i = 1, ..., n$ draw $Z_i^{(t+1)} \sim Multinomial(1, \pi_{i1}^{(t)}, ..., \pi_{ik^{(t)}}^{(t)})$, wherein

   $$\pi_{ij}^{(t)} = P\big(Z_{ij}^{(t)} = 1 | y_i, \mu_j^{(t)}, (\sigma_j^2)^{(t)}\big) \propto \omega_j S(y_i | \mu_j^{(t)}, (\sigma_j^2)^{(t)})$$

   (b) Generate $\boldsymbol{\omega}^{(t+1)}$ from the distribution given by (14).

   (c) For $j = 1, ..., k^{(t)}$ do

      i. Generate $\phi_j^{(t+1)}$ from the distribution given by (12) and do $(\sigma_j^2)^{(t+1)} = 1/\phi_j^{(t+1)}$.

      ii. For updating μ_j a *Metroplis-hastings* step is done, then

      - generate $\mu_j^{'} \sim Beta\left(\delta^{(t)}, \eta^{(t)}\right)$ where $(\delta^{(t)}, \eta^{(t)})$ is computed in A.1.1.

      - Compute
        $\alpha\left(\mu_j^{(t)}, \mu_j^{'}\right) = \min\left\{1, \frac{P\left(\mu_j^{'} | \mathbf{y}, Z, (\sigma_j^2)^{(t+1)}\right)}{P\left(\mu_j^{(t)} | \mathbf{y}, Z, (\sigma_j^2)^{(t+1)}\right)} \frac{Be\left(\mu_j^{(t)} | \delta^{(t+1)}, \eta^{(t+1)}\right)}{Be\left(\mu_j^{'} | \delta^{(t)}, \eta^{(t)}\right)}\right\}$
        where $Be(x|.)$ is the density of beta distribution evaluated at x.

      - Generate $u \sim Uniform(0, 1)$

      - If $\alpha\left(\mu_j^{(t)}, \mu_j^{'}\right) < u$ then $\mu_j^{(t+1)} = \mu_j^{'}$ else $\mu_j^{(t+1)} = \mu_j^{(t)}$.

   (d) For updating $k^{(t)}$, merge two component of the mixture into one or splinting one into two by using reversible-jump step.

A.1.1 Proposal distribution

We observed in the simulation process that the convergence of algorithm *Gibbs sampling* is affected by the acceptance of MR step in (2(c)ii). In order to improve the acceptance rate of $\mu_j^{'}$ ($j = 1, ..., k^{(t)}$ and $t = 0, 1, 2, ...$), in the *Metropolis-Hastings* step (2(c)ii) of the algorithm, we adopt

a beta distribution as the proposal distribution where the parameters $\delta^{(t)}$ and $\eta^{(t)}$ are obtained by solving

$$\begin{cases} \mu_j^{(t)} = \frac{\delta^{(t)}}{\delta^{(t)}+\eta^{(t)}} \\ \psi^{(t)} = \frac{\delta^{(t)}\eta^{(t)}}{(\delta^{(t)}+\eta^{(t)})^2(\delta^{(t)}+\eta^{(t)}+1)} \end{cases} \tag{17}$$

where $\mu_j^{(t)}$ and $\psi^{(t)}$ is the mean and variance of beta distribution with parameters $\delta^{(t)} > 0$ and $\delta^{(t)} > 0$. Then we have

$$\begin{cases} \eta^{(t)} = \delta^{(t)} \left(\frac{1}{\mu_j^{(t)}} - 1 \right) \\ \delta^{(t)} = (\mu_j^{(t)})^2 \left(\frac{1-\mu_j^{(t)}}{\psi^{(t)}} - \frac{1}{\mu_j^{(t)}} \right) \end{cases} \tag{18}$$

The positivity of $\delta^{(t)}$ and $\delta^{(t)}$ is secured by making $\psi^{(t)} < \mu^{(t)}(1-\mu^{(t)})$ leading to $\psi^{(t)} = \mu_j^{(t)}(1-\mu_j^{(t)})\times\tau$, with $0 < \tau < 1$.