Brazilian Journal of Probability and Statistics
2020, Vol. 34, No. 2, 323–344
https://doi.org/10.1214/18-BJPS425

A Bayesian sparse finite mixture model for clustering data from a heterogeneous population

Erlandson F. Saraiva[a], **Adriano K. Suzuki**[b] **and Luís A. Milan**[c]

[a]*Universidade Federal de Mato Grosso do Sul*
[b]*Unversidade de São Paulo*
[c]*Universidade Federal de São Carlos*

Abstract. In this paper, we introduce a Bayesian approach for clustering data using a sparse finite mixture model (SFMM). The SFMM is a finite mixture model with a large number of components k previously fixed where many components can be empty. In this model, the number of components k can be interpreted as the maximum number of distinct mixture components. Then, we explore the use of a prior distribution for the weights of the mixture model that take into account the possibility that the number of clusters $k_{\mathbf{c}}$ (e.g., nonempty components) can be random and smaller than the number of components k of the finite mixture model. In order to determine clusters we develop a MCMC algorithm denominated Split-Merge allocation sampler. In this algorithm, the split-merge strategy is data-driven and was inserted within the algorithm in order to increase the mixing of the Markov chain in relation to the number of clusters. The performance of the method is verified using simulated datasets and three real datasets. The first real data set is the benchmark galaxy data, while second and third are the publicly available data set on Enzyme and Acidity, respectively.

1 Introduction

The main goal of the clustering analysis is to group the observed data into homogeneous clusters. The first works on clustering are due Sneath (1957) and Sokal and Michener (1958). Since the publication of these articles the clustering problem has been studied by many authors as Hartigan and Wong (1978), Binder (1978), Anderson (1985), Banfield and Raftery (1993), Bensmail et al. (1997) and Witten and Tibshirani (2010). According to Bouveyron and Brunet (2013), more and more scientific fields requires clustering data with the aim of understanding the phenomenon of interest.

Usual clustering approaches are distance-based, such as k-means (MacQueen (1967)) and hierarchical clustering (Ward (1963)). Although simple and visually appealing these methods can be implemented only for cases where the number of clusters are known. Besides, according to Oh and Raftery (2007) “these methods are not based on standard principles of statistical inference and they do not provide a statistically based method for choosing the number of clusters”.

Under a probabilistic approach McLachlan and Basford (1988), Banfield and Raftery (1993), McLachlan and Peel (2000) and Fraley and Raftery (2002) proposed the use of a finite mixture model in which each component of the mixture represents a cluster. In these methods, partitions are determined by the EM (expectation-maximization) algorithm and the number of components of the mixture are, in general, determined comparing fitted models with different number of components using some model selection criterion, such that, AIC (Akaike (1974), Bozdogan (1987)) or BIC (Schwarz (1978)). A similar strategy is adopted in

Key words and phrases. Mixture model, Bayesian approach, Gibbs sampling, Metropolis–Hastings, split-merge update.

Received April 2018; accepted November 2018.

the Bayesian approach, considering the DIC (Spiegelhalter et al. (2002)) as model selection criterion, see, for example, Celeux, Hurn and Robert (2000).

This can be seen as a drawback to be overcome, since in practice it may be very tedious to fit several models and afterwards compare them according to a model selection criterion. Also, in many cases the estimation depends on iterative methods which may not converge imposing additional difficulties to the process. Thus, a practical and efficient computational algorithm to estimate the number of cluster jointly with the component-specific parameters is desirable. Under this scenario, the Bayesian approach has been successful, in special, due the reversible-jump Markov Chain Monte Carlo algorithm proposed by Richardson and Green (1997) in the context of Gaussian mixture models. However, one difficulty frequently encountered for implementing a reversible-jump algorithm is the construction of efficient transitions proposals that lead to a reasonable acceptance rate.

In this paper, we consider a sparse finite mixture model for clustering data. This model is a finite mixture model with k components, previously fixed as a large number, where many components can be empty. The large value assumed for k can be interpreted as the maximum number of distinct mixture components. The term sparse refers refers to the existence of many empty components. The main motivation to consider this model is the fact that a finite mixture model is a population model, then given an observed sample $\mathbf{y}$ not all k components may have observations in the sample and we may have empty components. In addition, assuming this model, we avoid the need to use the reversible-jump MCMC method. It allowed us to implement the split-merge movements using the observed data; instead of to use a transition function as in a RJMCMC.

Thus, we assume that the number of clusters $k_{\mathbf{c}}$ (*i.e.*, number of non empty components) is an unknown quantity, but smaller than k. To estimate $k_{\mathbf{c}}$ jointly with the other parameters of interest, we consider a Bayesian approach. In this approach, we set up a symmetric Dirichlet prior distribution with hyperparameter γ on the weights of the mixture. So, using the Dirichlet integral we integrate out the weights and derive the prior and posterior allocations probabilities.

In order to simulate from joint posterior distribution of the parameters of interest, we develop a MCMC algorithm, the Split-Merge Allocation sampler (SMAS). This algorithm consist of three steps. In the first step, parameters of the not-empty components (e.g., clusters) are updated from its conditional posterior distribution and parameters of the empty components are updated from prior distribution. In the second step, a Gibbs sampling is performed to update the latent indicator variables using the posterior allocations probabilities. And the third step consist of a split-merge step that change the number of clusters on the neighbourhood ± 1, respectively. This step was inserted within the algorithm in order to allow a major change in configuration of the latent variables in a single iteration of the algorithm, and consequently increasing the mixing of the Markov chain in relation to the number of clusters.

The split and merge movements are developed in a way that they are reversible. In a split move each observation is allocated to one of two new partitions based on probabilities which are calculated according to marginal likelihood function from observations previously allocated to two new partitions.

Given the proposed allocation of observations, the candidate-values for component parameters are generated from the conditional posterior distributions. Choosing the posterior density as candidate-generating density, the likelihood ratio and the ratio of prior densities from the Metropolis–Hastings acceptance probability cancels the corresponding term in the posterior density, simplifying the computation of the acceptance probability. Thus, the proposed algorithm performs a standard Metropolis–Hastings update with an acceptance probability which depends only on data associated with component(s) selected for a split or a merge.

Our approach does not require the specification of transition functions to estimate the number of clusters jointly with the component-specific parameters and when a new cluster is

Table 1 *Mathematical notation*

Notation	Description
k	Number of components
$k_{\mathbf{c}}$	Number of clusters
θ_j	Parameter of the jth component, for $j = 1, \dots, k$
$\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$	The whole vector of parameters
w_j	Weight of the jth component, for $j = 1, \dots, k$
Y_i	The ith sampled value, for $i = 1, \dots, n$
c_i	The ith indicator variable, for $i = 1, \dots, n$
$\mathbf{y} = (y_1, \dots, y_n)$	The vector of independent observations
$\mathbf{c} = (c_1, \dots, c_n)$	The vector of latent indicator variables
$\boldsymbol{\Phi} = (\boldsymbol{\theta}, \mathbf{c}, k_{\mathbf{c}})$	Current state
$\boldsymbol{\Phi}^{\mathrm{sp}} = (\boldsymbol{\theta}^{\mathrm{sp}}, \mathbf{c}^{\mathrm{sp}}, k_{\mathbf{c}}^{\mathrm{sp}})$	Split state
$\boldsymbol{\Phi}^{\mathrm{mg}} = (\boldsymbol{\theta}^{\mathrm{mg}}, \mathbf{c}^{\mathrm{mg}}, k_{\mathbf{c}}^{\mathrm{mg}})$	Merge state

created, through a split or a merge movement, it determines a new partition in the data set. This is due to the way that we implement the split-merge movements which is data-driven instead of parameter based.

In order verify the performance of the SMAS, we develop a simulation study considering data sets generated with a different number of clusters. We also verify its sensibility to the choice of the value of hyperparameter γ of the Dirichlet prior distribution. To do this, we explore two scenery. In the first we set up $\gamma = r$, where r is value of a grid $\mathbb{G}$. In the second, we consider one more hierarchical level with a prior distribution on γ and estimating it from its posterior distribution. For each artificial dataset, we present the performance of SMAS in terms of posterior probability for number of clusters, convergence, mixing and autocorrelation.

We also apply the methodology to three real data sets. The first is the benchmark data on velocities from distant galaxies diverging each other, previously analyzed by Roeder and Wasserman (1997), Escobar and West (1995), Richardson and Green (1997), Stephens (2000) and Saraiva, Louzada and Milan (2014), available in R software. The second and the third one are Enzyme and Acidity datasets extracted from website https://people.maths.bris.ac.uk/~mapjg/mixdata.

The remainder of the paper is structured as follows. In Section 2, we describe the sparse finite mixture model for clustering data. In Section 3, we introduce the hierarchical Bayesian approach and develop the SMAS algorithm. In Section 4, the proposed sampler is applied to simulated data sets to access its performance and to real data sets to illustrate its use. Section 5 concludes the paper with final remarks. Additional details are provided in the supplementary material (Saraiva, Suzuki and Milan (2019)), denoted by prefix "SM" when referred to in this paper. Table 1 presents the main notations used throughout the article.

2 Mixture model for clustering

Consider a population composed by k subpopulations, such that, the sampling units are homogeneous with respect to the characteristic under study within the subpopulation and heterogeneous among the subpopulations. Let $w_1, \dots, w_k$ be the relative frequencies of each subpopulation in relation to the overall population, for $0 \leq w_j \leq 1$ and $\sum_{j=1}^{k} w_j = 1$. Assume that each subpopulation j is modeled by a probability distribution $F(\theta_j)$ indexed by parameter θ_j (scalar or vector), for $j = 1, \dots, k$.

Suppose that the sampling process consists of choosing a subpopulation j with probability w_j and then sample a Y_i value of this subpopulation, for $j = 1, \dots, k$ and $i = 1, \dots, n$ where

n is the sample size. Then we can represent each sample unit by the pair (Y_i, c_i), where c_i is an indicator variable that assume a value of the set $\{1, \dots, k\}$ with probabilities $\{w_1, \dots, w_k\}$, respectively. Thus, we have that

$$(Y_i | c_i = j, \theta_j) \sim F(\theta_j) \quad \text{and} \quad P(C_i = j | \mathbf{w}) = w_j, \tag{2.1}$$

where $\mathbf{w} = (w_1, \dots, w_k)$, for $i = 1, \dots, n$ and $j = 1, \dots, k$.

In many practical problems such as clustering problems indicator variables are non-observable (also known as latent variables). The probability of ith observation coming from subpopulation j is w_j, for $i = 1, \dots, n$ and $j = 1, \dots, k$. The marginal probability density function for $Y_i = y_i$ is given by

$$f(y_i | \boldsymbol{\theta}, \mathbf{w}) = \sum_{j=1}^{k} w_j f(y_i | \theta_j), \tag{2.2}$$

where $f(y_i|\theta_j)$ is the probability density function of $F(\theta_j)$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ is the whole vector of parameters and $\mathbf{w} = (w_1, \dots, w_k)$ are the weights. Model (2.2) is denominated in the literature by finite mixture model. In this model each probability density function, $f(\cdot|\theta_j)$, corresponds to a component of the mixture.

The finite mixture model is a natural probabilistic approach for clustering. However, as model (2.2) is a population model, then given an observed sample $\mathbf{y} = (y_1, \dots, y_n)$ not all k components may have observations in the sample and we may have empty components, *that is*, fewer than k components are in the sample (Fruhwirth-Schnatter (2017)). In this case, we have that the number of clusters (*i.e.*, non-empty components) is smaller than the number of components k. Walli, Frhwirth-Schnatter and Grn (2016) and Fruhwirth-Schnatter (2017) call this model by sparse finite mixture model.

Our interest is to infer about the number of clusters $k_{\mathbf{c}}$ and the latent variables $\mathbf{c}$ jointly with the components parameters. Due to this, we consider $\mathbf{c}$ as "parameters" to be estimated in model (2.2).

2.1 Joint distribution for complete data

Let $(\mathbf{y}, \mathbf{c})$ be the complete data, where $\mathbf{y} = (y_1, \dots, y_n)$ is the vector of independent observations and $\mathbf{c} = (c_1, \dots, c_n)$ is the vector of latent indicator variables, with $\mathbf{y}$ and $\mathbf{c}$ being paired. Consider $k_{\mathbf{c}}$ the number of clusters defined by the configuration $\mathbf{c}$, $k_{\mathbf{c}} \leq k$.

From model (2.1), we have that the joint probability for the latent indicator variables $\mathbf{c}$ is given by

$$\pi(\mathbf{C} = \mathbf{c} | \mathbf{w}, k) = \prod_{j=1}^{k} w_j^{n_j}, \tag{2.3}$$

where n_j is the number of observations y_i's allocated to subpopulation j, for $j = 1, \dots, k$.

The distribution of the number of observations assigned to each component, $n_1, \dots, n_k$, called the occupation number, is multinomial

$$(n_1, \dots, n_k | n, \mathbf{w}) \sim \text{Multinomial}(n, \mathbf{w}), \tag{2.4}$$

where $n = n_1 + \cdots + n_k$.

Consider $D_j = \{y_i; c_i = j\}$ be the set of observations allocated to component j, for $j = 1, \dots, k$. Without loss of generality, assume that clusters are labelled from 1 to $k_{\mathbf{c}}$ and that the empty components are labelled from $k_{\mathbf{c}} + 1$ to k. Thus, the joint distribution for complete data $(\mathbf{y}, \mathbf{c})$ conditional on mixing proportions $\mathbf{w}$, component parameters $\boldsymbol{\theta}$ and number of components k is

$$\begin{aligned} P(\mathbf{Y}=\mathbf{y}, \mathbf{C}=\mathbf{c}|\mathbf{w}, \boldsymbol{\theta}, k) &= P(\mathbf{Y}=\mathbf{y}|\mathbf{c}, \mathbf{w}, \boldsymbol{\theta}, k)\pi(\mathbf{C}=\mathbf{c}|\mathbf{w}, k) \\ &= \prod_{j=1}^{k}\prod_{i=1}^{n}[w_j f(y_i|\theta_j)]^{\mathbb{I}_{c_i}(j)} \\ &= \prod_{j=1}^{k_{\mathbf{c}}} w_j^{n_j}\left(\prod_{D_j} f(y_i|\theta_j)\right), \end{aligned} \tag{2.5}$$

where $\mathbb{I}_{c_i}(j) = 1$ if $c_i = j$ and $\mathbb{I}_{c_i}(j) = 0$ otherwise, for $i = 1, \dots, n$ and $j = 1, \dots, k$.

At this point some remarks about the label switching are necessary. Note that, the cluster labels $j = 1, \dots, k_{\mathbf{c}}$ are not uniquely determined and a permutation of the labels would lead to the same model. Since our interest lies in inferences on partitions, the non-identifiability of labels would cause a problem in posterior computation and allocation probabilities are useless for partitioning the observations (Stephens (2000)). Therefore, we impose restrictions on the class of component means of the clusters to get identifiability, *that is*, we assume that $\mu_1, \dots, \mu_{k_{\mathbf{c}}}$ are the component means for clusters and that $\mu_1 < \cdots < \mu_{k_{\mathbf{c}}}$. However, it does not prevent the MCMC algorithm described in the next section from being applicable to other labelling. For further discussion and additional references about label switching, see Stephens (2000), Jasra, Holmes and Stephens (2005) and their references.

3 Bayesian approach

In order to estimate $\mathbf{c}$ and the number of clusters $k_{\mathbf{c}}$ jointly with the component parameters $\boldsymbol{\theta}$, we consider the following hierarchical Bayesian model

$$\begin{aligned} Y_i|c_i = j, \boldsymbol{\theta}, k &\sim F(\theta_j), \\ c_i|\mathbf{w}, k &\sim \text{Discrete}(w_1, \dots, w_k), \\ \theta_j &\sim G(\eta_j), \\ \mathbf{w}|\gamma, k &\sim \text{Dirichlet}(\gamma), \end{aligned} \tag{3.1}$$

for $i = 1, \dots, n$, where $G(\eta_j)$ is the *a priori* distribution for component parameters θ_j, η_j (scalar or vector) are the hyperparameters, for $j = 1, \dots, k$, and Dirichlet(γ) represents the symmetric Dirichlet distribution with parameter γ, $\gamma > 0$, and probability density function given by

$$\pi(\mathbf{w}|\gamma, k) = \frac{\Gamma(k\gamma)}{[\Gamma(\gamma)]^k}\prod_{j=1}^{k} w_j^{\gamma-1}. \tag{3.2}$$

The choice of the prior distribution $G(\eta_j)$ for the parameter θ_j depends on the probability distribution $F(\theta_j)$ assumed for Y_i, for $i = 1, \dots, n$ and $j = 1, \dots, k$. In Section 4, we discuss the choice of $G(\eta_j)$ in the context in which $F(\theta_j)$ is given by a Gaussian distribution with mean μ_j and variance σ_j^2, *i.e.*, $\theta_j = (\mu_j, \sigma_j^2)$, for $j = 1, \dots, k$.

It is important to note that depending on the value fixed for the hyperparameter γ the Dirichlet sampling may generate some weights $w_j = 0$ and consequently the multinomial sampling given in (2.4) lead to partitions with $n_j = 0$, for some $j \in \{1, ..., k\}$. According to Walli, Frhwirth-Schnatter and Grn (2016) and Fruhwirth-Schnatter (2017) this happens when we consider a small value for hyperparameter γ.

In order to verify the sensibility of the results to the choice of the value for the hyperparameter γ we explore two scenery. First, we set up $\gamma = r$, where r is a value of the set $\mathbb{G} = \{R_1, R_2\}$ in which R_1 is a grid from 0.01 to 0.09 with a fixed increment step equals to 0.01 and R_2 is a grid from 0.1 to 2 with a fixed increment step equals to 0.1. In the second scenery, we consider one more hierarchical level with a gamma prior distribution on γ, $\gamma \sim \Gamma(a, b)$, for $a, b > 0$. The parametrization of the gamma distribution is so that the expected value is a/b and the variance is a/b^2.

As our main interest lies on configuration $\mathbf{c}$, then we integrate out the mixing proportion. Taking the product of equations in (2.3) and (3.2) and integrating out the mixing proportions, we can write the joint probability of $\mathbf{c}$ given γ and k as

$$\pi(\mathbf{C} = \mathbf{c}|\gamma, k) = \frac{\Gamma(k\gamma)}{[\Gamma(\gamma)]^k \Gamma(n + k\gamma)} \prod_{j=1}^{k} \Gamma(n_j + \gamma). \tag{3.3}$$

The joint probability for complete data $(\mathbf{y}, \mathbf{c})$ given $\boldsymbol{\theta}$, γ and k is given by

$$\begin{aligned} P(\mathbf{Y} = \mathbf{y}, \mathbf{C} = \mathbf{c}|\boldsymbol{\theta}, \gamma, k) &= P(\mathbf{Y} = \mathbf{y}|\mathbf{c}, \boldsymbol{\theta}, k)\pi(\mathbf{C} = \mathbf{c}|\gamma, k) \\ &= \left[\prod_{j=1}^{k_{\mathbf{c}}} \left(\prod_{D_j} f(y_i|\theta_j)\right)\right]\pi(\mathbf{C} = \mathbf{c}|\gamma, k), \end{aligned} \tag{3.4}$$

where $\pi(\mathbf{C} = \mathbf{c}|\gamma, k)$ is given in (3.3).

Using the Dirichlet integral and keeping all but a single indicator variable fixed, the conditional distribution for a single latent indicator variable c_i given all others, denoted by $\mathbf{c}_{-i} = (c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_n)$, is given by

$$\pi(C_i = j|\mathbf{c}_{-i}, \gamma) = \frac{n_{j,-i} + \gamma}{n + k\gamma - 1}, \tag{3.5}$$

where $n_{j,-i}$ is the number of observations in D_j except the observation y_i, for $i = 1, \ldots, n$ and $j = 1, \ldots, k$.

From (3.5), the probability of observation y_i to be assigned to one empty component is

$$\pi(C_i = j^*|\mathbf{c}_{-i}, \gamma) = \frac{\gamma}{n + k\gamma - 1}, \tag{3.6}$$

for $j^* \in \{k_{\mathbf{c}_{-i}} + 1, \ldots, k\}$, where $k_{\mathbf{c}_{-i}}$ is the number of clusters defined by configuration $\mathbf{c}_{-i}$, for $i = 1, \ldots, n$. The expression in (3.6) is the probability of a observation y_i to define a new cluster, for $i = 1, \ldots, n$.

3.1 Posterior distribution

Using the Bayes theorem, the joint posterior distribution upon which inference is based is given by

$$\pi(\boldsymbol{\theta}, \mathbf{c}|\mathbf{y}, \gamma, k) \propto P(\mathbf{Y} = \mathbf{y}|\mathbf{c}, \boldsymbol{\theta}, k)\pi(\mathbf{C} = \mathbf{c}|\gamma, k)\pi(\boldsymbol{\theta}|k), \tag{3.7}$$

where $P(\mathbf{Y} = \mathbf{y}|\mathbf{c}, \boldsymbol{\theta}, k)\pi(\mathbf{C} = \mathbf{c}|\gamma, k) = L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{c}, \gamma, k)$ is the complete-data likelihood function for $\boldsymbol{\theta}$, which is equal to the sampling distribution given in (3.4), regarded as a function of the unknown parameters $\boldsymbol{\theta}$, and $\pi(\boldsymbol{\theta}|k) = \prod_{j=1}^{k} \pi_G(\theta_j|\eta_j)$ is the joint prior distribution for $\boldsymbol{\theta}$ with $\pi_G(\theta_j|\eta_j)$ being the probability density function of the prior distribution $G(\eta_j)$, for $j = 1, \ldots, k$.

In order to sample from the joint posterior distribution in (3.7) and estimate parameters of interest, we propose the Split-Merge allocation sampler algorithm (SMAS). This is a MCMC algorithm that makes use of the following three moves types.

(a) update the parameters $\boldsymbol{\theta}$;
(b) update the allocation indicators $\mathbf{c}$;
(c) split one cluster into two, or combine two clusters into one.

3.1.1 *Updating the parameters.* Conditional on configuration $\mathbf{c}$, we have $k_{\mathbf{c}}$ clusters and $(k - k_{\mathbf{c}})$ empty components. The conditional posterior distribution for θ_j is given by

$$\pi(\theta_j|\mathbf{y}, \mathbf{c}, k) \propto L(\theta_j|D_j)\pi_G(\theta_j|\eta_j), \tag{3.8}$$

where

$$L(\theta_j|D_j) = \begin{cases} \prod_{D_j} f(y_i|\theta_j), & \text{if } D_j \neq \varnothing; \\ 1, & \text{if } D_j = \varnothing \end{cases} \tag{3.9}$$

is the likelihood function for component j, for $i = 1, \dots, n$ and $j = 1, \dots, k$.

Thus, we update component parameters according to Algorithm 1.

Algorithm 1. Let the state of the Markov chain consist of $\mathbf{c} = (c_1, \dots, c_n)$ and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$. Conditional on configuration $\mathbf{c}$, update component parameters $\boldsymbol{\theta}$ as follows:

(1) For clusters j, $j = 1, \dots, k_{\mathbf{c}}$, generate θ_j from posterior distribution given in (3.8);
(2) For empty components j, $j = k_{\mathbf{c}} + 1, \dots, k$, generate θ_j from prior distribution, $\theta_j \sim \pi_G(\theta_j)$;
(3) Accept the updated values, $\boldsymbol{\theta}^{\text{updated}}$, only if adjacency condition for component parameters of the clusters is met, *that is*, if $\mu_1^{\text{updated}} < \cdots < \mu_{k_{\mathbf{c}}}^{\text{updated}}$. Otherwise, keep $\boldsymbol{\theta}^{\text{updated}} = \boldsymbol{\theta}$;

3.1.2 *Updating the allocation indicators.* Given the component parameters $\boldsymbol{\theta}$, the conditional posterior probability for $C_i = j$ is given by

$$\pi(C_i = j|y_i, \theta_j, \mathbf{c}_{-i}, \gamma) = \frac{n_{j,-i} + \gamma}{n + k\gamma - 1} f(y_i|\theta_j), \tag{3.10}$$

for $j = 1, \dots, k_{\mathbf{c_{-i}}}$ and $i = 1, \dots, n$.

We need now to define the probability of an observation y_i to be allocated to an empty component, *that is*, to create a new cluster. In order to probability of creating a new cluster does not depend of the parameter value θ_{j^*}, which is generated from prior distribution, we integrate parameters out for this case. Thus, the conditional posterior probability for $C_i = j^*$ is

$$\pi\big(C_i = j^*|y_i, \mathbf{c}_{-i}, \gamma, \eta_{j^*}\big) = \frac{\gamma}{n + k\gamma - 1} \int f(y_i|\theta_{j^*})\pi_G(\theta_{j^*}|\eta_{j^*})\, d\theta_{j^*} \tag{3.11}$$

for $j^* \in \{k_{\mathbf{c}_{-i}} + 1, \dots, k\}$ and $i = 1, \dots, n$.

Let $\mathbf{P_i} = (\mathbf{P}_{\mathbf{c}_{-i}}, \mathbf{P}_{0_i})$ be the vector of allocation probabilities for the ith observation, where $\mathbf{P}_{\mathbf{c}_{-i}} = (P_1, \dots, P_{k_{\mathbf{c}_{-i}}})$ for P_j given in (3.10) and $P_0 = (P_{k_{\mathbf{c}_{-i}}+1}, \dots, P_k)$ for P_{j^*} given in (3.11), for $j = 1, \dots, k_{\mathbf{c_{-i}}}$, $j^* = k_{\mathbf{c}_{-i}} + 1, \dots, k$ and $i = 1, \dots, n$.

Thus, we update the latent indicator variables according to Algorithm 2.

Algorithm 2. Let the state of the Markov chain consist of $\mathbf{c} = (c_1, \dots, c_n)$ and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$. Conditional on $\boldsymbol{\theta}$, update $\mathbf{c} = (c_1, \dots, c_n)$ as follows. For $i = 1, \dots, n$:

(1) remove c_i from current state $\mathbf{c}$, obtaining $\mathbf{c}_{-i}$ and $k_{\mathbf{c}_{-i}}$;
(2) calculate the vector of allocation probabilities $\mathbf{P}_i$;
(3) generate an auxiliary variable

$$\mathbf{Z}_i = (Z_{i1}, \dots, Z_{ik}) \sim \text{Multinomial}(1, \mathbf{P}_i);$$

(4) If $Z_{ij} = 1$, for $j \in \{1, \dots, k_{\mathbf{c}_{-i}}\}$, set up $c_i = j$ and do $n_j = n_{j,-i} + 1$;

(5) If $Z_{ij^*} = 1$ do $n_{j^*} = 1$ and $k_{\mathbf{c}} = k_{\mathbf{c}_{-i}} + 1$. Generate a value for the component parameter θ_{j^*} of the new cluster from the posterior distribution $\pi(\theta_{j^*}|y_i)$. Relabel the $k_{\mathbf{c}}$ clusters in order to maintain the adjacency condition. If the component mean μ_{j^*} of the new cluster is so that:
   (a) $\mu_{j^*} = \min_{1\le j\le k_{\mathbf{c}}} \mu_j$, then do $j^* = 1$ and relabel all other clusters doing $j + 1$;
   (b) $\mu_{j^*} = \max_{1\le j\le k_{\mathbf{c}}} \mu_j$, then do $j^* = k_{\mathbf{c}}$ and keep all other clusters labels;
   (c) $\mu_j < \mu_{j^*} < \mu_{j+1}$, for $j \ne \{1, k_{\mathbf{c}}\}$, then do $j^* = j + 1$ and relabel all other clusters $j' \ge j + 1$ doing $j' = j' + 1$.

At this point, the estimation procedure could be done by iterating between Algorithms 1 and 2. But, Algorithm 2 may be inefficient in situations where clusters have near means. This happens because the algorithm updates only one latent indicator variable at a time. Consequently, we may have a poor exploration of observation clusters and the algorithm may be trapped in local modes. This is the reason why we insert a split-merge step within the MCMC algorithm in order to increase the mixing on $k_{\mathbf{c}}$.

3.2 Split-merge movements

Let $\boldsymbol{\Phi} = (\boldsymbol{\theta}, \mathbf{c}, k_{\mathbf{c}})$ be the current state of the MCMC algorithm and $\boldsymbol{\Phi}^{\mathrm{sp}} = (\boldsymbol{\theta}^{\mathrm{sp}}, \mathbf{c}^{\mathrm{sp}}, k_{\mathbf{c}}^{\mathrm{sp}})$ and $\boldsymbol{\Phi}^{\mathrm{me}} = (\boldsymbol{\theta}^{\mathrm{me}}, \mathbf{c}^{\mathrm{me}}, k_{\mathbf{c}}^{\mathrm{me}})$ be the proposal states obtained by split and merge movements, respectively.

As dimension of the parametric space of $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ is fixed, the acceptance probability for a split or a merge movement is given by the Metropolis–Hastings acceptance probability (Chib and Greenberg (1995)), *i.e.*, $\Psi[\boldsymbol{\Phi}^*|\boldsymbol{\Phi}] = \min(1, A^*)$, where

$$A^* = \frac{P(\mathbf{y}|\mathbf{c}^*, \boldsymbol{\theta}^*, k)}{P(\mathbf{y}|\mathbf{c}, \boldsymbol{\theta}, k)} \frac{\pi(\mathbf{c}^*|\gamma, k)}{\pi(\mathbf{c}|\gamma, k)} \frac{\pi(\boldsymbol{\theta}^*|\eta, k)}{\pi(\boldsymbol{\theta}^*|\eta, k)} \frac{q[\boldsymbol{\Phi}|\boldsymbol{\Phi}^*]}{q[\boldsymbol{\Phi}^*|\boldsymbol{\Phi}]}, \tag{3.12}$$

where "$*$" means either a split or a merge, $q[\cdot]$ is the transition proposal which is obtained by a split or a merge depending on the type of proposal, $P(\mathbf{y}|\mathbf{c}, \boldsymbol{\theta}, k)\pi(\mathbf{c}|\gamma, k)$ is the complete-data likelihood function which is equal to the sampling distribution given in (3.4), regarded as a function of the unknown parameters $\boldsymbol{\theta}$, and $\pi(\boldsymbol{\theta}|k) = \prod_{j=1}^{k} \pi_G(\theta_j|\eta_j)$ is the joint prior distribution for $\boldsymbol{\theta}$.

3.2.1 *Split movement.* Let $P_{\mathrm{sp}|k_{\mathbf{c}}}$ and $P_{\mathrm{me}|k_{\mathbf{c}}}$ be the probabilities of proposing a split and a merge, respectively, with $P_{\mathrm{sp}|k_{\mathbf{c}}} + P_{\mathrm{me}|k_{\mathbf{c}}} = 1$. These probabilities depend on $k_{\mathbf{c}}$ because if $k_{\mathbf{c}} = 1$, we can propose only a split, $P_{\mathrm{sp}|k_{\mathbf{c}}} = 1$; in the other hand, if $k_{\mathbf{c}} = k$ we can propose only a merge, $P_{\mathrm{me}|k_{\mathbf{c}}} = 1$. For $2 \le k_{\mathbf{c}} \le (k - 1)$ we use $P_{\mathrm{sp}|k_{\mathbf{c}}} = P_{\mathrm{me}|k_{\mathbf{c}}} = 1/2$.

Provided we choose a split, consider $\mathbb{C}_2$ be the number of clusters with $n_j \ge 2$. Select a component D_j, with $n_j > 2$, with probability $P_{j|\mathbb{C}_2} = \frac{1}{\mathbb{C}_2}$ and propose a split of observations $y_i \in D_j$ in two new sets D_{j_1} and D_{j_2} as follows:

(i) Let y_{h_1} and y_{h_2} be the minimum and the maximum value of the set D_j, respectively;
(ii) Do $D_{j_1} = \{y_{h1}\}$, $D_{j_2} = \{y_{h_2}\}$ and $n_{j_1} = n_{j_2} = 1$;
(iii) For $i = 1, \dots, n$ do the following:
   (a) if $c_i = j$ and $y_i \notin \{y_{h_1}, y_{h_2}\}$, then allocate y_i in D_{j_1} with probability

$$P_{j_1}(y_i) = \frac{n_{j_1} \int f(y_i|\theta_{j_1})\pi(\theta_{j_1}|D_{j_1})\, d\theta_{j_1}}{n_{j_1} \int f(y_i|\theta_{j_1})\pi(\theta_{j_1}|D_{j_1})\, d\theta_{j_1} + n_{j_2} \int f(y_i|\theta_{j_2})\pi(\theta_{j_2}|D_{j_2})\, d\theta_{j_2}},$$

   where $\pi(\theta_m|D_m)$ is the posterior distribution for θ_m given D_m, for $m = j_1, j_2$;

(b) Generate an indicator variable $\mathbb{I}_i \sim \text{Bernoulli}(P_{j_1}(y_i))$. If $\mathbb{I}_i = 1$, $y_i \in D_{j_1}$. Then, do $D_{j_1} = \{D_{j_1}\} \cup \{y_i\}$, $n_{j_1} = n_{j_1} + 1$ and $c_i^{\text{sp}} = j_1$. Otherwise, do $D_{j_2} = \{D_{j_2}\} \cup \{y_i\}$, $n_{j_2} = n_{j_2} + 1$ and $c_i^{\text{sp}} = j_2$;

(c) If $c_i = j$ and $y_i = y_{h_1}$, then do $P_{j_1}(y_i) = 1$ and $c_i^{\text{sp}} = j_1$. If $c_i = j$ and $y_i = y_{h_2}$, then do $P_{j_1}(y_i) = 0$ and $c_i^{\text{sp}} = j_2$;

We fix the new labels as $j_1 = j$, $j_2 = j + 1$ and for all other labels $j' > j$ we do $j' = j' + 1$, for $j \in \{1, \dots, k\}$. Thus, we have a new configuration $\mathbf{c}^{\text{sp}} = (c_1^{\text{sp}}, \dots, c_n^{\text{sp}})$ with $k_{\mathbf{c}}^{\text{sp}} = k_{\mathbf{c}} + 1$ clusters where

(i) for all $c_i \in \mathbf{c}$, so that, $c_i = j$, do $c_i^{\text{sp}} = j_1$ or $c_i^{\text{sp}} = j_2$ according to step (iii) above;
(ii) for all $c_i \in \mathbf{c}$, so that $c_i = j'$ for $j' < j$, do $c_i^{\text{sp}} = c_i$;
(iii) for all $c_i \in \mathbf{c}$, so that $c_i = j'$ for $j' > j$, do $c_i^{\text{sp}} = c_i + 1$;

for $i = 1, \dots, n$, $j \in \{1, \dots, k_{\mathbf{c}}\}$ and $j' \in \{1, \dots, k\} \setminus \{j_1, j_2\}$.

The probability of configuration D_{j_1} and D_{j_2} is

$$P_{\text{alloc}} = \prod_{y_i \in D_{j_1}} P_{j_1}(y_i) \prod_{y_i \in D_{j_2}} (1 - P_{j_1}(y_i)).$$

Conditional on D_{j_1} and D_{j_2}, generate candidate-values $\theta_{j_1}^{\text{sp}}$ and $\theta_{j_2}^{\text{sp}}$ for parameters θ_{j_1} and θ_{j_2} from posterior distributions $\pi(\theta_{j_1}|D_{j_1})$ and $\pi(\theta_{j_2}|D_{j_2})$, respectively. Thus, we have a new vector of parameters $\boldsymbol{\theta}^{\text{sp}} = (\theta_1^{\text{sp}}, \dots, \theta_{k_{\mathbf{c}}}^{\text{sp}}, \theta_{k_{\mathbf{c}}+1}^{\text{sp}}, \dots, \theta_k^{\text{sp}})$, where

(i) for all $\theta_{j'}^{\text{sp}} \in \boldsymbol{\theta}^{\text{sp}}$, so that $j' < j_1$, do $\theta_{j'}^{\text{sp}} = \theta_{j'}$;
(ii) for all $\theta_{j'}^{\text{sp}} \in \boldsymbol{\theta}^{\text{sp}}$, so that $j' > j_2$, do $\theta_{j'}^{\text{sp}} = \theta_{j'-1}$;

for $\theta_{j'} \in \boldsymbol{\theta}$ and $j' \in \{1, \dots, k\} \setminus \{j_1, j_2\}$.

Now we must check if the adjacency condition is met in the split proposal, *that is*, if component means for clusters are in increasing numerical order, $\mu_{j_1-1}^{\text{sp}} < \mu_{j_1}^{\text{sp}} < \mu_{j_2}^{\text{sp}} < \mu_{j_2+1}^{\text{sp}}$. In the case where it is not, the proposal is rejected because the movement may not be reversible by the merge proposal.

If the adjacency condition is met, we have a new configuration $\mathbf{c}^{\text{sp}}$ and a new set of parameters $\boldsymbol{\theta}^{\text{sp}}$ with $k_{\mathbf{c}}^{\text{sp}} = k_{\mathbf{c}} + 1$ clusters. This transition proposal is denoted by $\boldsymbol{\Phi}^{\text{sp}}|\boldsymbol{\Phi}$ and its probability is given by

$$q[\boldsymbol{\Phi}^{\text{sp}}|\boldsymbol{\Phi}] = P_{\text{sp}|k_{\mathbf{c}}} P_{j|\mathbb{C}_2} P_{\text{alloc}} \pi(\theta_{j_1}|D_{j_1}) \pi(\theta_{j_2}|D_{j_2}), \tag{3.13}$$

where $\pi(\theta_m|D_m)$ is the posterior density for θ_m, for $m = j_1, j_2$.

3.2.2 *Merge movement.* We deal with merge a reverse of split movement. This movement is initialized choosing sets D_{j_1} and D_{j_2}.

We establish a criterion which merges clusters adjacent in relation to the current values of their means. This is due to the adjacency condition assumed in the split movement. Thus, the probability of selecting D_{j_1} and D_{j_2} for a merge is

$$P_{j_1,j_2} = P_{j_1} P_{j_2|j_1} + P_{j_2} P_{j_1|j_2} = \begin{cases} 1, & \text{if } k_{\mathbf{c}} = 2; \\ \dfrac{3}{2k_{\mathbf{c}}}, & \text{if } k_{\mathbf{c}} > 2 \text{ and } j_1 = 1 \text{ or } j_2 = k_{\mathbf{c}}; \\ \dfrac{1}{k_{\mathbf{c}}}, & \text{if } k_{\mathbf{c}} > 2 \text{ and } j_1 \neq 1 \text{ or } j_2 \neq k_{\mathbf{c}}; \end{cases}$$

where P_{b_1} is the probability of choosing cluster b_1 and $P_{b_2|b_1}$ is the conditional probability of choosing cluster b_2 given the previous choice of b_1.

After selecting clusters D_{j_1} and D_{j_2}, we join them in a single cluster D_j, *i.e.*, we do $D_j = \{D_{j_1}\} \cup \{D_{j_2}\}$. We fix the new labels as $j = j_1$ and for all other labels j' with $j' \geq j_2$ we do $j' = j' - 1$. Thus, we have a new configuration $\mathbf{c}^{\text{me}} = (c_1^{\text{me}}, \ldots, c_n^{\text{me}})$ with $k_{\mathbf{c}}^{\text{me}} = k_{\mathbf{c}} - 1$ clusters, where

(a) for all $c_i \in \mathbf{c}$, so that $c_i = j'$ and $j' \leq j_1$, do $c_i^{\text{mg}} = c_i$;
(b) for all $c_i \in \mathbf{c}$, so that $c_i = j'$ and $j' \geq j_2$, do $c_i^{\text{mg}} = c_i - 1$;

for $i = 1, \ldots, n$, $j' = 1, \ldots, k_{\mathbf{c}}$ and $j \in \{1, \ldots, k_{\mathbf{c}} - 1\}$.

Conditional on D_j, generate the candidate-value θ_j^{mg} for parameter θ_j from posterior distribution $\pi(\theta_j|D_j)$. This determine a new vector of parameters $\boldsymbol{\theta}^{\text{me}} = (\theta_1^{\text{me}}, \ldots, \theta_{k_{\mathbf{c}-1}}^{\text{me}}, \theta_{k_{\mathbf{c}}}^{\text{me}}, \ldots, \theta_k^{\text{me}})$, where

(i) for all $\theta_{j'}^{\text{me}} \in \boldsymbol{\theta}^{\text{me}}$, so that $j' < j_1$, do $\theta_{j'}^{\text{me}} = \theta_{j'}$;
(ii) for all $\theta_{j'}^{\text{me}} \in \boldsymbol{\theta}^{\text{me}}$, so that $j' \geq j_2$, do $\theta_{j'}^{\text{me}} = \theta_{j'+1}$, for $\theta_{j'} \in \boldsymbol{\theta}$ and $j' \in \{1, \ldots, k-1\} \setminus \{j_1, j_2\}$;
(iii) in order to complete $\boldsymbol{\theta}^{\text{me}}$, generate θ_k^{me} from prior distribution, $\theta_k \sim \pi_G(\theta_k)$.

Here we also must check if the adjacency condition is met, *i.e.*, $\mu_{j-1}^{\text{me}} < \mu_j^{\text{me}} < \mu_{j+1}^{\text{me}}$. In the case where it is not, the proposal is rejected because the movement may not be reversible by the split proposal.

If the adjacency condition is met, the merge proposal determine the new configuration $\mathbf{c}^{\text{me}}$ and a new vector of parameters $\boldsymbol{\theta}^{\text{me}}$ with $k_{\mathbf{c}}^{\text{me}} = k_{\mathbf{c}} - 1$ clusters. This transition proposal is denoted by $\boldsymbol{\Phi}^{\text{me}}|\boldsymbol{\Phi}$ and its probability is given by

$$q\big[\boldsymbol{\Phi}^{\text{me}}|\boldsymbol{\Phi}\big] = P_{\text{me}|k_{\mathbf{c}}} P_{j_1,j_2} \pi(\theta_j|D_j)\pi_G(\theta_k|\eta_k). \tag{3.14}$$

Defined the transition probabilities for split and merge movements, it is important to note that, given the current state $\boldsymbol{\Phi}$, the probability of proposing a split of the cluster D_j in D_{j_1} and D_{j_2} (*i.e.*, the split state $\boldsymbol{\Phi}^{\text{me}}$) is equivalent to being in the state with D_{j_1} and D_{j_2} merged in D_j (*i.e.*, the merge state $\boldsymbol{\Phi}^{\text{me}}$) and proposing the move back to the current state $\boldsymbol{\Phi}$. In terms of transition probability this means that

$$q\big[\boldsymbol{\Phi}^{\text{sp}}|\boldsymbol{\Phi}\big] = q\big[\boldsymbol{\Phi}|\boldsymbol{\Phi}^{\text{me}}\big] = P_{\text{sp}|k_{\mathbf{c}}^{\text{me}}} P_{j|\mathbb{C}_2^{\text{me}}} P_{\text{alloc}} \pi(\theta_{j_1}|D_{j_1})\pi(\theta_{j_2}|D_{j_2}). \tag{3.15}$$

Besides, it is also important to note that, in the merge proposal there is only one way to merge the observations from two components in one component. Thus, we need to calculate the corresponding probability, P_{alloc}, of generating the current split state from the proposed merge state. This is done in the same way as the split proposal, but now considering the known of the current split state. Analogously to (3.15), we have that

$$q\big[\boldsymbol{\Phi}^{\text{me}}|\boldsymbol{\Phi}\big] = q\big[\boldsymbol{\Phi}|\boldsymbol{\Phi}^{\text{sp}}\big] = P_{\text{me}|k_{\mathbf{c}}^{\text{sp}}} P_{j_1,j_2} \pi(\theta_j|D_j)\pi_G(\theta_k|\eta_k). \tag{3.16}$$

3.2.3 *Acceptance probability.* From equation (3.12), the acceptance probability for a split movement is $\Psi[\boldsymbol{\Phi}^{\text{sp}}|\boldsymbol{\Phi}] = \min(1, A^{\text{sp}})$, where

$$A^{\text{sp}} = \frac{P(\mathbf{y}|\mathbf{c}^{\text{sp}}, \boldsymbol{\theta}^{\text{sp}}, k)}{P(\mathbf{y}|\mathbf{c}, \boldsymbol{\theta}, k)} \frac{\pi(\mathbf{c}^{\text{sp}}|\gamma, k)}{\pi(\mathbf{c}|\gamma, k)} \frac{\pi(\boldsymbol{\theta}^{\text{sp}}|k)}{\pi(\boldsymbol{\theta}|k)} \frac{q[\boldsymbol{\Phi}|\boldsymbol{\Phi}^{\text{sp}}]}{q[\boldsymbol{\Phi}^{\text{sp}}|\boldsymbol{\Phi}]}.$$

The ratio of the joint probabilities of $\mathbf{y}$ conditional on latent indicator variables and component parameters is given by

$$\frac{P(\mathbf{y}|\mathbf{c}^{\text{sp}}, \boldsymbol{\theta}^{\text{sp}}, k)}{P(\mathbf{y}|\mathbf{c}, \boldsymbol{\theta}, k)} = \frac{L(\theta_{j_1}^{\text{sp}}|D_{j_1})L(\theta_{j_1}^{\text{sp}}|D_{j_1})}{L(\theta_{j_1}^{\text{sp}}|D_{j_1})}, \tag{3.17}$$

where $L(\theta_m|D_m)$ is given in (3.9), for $m = j, j_1, j_2$.

From (3.3), the ratio of the joint prior probability for latent indicator variables is

$$\frac{\pi(\mathbf{c}^{\text{sp}}|\gamma,k)}{\pi(\mathbf{c}|\gamma,k)} = \frac{\Gamma(n_{j_1}+\gamma)\Gamma(n_{j_2}+\gamma)}{\Gamma(n_j+\gamma)\Gamma(\gamma)}. \tag{3.18}$$

The prior distributions ratio for component parameters is

$$\frac{\pi(\boldsymbol{\theta}^{\text{sp}}|k)}{\pi(\boldsymbol{\theta}|k)} = \frac{\prod_{j=1}^{k}\pi_G(\theta_j^{\text{sp}}|\eta_j)}{\prod_{j=1}^{k}\pi_G(\theta_j|\eta_j)} = \frac{\pi_G(\theta_{j1}^{\text{sp}}|\eta_{j1})\pi_G(\theta_{j2}^{\text{sp}}|\eta_{j2})}{\pi_G(\theta_j|\eta_j)\pi_G(\theta_k|\eta_k)}. \tag{3.19}$$

From (3.13) and (3.16), the transition probability ratio for the split proposal is

$$\frac{q[\boldsymbol{\Phi}|\boldsymbol{\Phi}^{\text{sp}}]}{q[\boldsymbol{\Phi}^{\text{sp}}|\boldsymbol{\Phi}]} = \frac{P_{\text{me}|k_{\mathbf{c}}^{\text{sp}}}}{P_{\text{sp}|k_{\mathbf{c}}}}\frac{P_{j_1,j_2}}{P_{j|\mathbb{C}_2}}\frac{1}{P_{\text{alloc}}}\frac{\pi(\theta_j|D_j)\pi_G(\theta_k|\eta_k)}{\pi(\theta_{j1}^{\text{sp}}|D_{j1})\pi(\theta_{j2}^{\text{sp}}|D_{j2})} = \frac{Q^{\text{sp}}P^r}{P_{\text{alloc}}}, \tag{3.20}$$

where

$$Q^{\text{sp}} = \frac{P_{\text{me}|k_{\mathbf{c}}^{\text{sp}}}}{P_{\text{sp}|k_{\mathbf{c}}}}\frac{P_{j_1,j_2}}{P_{j|\mathbb{C}_2}} = \begin{cases} \dfrac{1}{2}, & \text{if } k_{\mathbf{c}} = 1; \\ \left(\dfrac{1}{2}\right)^{1-\mathbb{I}_{\mathbf{c}}(k-1)}\dfrac{3\mathbb{C}_2}{k_{\mathbf{c}+1}}, & \text{if } k_{\mathbf{c}} \in \mathbb{K}_1; \\ 2^{\mathbb{I}_{\mathbf{c}}(k-1)}\dfrac{\mathbb{C}_2}{k_{\mathbf{c}}+1}, & \text{if } k_{\mathbf{c}} \in \mathbb{K}_2; \end{cases}$$

$\mathbb{I}_{\mathbf{c}}(k-1) = 1$ if $k_{\mathbf{c}} = k-1$ and $\mathbb{I}_{\mathbf{c}}(k-1) = 0$ otherwise, $\mathbb{K}_1 = \{2 \le k_{\mathbf{c}} \le k-1;$ and $j_1 = 1$ or $j_2 = k_{\mathbf{c}}\}$, $\mathbb{K}_2 = \{2 \le k_{\mathbf{c}} \le k-1$ and $j_1 \neq 1$ or $j_2 \neq k_{\mathbf{c}}\}$ and P^r is the posterior densities ratio, given by

$$\begin{aligned} &\frac{\pi(\theta_j|D_j)\pi_G(\theta_k)}{\pi(\theta_{j_1}|D_{j_1})\pi(\theta_{j_2}|D_{j_2})} \\ &\quad = \frac{L(\theta_j|D_j)}{L(\theta_{j_1}|D_{j_1})L(\theta_{j_2}|D_{j_2})}\frac{\pi_G(\theta_j|\eta_j)\pi_G(\theta_k|\eta_k)}{\pi_G(\theta_{j_1}|\eta_{j_1})\pi_G(\theta_{j_2}|\eta_{j_2})}\frac{\mathbf{I}(D_{j_1})\mathbf{I}(D_{j_!})}{\mathbf{I}(D_j)}, \end{aligned}$$

where $\mathbf{I}(D_m) = \int L(\theta_m|D_m)\pi_G(\theta_m|\eta_m)\,d\theta_m$ is the normalizing constant, for $m = j, j_1, j_2$.

Multiplying (3.17), (3.18), (3.19) and (3.20), a split is accepted with probability $\Psi[\boldsymbol{\Phi}^{\text{sp}}|\boldsymbol{\phi}] = \min(1, A^{\text{sp}})$ for

$$A^{\text{sp}} = \frac{\mathbf{I}(D_{j_1})\mathbf{I}(D_{j_2})}{\mathbf{I}(D_j)}\frac{\Gamma(n_{j_1}+\gamma)\Gamma(n_{j_2}+\gamma)}{\Gamma(n_j+\gamma)}\frac{Q^{\text{sp}}}{P_{\text{alloc}}}.$$

Similarly, the acceptance probability for a merge is $\Psi[\boldsymbol{\Phi}^{\text{me}}|\boldsymbol{\Phi}] = \min(1, A^{\text{me}}) = \frac{1}{A^{\text{sp}}}$.

3.2.4 *Split-merge allocation sampler.* Now the split-merge procedure is described as an algorithm denominated by Split-Merge allocation sampler (SMAS).

SMAS Algorithm. Initialize with a configuration $\mathbf{c}^{(0)}$. For lth iteration of the algorithm do:

(i) Update the component parameters $\boldsymbol{\theta}$ using Algorithm 1;
(ii) Update the indicator variables $\mathbf{c}$ using Algorithm 2;
(iii) Choose between split or merge with probabilities $P_{\text{sp}|k_{\mathbf{c}}}$ and $P_{\text{me}|k_{\mathbf{c}}}$;
(iv) Accept the proposal with probability $\Psi[\boldsymbol{\Phi}^*|\boldsymbol{\Phi}]$, where "$*$" is either a sp or a me;
  (a) If a split proposal is accepted, do $k_{\mathbf{c}}^{(l)} = k_{\mathbf{c}}^{(l-1)} + 1$;
  (b) If a merge proposal is accepted, do $k_{\mathbf{c}}^{(l)} = k_{\mathbf{c}}^{(l-1)} - 1$;
  (c) Otherwise, maintain $k_{\mathbf{c}}^{(l)} = k_{\mathbf{c}}^{(l-1)}$;

In order to estimate the number of clusters, we discard the first B iterations as a burn in and calculate $N(k_{\mathbf{c}} = j)$, the number of times that $k_{\mathbf{c}} = j$ in the $L - B$ iterations, for $j \in \{1, \ldots, k\}$. Let $\tilde{P}(k_{\mathbf{c}} = j) = N(k_{\mathbf{c}} = j)/(L - B)$ be the estimated posterior probability for $k_{\mathbf{c}} = j$, for $j = 1, \ldots, k$. Thus, $\tilde{k}_{\mathbf{c}} = \operatorname{argmax}_{1 \leq j \leq k}(\tilde{P}(k_{\mathbf{c}} = j))$ is the estimate for the number of clusters.

Conditional on $\tilde{k}_{\mathbf{c}}$, consider

(i) $L_{\tilde{k}_{\mathbf{c}}} = \sum_{l=B+1}^{L} \mathbb{I}_{k_{\mathbf{c}}^{(l)}}(\tilde{k}_{\mathbf{c}})$ be the number of iterations for which $k_{\mathbf{c}} = \tilde{k}_{\mathbf{c}}$ in $L - B$ iterations, where $\mathbb{I}_{k_{\mathbf{c}}^{(l)}} = 1$ if in l-iteration $k_{\mathbf{c}}^{(l)} = \tilde{k}_{\mathbf{c}}$ and $\mathbb{I}_{k_{\mathbf{c}}^{(l)}} = 0$ otherwise;

(ii) $N_{ij} = \sum_{l=B+1}^{L} \mathbb{I}_{c_i^{(l)}}(j)\mathbb{I}_{k_{\mathbf{c}}^{(l)}}$ be the number of times that observation y_i is allocated in component j in $L_{\tilde{k}_{\mathbf{c}}}$ iterations, where $\mathbb{I}_{c_i^{(l)}}(j) = 1$ if in lth iteration $c_i = j$ and $\mathbb{I}_{c_i^{(l)}}(j) = 0$ otherwise, for $i = 1, \ldots, n$ and $j = 1, \ldots, k$.

Let $\tilde{P}(c_i = j) = \frac{N_{ij}}{L_{\tilde{k}_{\mathbf{c}}}}$ be the posterior probability that the observation y_i belongs to cluster j, for $j = 1, \ldots, \tilde{k}_{\mathbf{c}}$. If $\tilde{P}(c_i = j) = \max_{1 \leq j' \leq \tilde{k}_{\mathbf{c}}}(\tilde{P}(c_i = j'))$, then we consider that y_i belongs to cluster j, for $i = 1, \ldots, n$ and $j = 1, \ldots, \tilde{k}_{\mathbf{c}}$.

We estimate parameters θ_j, $j = 1, \ldots, \tilde{k}_{\mathbf{c}}$, considering the average of the generated values, *that is*,

$$\tilde{\theta}_j | \tilde{k}_{\mathbf{c}} = \frac{1}{L_{\tilde{k}_{\mathbf{c}}}} \sum_{l=B+1}^{L} \theta_j^{(l)} \mathbb{I}_{k_{\mathbf{c}}^{(l)}}(\tilde{k}_{\mathbf{c}}).$$

4 Data analysis

We illustrate the performance of the proposed method by using simulated data and three real data sets. Following Richardson and Green (1997), we model these datasets by assuming an univariate normal mixture model. From model (2.2), $f(y_i|\theta_j)$ is the density of a normal distribution with mean μ_j and variance σ_j^2 and $\theta_j = (\mu_j, \sigma_j^2)$, for $j = 1, \ldots, k$.

In order to explore the fully conjugation, we consider the following prior distributions for component parameters $\theta_j = (\mu_j, \sigma_j^2)$,

$$\mu_j | \sigma_j^2, \mu_0, \lambda \sim \mathcal{N}\left(\mu_0, \frac{\sigma_j^2}{\lambda}\right) \quad \text{and} \quad \sigma_j^{-2} | \alpha, \beta \sim \Gamma(\alpha, \beta)$$

where μ_0, λ, α and β are hyperparameters, $\mathcal{N}(\mu_0, \frac{\sigma_j^2}{\lambda})$ represents the normal distribution with mean μ_0 and variance $\frac{\sigma_j^2}{\lambda}$ and $\Gamma(\alpha, \beta)$ represents the Gamma distribution with location parameter α and scale parameter β. The parametrization of the Gamma distribution is so that the mean is α/β and the variance is α/β^2. These prior distributions are also used by Casella, Robert and Wells (2000), Nobile and Fearnside (2007) and Saraiva et al. (2016).

Since it may be unrealistic to assume the availability of strong prior information regarding component parameters θ_j in practice, we specify the hyperparameters values according to guidelines of Richardson and Green (1997) and Saraiva et al. (2016). Thus, we set $\mu_0 = \varepsilon$, $\alpha = 2$ and $\beta = 0.2/10R^2$, where ε is the midpoint of the observed variation interval of the data and R is the length of this interval. We also fix the hyperparameter $\lambda = 10^{-2}$ in order to get a prior distribution for component means with large variance. The normalizing constant $\mathbf{I}(D_j)$ present in the acceptance probability for split-merge is presented in Section S-1 of the SM.

4.1 Simulated data sets

Consider the population model given in (2.2) with $k = 10$ components. To simulate the data sets, we set up the number of clusters $k_{\mathbf{c}}$ and the component parameters for the $k_{\mathbf{c}}$ clusters according to the specified in Table 2.

In the set up A_1 the two clusters have equal variances and weights while A_2 has different variances and weights. In A_3 the three clusters have equal variances and similar weight and A_4 has different variances and weights. In A_5 we consider four clusters with same weights values and two clusters with higher variances in relation to the other two, and in A_6 we consider five clusters with different weights and one cluster with large variance in relation to the others.

The procedure for generating the data sets is

(i) For $i = 1, \ldots, n$, generate $U_i \sim \mathcal{U}(0, 1)$; if $\sum_{j'=1}^{j-1} w_j < u_i \leq \sum_{j'=1}^{j} w_j$, generate $Y_i \sim \mathcal{N}(\mu_j, \sigma_j^2)$, with fixed parameter values according to Table 2, for $w_0 = 0$ and $j = 1, \ldots, k_{\mathbf{c}}$.

(ii) In order to record from which component each observation is generated from we define $G = (G_1, \ldots, G_n)$ such that $G_i = j$ if $Y_i \sim \mathcal{N}(\mu_j, \sigma_j^2)$, for $i = 1, \ldots, n$ and $j = 1, \ldots, k_{\mathbf{c}}$.

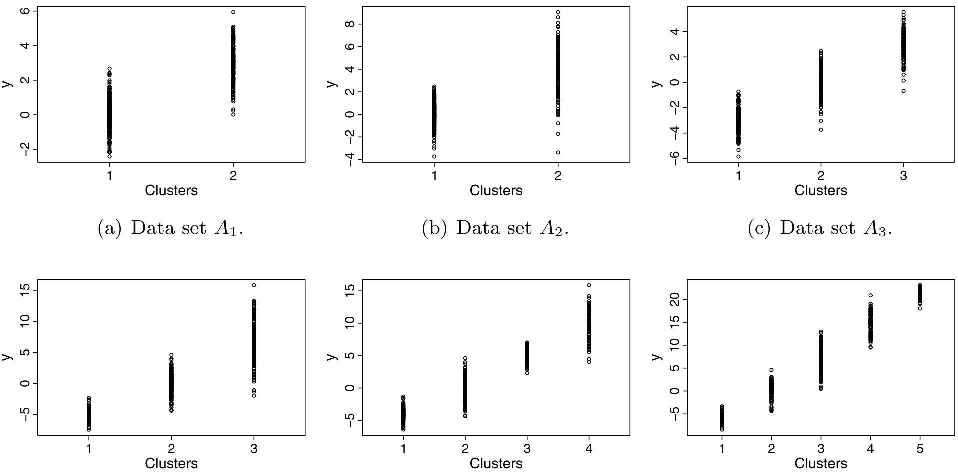
In the remainder of the paper, we simulate datasets with sizes $n = 500$; samples with size $n = 1000$ also are simulated and results are presented in Section S-3 of the SM. Figure 1 shows the values generated by cluster for datasets A_1 to A_6 for $n = 500$.

We apply the proposed SMAS algorithm fixing $L = 110{,}000$ iterations and a burn in of $B = 10{,}000$. We also consider a sample of one draw for every 20 obtaining a sequence of 5000 cases. These values were enough for reliable results.

For all values $\gamma = r$ for $r \in R_1$, the SMAS poses maximum posterior probability on the $k_{\mathbf{c}}$ true value. These results are presented in Section S-2 of the SM. For $\gamma = r$ for $r \in R_2$

Table 2 *Number of clusters and parameter values used for simulating the datasets*

Artificial data set	Number of clusters	Parameter values				
A_1	$k_{\text{true}} = 2$	$\mu_1 = 0,$ $\sigma_1 = 1,$ $w_1 = 0.50,$	$\mu_2 = 3,$ $\sigma_2 = 1,$ $w_2 = 0.50,$			
A_2	$k_{\text{true}} = 2$	$\mu_1 = 0,$ $\sigma_1 = 1,$ $w_1 = 0.70,$	$\mu_2 = 4,$ $\sigma_2 = 2,$ $w_2 = 0.30,$			
A_3	$k_{\text{true}} = 3$	$\mu_1 = -3,$ $\sigma_1 = 1,$ $w_1 = 0.30,$	$\mu_2 = 0,$ $\sigma_2 = 1,$ $w_2 = 0.40,$	$\mu_3 = 3,$ $\sigma_3 = 1,$ $w_3 = 0.30,$		
A_4	$k_{\text{true}} = 3$	$\mu_1 = -5,$ $\sigma_1 = 1,$ $w_1 = 0.20,$	$\mu_2 = 0,$ $\sigma_2 = 2,$ $w_2 = 0.30,$	$\mu_3 = 7,$ $\sigma_3 = 3,$ $w_3 = 0.50,$		
A_5	$k_{\text{true}} = 4$	$\mu_1 = -4,$ $\sigma_1 = 1,$ $w_1 = 0.25,$	$\mu_2 = 0,$ $\sigma_2 = 2,$ $w_2 = 0.25,$	$\mu_3 = 5,$ $\sigma_3 = 1,$ $w_3 = 0.25,$	$\mu_4 = 10,$ $\sigma_4 = 2,$ $w_4 = 0.25,$	
A_6	$k_{\text{true}} = 5$	$\mu_1 = -6,$ $\sigma_1 = 1,$ $w_1 = 0.15,$	$\mu_2 = 0,$ $\sigma_2 = 2,$ $w_2 = 0.20,$	$\mu_3 = 7,$ $\sigma_3 = 3,$ $w_3 = 0.30,$	$\mu_4 = 15,$ $\sigma_4 = 2,$ $w_4 = 0.20,$	$\mu_5 = 21,$ $\sigma_5 = 1,$ $w_5 = 0.15$



(a) Data set A_1. (b) Data set A_2. (c) Data set A_3.

(d) Data set A_4. (e) Data set A_5. (f) Data set A_6.

Figure 1 *Values generated by cluster for datasets A_1 to A_6 with $n = 500$.*

the SMAS tends to overestimate the number of clusters when we increase the r value. These results are presented by Figure 2 which show the posterior probability for $k_{\mathbf{c}}$ true value and the $k_{\mathbf{c}}$ values with maximum posterior probability for each $\gamma = r \in R_2$. As one can note, for large values of γ the method overestimate the number of cluster. For datasets A_1, A_2 and A_4, the maximum posterior on $k_{\mathbf{c}}$ true value is obtained only for $\gamma \in \{0.1, 0.2\}$; while the maximum posterior on $k_{\mathbf{c}}$ true value for datasets A_3 and A_5 is obtained for $\gamma \in \{0.1, 0.2, 0.3\}$ and for dataset A_6 for $\gamma \in \{0.1, 0.2, 0.3, 0.4\}$. Besides, for all simulated cases, the posterior probability on $k_{\mathbf{c}}$ true value goes to zero when γ approaches to 2.

Motivated by the sensibility to the choice of the γ we develop a second approach considering for γ a Gamma prior distribution, $\gamma \sim \Gamma(a, b)$, $a, b > 0$.

In this approach, rather than the value of γ it is the choice of the hyperparameters a and b that are influential on the posterior probability for $k_{\mathbf{c}}$. To verify the sensibility in relation to the choice of the hyperparameters values, we consider:

(i) $a = b = 1$, obtaining $E(\gamma) = 1$ and $\mathrm{Var}(\gamma) = 1$. This prior represents the belief that the weights have Uniform distribution over the simplex;

(ii) $a = 2$ and $b = 4$, obtaining $E(\gamma) = 0.5$ and $\mathrm{Var}(\gamma) = 0.125$. This prior was suggested by Escobar and West (1995) in the context of the Dirichlet process mixture model.

In order to simulate values from conditional posterior distribution of γ, $\pi(\gamma|\mathbf{c}, \mathbf{y}, \boldsymbol{\theta}, k)$, we implement a random walk Metropolis (RWM) algorithm. We consider the candidate-generating density $q[\gamma^*|\gamma]$, where γ^* is a candidate value, with Uniform distribution centered on the current value of γ, *i.e.*, $\mathcal{U}(\gamma - \epsilon, \gamma + \epsilon)$. We set up $\epsilon = 0.05$.

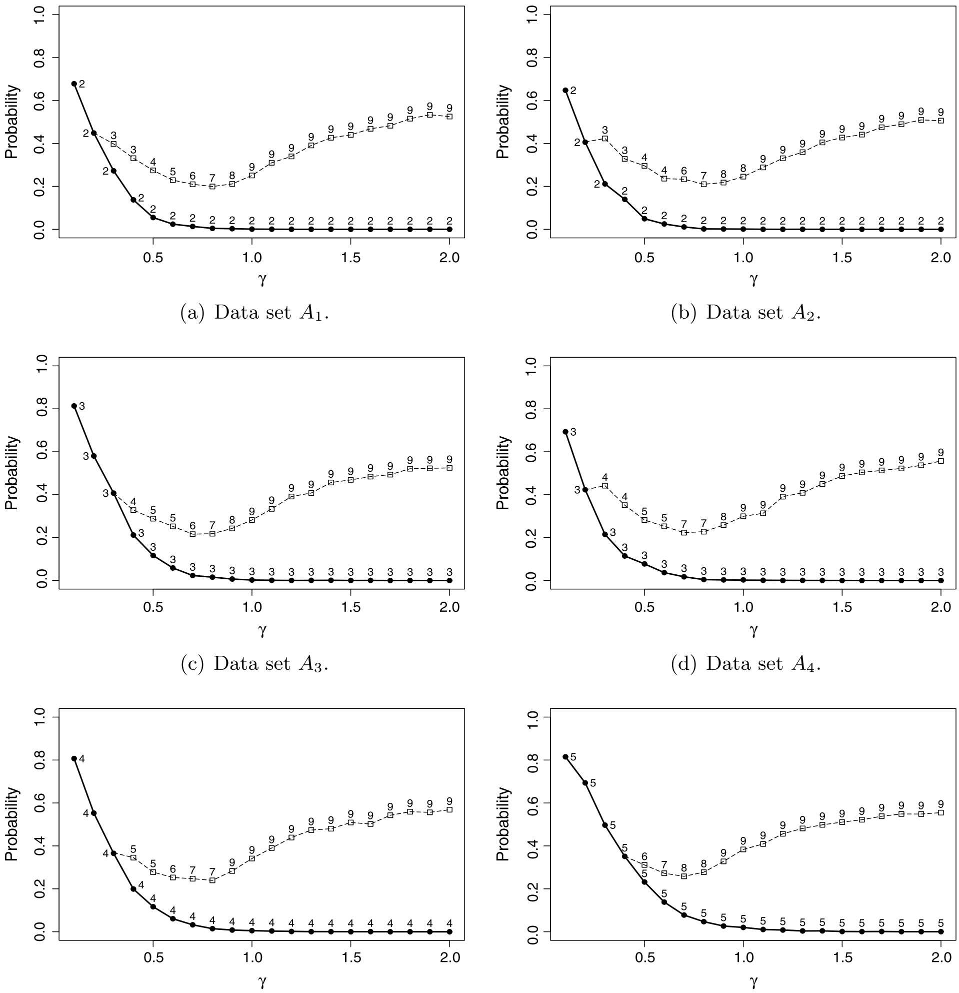
The candidate value γ^* is accepted with probability $\Psi(\gamma^*|\gamma) = \min(1, A_\gamma)$, where $A_\gamma = \frac{\pi(\mathbf{c}|\gamma^*,k)}{\pi(\mathbf{c}|\gamma,k)} \frac{\pi(\gamma^*)}{\pi(\gamma)}$, since the proposal kernels from numerator and denominator cancel, $\pi(\mathbf{c}|\gamma, k)$ is given in (3.3) and $\pi(\gamma)$ is the density of the Gamma prior for γ.

This algorithm is implemented as follows.

RWM Algorithm. Let the state of the Markov chain be $\mathbf{c} = (c_1, \ldots, c_n)$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ and γ. Conditional on $\mathbf{c}$ and $\boldsymbol{\theta}$, update γ at the lth iteration of the algorithm as follows.

(i) Generate $\gamma^* \sim \mathcal{U}(\gamma - \epsilon, \gamma + \epsilon)$;

(ii) Calculate $\Psi(\gamma^*|\gamma) = \min(1, A_\gamma)$;



(a) Data set A_1.

(b) Data set A_2.

(c) Data set A_3.

(d) Data set A_4.

(e) Data set A_5.

(f) Data set A_6.

Figure 2 *Posterior probability on* $k_{\mathbf{c}}$ *true value and* $k_{\mathbf{c}}$ *values with maximum posterior probability for* $n = 500$. *"●" represents the posterior probability for the* $k_{\mathbf{c}}$ *true value and "□" represents the* $k_{\mathbf{c}}$ *value with maximum posterior probability.*

(iii) Generate $U \sim \mathcal{U}(0, 1)$. If $u \leq \Psi(\gamma^*|\gamma)$ accept γ^* doing $\gamma^{(l)} = \gamma^*$. Otherwise, do $\gamma^{(l)} = \gamma^{(l-1)}$.

Estimate γ by the average of the simulated values, *i.e.*,

$$\tilde{\gamma} = \frac{1}{L_{\tilde{k}_{\mathbf{c}}}} \sum_{l=B+1}^{L} \gamma_j^{(l)}.$$

Tables 3 and 4 show the estimates and the credibility intervals (95%) for γ and the estimates for posterior probabilities of $k_{\mathbf{c}}$ for datasets A_1 to A_6, for $\gamma \sim \Gamma(1, 1)$ and $\gamma \sim \Gamma(2, 4)$, respectively. All estimates for γ lead to the maximum posterior probability on the $k_{\mathbf{c}}$ true value (highlighted in bold). The prior $\Gamma(1, 1)$ on γ has led to higher posterior probability on the $k_{\mathbf{c}}$ than prior distribution $\Gamma(2, 4)$.

For each dataset the credibility intervals contain the most of the values $r \in \mathbb{G}$ that lead to maximum posterior probability on the $k_{\mathbf{c}}$ true value. For instance, for data sets A_1 and A_2, all values r that lead to maximum posterior probability on the $k_{\mathbf{c}}$ true value, except $r = 0.2$, belong to the estimated credibility intervals.

Table 3 *Estimates for γ and estimated posterior probabilities for $k_{\mathbf{c}}$, with $\gamma \sim \Gamma(1, 1)$*

Data set	$k_{\mathbf{c}}$ true	Estimates for γ	Probabilities for $k_{\mathbf{c}}$ values							
			1	2	3	4	5	6	7	≥ 8
A_1	2	0.0432 (0.0051, 0.1318)	0.0000	**0.8948**	0.0968	0.0072	0.0010	0.0002	0.0000	0.0000
A_2	2	0.0434 (0.0048, 0.1323)	0.0000	**0.8632**	0.1218	0.0136	0.0012	0.0002	0.0000	0.0000
A_3	3	0.0734 (0.0148, 0.1883)	0.0000	0.0012	**0.8512**	0.1258	0.0182	0.0028	0.0006	0.0002
A_4	3	0.0810 (0.0151, 0.2145)	0.0000	0.0000	**0.7510**	0.2136	0.0308	0.0044	0.0002	0.0000
A_5	4	0.1259 (0.0298, 0.3072)	0.0000	0.0004	0.0098	**0.7502**	0.1972	0.0340	0.0070	0.0014
A_6	5	0.1899 (0.0495, 0.4719)	0.0000	0.0000	0.0002	0.0570	**0.6782**	0.1926	0.0486	0.0234

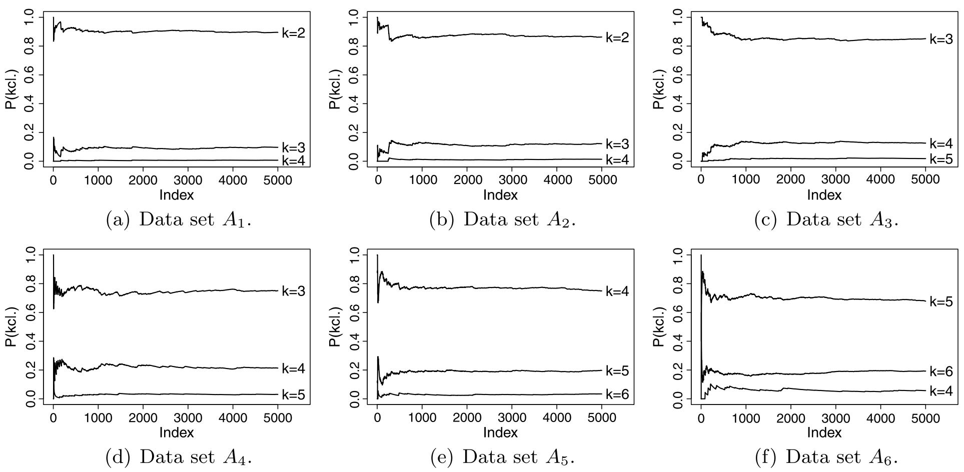
Table 4 *Estimates for γ and estimated posterior probabilities for $k_{\mathbf{c}}$, with $\gamma \sim \Gamma(2, 4)$*

Data set	$k_{\mathbf{c}}$ true	Estimates for γ	Probabilities for $k_{\mathbf{c}}$ values							
			1	2	3	4	5	6	7	≥ 8
A_1	2	0.0642 (0.0119, 0.1710)	0.0000	**0.7768**	0.1986	0.0216	0.0028	0.0020	0.0000	0.0000
A_2	2	0.0627 (0.0124, 0.1639)	0.0000	**0.8120**	0.1634	0.0230	0.0016	0.0000	0.0000	0.0000
A_3	3	0.0947 (0.0239, 0.2166)	0.0000	0.0014	**0.8262**	0.1498	0.0186	0.0040	0.0000	0.0000
A_4	3	0.1021 (0.0249, 0.2537)	0.0000	0.0000	**0.6910**	0.2540	0.0478	0.0062	0.0010	0.0000
A_5	4	0.1471 (0.0414, 0.3349)	0.0000	0.0000	0.0132	**0.6942**	0.2246	0.0556	0.0082	0.0042
A_6	5	0.2178 (0.0601, 0.5331)	0.0000	0.0000	0.0000	0.0604	**0.6324**	0.2142	0.0666	0.0264

4.1.1 *Performance of the SMAS.* Now we verify empirically the convergence of the sequence of the posterior probability for $k_{\mathbf{c}}$ across iterations and the capacity to move for different values of $k_{\mathbf{c}}$ in the course of the iterations and estimated autocorrelation function (acf).

We present performance of SMAS using the results obtained with $\gamma = \tilde{\gamma}$ where $\tilde{\gamma}$ is the estimated value for γ given in Table 3. For other values of γ (estimates given in Table 4 and fixed as r for $r \in \mathbb{G}$) with maximum posterior probability on the $k_{\mathbf{c}}$ true value, the results are similar.

Figure 3 shows the plots of $P(k_{\mathbf{c}}|\cdot)$ estimates across the iterations. To maintain a good visualization we display in each graphic only the three higher $P(k_{\mathbf{c}}|\cdot)$ estimates. The number of iterations and burn in seems to be adequate to achieve stability for the posterior probabilities of $k_{\mathbf{c}}$.

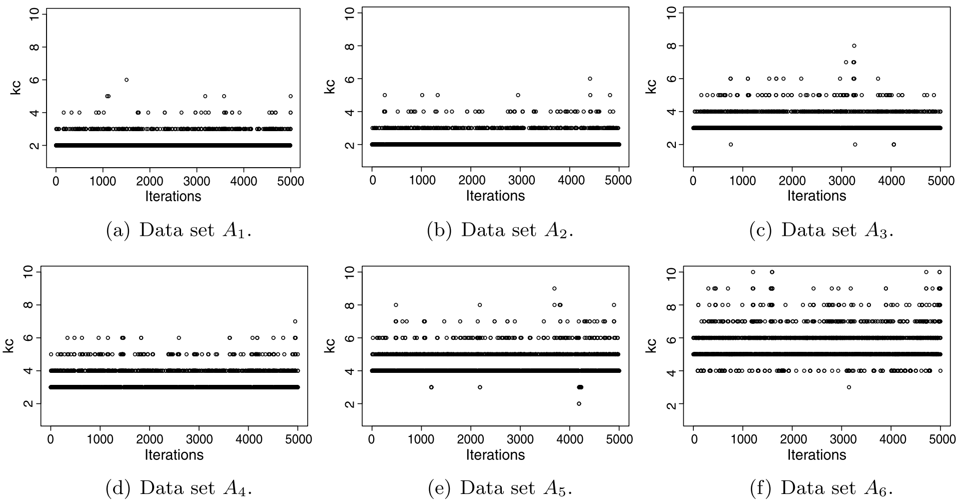


(a) Data set A_1.
(b) Data set A_2.
(c) Data set A_3.
(d) Data set A_4.
(e) Data set A_5.
(f) Data set A_6.

Figure 3 *A posteriori probability for $k_{\mathbf{c}}$ across iterations for thinned sequence.*

Table 5 *Percentages of split and merge movements accepted*

Movement and γ	Data set					
	A_1	A_2	A_3	A_4	A_5	A_6
split	1.2391	1.2297	1.8985	1.2273	1.9273	4.7576
merge	1.2455	1.2333	1.9091	1.2191	1.9333	4.8424
γ	59.9394	59.8303	72.4455	74.0788	81.7727	87.0758

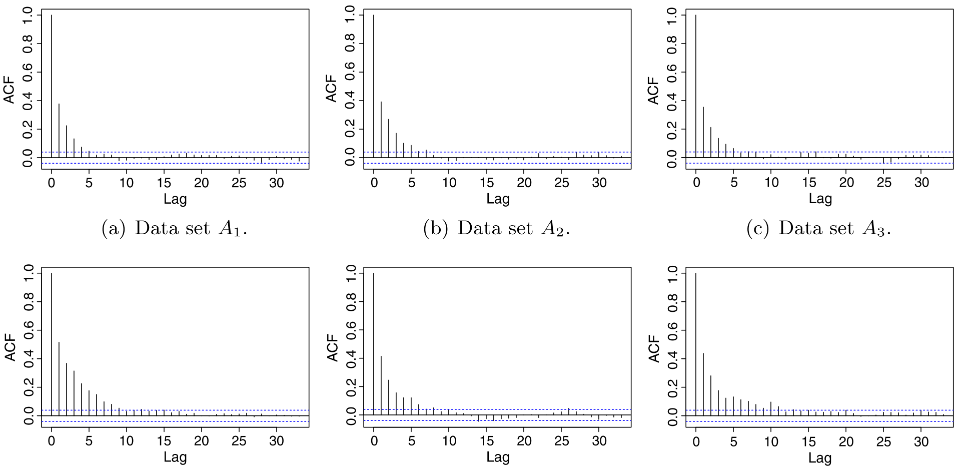


(a) Data set A_1.
(b) Data set A_2.
(c) Data set A_3.
(d) Data set A_4.
(e) Data set A_5.
(f) Data set A_6.

Figure 4 *Sampled $k_{\mathbf{c}}$ values for thinned sequence.*

Figure 4 shows the sampled $k_{\mathbf{c}}$ values in the course of iterations. The algorithm mix well over $k_{\mathbf{c}}$ and remains, in the most of iterations, around the target $k_{\mathbf{c}}$ true values. The sampled $k_{\mathbf{c}}$ values do not present significant autocorrelation, as showed by Figure 5.

Table 5 shows the percentage of split-and-merge moves accepted and the percentage of γ values accepted. Theses percentages show us that split-merge moves are proposed and accepted in a balancing way. The high percentages of accepted values for γ is due the way that we implement the RWM, assuming a small $\varepsilon = 0.05$ value.



(a) Data set A_1.

(b) Data set A_2.

(c) Data set A_3.

(d) Data set A_4.

(e) Data set A_5.

(f) Data set A_6.

Figure 5 *Estimated autocorrelation for thinned sequence.*

Table 6 *Estimated probabilities for* $k_{\mathbf{c}}$, *Galaxy data set, with* $\gamma \sim \Gamma(1,1)$

	$k_{\mathbf{c}}$ values					
	≤ 2	3	4	5	6	≥ 7
$P(k_{\mathbf{c}}\mid\cdot)$	0.0000	**0.8958**	0.0984	0.0050	0.0008	0.0000

Figure 2 in Section S-2.2 of the SM shows the simulated values and the clusters identified by the method. The clusters were satisfactorily identified. Section S-2.2 of the SM also presents the estimates for components of each cluster (Tables 1 and 2) and the histogram of the observed data and the estimated density function (Figure 3).

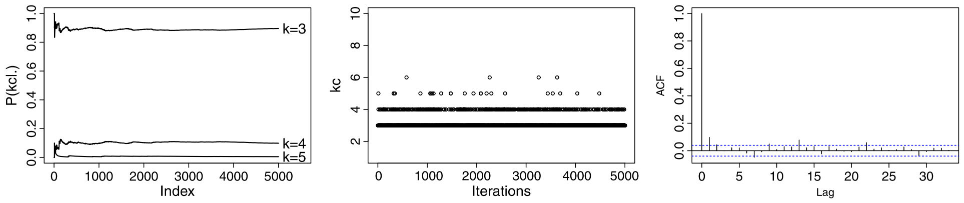
4.2 Galaxy data set

We now apply the proposed methodology to the well-known galaxy dataset, previously analyzed by Roeder and Wasserman (1997), Escobar and West (1995), Richardson and Green (1997), Stephens (2000), among others. The data set refers to velocity (in 103 km/s) from distant galaxies diverging from our own. The sample size is $n = 82$ observations. For more details on this dataset, see Escobar and West (1995) and Richardson and Green (1997).

We consider the galaxy velocities as realizations from a mixture of k normal distributions and the Bayesian model in (3.1) with $k = 10$. We use the same hyperparameters specification used for artificial datasets. To estimate the hyperparameter γ we consider the hierarchical approach setting up $\gamma \sim \Gamma(a, b)$, for $a = b = 1$. The number of iterations, burn in and thin value are the same used in previous analysis.

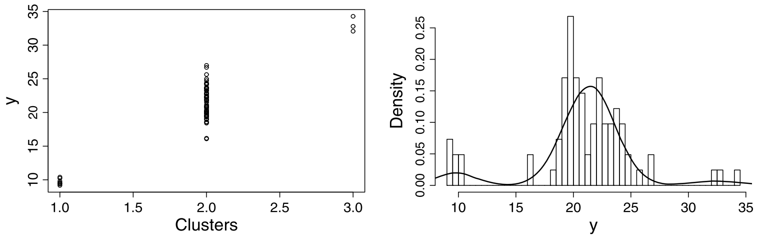
Considering the SMAS algorithm, the estimate and the credibility interval (95%) for γ are 0.1029 and $(0.0182, 0.2696)$, respectively. The estimated posterior probabilities for $k_{\mathbf{c}}$, $1 \leq k_{\mathbf{c}} \leq 10$, are presented in Table 6. The maximum posterior is at $k_{\mathbf{c}} = 3$ with $P(k_{\mathbf{c}} = 3|\cdot) = 0.8958$. For the sake of comparison, estimates of k for this data set range from 3 and 4 for Roeder and Wasserman (1997) and Stephens (2000), from 5 to 7 for Richardson and Green (1997) and just 7 for Escobar and West (1995).

Figure 6 shows the performance of the SMAS. Similar to simulation results, the SMAS sampler mixes well over $k_{\mathbf{c}}$ and shows a satisfactorily stability for probabilities of $k_{\mathbf{c}}$. The sampled $k_{\mathbf{c}}$ values do not present significant autocorrelation. The acceptance ratio for split



(a) Data set A_1. (b) Data set A_2. (c) Data set A_3.

Figure 6 *Performance of SMAS for Galaxy data set.*



(a) Galaxy data set. (b) Galaxy data set.

Figure 7 *Clusters and estimated density function by SMAS for Galaxy data set.*

Table 7 *Estimated probabilities for k_c*

Data set	Estimate for γ	k_c values					% accepted values		
		1	2	3	4	≥ 5	split	merge	γ
Enzyme	0.0862 (0.0162, 0.2439)	0.0000	0.0136	**0.7874**	0.1766	0.0224	2.1761	1.9181	75.9424
Acidity	0.0794 (0.0077, 0.2483)	0.0000	**0.5096**	0.3316	0.1282	0.0304	3.8545	3.8394	72.6606

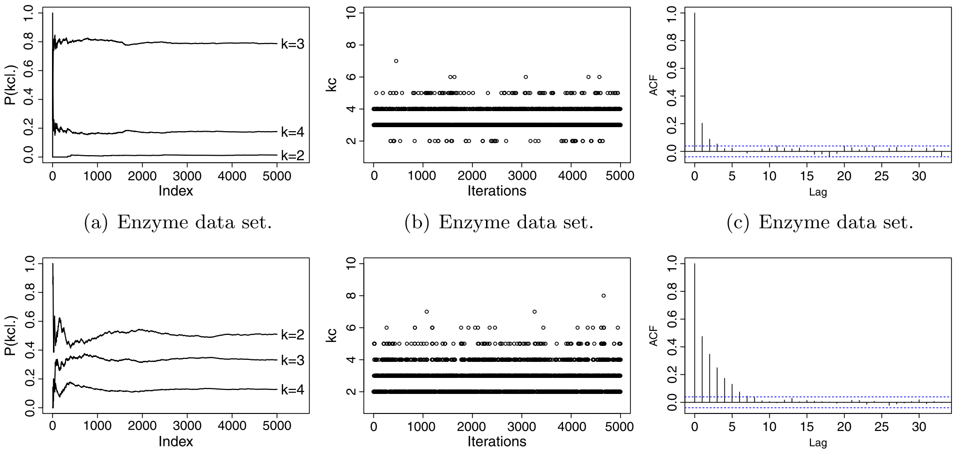
and merge moves are 1.3818% and 1.4364%, respectively. The acceptance ratio for generated γ values is 80.1695%.

Figure 7(a) shows clusters identified conditional on the estimate $\tilde{k}_c = 3$. Figure 7(b) shows the histogram of the observed data and the estimated density function. As noted by Roeder and Wasserman (1997) and Stephens (2000), the multimodality of the velocities indicates the presence of superclusters of galaxies surrounded by large voids, each mode representing a cluster as it moves away at its own speed.

4.3 Enzyme and acidity datasets

Consider now the Enzyme and Acidity datasets downloaded from the website https://people.maths.bris.ac.uk/~mapjg/mixdata. The Enzyme dataset re- fers to enzymatic activity in the blood, for an enzyme involved in the metabolism of carcinogenic substances, among a group of 245 unrelated individuals; and the acidity dataset refers to an acidity index measured in a sample of 155 lakes in north-central Wisconsin.

The two datasets have been analyzed with SMAS using the same hyperparameters specification, the number of iterations, burn in size and thin value used for Galaxy dataset. Table 7 shows the estimates for γ and for posterior probability of k_c, for $1 \leq k_c \leq 10$. For Enzyme dataset the maximum posterior is at $k_c = 3$ with $P(k_c = 3|\cdot) = 0.78744$; while for Acidity dataset the maximum posterior is at $k_c = 2$ with $P(k_c = 2|\cdot) = 0.5096$. This Table also



(a) Enzyme data set. (b) Enzyme data set. (c) Enzyme data set.

(d) Acidity data set. (e) Acidity data set. (f) Acidity data set.

Figure 8 *Performance of SMAS for Enzyme and Acidity data sets.*

present the percentages of accepted values for split-merge movements and for γ values generated. For both datasets, the split-merge moves are proposed and accepted in a balancing way. Section S-4 of the SM presents estimates for component parameters of the clusters, the graphics of clusters identified conditional on the estimate $\tilde{k}_{\mathbf{c}}$ and the histogram of the observed data and the estimated density function.

Figure 8 shows the performance of the SMAS for both datasets. As one can note, the SMAS sampler mixes well over $k_{\mathbf{c}}$ and shows a satisfactorily stability for probabilities of $k_{\mathbf{c}}$.

5 Final remarks

In clustering analysis when the number of clusters is unknown the analyst has to estimate the number of clusters and the parameters conditional on the number of cluster. In this paper, we consider a sparse finite mixture model in order to accommodate the possibility of the number of cluster $k_{\mathbf{c}}$ to be smaller than the number of components k of the mixture model. We develop a Bayesian approach to estimate $k_{\mathbf{c}}$ and the parameters of interest jointly.

The estimation procedure was carried out through the SMAS-MCMC algorithm. The SMAS is essentially a Metropolis–Hastings within Gibbs sampling with a split-merge step. The split-merge step was inserted within the algorithm in order to increase the mixing of the Markov chain in relation to the number of cluster $k_{\mathbf{c}}$. Due to the way that we implement the split-merge strategy, these proposals determines a new partition in the observed data set. This is one factor which improves the efficiency of the method in identifying clusters.

In order to verify the performance of SMAS, we developed a simulation study considering a mixture of univariate normal distributions. The simulation study show us that the estimates of the components are sensible to the choice of the value for hyperparameter γ of the prior Dirichlet distributions; and that the approach considering γ as being an unknown quantity and estimated from the data is preferable. The values used for hyperparameters a and b ($a = b = 1$ and $a = 2$ and $b = 4$) of the Gamma prior distribution for γ lead to an estimate of γ with maximum posterior on the $k_{\mathbf{c}}$ true value. Thus we recommend the use of these values for the hyperparameters a and b.

We also apply the SMAS to three real datasets. Results from simulated and real data sets show that SMAS is an effective alternative for joint estimation of $k_{\mathbf{c}}$, identification of clusters and estimation of parameters. A practical differential of the proposed algorithm is its simplicity to implement in softwares like *R* (the Comprehensive R Archive Network, http://cran.r-project.org). The source code used in data set analysis was developed in software *R* and is available upon request by emailing authors.

The SMAS algorithm was proposed here considering a Bayesian approach with conjugated prior distribution so that we could develop the split-merge movements using the marginal likelihood function and to use the posterior density as generating-candidate density. Extending the SMAS for nonconjugated cases and the generalization for the multivariate case are possible future developments of the method.

Acknowledgment

The first author acknowledges the Brazilian institution CNPq.

Supplementary Material

Supplement to "A Bayesian sparse finite mixture model for clustering data from a heterogeneous population" (DOI: 10.1214/18-BJPS425SUPP; .pdf). Additional details.

References

Akaike, H. A. (1974). New look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723. MR0423716 https://doi.org/10.1109/tac.1974.1100705

Anderson, J. J. (1985). Normal mixtures and the number of clusters problem. *Computational Statistics Quarterly* **2**, 3–14.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821. MR1243494 https://doi.org/10.2307/2532201

Bensmail, H., Celeux, G., Raftery, A. E. and Robert, C. P. (1997). Inference in model-based cluster analysis. *Statistics and Computing* **7**, 1–10.

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika* **65**, 31–38. MR0501592 https://doi.org/10.1093/biomet/65.1.31

Bouveyron, C. and Brunet, C. (2013). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis* **71**, 52–78. MR3131954 https://doi.org/10.1016/j.csda.2012.12.008

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrica* **52**, 345–370. MR0914460 https://doi.org/10.1007/BF02294361

Casella, G., Robert, C. and Wells, M. (2000). Mixture models, latent variables and partitioned importance sampling. Technical Report-2000-03, CREST, INSEE, Paris.

Celeux, G., Hurn, M. and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association* **95**, 957–970. MR1804450 https://doi.org/10.2307/2669477

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis–Hastings algorithm. *American Statistician* **49**, 327–335.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588. MR1340510

Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**. MR1951635 https://doi.org/10.1198/016214502760047131

Fruhwirth-Schnatter, S. (2017). From here to infinity-sparse finite versus Dirichlet process mixture in model-based clustering. https://arxiv.org/abs/1706.07194.

Hartigan, J. A. and Wong, M. A. (1978). Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics* **28**, 100–108. MR0405726

Jasra, A., Holmes, C. C. and Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science* **20**, 50–67. MR2182987 https://doi.org/10.1214/088342305000000016

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 281–297. Berkeley, CA: University of California Press. MR0214227

McLachlan, G. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker. MR0926484

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley Interscience. MR1789474 https://doi.org/10.1002/0471721182

Nobile, A. and Fearnside, A. T. (2007). Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing* **17**, 147–162. MR2380643 https://doi.org/10.1007/s11222-006-9014-7

Oh, M.-S. and Raftery, A. E. (2007). Model-based clustering with dissimilarities: A Bayesian approach. *Journal of Computational and Graphical Statistics* **16**, 559–585. MR2351080 https://doi.org/10.1198/106186007X236127

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* **59**, 731–792. MR1483213 https://doi.org/10.1111/1467-9868.00095

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixture of normals. *Journal of the American Statistical Association* **92**, 894–902. MR1482121 https://doi.org/10.2307/2965553

Saraiva, E. F., Louzada, F. and Milan, L. A. (2014). Mixture models with an unknown number of components via a new posterior split–merge MCMC algorithm. *Applied Mathematics and Computation* **244**, 959–975. MR3250635 https://doi.org/10.1016/j.amc.2014.07.032

Saraiva, E. F., Suzuki, A. K., Louzada, F. and Milan, L. A. (2016). Partitioning gene expression data by data-driven Markov chain Monte Carlo. *Journal of Applied Statistics* **43**, 1155–1173. MR3460559 https://doi.org/10.1080/02664763.2015.1092113

Saraiva, E. F., Suzuki, A. K. and Milan, L. A. (2019). Supplement to "A Bayesian sparse finite mixture model for clustering data from a heterogeneous population." https://doi.org/10.1214/18-BJPS425SUPP.

Schwarz, G. E. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464. MR0468014

Sneath, P. H. A. (1957). The application of computers to taxonomy. *Journal of General Microbiology* **17**, 201–206.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin* **38**, 1409–1438.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B* **64**, 583–616. MR1979380 https://doi.org/10.1111/1467-9868.00353

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* **62**, 795–809. MR1796293 https://doi.org/10.1111/1467-9868.00265

Walli, G. M., Frhwirth-Schnatter, S. and Grn, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing* **34**, 303–324. MR3439375 https://doi.org/10.1007/s11222-014-9500-2

Ward, J. H. (1963). Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association* **58**, 234–244. MR0148188

Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association* **105**, 713–726. MR2724855 https://doi.org/10.1198/jasa.2010.tm09415

E. F. Saraiva
Instituto de Matemática
Universidade Federal de Mato Grosso do Sul
Campo Grande
Brazil
E-mail: erlandson.saraiva@ufms.br

A. K. Suzuki
Departamento de Matemática Aplicada e Estatística
Unversidade de São Paulo
São Paulo
Brazil

L. A. Milan
Departamento de Estatística
Universidade Federal de São Carlos
São Carlos
Brazil