Latent class model with conditional dependency per modes to cluster categorical data

Matthieu Marbac, Christophe Biernacki, Vincent Vandewalle

Advances in Data Analysis and Classification manuscript No.
(will be inserted by the editor)

Latent class model with conditional dependency per modes to cluster categorical data

Matthieu Marbac · Christophe Biernacki · Vincent Vandewalle

Received: May 22th, 2015 / Accepted: date

Abstract We propose a parsimonious extension of the classical latent class model to cluster categorical data by relaxing the conditional independence assumption. Under this new mixture model, named Conditional Modes Model (CMM), variables are grouped into conditionally independent blocks. Each block follows a parsimonious multinomial distribution where the few free parameters model the probabilities of the most likely levels, while the remaining probability mass is uniformly spread over the other levels of the block. Thus, when the conditional independence assumption holds, this model defines parsimonious versions of the standard latent class model. Moreover, when this assumption is violated, the proposed model brings out the main intra-class dependencies between variables, summarizing thus each class with relatively few characteristic levels. The model selection is carried out by an hybrid MCMC algorithm that does not require preliminary parameter estimation. Then, the maximum likelihood estimation is performed via an EM algorithm only for the best model. The model properties are illustrated on simulated data and on three real data sets by using the associated R package `CoModes`[1]. The results show that this model allows to reduce biases involved by the conditional independence assumption while providing meaningful parameters.

Keywords categorical data · clustering · integrated complete-data likelihood · MCMC algorithm · mixture models · model selection.

Mathematics Subject Classification (2000) 62H30 · 62F15 · 62-07 · 62F07

[1] Downloadable at https://r-forge.r-project.org/R/?group_id=1809

1 Introduction

Clustering (Jajuga et al., 2002) is a worthwhile method for analysing complex data sets. It aims to extract the main information from the data by grouping individuals into *homogeneous* classes. This paper focuses on the cluster analysis of categorical variables which occurs in many different fields (biology, sociology, marketing...). Clustering methods can be split into two approaches: the *geometric* ones based on distances between individuals and the *probabilistic* ones which model the data generation.

Geometric approaches are generally simpler and faster than the probabilistic ones, but they are sensitive to the choice of the distance between individuals. For the categorical data analysis, either they define a metric in the initial variable space like the *k-means* (Huang et al., 2005), or they compute a metric on the factorial axes (Chavent et al., 2010). Note that lots of geometric approaches can be interpreted as probabilistic ones (Govaert, 2010) revealing probabilistic hidden assumptions. Moreover, the probabilistic approaches allow to solve difficult questions, like the class number selection, in a rigorous mathematical framework.

The probabilistic approaches define a *homogeneous* class as the subset of the individuals arisen from the same distribution. In this context, the most classical approach models the data distribution by a finite mixture model of parametric distributions (McLachlan and Peel, 2000). The resulting partition is meaningful since each class is described by the parameters of one mixture component.

The latent class model (Goodman, 1974) which assumes the conditional independence between variables (further referred as CIM for conditional independent model) is a popular probabilistic approach to cluster categorical data. Its interpretation is easy since classes are explicitly described by the probability of each modality for each variable. Moreover, the sparsity involved by the conditional independence assumption is a great advantage since it circumvents the curse of dimensionality. In practice, this model obtains good results in lots of applications (Hand and Yu, 2001). However, it leads to severe biases when its main assumption is violated (see Section 5), like an overestimation of the number of components (Van Hattum and Hoijtink, 2009). Furthermore, the larger the number of variables, the higher the risk to observe conditionally correlated variables in a data set, and consequently the higher the risk to involve such biases by using CIM.

Different models relax the conditional independence assumption. Among them, the *multilevel latent class model* (Vermunt, 2003) assumes that conditional dependency between the observed variables can be explained by other unobserved variables. Another approach considers the intra-class dependencies by using a single latent continuous variable and a probit function (Qu et al., 1996). Recently proposed, the *mixture of latent trait analyzers* (Gollini and Murphy, 2014; Bartholomew et al., 2011) is a good challenger for CIM. It assumes that the distribution of the observed variables depends on many

latent variables: one categorical variable (the class) and many continuous latent variables (modeling the intra-class dependencies between the observed categorical variables). However, its parameter inference is not so easy, so the authors advise to use a variational approach but they also propose to use the Gauss-Hermite quadrature to assess the log-likelihood. Although this model is very flexible, intra-class dependency is hardly interpretable since the intra-class correlations are interpreted throughout relationships with unobserved continuous variables.

The log-linear models' (Agresti, 2002) purpose is to model the individual log-probability by selecting interactions between variables. Thus, the most general mixture model is the *log-linear mixture model* where all the kinds of interactions can be considered. It has been used for a long time (Hagenaars, 1988) and it obtains good results in many applications (Espeland and Handelman, 1989; Van Hattum and Hoijtink, 2009). However this model family is huge and the model selection stays a real challenge. In the literature, authors either fix the considered interactions in advance or they they perform a deterministic search like the *forward* method which is sub-optimal. Furthermore, the number of parameters increases with the conditional modality crossings, so there is an over-fitting risk and interpretation becomes harder.

This paper presents the *conditional modes model* (referred by CMM) which groups the variables into *conditionally independent blocks* in order to consider the main conditional dependencies. Note that this idea was introduced in the Multimix software by Jorgensen and Hunt (1996) to cluster continuous and categorical data. Indeed, this software splits the variables into conditionally independent blocks but each block is allowed to contain one categorical variable at most (and a free number of continuous variables). Moreover, the categorical variable of a block is modelled by a full multinomial distribution. The CMM approach is developed only for categorical variables and considers blocks with more than one categorical variable. Moreover, each block of CMM follows a multinomial distribution per modes which assumes that few levels, named *modes*, are characteristic whereas the other ones follow a uniform distribution. Thus, the resulting multinomial distribution is parsimonious since its free parameters are limited to these few modes. Finally, the Multimix software carries out the model selection with a forward procedure while the model selection of CMM is achieved by a hybrid MCMC algorithm, inheriting thereby some optimality properties.

The resulting model is a good challenger for CIM since it preserves the sparsity and avoids many biases by considering the main conditional correlations. This model can also be interpreted as a parsimonious version of the log-linear mixture model. Indeed, the repartition of the variables into blocks defines the considered interactions while the multinomial distribution per modes defines specific intra-block interactions. In this way, it produces meaningful classes since the intra-class dependencies are brought out at two complementary levels: the block variable interaction level and the associated mode interaction level (through both locations and probabilities). Note that even if CMM can

model some intra-class dependencies, it can also assume the conditional independence between variables. In this way, it generalizes lots of the parsimonious versions of CIM (Celeux and Govaert, 1991).

For a fixed number of components, the model selection (repartition of the variables into blocks and mode numbers) is the most challenging problem since the number of competing models is huge. Therefore, the model selection is carried out by an MCMC algorithm whose the mode of the stationary distribution corresponds to the model having the highest posterior probability. This algorithm performs a random walk in the model space and requires the computation of the integrated complete-data likelihood. This quantity is not approximated by BIC-like methods since their results are biased (see our numerical experiments). Indeed, the integrated complete-data likelihood is accessible and non ambiguous through weakly informative conjugate prior. This approach provides an efficient model selection in a reasonable computational time since the parameters are estimated via an EM algorithm only for the single selected model. Thus, this approach is a possible answer to the combinatorial model selection problem which is known to be a real challenge for a log-linear mixture model.

This paper is organized as follows. Section 2 presents the new mixture model. Section 3 presents the hybrid MCMC algorithm which performs the model selection. Section 4 presents the EM algorithm used to perform the maximum likelihood inference. In Section 5, numerical experiments show the relevance of the proposed criterion for model selection and the properties of the estimation algorithms. Section 6 presents three clusterings of real data sets performed by the R package CoModes. A conclusion is given in Section 7.

2 Conditional modes model

2.1 Mixture model of conditionally independent blocks

Observations $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ consist of n individuals $\mathbf{x}_i = (\boldsymbol{x}_i^1, \ldots, \boldsymbol{x}_i^p)$ described by p categorical variables, called thereafter initial variables.

Running example The Alzheimer data set (Moran et al., 2004) indicates the presence (yes) or absence (no) of six symptoms of Alzheimer's disease (AD) for $n = 240$ patients diagnosed with early onset AD conducted in the Mercer Institute in St. James's Hospital, Dublin. The $p = 6$ binary symptoms are, in this order: *hallucination* (Hal), *activity* (Act), *aggression* (Agg), *agitation* (Agi), *diurnal* (Diu) and *affective* (Aff).

The *conditional modes model* (CMM) assumes that individuals arise independently from a mixture with g components where variables are grouped into d *conditionally independent blocks of variables*. The repartition of the variables into the d blocks[2] is defined by the partition $\boldsymbol{s} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_d)$ of

[2] Note that the repartition of the variables into blocks is identical between classes. This choice is motivated by reasons of identifiability and interpretation that we will detail later.

$\{1, \ldots, p\}$. The new variable resulting from the concatenation of the initial variables affiliated to block j (*i.e.* $\{\boldsymbol{x}_i^b; b \in \boldsymbol{s}_j\}$) is itself a categorical variable whose levels are defined by the Cartesian product of the variables affiliated to block j. This new (block dependent) categorical variable defined for block j is denoted by $\mathbf{x}_i^j = (\mathrm{x}_i^{jh}; h = 1, \ldots, \mathrm{m}_j)$, such as $\mathrm{x}_i^{jh} = 1$ if individual i has level h and $\mathrm{x}_i^{jh} = 0$ otherwise, where m_j is the number of levels for block j. Since this mapping of the variables is bijective, defining a probability on $\mathbf{x}_i = (\boldsymbol{x}_i^1, \ldots, \boldsymbol{x}_i^p)$ is equivalent to defining a probability on $\mathbf{x}_i = (\mathbf{x}_i^1, \ldots, \mathbf{x}_i^d)$. The running example below provides an illustration of the relations between the initial variables $\boldsymbol{x}_i^b$ and the new (block dependent) variables $\mathbf{x}_i^j$.

Running example If $\boldsymbol{s} = (\{1\}, \{2\}, \{3, 5\}, \{4\}, \{6\})$, then the mixture components consider the dependency between the variable *aggression* and the variable *diurnal* (initial variables 3 et 5) while the other variables are conditionally independent. The repartition of the variables into blocks given by $\boldsymbol{s}$ provides five new categorical (block dependent) variables $\mathbf{x}_i^j$ (with $j = 1, \ldots, 5$). Only one variable is affiliated to blocks 1, 2, 4 and 5, so the variables of these blocks are binary ones, the variable of block 3 takes 4 levels described in Table 1.

observed variables		block 3
$\boldsymbol{x}_i^3$	$\boldsymbol{x}_i^5$	$\mathbf{x}_i^3$
yes	yes	level 1
yes	no	level 2
no	yes	level 3
no	no	level 4

Table 1: Example of the new 4-level variable associated to block 3 and composed of two initial binary variables.

The distribution of the (variables associated to the) block j under component k has u_{kj} degrees of freedom and is parametrized by $(\boldsymbol{\delta}_{kj}, \boldsymbol{\alpha}_{kj})$, k designating the index of the component among the g ones. We detail in Section 2.2 the exact expression of this distribution and in particular the specific definition of its two important parameters $\boldsymbol{\delta}_{kj}$ and $\boldsymbol{\alpha}_{kj}$. The probability distribution function (pdf) of CMM is

$$p(\mathbf{x}_i|\boldsymbol{m}, \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \prod_{j=1}^{d} p(\mathbf{x}_i^j|u_{kj}, \boldsymbol{\delta}_{kj}, \boldsymbol{\alpha}_{kj}), \tag{1}$$

where $\boldsymbol{m} = (g, \boldsymbol{s}, \boldsymbol{u})$ specifies the model with $\boldsymbol{u} = (u_{kj}; k = 1, \ldots, g; j = 1, \ldots, d)$, and where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\delta}, \boldsymbol{\alpha})$ denotes the parameters with $\boldsymbol{\delta} = (\boldsymbol{\delta}_{kj}; k = 1, \ldots, g; j = 1, \ldots, d)$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{kj}; k = 1, \ldots, g; j = 1, \ldots, d)$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ groups the mixture proportions with $0 < \pi_k \leq 1$ and $\sum_{k=1}^{g} \pi_k = 1$.

2.2 Block of multinomial distribution per modes

The CMM model considers that each block follows a so-called *multinomial distribution per modes*. This distribution has only few free parameters corresponding to its *modes*, while the other parameters are equal. Thus, the free parameters are those of the levels having the greatest probabilities whereas uniformity holds for non-mode levels. The distribution of block j for component k has u_{kj} degrees of freedom (so u_{kj} modes) with $0 \leq u_{kj} < \mathrm{m}_j$, and its mode locations are defined by the discrete parameter $\boldsymbol{\delta}_{kj} = \{\delta_{kjh}; h = 1, \ldots, u_{kj}\}$ with $\delta_{kjh} \in \{1, \ldots, \mathrm{m}_j\}$ and $\delta_{kjh} \neq \delta_{kjh'}$ if $h \neq h'$. Its probabilities are given by $\boldsymbol{\alpha}_{kj} = (\alpha_{kjh}; h = 1, \ldots, u_{kj}+1)$ where α_{kjh} denotes the probability of mode h for $h = 1, \ldots, u_{kj}$ and where $\alpha_{kju_{kj}+1}$ corresponds to the remaining probability mass. So, $\boldsymbol{\alpha}_{kj}$ is defined on a truncated simplex denoted by $S(u_{kj}, \mathrm{m}_j)$ with

$$S(u_{kj}, \mathrm{m}_j) = \left\{ \boldsymbol{\alpha}_{kj} : \sum_{h=1}^{u_{kj}+1} \alpha_{kjh} = 1 \text{ and for } 1 \leq h \leq u_{kj},\ \alpha_{kjh} \geq \frac{\alpha_{kju_{kj}+1}}{\mathrm{m}_j - u_{kj}} > 0 \right\}. \quad (2)$$

Therefore, the pdf of block j for component k is

$$p(\mathbf{x}_i^j | u_{kj}, \boldsymbol{\delta}_{kj}, \boldsymbol{\alpha}_{kj}) = \left(\prod_{h=1}^{u_{kj}} (\alpha_{kjh})^{\mathrm{x}_i^{j(k,h)}} \right) \left(\frac{\alpha_{kju_{kj}+1}}{\mathrm{m}_j - u_{kj}} \right)^{1-\sum_{h \in \boldsymbol{\delta}_{kj}} \mathrm{x}_i^{j(k,h)}}, \quad (3)$$

where the implicit notation $\mathrm{x}_i^{j(k,h)} = \mathrm{x}_i^{j\delta_{kjh}}$ is used.

Running example We consider the CMM model defined by $\boldsymbol{m} = (g, \boldsymbol{s}, \boldsymbol{u})$ where $g = 2$, $\boldsymbol{s} = (\{1\}, \{2\}, \{3, 5\}, \{4\}, \{6\})$ and where $\boldsymbol{u}$ is defined by Table 2. For instance, for component 1, uniformity holds in block 2 since $u_{kj} = 0$

	Block 1	Block 2	Block 3	Block 4	Block 5
Component 1	1	0	1	1	0
Component 2	1	1	1	0	1

Table 2: Example of the number of modes for the Alzheimer data set.

while the distribution of block 3 has one mode and can be summarized by Table 3.

block 3	observed variables		probability
$\mathbf{x}_i^3$	aggression	diurnal	α_{13h}
level 4	no	no	0.82
level 1	yes	yes	0.06
level 2	yes	no	0.06
level 3	no	yes	0.06

Table 3: Summary of the distribution of block 3 for component 1.

2.3 Model characteristics

The proposed model has two levels of interpretation. First, the intra-class dependencies of variables (equal over the mixture components) are brought out by the repartition of the variables into blocks given by $\boldsymbol{s}$. Second, the intra-class and intra-block dependencies are summarized by the modes (possibly different over the mixture components) where the locations are given by $\boldsymbol{\delta}_{kj}$ and where the probabilities are given by $\boldsymbol{\alpha}_{kj}$. The modes are interpreted as an over-contribution in comparison to the uniform distribution, since the probability of each mode is greater than the probability of the other locations (see constraints of (2)). A shorter summary for each distribution is also available by using the following two compact terms defined on $[0, 1]$ and which reflect respectively the *complexity* and the *strength* of the intra-class and intra-block dependencies:

$$\bar{u}_{kj} = \frac{u_{kj}}{\mathrm{m}_j - 1} \text{ and } \alpha_{kj\bullet} = \sum_{h=1}^{u_{kj}} \alpha_{kjh} \tag{4}$$

Thus, the smaller is $\bar{u}_{kj}$ and the larger is $\alpha_{kj\bullet}$, the more massed in few characteristic levels is the distribution.

Running example For block 3 of component 1, $\bar{u}_{13} = 1/3$ and $\alpha_{13\bullet} = 0.82$, so this block distribution is strongly concentrated on one level.

The CIM model and its parsimonious versions implemented in Rmixmod (Lebret et al., 2014) belong to the family of CMM. Indeed, if each block of CMM is composed by a single observed variable then the CMM model considers independence between all the variables within class (often called local independence). Table 4 presents the link between both model families (notations $[\varepsilon_k^{jh}]$ and $[\varepsilon_k^{j}]$ designate models used in Lebret et al. (2014); see more details in this paper). It is also very interesting to notice that CMM

Rmixmod model	CMM				
	d	$\boldsymbol{s}$	u_{kj}	$\bar{u}_{kj}$	$\alpha_{kj\bullet}$
Free $[\varepsilon_k^{jh}]$	p	$(\{1\}, \ldots, \{p\})$	$\mathrm{m}_j - 1$	1	$[1 - \frac{1}{\mathrm{m}_j}; 1]$
Constrained $[\varepsilon_k^{j}]$	p	$(\{1\}, \ldots, \{p\})$	1	$\frac{1}{\mathrm{m}_j - 1}$	$[\frac{1}{\mathrm{m}_j}; 1]$

Table 4: Link between models of Rmixmod and CMM.

can be more parsimonious than the locally independent model although it takes into account the conditional dependencies. Finally, note that a sufficient condition for the generic identifiability of CMM (see Appendix A) results from the property of the generic identifiability of CIM (Allman et al., 2009).

3 Model selection by a hybrid MCMC algorithm

3.1 Model selection by integrated likelihood

The number of mixture components for the competing models is usually bounded by a value $g_{\max}$. The goal of model selection is to find, among the set of competing models $\mathcal{M}$, the model $\hat{\boldsymbol{m}}$ having the largest posterior probability. Thus,

$$\hat{\boldsymbol{m}} = \underset{\boldsymbol{m}\in\mathcal{M}}{\arg\max}\, p(\boldsymbol{m}|\mathbf{x}). \tag{5}$$

We denote by $\mathcal{M}_g$ the space of the models having g components, therefore

$$\hat{\boldsymbol{m}} = \underset{g\in\{1,\dots,g_{\max}\}}{\arg\max}\; p(\boldsymbol{m}_g|\mathbf{x}) \text{ with } \boldsymbol{m}_g = \underset{\boldsymbol{m}\in\mathcal{M}_g}{\arg\max}\, p(\boldsymbol{m}|\mathbf{x}). \tag{6}$$

By assuming that the model prior follows a uniform distribution,

$$p(\boldsymbol{m}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{m}) \text{ with } p(\mathbf{x}|\boldsymbol{m}) = \int p(\mathbf{x}|\boldsymbol{m},\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{m})d\boldsymbol{\theta}, \tag{7}$$

where $p(\mathbf{x}|\boldsymbol{m},\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{m},\boldsymbol{\theta})$ is the likelihood function and $p(\boldsymbol{\theta}|\boldsymbol{m})$ is the parameter distribution of the parameters. We assume independence between the prior distributions of the parameters, so

$$p(\boldsymbol{\theta}|\boldsymbol{m}) = p(\boldsymbol{\pi}|\boldsymbol{m}) \prod_{k=1}^{g}\prod_{j=1}^{d} p(\boldsymbol{\delta}_{kj}|\boldsymbol{m})p(\boldsymbol{\alpha}_{kj}|\boldsymbol{m}). \tag{8}$$

The conjugate prior distribution of a multinomial distribution is a Dirichlet distribution. Thus, the Jeffreys non informative prior of the mixture proportions is the Dirichlet distribution $\mathcal{D}_g(\frac{1}{2},\dots,\frac{1}{2})$. Moreover, $\boldsymbol{\delta}_{kj}|\boldsymbol{m}$ follows a uniform distribution over the subset of $\{1,\dots,\mathrm{m}_j\}$ of size u_{kj}. Finally, $\boldsymbol{\alpha}_{kj}|\boldsymbol{m}$ follows a Dirichlet distribution $\mathcal{D}_{u_{kj}+1}(1,\dots,1)$ restricted on the space $S(u_{kj},\mathrm{m}_j)$. Note that parameters of this last distribution are chosen equal to one in order to facilitate the computation of the integrated complete-data likelihood (see Section 3.4).

The number of competing models is huge, so an exhaustive approach is not doable. Therefore, the estimation of $\boldsymbol{m}_g$ is performed by an MCMC algorithm for $g = 1,\dots,g_{\max}$.

3.2 Model selection via an ideal Gibbs sampler

For any g, the estimation of $\boldsymbol{m}_g$ is equivalent to the estimation of the couple $(\boldsymbol{s},\boldsymbol{u})$ maximizing $p(\boldsymbol{s},\boldsymbol{u}|\mathbf{x},g)$. This aim can be achieved by a Gibbs sampler having $p(\boldsymbol{s},\boldsymbol{u}|\mathbf{x},g)$ as marginal stationary distribution. Thus, we return the most generated couple $(\boldsymbol{s},\boldsymbol{u})$.

We introduce the instrumental variable $\mathbf{z} = (\mathbf{z}_i; i = 1,\dots,n)$ which indicates the class membership of the individuals, where $\mathbf{z}_i = (\mathrm{z}_{ik}; k = 1,\dots,g)$

with $\mathrm{z}_{ik} = 1$ if individual i belongs to component k and $\mathrm{z}_{ik} = 0$ otherwise. The Gibbs sampler alternates between the conditional sampling of the partition $\mathbf{z}$ and the conditional sampling of the couple $(\boldsymbol{s}, \boldsymbol{u})$. Thus, its stationary distribution is $p(\boldsymbol{s}, \boldsymbol{u}, \mathbf{z}|\mathbf{x}, g)$. Its iteration $[q]$ is written as

$$\mathbf{z}^{[q+1]} \sim p(\mathbf{z}|g, \boldsymbol{s}^{[q]}, \boldsymbol{u}^{[q]}, \mathbf{x}) \tag{9}$$

$$(\boldsymbol{s}^{[q+1]}, \boldsymbol{u}^{[q+1]}) \sim p(\boldsymbol{s}, \boldsymbol{u}|g, \mathbf{x}, \mathbf{z}^{[q+1]}). \tag{10}$$

Both previous steps are performed with difficulty because independence between individuals does not hold in (9) and because the space of $(\boldsymbol{s}, \boldsymbol{u})$ is huge in (10). Thus, an hybrid algorithm derived from this Gibbs sampler is used. The mode of its marginal stationary distribution stays located at $\boldsymbol{m}_g$. The samplings from (9) and (10) are detailed in Section 3.3 and Section 3.4 respectively.

3.3 Gibbs algorithm for partition sampling

In this section, we show that the sampling from (9) can be achieved by a Gibbs sampler which alternates between both following samplings: $\mathbf{z}|\boldsymbol{m}, \mathbf{x}, \boldsymbol{\theta}$ and $\boldsymbol{\theta}|\boldsymbol{m}, \mathbf{x}, \mathbf{z}$. Both previous samplings can be easily defined by introducing the complete-data likelihood function, which is written as

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}, \boldsymbol{\theta}) = \prod_{k=1}^{g} \pi_k^{n_k} \prod_{j=1}^{d} \prod_{h=1}^{u_{kj}} (\alpha_{kjh})^{n_{kj(h)}} \left(\frac{\alpha_{kju_{kj}+1}}{\mathrm{m}_j - u_{kj}} \right)^{\bar{n}_{kj}} \mathbb{1}_{\{\boldsymbol{\alpha}_{kj} \in S(u_{kj}; \mathrm{m}_j)\}}, \tag{11}$$

where $n_k = \sum_{i=1}^{n} \mathrm{z}_{ik}$, $n_{kj(h)} = \sum_{i=1}^{n} \mathrm{z}_{ik} \mathrm{x}_i^{j(k,h)}$ and $\bar{n}_{kj} = n_k - \sum_{h=1}^{u_{kj}} n_{kj(h)}$. The sampling from $\mathbf{z}|\boldsymbol{m}, \mathbf{x}, \boldsymbol{\theta}$ is performed by n independent samplings of the multinomial distributions $\mathcal{M}_g(t_{i1}(\boldsymbol{\theta}|\boldsymbol{m}), \ldots, t_{ig}(\boldsymbol{\theta}|\boldsymbol{m}))$ where

$$t_{ik}(\boldsymbol{\theta}|\boldsymbol{m}) = \frac{\pi_k \prod_{j=1}^{d} p(\mathbf{x}_i^j|u_{kj}, \boldsymbol{\delta}_{kj}, \boldsymbol{\alpha}_{kj})}{\sum_{k'=1}^{g} \pi_{k'} \prod_{j=1}^{d} p(\mathbf{x}_i^j|u_{k'j}, \boldsymbol{\delta}_{k'j}, \boldsymbol{\alpha}_{k'j})}. \tag{12}$$

The sampling from $\boldsymbol{\theta}|\boldsymbol{m}, \mathbf{x}, \mathbf{z}$ is performed as follows. The distribution of $\boldsymbol{\pi}|\boldsymbol{m}, \mathbf{x}, \mathbf{z}$ is the Dirichlet $\mathcal{D}_g\left(\frac{1}{2} + n_1, \ldots, \frac{1}{2} + n_g\right)$ and the posterior distribution of $(\boldsymbol{\delta}_{kj}, \boldsymbol{\alpha}_{kj})$ is written as

$$p(\boldsymbol{\delta}_{kj}, \boldsymbol{\alpha}_{kj}|\boldsymbol{m}, \mathbf{x}, \mathbf{z}) = p(\boldsymbol{\delta}_{kj}|\boldsymbol{m}, \mathbf{x}, \mathbf{z}) p(\boldsymbol{\alpha}_{kj}|\boldsymbol{m}, \mathbf{x}, \mathbf{z}, \boldsymbol{\delta}_{kj}). \tag{13}$$

The distribution of $\boldsymbol{\delta}_{kj}|\boldsymbol{m}, \mathbf{x}, \mathbf{z}$ is a multinomial one with too many values to be computable. Let the set $\boldsymbol{\delta}_{kj}^{\star} = \{\delta_{kjh}^{\star}; h = 1, \ldots, u_{kj}\}$ containing the indices of the u_{kj} largest values of $n_{kjh} = \sum_{i=1}^{n} \mathrm{z}_{ik} \mathrm{x}_i^{jh}$ such as

$$\forall h \in \boldsymbol{\delta}_{kj}^{\star}, \forall h' \in \{1, \ldots, \mathrm{m}_j\} \setminus \boldsymbol{\delta}_{kj}^{\star}, \quad n_{kjh} \geq n_{kjh'}. \tag{14}$$

We assume now that the difference between the mode probabilities and the non-mode probabilities are significant. So, the distribution of $\boldsymbol{\delta}_{kj}|\boldsymbol{m}, \mathbf{x}, \mathbf{z}$ can

be approximated by a Dirac in $\boldsymbol{\delta}^\star_{kj}$. This approximation is strengthened by the fast convergence speed of the discrete parameters (Choirat and Seri, 2012). Finally, $\boldsymbol{\alpha}_{kj}|\boldsymbol{m},\mathbf{x},\mathbf{z},\boldsymbol{\delta}^\star_{kj}$ follows the Dirichlet distribution $\mathcal{D}_{u_{kj}+1}\Big(1+n^\star_{kj(1)},\dots,1+n^\star_{kj(u_{kj})},1+\bar{n}^\star_{kj}\Big)$ truncated on $S(u_{kj};\mathrm{m}_j)$, where $n^\star_{kj(h)}=n_{kj\delta^\star_{kjh}}$ and where $\bar{n}^\star_{kj}=n_k-\sum_{h=1}^{u_{kj}} n^\star_{kj(h)}$.

3.4 MCMC algorithm for model sampling

The sampling from (10) requires the computation of the integrated complete-data likelihood $p(\mathbf{x},\mathbf{z}|\boldsymbol{m})=\int p(\mathbf{x},\mathbf{z}|\boldsymbol{m},\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{m})d\boldsymbol{\theta}$, since

$$p(\boldsymbol{s},\boldsymbol{u}|g,\mathbf{x},\mathbf{z})\propto p(\mathbf{x},\mathbf{z}|\boldsymbol{m}). \tag{15}$$

The integrated complete-data likelihood is written as

$$p(\mathbf{x},\mathbf{z}|\boldsymbol{m})=\frac{\Gamma(\frac{g}{2})\prod_{k=1}^{g}\Gamma(n_k+\frac{1}{2})}{\Gamma(\frac{1}{2})^g\Gamma(n+\frac{g}{2})}\prod_{k=1}^{g}\prod_{j=1}^{d}I_{kj}(u_{kj}), \tag{16}$$

where $I_{kj}(u_{kj})$ is the integral related to block j of component k defined by

$$I_{kj}(u_{kj})=\sum_{\delta_{kj}}\int_{S(u_{kj},\mathrm{m}_j)}\left(\prod_{h=1}^{u_{kj}}(\alpha_{kjh})^{n_{kj\delta_{kj}(h)}}\right)\left(\frac{\alpha_{kju_{kj}+1}}{\mathrm{m}_j-u_{kj}}\right)^{n_k-\sum_{h\in\delta_{kj}}n_{kjh}}d\boldsymbol{\alpha}_{kj}. \tag{17}$$

The integral $I_{kj}(u_{kj})$ has not a closed form. A BIC-like method could evaluate it by approximating the sum over the discrete parameters with its largest term and by approximating the integral with Laplace method. However, we propose an alternative which still neglects the sum over the discrete parameters of the modes locations but which now performs the exact computation on the continuous parameters. This approach is more precise and avoids the bias of the BIC criterion (see Section 5). The value of $I_{kj}(u_{kj})$ is finally approximated by (proof is in Appendix B)

$$I_{kj}(u_{kj})\approx\left(\frac{1}{m_j-u_{kj}}\right)^{\bar{n}^\star_{kj}}\prod_{h=1}^{u_{kj}}\frac{Bi\left(\frac{1}{\mathrm{m}_j-h+1};n^\star_{kj(h)}+1;\bar{n}^\star_{kjh}+1\right)}{\mathrm{m}_j-h}, \tag{18}$$

where $\bar{n}^\star_{kjh}=n_k-\sum_{h'=1}^{h}n^\star_{kj(h')}$ and $Bi(x;a,b)=B(1;a,b)-B(x;a,b)$, $B(x;a,b)$ denoting the incomplete beta function defined by $B(x;a,b)=\int_0^x w^a(1-w)^b dw$.

It is not doable to compute the integrated complete-data likelihood for each $(\boldsymbol{s},\boldsymbol{u})$ since the number of competing models is too huge. Thus, the sampling from (10) is performed by an MCMC algorithm, detailed in Appendix C, which has $p(\boldsymbol{s},\boldsymbol{u}|g,\mathbf{x},\mathbf{z})$ as stationary distribution and which also requires the computation of integral I_{kj} defined by (17).

4 Maximum likelihood estimate

When model $\hat{\boldsymbol{m}}$ has been assessed, its parameters $\hat{\boldsymbol{\theta}}_{\hat{\boldsymbol{m}}}$ maximizing the likelihood function have to be obtained

$$\hat{\boldsymbol{\theta}}_{\hat{\boldsymbol{m}}} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x}|\hat{\boldsymbol{m}}, \boldsymbol{\theta}). \tag{19}$$

The direct optimization of the likelihood function involves to solve equations having no analytical solution. So, the parameter estimation is performed via an EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 1997), which is often simple and efficient for missing data. This iterative algorithm alternates between two steps: the computation of the complete-data log-likelihood conditional expectation (E step) and its maximization (M step). Its iteration $[r]$ is written as:
E step: computation of the conditional probabilities $t_{ik}^{[r]} = t_{ik}(\boldsymbol{\theta}^{[r]}|\hat{\boldsymbol{m}})$.
M step: maximization of the complete-data log-likelihood

$$\pi_k^{[r+1]} = \frac{t_{\bullet k}^{[r]}}{n},\ \boldsymbol{\delta}_{kj}^{[r+1]} = \underset{\boldsymbol{\delta}_{kj} \text{ s.c. } \mathrm{card}(\boldsymbol{\delta}_{kj})=\hat{u}_{kj}}{\arg\max} \sum_{h\in\boldsymbol{\delta}_{kj}} \sum_{i=1}^{n} t_{ik}^{[r]} \hat{\mathrm{x}}_i^{jh},$$

$$\alpha_{kjh}^{[r+1]} = \begin{cases} \frac{1}{t_{\bullet k}^{[r]}} \sum_{i=1}^{n} t_{ik}^{[r]} \hat{\mathrm{x}}_i^{j\delta_{kjh}^{[r+1]}} & \text{if } h = 1, \ldots, \hat{u}_{kj} \\ 1 - \sum_{h=1}^{\hat{u}_{kj}} \alpha_{kjh}^{[r+1]} & \text{if } h = \hat{u}_{kj}+1, \end{cases}$$

where $\hat{\mathbf{x}}_i^j = (\hat{\mathrm{x}}_i^{jh}; h = 1, \ldots, \mathrm{m}_j)$ is the variable resulting from the concatenation of the variables affiliated into block j by $\hat{\boldsymbol{s}}$ and where $t_{\bullet k}^{[r]} = \sum_{i=1}^{n} t_{ik}^{[r]}$.

5 Simulations

5.1 Challenge of the mode number selection: Integrated likelihood vs BIC

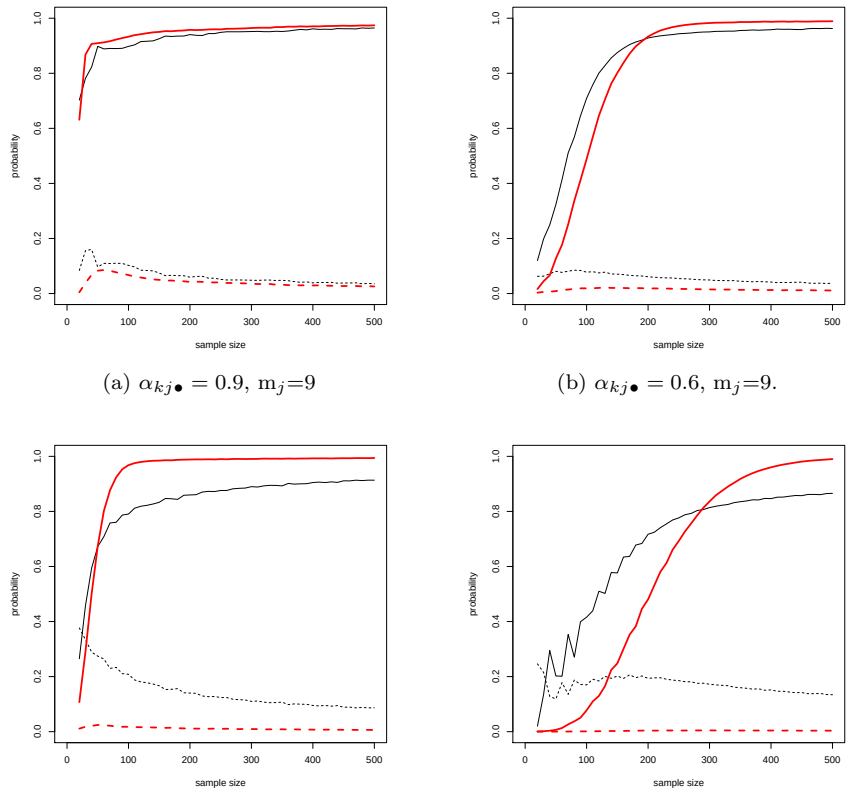
Experiment design This experiment compares two approximations of $I_{kj}(u_{kj})$: the proposed method defined by (18) and the BIC method (Schwarz, 1978) defined by

$$\mathrm{BIC}(I_{kj}(u_{kj})|\mathbf{z}) = \sum_{i=1}^{n}\sum_{k=1}^{g} \mathrm{z}_{ik} \ln\Big(p(\mathbf{x}_i^j|u_{kj}, \boldsymbol{\delta}_{kj}^{\star}, \boldsymbol{\alpha}_{kj}^{\star})\Big) - \frac{u_{kj}}{2}\ln n_k, \tag{20}$$

where $\alpha_{kjh}^{\star} = \frac{1}{n_k}\sum_{i=1}^{n} \mathrm{z}_{ik}\mathrm{x}_i^{j\delta_{kjh}^{\star}}$ if $h = 1, \ldots, u_{kj}$, and $\alpha_{kju_{kj}+1}^{\star} = 1 - \sum_{h=1}^{\hat{u}_{kj}} \alpha_{kjh}^{\star}$. Note the approximations of $I_{kj}(u_{kj})$ should be precise, since this quantity is used in the Gibbs sampler carrying out the model selection.

Samples composed of n i.i.d individuals are drawn from a multinomial distribution per modes $\mathcal{M}_{\mathrm{m}_j}(\frac{\alpha_{kj\bullet}}{3}, \frac{\alpha_{kj\bullet}}{3}, \frac{\alpha_{kj\bullet}}{3}, \frac{1-\alpha_{kj\bullet}}{\mathrm{m}_j-3}, \ldots, \frac{1-\alpha_{kj\bullet}}{\mathrm{m}_j-3})$ with m_j levels and three modes having probability $\frac{\alpha_{kj\bullet}}{3}$. For different sizes of sample, 10^5 samples are generated with different values of $(\alpha_{kj\bullet}, \mathrm{m}_j)$.

Results Figure 1 gives a comparison between (18) and the BIC-like approximation for the selection of the number of modes. The proposed criterion outperforms the BIC criterion in the four studied situations for large sample sizes. Indeed, its asymptotic behavior is better than the BIC criterion, since it rarely overestimates the mode number and since its variability is smaller than the BIC criterion. Since exact criteria are more efficient than approximated criteria (see Biernacki et al. (2010)), this behavior was expected because the proposed criterion is closer to the exact value of the integrated likelihood than the BIC criterion. Indeed, the approximation of the integrated likelihood made by the proposed criterion is only made on the discrete parameters while the BIC criterion performs an approximation on both the discrete and the continuous parameters. We present now more specific comments.



(a) $\alpha_{kj\bullet} = 0.9$, m_j=9

(b) $\alpha_{kj\bullet} = 0.6$, m_j=9.

(c) $\alpha_{kj\bullet} = 0.6$, m_j=18.

(d) $\alpha_{kj\bullet} = 0.3$, m_j=27.

Fig. 1: Probability that the BIC criterion selects the true mode number (plain black line) and overestimates it (dotted black line) and probability that the proposed approach selects the true mode number (plain red line) and overestimates it (dotted red line).

In case (a), modes have a large probability mass and they are easily detected since there are few levels. Both criteria are efficient since they find the true number of modes with a probability close to one even for small samples. When the mode probabilities decrease (case (b)), it is more difficult to identify them. In such a case, the BIC criterion selects more often the true mode number with a moderate overestimation risk, for the small samples (size lower than 150), while the proposed approach can underestimate the number of modes. When the sample size is larger than 200, the proposed approach obtains better results since it finds a true number of modes almost always while the BIC criterion keeps an overestimation risk.

When the number of levels increases (case (c)), the problem is more complex and our approximation of the integrated complete-data likelihood given by (18) shows more clearly its interest. Indeed, the BIC criterion is strongly biased even for a large sample while the proposed approach almost always finds the true number of modes when the sample size is larger than 100. Finally, note that in the more complex situations like in case (d) (few probability mass for the modes and large number of levels), the approximation defined by (18) underestimates the mode number when the sample size is small. However, it converges to the true mode number when the sample size increases whereas the BIC criterion keeps significant bias even for a large data set.

5.2 Estimation accuracy with well specified model

Experiment design During this experiment, we highlight the relevance of the selected model $\hat{\boldsymbol{m}}$ and of its parameter estimate $\hat{\boldsymbol{\theta}}_{\hat{\boldsymbol{m}}}$. For different values of $n = (50, 100, 200, 400, 800)$, 25 data sets of six variables with three levels each are sampled from a bi-component CMM with the following parameters:

$$\boldsymbol{s} = (\{1,2\},\{3,4\},\{5,6\}),\ u_{kj} = 2,\ \forall (k,j)$$

$$\boldsymbol{\pi} = (0.5, 0.5),\ \boldsymbol{\alpha}_{kj} = (0.4, 0.4, 0.2),\ \boldsymbol{\delta}_{1j} = \{2,4\} \text{ and } \boldsymbol{\delta}_{2j} = \{6,8\}.$$

The estimation of the CMM model is performed with the R package CoModes (using the default option). Finally, the computing times are obtained on a 30 cores Intel(R) Xeon(R) CPU E5-4627 v2 @ 3.30GHz.

Results Table 5 presents the simulation results. The model selection works well since the true model is selected with a probability tending to one when n grows toward infinity. Moreover, the Kullback-Leibler divergence vanishes also when the sample size increases. Thus, the estimated distribution converges to the true one.

5.3 Estimation accuracy with misspecified model

Experiment design During this experiment, we underline the robustness of CMM. Thus, 25 samples of size 100 are generated by the following bi-component

n		50	100	200	400	800
Model accuracy	$\hat{\boldsymbol{s}} = \boldsymbol{s}$	100	100	100	100	100
	$\hat{\boldsymbol{u}} = \boldsymbol{u}$	36	60	80	84	90
Parameter accuracy	KL mean	0.38	0.20	0.07	0.03	0.03
	KL sd	0.68	0.23	0.02	0.03	0.01
Time (min)		0.85	0.95	1.26	1.56	1.83

Table 5: Percentage (over the 25 samples) where $\hat{\boldsymbol{s}}$ corresponds to $\boldsymbol{s}$ and where $\hat{\boldsymbol{u}}$ corresponds to $\boldsymbol{u}$. Mean and standard deviation (sd) of the Kullback-Leibler (KL) divergence of the true model from the estimated distribution.

mixture model of dimension six where the intra-class dependencies are different for both components

$$p(\mathbf{x}_i;\boldsymbol{\theta}) = 0.5\prod_{h=1}^{3} p(\boldsymbol{x}_i^{2h-1},\boldsymbol{x}_i^{2h};\boldsymbol{\theta}) + 0.5\, p(\boldsymbol{x}_i^{1};\boldsymbol{\theta})p(\boldsymbol{x}_i^{6};\boldsymbol{\theta})\prod_{h=1}^{2} p(\boldsymbol{x}_i^{2h},\boldsymbol{x}_i^{2h+1};\boldsymbol{\theta}), \tag{21}$$

with $p(\boldsymbol{x}_i^{j},\boldsymbol{x}_i^{j+1};\boldsymbol{\theta}) = p(\boldsymbol{x}_i^{j};\boldsymbol{\theta})\big(\lambda\mathbb{1}_{\{\boldsymbol{x}_i^{j+1}=\boldsymbol{x}_i^{j}\}}+(1-\lambda)p(\boldsymbol{x}_i^{j+1};\boldsymbol{\theta})\big)$ and with $\boldsymbol{x}_i^{j} \sim \mathcal{M}_3(1/3,1/3,1/3)$. Thus, when $\lambda = 0$, the sample is generated by a uniform distribution and classes are confused. The larger is the tuning parameter λ, the larger are the intra-class dependencies and the class separation.

Results Table 6 presents the Kullback-Leibler divergence of the model defined by (21) from CMM and CIM with their best number of classes (with $g_{\max} = 4$).

λ		0.2	0.4	0.6	0.8
CMM	KL-mean	0.08	0.24	0.44	0.64
	KL-sd	0.02	0.01	0.07	0.11
	$g = 1$	0.89	0.78	0.18	0.00
	$g = 2$	0.11	0.22	0.72	0.68
	$g = 3$	0.00	0.00	0.05	0.24
	$g = 4$	0.00	0.00	0.05	0.08
CIM	KL-mean	0.11	0.27	0.59	1.09
	KL-sd	0.02	0.02	0.02	0.15
	$g = 1$	1.00	1.00	0.89	0.44
	$g = 2$	0.00	0.00	0.11	0.56
	$g = 3$	0.00	0.00	0.00	0.00
	$g = 4$	0.00	0.00	0.00	0.00

Table 6: Mean (KL-mean) and standard deviation (KL-sd) of the Kullback-Leibler divergence of the model defined by (21) from CMM and CIM and means of the selected number of classes (g).

The larger is λ, the larger is the Kullback-Leibler divergence for both models. However, the flexibility of CMM allows to keep an acceptable value of the Kullback-Leibler divergence while this divergence grows dramatically faster

with CIM. Furthermore, when the classes are well separated (large value of λ), CMM finds more often the true class number than CIM.

6 Applications

This section illustrates the CMM model on three real data sets. Its results are obtained by the R package CoModes (using the default options) and are compared with the results of the CIM model obtained by the R package Rmixmod (Lebret et al. (2014), using the default options too). Since, the first example is a real clustering problem, the partitioning accuracy of the competing models cannot be evaluated. For the other two data sets, a partition among the individuals is known and is used to compare the partitioning accuracy of the competing models. Finally, the computing times of CMM are obtained on a 30 cores Intel(R) Xeon(R) CPU E5-4627 v2 @ 3.30GHz. The computing times of CIM is not provided since it is very low compared to CMM (CIM is implemented in C++ for Rmixmod, CMM is only implemented in R).

6.1 Alzheimer clustering

Data We consider the Alzheimer data set introduced in the running example (see Section 2). We recall that this data set describes 240 patients with six binary variables (absence:no, presence:yes) indicating relevant symptoms of the Alzheimer disease.

Model comparison Table 7 summarizes the clustering results obtained by the CIM and the CMM models on the six features of the Alzheimer data set. The CMM model obtains a greater value of the BIC criterion than the CIM model independently of the number of components. These results were expected since the CIM model is included in the CMM model. Thus, the CMM fits the data better than the CIM model. Both models select two components but the CMM model also considers intra-component dependencies between the variables *agression* and *diurnal*. Table 8 shows that modelling the intra-component dependencies and introducing parsimonious constraints on the multinomial distributions can strongly impact the partition.

		number of components					
		1	2	3	4	5	6
CIM	BIC	-789.37	**-785.13**	-801.86	-818.02	-834.94	-855.88
CMM	BIC	-779.22	**-772.57**	-785.38	-796.48	-810.81	-815.68
	Time	0.78	**1.16**	1.70	2.11	2.61	2.92

Table 7: Results of CMM and CIM on the Alzheimer data set: BIC criterion (BIC) and computing time in minutes (Time). Values of the best model are in bold.

	CIM-class 1	CIM-class 2
CMM-class 1	84	1
CMM-class 2	51	104

Table 8: Confusion matrix between the partitions resulting from the CMM and the CIM models obtained on the Alzheimer data set.

Best model interpretation The model selected by the BIC criterion is the model described in the running example (see Section 2): *i.e.* bi-component CMM model with $\boldsymbol{s} = (\{1\}, \{2\}, \{3, 5\}, \{4\}, \{6\})$. Its parameters are summarized in Table 9. The model splits the population into two equally balanced classes ($\pi_1 = 0.45$ and $\pi_2 = 0.55$). The class 1 groups the individuals having few symptoms. Indeed, all the modes in this class indicate an absence of symptoms. Note that in this class, two blocks which are composed with a single variable (*activity* and *affective*) follow a uniform distribution. The class 2 contains the individuals presenting more symptoms. Indeed, two modes indicate a presence of symptoms *activity* and *affective*. However, the modes of blocks 1 and 3 are the same for both classes. However, their probabilities in class 2 are smaller than in class 1 (*e.g.* $\alpha_{13\bullet} = 0.82$ and $\alpha_{13\bullet} = 0.44$). Thus, the CMM model overall splits the population between two classes according to the stage of the disease.

Class	Block 1		Block 2		Block 3			Block 4		Block 5	
	Hall	α_{kjh}	Act	α_{kjh}	Agg	Diu	α_{kjh}	Agi	α_{kjh}	Aff	α_{kjh}
1	no	0.94			no	no	0.82	no	0.86		
$\bar{u}_{1j}$ $\alpha_{1j\bullet}$	1.00	0.93	0.00	0.00	0.33		0.82	1.00	0.86	0.00	0.00
2	no	0.90	yes	0.78	no	no	0.44			yes	0.96
$\bar{u}_{2j}$ $\alpha_{2j\bullet}$	1.00	0.90	1.00	0.78	0.33		0.44	0.00	0.00	1.00	0.96

Table 9: Parameters of the best CMM model for the Alzheimer data set. For each class, the first row gives the mode description (names of the levels h of the block j associated to their probabilty α_{kjh}) and the last row gives the indices of the block distribution. We recall that $\bar{u}_{kj}$ and $\alpha_{kj\bullet}$ stand respectively for the complexity and the strength as defined in (4).

6.2 Seabirds clustering

Data The Seabirds data set is a biological data set describing 153 puffins (seabirds) by five plumage and external morphological characteristics presented in Table 10 (Bretagnolle, 2007). These seabirds are divided into three subspecies *dichrous* (84 birds), *lherminieri* (34 birds) and *Subalaris* (35 birds).

variables			levels		
collar	none	trace	dashed	almost continuous (al.cont.)	continuous
eyebrows	none	trace	visible (vis.)	pronounced (pron.)	
sub-caudal	white	black	black&white	BLACK&white	
border	none	few	many		
gender	male	female			

Table 10: Presentation of the five plumage and external morphological variables describing the puffins.

Model comparison Table 11 summarizes the clustering results obtained by the CIM and the CMM models on the five features of the Seabirds data set. Results show that the CMM model allows to fit the data distribution better than the CIM model since it obtains greater values of the BIC criterion. Note that the parameter estimation of the CIM model is computed immediately but that the estimation of the CMM model is performed in a reasonable computing time (less than ten minutes for the whole analysis). Finally, the values of the adjusted Rand Index (Hubert and Arabie, 1985) indicates that the partition resulting from the best CMM model (according to the BIC criterion) is closer to the partition defined by the subspecies than the partition resulting from the best CIM model. Table 12 presents the confusion matrices between the partitions resulting form the best CIM and CMM models and the subspecies. It shows that the *Subalaris* are more different than the two other subspecies. Indeed, both models assign all the *Subalaris* into the same class. Finally note that, the first principal correspondence axis allows to well-visualize the partition provided by the CMM model (see Figure 2).

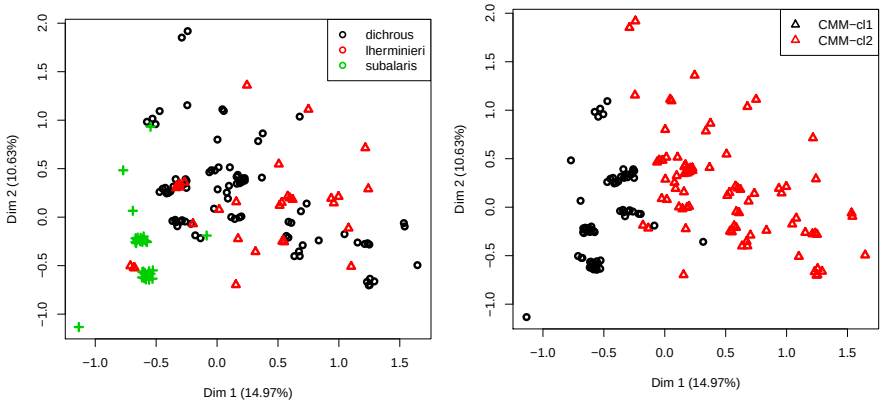
		number of components					
		1	2	3	4	5	6
CIM	BIC	-711.59	**-707.14**	-725.21	-749.64	-782.26	-811.15
	ARI	0.00	**0.13**	0.19	0.30	0.27	0.26
	BIC	-701.98	**-677.37**	-693.74	-697.08	-705.46	-716.05
CMM	ARI	0.00	**0.21**	0.10	0.07	0.04	0.04
	Time	0.56	**0.95**	1.30	1.64	1.93	2.22

Table 11: Results of CMM and CIM on the Seabirds data set: BIC criterion (BIC), adjusted Rand Index (ARI) and computing time in minutes (Time). Values of the best model are in bold.

Best model interpretation The best CMM model and its parameters are displayed in Table 13. Thus, the variables *collar, border* and *gender* are considered as conditionally independent while the intra-component dependencies between the variables *eyebrows* and *sub-caudal* are modelled. Note that the block composed by these last two variables is strongly discriminative, since its modes

	CMM		CIM	
	class 1	class 2	class 1	class 2
Dichrous	25	59	34	50
Lherminieri	11	23	12	22
Subalaris	35	0	35	0

Table 12: Confusion matrices between the subspecies and partitions with respect to the best CMM and CIM models.



(a) Partition defined by the subspecies

(b) Partition resulting from the CMM model.

Fig. 2: Seabirds on the first principal correspondence analysis map where an i.i.d. uniform noise on $[0, 0.1]$ has be added on both axes for each individual to improve visualization.

are different over the components and since they group a large probability mass for both components. Moreover, the variable *gender* follows a uniform distribution under each component (since there is an absence of mode), so it is not discriminative (the variable follows the same distribution over the mixture component). This results was expected since the gender should not provide any discriminative information regarding to the sub-specie of a seabird.

6.3 Acute inflammations clustering

Data The Acute data set (Czerniak and Zarzycki, 2003) describes 120 patients by six features. The first variable is continuous and indicates the temperature of the patient (Temp), so only its integer part is used in order to set into a categorical variable with seven levels (integer from 36 to 42). The remaining five binary variables indicate the occurrence of nausea (Nau), the lumbar pain

Class	Block 1		Block 2			Block 3		Block 4	
	col.	α_{kjh}	eye.	sub.	α_{kjh}	bor.	α_{kjh}	gen.	α_{kjh}
1	none	0.52	vis.	white	0.76	none	0.96		
	dash	0.24	pron.	white	0.10				
	al.cont.	0.21							
$\bar{u}_{1j}$ $\alpha_{1j\bullet}$	0.75	0.97	0.13		0.86	0.50	0.96	0.00	0.00
2	trace	0.46	vis.	black	0.43	none	0.74		
	dash	0.31	trace	white	0.21	few	0.25		
	al.cont.	0.18	trace	black	0.16				
$\bar{u}_{2j}$ $\alpha_{2j\bullet}$	0.75	0.95	0.20		0.80	1.00	0.99	0.00	0.00

Table 13: Parameters of the best CMM model for the seabirds data set. For each class, the first rows give the mode description (names of the levels h of the block j associated to their probabilty α_{kjh}) and the last row gives the indices of the block distribution. We recall that $\bar{u}_{kj}$ and $\alpha_{kj\bullet}$ stand respectively for the complexity and the strength as defined in (4).

(Lum), the urine pushing (Pus), the micturition pains (Mic) and the burning of urethra (Bur). Moreover, many patients have one of the following diseases of the urinary system: inflammation of urinary bladder (Uri) and Nephritis of renal pelvis origin (Ren). Therefore, there are three known partitions among the individuals: two two-group partitions defined by the variables Uri and Ren and one four-group partition defined by the Cartesian production between the variables Uri and Ren designed now by Uri/Ren.

Model comparison Table 14 summarizes the clustering results obtained by the CIM and the CMM models on the six features of the Acute data set. The CMM model fits better the distribution of the observed variables since its BIC criterion is greater than this one of the CIM model. About the partitioning accuracy, both models provide relevant partitions and their interpretations are complementary. The best CIM model produces a partition strongly similar to the partition in four groups defined by the variable *Uri/Ren* (ARI=0.85). On the other hand, the best CMM model produces a partition more similar to the two-group partition defined by the absence/presence of Nephritis of renal pelvis origin (ARI=0.56).

Best model interpretation According to the BIC criterion, the best model is the bi-component CMM model composed of two blocks where the first one only contains the variable *Temp* while the second one contains all the other variables. For both components, the block 1 follows a multinomial distribution with three modes while the block 2 follows a multinomial distribution with six modes. Thus, the model considers within-class dependencies between all variables, *Temp* excluded, and it requires only 19 parameters (while the bi-component CIM model needs 23 parameters). However, since this model has only two blocks, we cannot prove its identifiability (see Appendix A). Nevertheless, this model seems to be identifiable because the runs of the EM algorithm

		number of components					
		1	2	3	4	5	6
CIM	BIC	-623.23	-576.62	-551.37	**-533.46**	-546.89	-565.21
	ARI (Uri)	0.00	0.02	0.47	**0.41**	0.36	0.24
	ARI (Ren)	0.00	0.41	0.28	**0.50**	0.44	0.33
	ARI (Uri/Ren)	0.00	0.24	0.53	**0.85**	0.77	0.59
CMM	BIC	-497.28	**-447.54**	-458.48	-465.14	-474.85	-480.64
	ARI (Uri)	0.00	**0.06**	0.13	0.27	0.36	0.24
	ARI (Ren)	0.00	**0.56**	0.51	0.36	0.44	0.33
	ARI (Uri/Ren)	0.00	**0.33**	0.39	0.51	0.77	0.59
	Time	0.58	0.92	1.22	1.58	2.03	2.24

Table 14: Results of the Acute data set: BIC criterion (BIC), adjusted Rand Index (ARI) and computing time in minutes (Time). Values of the best models according to the BIC value are in bold.

achieving the best value of the log-likelihood function produce systematically the same estimates.

Class	Block 1		Block 2					
	Temp	α_{kjh}	Nau	Lum	Pus	Mic	Bur	α_{kjh}
1	41	0.47	no	yes	yes	no	yes	0.28
$\pi_1 = 0.51$	40	0.35	no	yes	yes	yes	no	0.18
	42	0.08	no	yes	no	yes	no	0.18
			no	no	no	no	no	0.18
			no	yes	yes	yes	yes	0.16
			no	yes	no	no	no	0.02
$\bar{u}_{1j}$ $\alpha_{1j\bullet}$	0.50	0.90			0.19			0.98
2	37	0.45	yes	yes	no	yes	yes	0.29
$\pi_2 = 0.49$	38	0.38	no	yes	no	no	no	0.29
	36	0.15	no	no	yes	no	no	0.15
			yes	yes	no	yes	no	0.15
			no	yes	yes	no	yes	0.08
			no	no	no	no	no	0.02
$\bar{u}_{2j}$ $\alpha_{2j\bullet}$	0.50	0.98			0.19			0.99

Table 15: Parameters of the best CMM model for the Acute data set. For each class, the first rows give the mode description (names of the levels h of the block j associated to their probabilty α_{kjh}) and the last row gives the indices of the block distribution. We recall that $\bar{u}_{kj}$ and $\alpha_{kj\bullet}$ stand respectively for the complexity and the strength as defined in (4).

Table 15 summarizes the block distribution and lists the modes of each block. Based on these results, the class 1 is characterized by individuals with a high temperature, often lumbar pain and no nausea. On the other hand, the class 2 contains only individuals with a temperature equal or less than 38C. Note that the model provides strongly different classes. Indeed, block 1 provides different modes for both classes and the probability mass of the modes

is high ($\alpha_{11\bullet} = 0.90$ and $\alpha_{21\bullet} = 0.98$). Moreover, the distributions of the two components for block 2 is mainly concentrated on the modes ($\alpha_{12\bullet} = 0.98$ and $\alpha_{22\bullet} = 0.99$) and only two modes appear in both components.

It is known (Czerniak and Zarzycki, 2003) that acute nephritis of renal pelvis origin begins with sudden fever, which reaches, and sometimes exceeds 40C. The fever is accompanied by shivers and one or both-side lumbar pains, which are sometimes very strong. Symptoms of acute inflammation of urinary bladder appear very often. Quite not infrequently there are nausea and vomiting and spread pains of whole abdomen. Therefore, the interpretation provided by the bi-component CMM model is relevant. As shown by Table 16, the class 1 mainly groups individuals with acute nephritis of renal pelvis origin.

	CMM	
	class1	class2
Ren - no	10	60
Ren - yes	50	0

Table 16: Confusion matrix between the acute nephritis of renal pelvis origin and the partition resuling from the best model.

7 Conclusion

In this article, we have presented a new mixture model (CMM) to cluster categorical data. Its strength is to relax the conditional independence assumption while staying parsimonious. A summary of the distribution of variables is given by both indices $\bar{u}_{kj}$ (complexity) and $\alpha_{kj\bullet}$ (strength) while each class can be summarized by the mode locations. As shown on the Seabirds application, CMM can improve the results of the classical latent class model even if the conditional independence assumption is true, thanks to its sparsity.

The combinatorial problems of the block detection and of the modes number selection is solved by a hybrid MCMC algorithm which uses the computation of the integrated complete-data likelihood and which does not require estimates. Thus, this approach can be used to select the interactions of the log-linear mixture model per block. The parameters are only estimated for the single selected model. The R package `CoModes` allows to perform the model selection and the parameter estimation. Both data sets presented in this article are included in this package. To efficiently reduce the computing time, the functions of this package will be soon implemented in C++.

However, the model selection becomes difficult if the data set has a large number of variables, since the number of competing models becomes large. Some constraints on the block variables repartition could also be added (for instance the number of variables into blocks could be limited at three variables).

Another solution could be to estimate the model by a forward/backward strategy but it is known that these methods are sub-optimal.

Finally, we imposed the equality of the repartition of the variables into blocks for all the classes. This property allows us to prove the generic identifiability of CMM. This lack of flexibility is counterbalanced by flexible block distribution. However, one could try to relax the class-equality of $\boldsymbol{s}$ with the model non-identifiability risk.

References

Agresti, A. (2002). *Categorical data analysis*, volume 359. John Wiley and Sons.

Allman, E., Matias, C., and Rhodes, J. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132.

Bartholomew, D., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*, volume 899. John Wiley & Sons.

Biernacki, C., Celeux, G., and Govaert, G. (2010). Exact and Monte Carlo calculations of integrated likelihoods for the latent class model. *Journal of Statistical Planning and Inference*, 140(11):2991–3002.

Bretagnolle, V. (2007). Personal communication. *source: Museum.*

Celeux, G. and Govaert, G. (1991). Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8(2):157–176.

Chavent, M., Kuentz, V., and Saracco, J. (2010). A partitioning method for the clustering of categorical variables. In *Classification as a Tool for Research*, pages 91–99. Springer Berlin Heidelberg.

Choirat, C. and Seri, R. (2012). Estimation in discrete parameter models. *Statistical Science*, 27(2):278–293.

Czerniak, J. and Zarzycki, H. (2003). Application of rough sets in the presumptive diagnosis of urinary system diseases. *Artifical Intelligence and Security in Computing Systems, ACS'2002 9th International Conference Proceedings*, pages 41–51.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Espeland, M. and Handelman, S. (1989). Using Latent Class Models to Characterize and Assess Relative Error in Discrete Measurements. *Biometrics*, 45(2):pp. 587–599.

Gollini, I. and Murphy, T. (2014). Mixture of latent trait analyzers for model-based clustering of categorical data. *Statistics and Computing*, 24(4):569–588.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.

Govaert, G. (2010). *Data analysis*, volume 136. John Wiley & Sons.

Hagenaars, J. (1988). Latent structure models with direct effects between indicators local dependence models. *Sociological Methods & Research*, 16(3):379–405.

Hand, D. and Yu, K. (2001). Idiot's Bayes Not So Stupid after All? *International Statistical Review*, 69(3):385–398.

Huang, J., Ng, M., Rong, H., and Li, Z. (2005). Automated variable weighting in k-means type clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(5):657–668.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Jajuga, K., Sokołowski, A., and Bock, H. (2002). *Classification, clustering and data analysis: recent advances and applications*. Springer Verlag.

Jorgensen, M. and Hunt, L. (1996). Mixture model clustering of data sets with categorical and continuous variables. In *Proceedings of the Conference ISIS*, volume 96, pages 375–384.

Kruskal, J. (1976). More factors than subjects, tests and treatments: An indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika*, 41(3):281–293.

Kruskal, J. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138.

Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G., and Govaert, G. (2014). Rmixmod: The R Package of the Model-Based Unsupervised, Supervised and Semi-Supervised Classification Mixmod Library. *Journal of Statistical Software*, in press.

McLachlan, G. and Krishnan, T. (1997). *The EM algorithm*. Wiley Series in Probability and Statistics: Applied Probability and Statistics, Wiley-Interscience, New York.

McLachlan, G. and Peel, D. (2000). *Finite mixutre models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics, Wiley-Interscience, New York.

Moran, M., Walsh, C., Lynch, A., Coen, R., Coakley, D., and Lawlor, B. (2004). Syndromes of behavioural and psychological symptoms in mild alzheimer's disease. *International Journal of Geriatric Psychiatry*, 19(4):359–364.

Qu, Y., Tan, M., and Kutner, M. (1996). Random Effects Models in Latent Class Analysis for Evaluating Accuracy of Diagnostic Tests. *Biometrics*, 52(3):pp. 797–810.

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

Van Hattum, P. and Hoijtink, H. (2009). Market Segmentation Using Brand Strategy Research: Bayesian Inference with Respect to Mixtures of Log-Linear Models. *Journal of Classification*, 26(3):297–328.

Vermunt, J. (2003). Multilevel latent class models. *Sociological methodology*, 33(1):213–239.

A Generic identifiability of CMM

When $d \geq 3$, CMM is generically identifiable (*i.e.* the parameter space where the model is not identifiable has a Lebesgue measure equal to zero). The demonstration is based on results of Allman et al. (2009) which use the Kruskal theorem (Kruskal, 1977, 1976). The demonstration is cut into three steps: reminder of the Kruskal results for the three-way tables, demonstration of the model generic identifiability when $d = 3$, then its extension when model has more than three blocks.

Kruskal results For a matrix M, the Kruskal rank of M, denoted by $\text{rank}_K\ M$ is the largest number I such that every set of I rows of M are linearly independent.
THEOREM 1 (Kruskal (Kruskal, 1977, 1976)). Let $I_j = \text{rank}_K\ M_j$. If

$$I_1 + I_2 + I_3 \geq 2g + 2,$$

then the tensor $[M_1, M_2, M_3]$ uniquely determines the M_j, up to simultaneous permutation and rescaling rows.

Generic identifiability of CMM with three blocks Let $k_0 = \underset{k}{\text{argmin}}\ u_{kj}$ and the matrix M_j where

$$M_j(k, h) = \alpha_{kjh}. \tag{22}$$

By denoting by $\xi_j = \min_k u_{kj} + 1$, generically, we have

$$\text{rank}_K\ M_j = \min(g, \xi_j).$$

COROLLARY 1 The parameters of CMM with three blocs are generically identifiable, up to label swapping, provided:

$$\min(g, \xi_1) + \min(g, \xi_2) + \min(g, \xi_3) \geq 2g + 2.$$

Generic identifiability of CMM with more than three blocks In the same way that Allman et al. (2009), we generalize the result with d blocks by observing that d blocks of categorical variables can be combined into three categorical variables. Thus, we can apply the Kruskal theorem.
COROLLARY 2 We consider a CMM with d blocks where $d \geq 3$. If there exists a tri-partition of the set $\{1, \ldots, d\}$ into three disjoint non empty subsets S_1, S_2 and S_3, such that $\gamma_i = \prod_{j \in S_i} \xi_j$ with

$$\min(g, \gamma_1) + \min(g, \gamma_2) + \min(g, \gamma_3) \geq 2g + 2, \tag{23}$$

then the model parameters are generically identifiable up to label swapping.

B Approximation of $I(u_{kj})$

First, we define a new parametrization of the block distribution facilitating the integrated complete-data likelihood computation and the prior distribution related to this new block parametrization. Second, we underline the relationship between the embedded models. We conclude by the integrated complete-data likelihood computation, which is the target result.

B.1 New parametrization of the block distribution

Without loss of generality, we assume that the elements of $\boldsymbol{\delta}_{kj}$ are ordered by decreasing values of the probability mass associated to them and we introduce the new parametrization of $\boldsymbol{\alpha}_{kj}$ denoted by $\boldsymbol{\varepsilon}_{kj} = (\varepsilon_{kjh}; h = 1, \ldots, u_{kj})$ where $\boldsymbol{\varepsilon}_{kj} \in \mathcal{E}_{kj} = \left]\frac{1}{m_j}; 1\right] \times \ldots \times \left]\frac{1}{\mathrm{m}_j - u_{kj}}; 1\right]$ and where ε_{kjh} is defined by

$$\varepsilon_{kjh} = \begin{cases} \alpha_{kj1} \text{ if } h = 1 \\ \frac{\alpha_{kjh}}{\prod_{h'=1}^{h-1}(1-\varepsilon_{kjh'})} \text{ otherwise.} \end{cases}$$

Each ε_{kjh} follows a truncated beta distribution on the interval $\left]\frac{1}{\mathrm{m}_j - h + 1}, 1\right]$, so

$$p(\boldsymbol{\varepsilon}_{kj}|\boldsymbol{\delta}_{kj}, \boldsymbol{m}) = \frac{\mathrm{m}_j}{\mathrm{m}_j - u_{kj}}. \tag{24}$$

The conditional probability of $\mathbf{x}^j = (\mathbf{x}_i^j; i = 1, \ldots, n)$ is

$$p(\mathbf{x}^j|\mathbf{z}, u_{kj}, \boldsymbol{\delta}_{kj}^\star, \boldsymbol{\varepsilon}_{kj}) = \prod_{h=1}^{u_{kj}} (\varepsilon_{kjh})^{n_{kj(h)}^\star} (1 - \varepsilon_{kjh})^{\bar{n}_{kjh}^\star}. \tag{25}$$

B.2 Relation between embedded models

Let the model with $u_{kj}^\ominus$ modes and the parameters $(\boldsymbol{\delta}_{kj}^{\star\ominus}, \boldsymbol{\varepsilon}_{kj}^\ominus)$ and the model with u_{kj} modes and the parameters $(\boldsymbol{\delta}_{kj}^\star, \boldsymbol{\varepsilon}_{kj})$ such as $u_{kj}^\ominus = u_{kj} - 1$ and such as the $u_{kj}^\ominus$ modes having the largest probabilities have the same locations $(\forall h \in \boldsymbol{\delta}_{kj}^{\star\ominus},\ h \in \boldsymbol{\delta}_{kj}^\star)$ and the same probability masses $(\varepsilon_{kjh}^\ominus = \varepsilon_{kjh},\ h < u_{kj})$. These embedded models follow this relation

$$\frac{p(\mathbf{x}^j|\mathbf{z}, u_{kj}, \boldsymbol{\delta}_{kj}^\star, \boldsymbol{\varepsilon}_{kj})}{p(\mathbf{x}^j|\mathbf{z}, u_{kj}^\ominus, \boldsymbol{\delta}_{kj}^{\star\ominus}, \boldsymbol{\varepsilon}_{kj}^\ominus)} = \frac{(\mathrm{m}_j - u_{kj} + 1)^{\bar{n}_{kju_{kj}-1}^\star - 1}}{(\mathrm{m}_j - u_{kj})^{\bar{n}_{kju_{kj}}^\star}} (\varepsilon_{u_{kj}})^{n_{kj(u_{kj})}^\star} (1 - \varepsilon_{u_{kj}})^{\bar{n}_{kju_{kj}}^\star}. \tag{26}$$

B.3 Integrated complete-data likelihood

The integrated complete-data likelihood is finally approximated, by neglecting the sum over the discrete parameters of the modes locations and by performing the exact computation on the continuous parameters, by

$$I_{kj}(u_{kj}) \approx \left(\frac{1}{m_j - u_{kj}}\right)^{\bar{n}_{kj}^\star} \prod_{h=1}^{u_{kj}} \frac{Bi\left(\frac{1}{\mathrm{m}_j - h + 1}; n_{kj(h)}^\star + 1; \bar{n}_{kjh}^\star + 1\right)}{\mathrm{m}_j - h}, \tag{27}$$

Proof If, for the model with $u_{kj} - 1$ modes, the best modes locations are known and given by $\boldsymbol{\delta}_{kj}^{\star\ominus}$ then the conditional probability of $\mathbf{x}^j$ for a model with u_{kj} modes is

$$p(\mathbf{x}^j|\mathbf{z}, u_{kj}, \boldsymbol{\delta}_{kj}^{\star\ominus}, \boldsymbol{\varepsilon}_{kj}) = \frac{1}{\mathrm{m}_j - u_{kj} + 1} \sum_{\boldsymbol{\delta}_{kj} \in \Delta_{kj}} p(\mathbf{x}^j|\mathbf{z}, u_{kj}, \boldsymbol{\delta}_{kj}, \boldsymbol{\varepsilon}_{kj}), \tag{28}$$

where $\Delta_{kj} = \{\boldsymbol{\delta}_{kj} : \boldsymbol{\delta}_{kj}^{\star\ominus} \subset \boldsymbol{\delta}_{kj} \text{ and } \mathrm{card}(\boldsymbol{\delta}_{kj}) = u_{kj}\}$. Thus, by approximating this sum by its maximum element, we obtain that

$$p(\mathbf{x}^j|\mathbf{z}, u_{kj}, \boldsymbol{\delta}_{kj}^{\star\ominus}, \boldsymbol{\varepsilon}_{kj}) = \frac{1}{\mathrm{m}_j - u_{kj} + 1} p(\mathbf{x}^j|\mathbf{z}, u_{kj}, \boldsymbol{\delta}_{kj}^\star, \boldsymbol{\varepsilon}_{kj}), \tag{29}$$

As $p(\mathbf{x}^j|\mathbf{z}, u_{kj}=0) = (m_j)^{-n_k}$, by applying recursively (26), we obtain that

$$p(\mathbf{x}^j|\mathbf{z}, u_{kj}, \boldsymbol{\varepsilon}_{kj}) \approx \left(\frac{1}{\mathrm{m}_j - u_{kj}}\right)^{\bar{n}^\star_{kju_{kj}}} \prod_{h=1}^{u_{kj}} \frac{(\varepsilon_{kjh})^{n^\star_{kj(h)}}(1-\varepsilon_{kjh})^{\bar{n}^\star_{kjh}}}{\mathrm{m}_j - h + 1}. \tag{30}$$

C Details on the model sampling

The sampling of $(\boldsymbol{s}^{[s+1]}, \boldsymbol{u}^{[s+1]})$ from (10) is performed in two steps. Firstly, a new repartition of the variables into blocks and the mode number of the modified blocks, respectively denoted by $\boldsymbol{s}^{[s+1]}$ and $\boldsymbol{u}^{[s+1/2]}$, are sampled by one iteration of a Metropolis-Hastings algorithm. Secondly, the mode number of each block is sampled by one MCMC iteration. Thus, the sampling of $\boldsymbol{m}^{[s+1]}$ is decomposed into the two following steps

$$(\boldsymbol{s}^{[s+1]}, \boldsymbol{u}^{[s+1/2]}) \sim p(\boldsymbol{s}, \boldsymbol{u}|g, \boldsymbol{s}^{[s]}, \boldsymbol{u}^{[s]}, \mathbf{x}, \mathbf{z}^{[s+1]}) \tag{31}$$

$$\boldsymbol{u}^{[s+1]} \sim p(\boldsymbol{u}|\boldsymbol{s}^{[s+1]}, \boldsymbol{u}^{[s+1/2]}, \mathbf{x}, \mathbf{z}^{[s+1]}). \tag{32}$$

Thus, this chain has $p(\boldsymbol{s}, \boldsymbol{u}|g, \mathbf{x}, \mathbf{z}^{[s+1]})$ as stationary distribution.

C.1 Metropolis-Hastings algorithm to sample from (31)

This sampling is performed by one iteration of the Metropolis-Hastings algorithm divided into two steps. Firstly, the proposal distribution $q(.;\boldsymbol{m}^{[s]})$ generates a candidate $\boldsymbol{m}^\star = (g, \boldsymbol{s}^\star, \boldsymbol{u}^\star)$. Secondly $\boldsymbol{m}^{[s+1]}$ is sampled according to the acceptance probability $\mu^{[s]}$ defined by

$$\mu^{[s]} = 1 \wedge \frac{p(\mathbf{x}, \mathbf{z}^{[s+1]}|\boldsymbol{m}^\star)}{p(\mathbf{x}, \mathbf{z}^{[s+1]}|\boldsymbol{m}^{[s]})} \frac{q(\boldsymbol{m}^{[s]};\boldsymbol{m}^\star)}{q(\boldsymbol{m}^\star;\boldsymbol{m}^{[s]})}. \tag{33}$$

Note that the computation of $\mu^{[s]}$ involves to compute the integrated complete-data likelihood defined by (17). The sampling of $\boldsymbol{m}^{[s+1/2]}$ is written as

$$\boldsymbol{m}^\star \sim q(.;\boldsymbol{m}^{[s]}) \tag{34}$$

$$\boldsymbol{m}^{[s+1/2]} = \begin{cases} \boldsymbol{m}^\star & \text{with a probability } \mu^{[s]} \\ \boldsymbol{m}^{[s]} & \text{otherwise.} \end{cases}$$

The proposal distribution $q(.;\boldsymbol{m}^{[s]})$ samples $\boldsymbol{m}^\star$ in two steps. The first step changes the block affectation of one variable. In practice, $\boldsymbol{s}^\star$ is uniformly sampled in $V(\boldsymbol{s}^{[s]}) = \{\boldsymbol{s} : \exists! b \text{ as } b \in \boldsymbol{s}_j^{[s]} \text{ and } b \notin \boldsymbol{s}_j\}$. The second step uniformly samples the mode numbers among all its possible values for the modified blocks while $u^\star_{kj} = u^{[s]}_{kj}$ for non-modified blocks (i.e. j as $\boldsymbol{s}_j^{[s]} = \boldsymbol{s}_j^\star$).

C.2 MCMC algorithm to sample $\boldsymbol{u}^{[s+1]}$

This step allows to increase or decrease the mode number of each block by one at each iteration. So, $u_{kj}^{[s+1]}$ is sampled according to $p(u_{kj}|\boldsymbol{m}^{[s+1/2]}, \mathbf{x}, \mathbf{z}^{[s+1]})$ defined by

$$p(u_{kj}|g, \boldsymbol{m}^{[s+1]}, u_{kj}^{[s+1/2]}, \mathbf{x}, \mathbf{z}^{[s+1]}) \propto \begin{cases} p(\mathbf{x}^j|\mathbf{z}^{[s+1]}, u_{kj}) & \text{if } |u_{kj} - u_{kj}^{[s+1/2]}| < 2 \\ & \text{and } u_{kj} \notin \{0, \mathrm{m}_j\}. \\ 0 & \text{otherwise.} \end{cases} \tag{35}$$