HAL open science

Model-based clustering for conditionally correlated categorical data

Matthieu Marbac, Christophe Biernacki, Vincent Vandewalle

To cite this version:

Matthieu Marbac, Christophe Biernacki, Vincent Vandewalle. Model-based clustering for conditionally correlated categorical data. [Research Report] RR-8232, 2013, pp.34. hal-00787757v1

HAL Id: hal-00787757

https://inria.hal.science/hal-00787757v1

Submitted on 12 Feb 2013 (v1), last revised 10 Jul 2014 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Model-based clustering for conditionally correlated categorical data

Matthieu Marbac, Christophe Biernacki, Vincent Vandewalle

RESEARCH REPORT
N° 8232
February 12, 2013
Project-Teams MΘDAL

ISSN 0249-6399 ISRN INRIA/RR--8232--FR+ENG

Model-based clustering for conditionally correlated categorical data

Matthieu Marbac*, Christophe Biernacki†, Vincent Vandewalle‡

Project-Teams MΘDAL

Research Report n° 8232 — February 12, 2013 — 34 pages

Abstract: An extension of the latent class model is proposed for clustering categorical data by relaxing the classical class conditional independence assumption of variables. In this model, variables are grouped into inter-independent and intra-dependent blocks in order to consider the main intra-class correlations. The dependence between variables grouped into the same block of a class is taken into account by mixing two extreme distributions, which are respectively the independence and the maximum dependence ones. In the conditionally correlated data case, this approach is expected to reduce biases involved by the latent class model and to produce a meaningful dependency model with few additional parameters. The parameters estimation by maximum likelihood is performed by an EM algorithm while a Gibbs algorithm is used for model selection to avoid combinatorial problems involved by the block structure search. Applications on sociological and biological data sets bring out the proposed model interest. These results strengthen the idea that the proposed model is meaningful and that biases induced by the conditional independence assumption of the latent class model are reduced.

Key-words: Clustering, categorical data, mixture model, correlation, EM algorithm, model selection, Gibbs algorithm.

* DGA & Inria Lille
† CNRS & Inria Lille & University Lille 1
‡ Inria Lille & EA 2694 University Lille 2

RESEARCH CENTRE
LILLE – NORD EUROPE

Parc scientifique de la Haute-Borne
40 avenue Halley - Bât A - Park Plaza
59650 Villeneuve d'Ascq

Classification non supervisée de données catégorielles conditionnellement corrélées

Résumé : Nous proposons une extension du modèle des classes latentes pour la classification non supervisée de données catégorielles conditionnellement corrélées. Dans ce modèle, les variables sont regroupées en blocs inter-indépendants et intra-dépendants dans le but de prendre en compte les principales corrélations intra-classes. La dépendance entre les variables d'un même bloc est prise en compte par un mélange de deux distributions extrêmes, qui sont celles d'indépendance et de dépendance maximale. Dans le cas de données conditionnellement corrélées, on s'attend à ce que cette approche réduise les biais induits par le modèle des classes latentes et qu'il produise un modèle de dépendance facilement interprétable nécessitant peu de paramètres supplémentaires. L'estimation de ces derniers par maximum de vraisemblance est effectuée par un algorithme EM alors qu'un algorithme de Gibbs, permettant de résoudre les problèmes combinatoires dus à la recherche des blocs, est utilisé pour la sélection de modèle. Des applications sur des données sociologiques et biologiques permettent de mettre en avant l'intérêt du modèle proposé. Leurs résultats confortent l'idée que celui-ci est facilement interprétable et qu'il réduit les biais du modèle des classes latentes dus à l'hypothèse d'indépendance conditionnelle.

Mots-clés : Classification automatique, données qualitatives, modèle de mélange, corrélation, algorithme EM, sélection de modèle, algorithme de Gibbs.

1 Introduction

Currently, the practitioner is confronted with an abundance of data (many individuals are described by an increasing number of variables). However, it is difficult for a human to analyze directly the principal information of the data because of their size. In this context, clustering (Gordon [13]; Jajuga *et al.* [17]) provides a partition of individuals (or simultaneously of individuals and variables, Govaert & Nadif [14]) in order to facilitate practitioner analysis. But with the increasing number of variables, the risk of observing correlated descriptors, even within the same class, is growing. The practitioner who is confronted with this phenomenon can then opt for two approaches. The first one is to perform a variables selection step to keep only uncorrelated data, but with the risk of losing some information. The second one is to apply, on the set of variables, a model that takes into account this correlation. In the quantitative case, the Gaussian mixture models (Celeux & Govaert [6]) take into account the intra-class correlation and are particularly informative due to a confidence zone around the class center which can be easily deduced from the parameters. With few meaningful parameters, classes are then summarized so that intra-class correlations are often modeled (Biernacki *et al.* [3]), even when the number of variables is small. In the categorical case, this is the latent class model also known as naive Bayes which is traditionally used (Goodman [12]; Celeux & Govaert [5]). Classes are explicitly described by the probability of each modality for each variable under the conditional independence assumption. The sparsity, caused by this assumption, is a great advantage since it limits the curse of dimensionality but it entails biases when the data are intra-class correlated (partition estimation, class number...). It is necessary to develop a model for qualitative variables that takes into account this correlation in order to have similar clustering tools than those applied on the quantitative data. This is especially important since the categorical variables are often more numerous than the quantitative ones as they are often less informative which increases the risk of presence of intra-class correlation.

Some methods take into account the intra-class correlation as mixtures of Bayesian networks (Cheng & Greiner [7]). Conditionally on each class, a directed acyclic graph is constructed with a set of nodes representing each variable. Even without mixture, the networks estimation is complex if no constraints are added. By constraining the network to be a tree, the model selection and the parameters estimation are easily performed and the correlation modeling has large flexibility. The extension of the dependency tree of Chow & Liu [8] was done by Friedman *et al.* [11] for the supervised classification and by Meila & Jordan [24] for the clustering. However the main problem of these models is that they require too many parameters.

The most general mixture model is the log-linear mixture model as it may consider all forms of interactions (Agresti [1] and Bock [4]). The log-linear models purpose is to model the individuals log-probability by selecting interactions between variables. However this approach suffers from combinatorial problems in model selection since all the variables interactions are allowed. Furthermore, the parameters number becomes too large when the selected interactions number increases. Indeed, the number of models increases exponentially fast with the variables number so it is necessary to impose parsimonious constraints. The latent class model can be seen as a particular log-linear mixture model, where interactions are disregarded. Our aim is to present a version of this model which takes into account the interactions of order one or more, while limiting the number of parameters to be estimated.

We propose to extend the classical latent class model for categorical data (noted CIM for Conditional Independent Model), by a new latent class model which relaxes the variables conditional independence assumption and noted CCM for Conditional Correlated Model. Variables are grouped into blocks conditionally independent given the class. Each block follows a particular dependence distribution which is a bi-component mixture of an independence and a maximal

dependence distribution according to the Cramer's V criterion. This block distribution can be interpreted as a parsimonious version of a two-levels log-linear model. The first level corresponds to group in the same block the variables which are conditionally dependent, so it defines the variables interactions. The strength of the variables correlation is reflected by the proportion of the distribution of maximum dependence compared to that of the independence distribution. The second level of sparsity is done by the constraints of the maximum dependence distribution of the block which limits the allowed interactions. Since the maximum dependence distribution requires only few parameters, the proposed model remains parsimonious, while it provides easily interpretable parameters.

This paper is organized as follows. Section 2 reminds the latent class model principle. Section 3 presents the new mixture model taking into account the intra-class correlations. Section 4 is devoted to its parameter estimation. Section 5 presents a Gibbs algorithm for avoiding combinatorial difficulties inherent to block selection. Section 6 presents results on simulated data and Section 7 presents three applications to real data clustering. A conclusion is given in Section 8.

2 Classical models

2.1 Latent class model: intra-class independence of variables

Observations to be classified are described with d discrete variables $\boldsymbol{x} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^d)$ defined on the probabilistic space $\mathcal{X}$. Each variable j has m_j response levels with $m_j \geq 2$ and is written $\boldsymbol{x}^j = (x^{j1}, \ldots, x^{jm_j})$ where $x^{jh} = 1$ if the variable j takes the modality h and $x^{jh} = 0$ otherwise. In the standard latent class model denoted by CIM (Goodman [12]; Celeux & Govaert [5]), the variables are assumed to be *conditionally independent* knowing the latent cluster. Furthermore data are supposed to arise independently from a mixture of g multivariate multinomial distributions with probability distribution function (pdf):

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \mathring{p}(\boldsymbol{x};\boldsymbol{\alpha}_k) \qquad \text{with} \qquad \mathring{p}(\boldsymbol{x};\boldsymbol{\alpha}_k) = \prod_{j=1}^{d}\prod_{h=1}^{m_j} (\alpha_k^{jh})^{x^{jh}}, \tag{1}$$

with $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_g, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_g)$, π_k being the proportion of the component k in the mixture where $\pi_k > 0$ and $\sum_{k=1}^{g} \pi_k = 1$, and $\boldsymbol{\alpha}_k = (\alpha_k^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$ where α_k^{jh} denotes the probability that the variable j has level h if the object is in cluster k and respects the two following constraints: $\alpha_k^{jh} > 0$ and $\sum_{h=1}^{m_j} \alpha_k^{jh} = 1$.

The classical latent class model is much more parsimonious than the saturated log-linear model, which requires $(\prod_j m_j) - 1$ parameters, since it only requires ν_{CIM} parameters with:

$$\nu_{\text{CIM}} = (g-1) + g\sum_{k=1}^{g}(m_j - 1). \tag{2}$$

Its maximum likelihood estimators are easily computed by an EM algorithm (McLachlan & Krishnan [22]). In the clustering case, the mixture identifiability up to a permutation of the class is generally necessary (McLachlan & Peel [23]). However, there are mixtures, such as the products of Bernoulli distributions, which are not identifiable but produce good results in applications. Thus the notion of generic identifiability has been introduced by Allman *et al.* [2]: a model is generically identifiable if it is identifiable except for a subset of the parameter space with Lebesgue measure zero. This approach reduces the restrictive aspect of the identifiability notion and justifies the applications of this useful model.

2.2 Latent class model extension: intra-class independence of blocks

Despite its simplicity, the latent class model leads to good results in many situations (Hand & Yu [15]; Eaves *et al.* [10]; Keel *et al.* [19]). However, in the case of intra-correlated variables it can lead to severe biases in the partition estimation and also it may overestimate the components number. In order to reduce these biases, a classical extension of the latent class model was introduced by Jorgensen & Hunt [18] for conditionally correlated mixed data. It considers that *conditionally* on the class k, variables are grouped into B_k *independent blocks* and each block follows a specific distribution. The blocks repartition of the variables determines a partition $\boldsymbol{\sigma}_k = (\boldsymbol{\sigma}_{k1}, \ldots, \boldsymbol{\sigma}_{kB_k})$ of $\{1, \ldots, d\}$ in B_k disjoint non-empty subsets where $\boldsymbol{\sigma}_{kb}$ represents the subset b of variables in the partition $\boldsymbol{\sigma}_k$. This partition defines $\boldsymbol{x}^{\{kb\}} = (\boldsymbol{x}^{\{kb\}j}; j = 1, \ldots, d^{\{kb\}})$ which is the variables subset of $\boldsymbol{x}$ associated to $\boldsymbol{\sigma}_{kb}$ where $d^{\{kb\}} = \dim(\boldsymbol{x}^{\{kb\}})$ is the number of variables in the block b of the component k and $\boldsymbol{x}^{\{kb\}j} = (x^{\{kb\}jh}; h = 1, \ldots, m_j^{\{kb\}})$ corresponds to the variable j of the block b for the component k with $x^{\{kb\}jh} = 1$ if the individual takes the modality h for the variable $\boldsymbol{x}^{\{kb\}j}$ and $x^{\{kb\}jh} = 0$ otherwise and where $m_j^{\{kb\}}$ represents the modalities number of $\boldsymbol{x}^{\{kb\}j}$. Note that different variables repartitions in blocks are allowed for each component and they are grouped into $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_g)$. For each component k, each block b follows a specific parametric distribution noted $p(\boldsymbol{x}^{\{kb\}}; \boldsymbol{\theta}_{kb})$ where $\boldsymbol{\theta}_{kb}$ are the parameters of this distribution. The model pdf can be written as:

$$p(\boldsymbol{x}; \boldsymbol{\sigma}, \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}; \boldsymbol{\sigma}_k, \boldsymbol{\theta}_k) \qquad \text{with} \qquad p(\boldsymbol{x}; \boldsymbol{\sigma}_k, \boldsymbol{\theta}_k) = \prod_{b=1}^{B_k} p(\boldsymbol{x}^{\{kb\}}; \boldsymbol{\theta}_{kb}), \tag{3}$$

where $\boldsymbol{\theta}$ is redefined as $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_g, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)$ with $\boldsymbol{\theta}_k = (\boldsymbol{\theta}_{k1}, \ldots, \boldsymbol{\theta}_{kB_k})$. Figure 1 is an example of the distribution with conditional independent blocks for a mixture with two components described by five variables. Blank cells indicate that the intra-class correlation is neglected and black cells indicate that this correlation is taken into account. Note that the classical latent class model with conditional independence, would be represented by white cells off the diagonal and black on the latter.

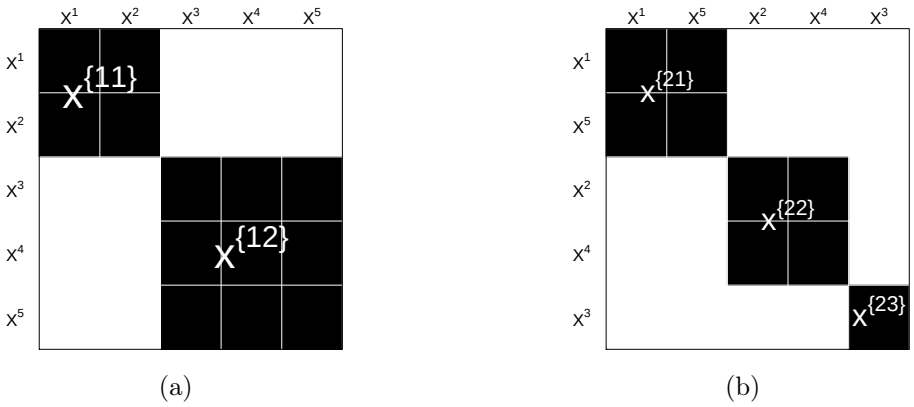


Fig 1: example of the distribution with conditional independent blocks for a mixture with $g = 2$ and $d = 5$, (a) $k = 1$, $B_1 = 2$ and $\boldsymbol{\sigma}_1 = (\{1, 2\}, \{3, 4, 5\})$, (b) $k = 2$, $B_2 = 3$ and $\boldsymbol{\sigma}_2 = (\{1, 5\}, \{2, 4\}, \{3\})$.

This approach is very general, since any distribution can be chosen for each block as soon as it is different from the distribution of independence. The mixture model by conditional independent

blocks is a parsimonious version of the log-linear mixture model. Indeed, the distribution of variables in blocks determines which interactions will be estimated. Interactions between variables of different blocks will be zero and those between variables of the same block can be modeled by the specific distribution of the block. The limiting case of this model where $B_k = d$ for each class is equivalent to the latent class model with the conditional independence assumption.

The generic identifiability of mixture model per conditionally independent blocks is demonstrated, under specific constraints, by using Theorem 4 of Allman *et al.* [2] by assuming that the distribution of each block is itself identifiable. This demonstration is made in Appendix A.

3 Block model extension: intra-block parsimonious distribution

The goal is now to define a parsimonious distribution for each block that takes into account the correlation between variables. Furthermore, the parameters of the block distribution have to be meaningful for the practitioner. In this context, it is proposed to model the distribution of each block by a mixture of the extreme distributions according to the Cramer's V criterion computed on all the couples of variables. It results in a bi-component mixture between an independence distribution and a maximum dependence distribution which can be easily interpreted by the user. The maximum dependence distribution is introduced before to detail this new mixture distribution of each block. The resulting Conditional Correlated Model, denoted CCM, is also defined as a block model extension of the latent class model where the block distribution is modeled by this bi-component mixture.

Remark: without loss of generality, afterwards, the variables are considered as ordered by decreasing number of modalities in each block: $\forall (k,b)\ m_j^{\{kb\}} \geqslant m_{j+1}^{\{kb\}}$ where $j = 1, \ldots, d^{\{kb\}} - 1$.

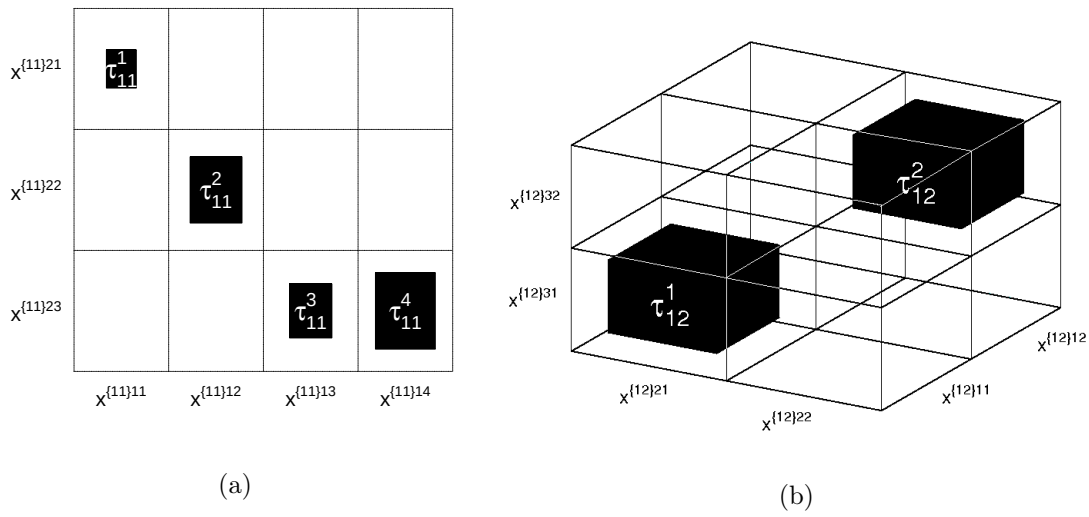
3.1 Maximum dependence distribution

The maximum dependence distribution is defined as the "opposite" distribution of independence according to the Cramer's V criterion computed on all the couples of variables since this latter minimizes this criterion while the maximum dependence distribution maximizes it. Under this distribution, the modality knowledge of one variable provides the maximum information on all the following variables. Note that it is a non-reciprocal functional link between variables. Indeed if $\boldsymbol{x}^{\{kb\}}$ arises from this distribution, the knowledge of the variable having the most number of modalities determines exactly the others but not necessarily the opposite. So this distribution defines successive surjections from the space of $x^{\{kb\}j}$ to the space of $x^{\{kb\}j+1}$ with $j = 1, \ldots, d^{\{kb\}} - 1$ (remind that variables are ordered by decreasing number of modalities in each block). In fact, it is a reciprocal functional link only when $m_j^{\{kb\}} = m_{j+1}^{\{kb\}}$. Since the first variable determines the other ones, this distribution is defined by a product between the multinomial distribution of the first variable parametrized by $\boldsymbol{\tau}_{kb} = (\tau_{kb}^h; h = 1, \ldots, m_1^{\{kb\}})$ with $\tau_{kb}^h \geqslant 0$ and $\sum_{h=1}^{m_1^{\{kb\}}} \tau_{kb}^h = 1$, and the product between the conditional distributions defined as specific multinomial distributions. So, if $x^{\{kb\}1h} = 1$, then $\forall j = 2, \ldots, d^{\{kb\}}$ $\boldsymbol{x}^{\{kb\}j}$ follows a multinomial distribution parametrized by $\boldsymbol{\delta}_{kb}^{hj} = (\delta_{kb}^{hjh'}; h' = 1, \ldots, m_j^{\{kb\}})$ with the following constraints defining the successive surjections: $\delta_{kb}^{hjh'} \in \{0,1\}$, $\sum_{h'=1}^{m_j^{\{kb\}}} \delta_{kb}^{hjh'} = 1$ (multinomial distribution) and $\sum_{h=1}^{m_1^{\{kb\}}} \delta_{kb}^{hjh'} \geq 1$ (surjections). By noting $\boldsymbol{\delta}_{kb} = (\boldsymbol{\delta}_{kb}^{hj}; h = 1, \ldots, m_1^{\{kb\}}; j = 2, \ldots, d^{\{kb\}})$,

the distribution of maximum dependence distribution is then defined as:

$$\acute{p}(\boldsymbol{x}^{\{kb\}};\boldsymbol{\tau}_{kb},\boldsymbol{\delta}_{kb}) = p(\boldsymbol{x}^{\{kb\}1};\boldsymbol{\tau}_{kb}) \prod_{j=2}^{d^{\{kb\}}} p(\boldsymbol{x}^{\{kb\}j}|\boldsymbol{x}^{\{kb\}1};\{\boldsymbol{\delta}_{kb}^{hj}\}_{h=1,\ldots,m_1^{\{kb\}}})$$
$$= \prod_{h=1}^{m_1^{\{kb\}}} \left(\tau_{kb}^h \prod_{j=2}^{d^{\{kb\}}} \prod_{h'=1}^{m_j^{\{kb\}}} (\delta_{kb}^{hjh'})^{x^{\{kb\}jh'}}\right)^{x^{\{kb\}1h}}. \quad (4)$$

Figure 2 shows two examples of the maximum dependence distributions. The probabilities of the joint distribution are represented by the area of gray boxes. Notice that $\boldsymbol{\delta}_{kb}$ defines the position where the probabilities are non zero (location of a dark boxes) and $\boldsymbol{\tau}_{kb}$ defines the probabilities of this non zero cells (area of the dark boxes).



(a) (b)

Fig 2: Two examples of the maximum dependence distributions for the first component of the mixture illustrated by Figure 1(a). (a) The first block is displayed with $m^{\{11\}1} = 4$, $m^{\{11\}2} = 3$, $\delta_{11}^{h1h} = 1$ for $h = 1, 2, 3$, $\delta_{11}^{413} = 1$ and $\boldsymbol{\tau}_{11} = (0.1, 0.3, 0.2, 0.4)$; (b) The second block is displayed with $m^{\{12\}1} = m^{\{12\}2} = m^{\{12\}3} = 2$, $\delta_{12}^{hjh'} = 1$ iff $(h = h')$ and $\boldsymbol{\tau}_{12} = (0.5, 0.5)$.

A sufficient condition of identifiability is to impose $\forall h\ \tau_{kb}^h > 0$. This distribution has very little interest in itself because it is so unrealistic that it can almost never be used alone. We will see in the next section how to use it in a more efficient way.

3.2 New block distribution: mixture of two extreme distributions

It is proposed to model the distribution of each block by a bi-components mixture between an *independence* distribution and a *maximum dependence* distribution. For block b of component k, the block distribution is modeled by:

$$p(\boldsymbol{x}^{\{kb\}};\boldsymbol{\theta}_{kb}) = (1-\rho_{kb})\mathring{p}(\boldsymbol{x}^{\{kb\}};\boldsymbol{\alpha}_{kb}) + \rho_{kb}\acute{p}(\boldsymbol{x}^{\{kb\}};\boldsymbol{\tau}_{kb},\boldsymbol{\delta}_{kb}), \quad (5)$$

where $\boldsymbol{\theta}_{kb} = (\rho_{kb}, \boldsymbol{\alpha}_{kb}, \boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb})$, ρ_{kb} being the proportion of the maximum dependence distribution in this mixture with $0 \leq \rho_{kb} \leq 1$. The proposed model requires little additional parameters

compared with the conditional independence model. In addition, it is easily interpretable as explained in the next paragraph. Note that the limiting case where $\rho_{kb} = 0$ defines the block distribution by the independence one. In this particular case, the parameters of the maximum dependence distribution are no longer defined.

Under this distribution the maximum dependence distribution proportion reflects the deviation from independence under the assumption that the other allowed distribution is the maximum dependence distribution. The parameter ρ_{kb} gives an indicator of the *inter-variables correlation* of the block. It is not here a dependence by pair of variables but a dependence on all variables of the block. Furthermore, it stays bounded when the number of variables is larger than two while the Cramer's V is non upper-bounded in this case. The *intra-variables dependencies* between the variables are defined by $\boldsymbol{\delta}_{kb}$. The strength of these dependencies is explained by $\boldsymbol{\tau}_{kb}$ since it gives the *weight of the over-represented modalities crossing* compared with the independence distribution.

We interpreted before the distribution with independent blocks as a parsimonious version of the log-linear mixture model because it determines the modeled interactions. By choosing the proposed distribution for blocks, a second level of parsimony is added. Indeed, among the interactions allowed by this distribution with independent blocks, only those corresponding to the maximum dependence distribution will be modeled. Other interactions are considered as null.

Properties:

- The whole proposed model, denoted CCM (Conditional Correlated Model), stays parsimonious compared with the latent class model since, for each block with at least two variables, the additional parameters number depends only on the modalities number of the first block variable and not on the number of its variables. By reminding that ν_{CIM} is the number of parameters for the classical latent class model (CIM) defined by Equation (2), the parameters number of the conditional correlated model is denoted ν_{CCM} with:

$$\nu_{\text{CCM}} = \nu_{\text{CIM}} + \sum_{\{(k,b)|d^{\{kb\}}>1\}} m_1^{\{kb\}}. \tag{6}$$

  Table 1 presents the continuous parameters numbers for each model with two components in different situations. Note that the cases 1 and 2 define different block repartitions, however their parameters numbers are equal.

d	m	ν_{CIM}	ν_{CCM} case 1 and 2	ν_{CCM} case 3
4	3	17	23	29
	5	33	43	53
6	3	25	31	37
	5	49	59	69

Table 1: Continuous parameters numbers for two different dimensions and for each three situations are studied where the variables have the same modalities number m: case 1: $B_1 = B_2 = 1$; case 2: $B_1 = B_2 = 2$, $\forall k\ d^{\{k1\}} = d - 1$ and $d^{\{k2\}} = 1$; case 3: $B_1 = B_2 = 2$ and $\forall (k, b)\ d^{\{kb\}} = \frac{d}{2}$.

- The proposed distribution is identifiable under the condition that the block is composed by at least three variables ($d^{\{kb\}} > 2$) or that the modalities number of the last variable of

the block is more than two ($m_2^{\{kb\}} > 2$). This result is demonstrated in Appendix B. The parameter ρ_{kb} is a new indicator allowing to measure the correlation between variables, not limited to correlation by couple of variables. In case where the identifiability conditions could not be met, we distinguish two cases. If $d^{\{kb\}} = 1$, then the block b contains only one variable, the proposed model is reduced to model a multinomial distribution, $\rho_{kb} = 0$ and the maximum dependence distribution is not defined. If $d^{\{kb\}} = 2$ and $m_2^{\{kb\}} = 2$ then, a new constraint is added. In order to have the most meaningful parameters, the chosen value of ρ_{kb} is the largest value maximizing the log-likelihood. This additional constraint does not falsify the definition of ρ_{kb} as an indicator of the dependence strength between the variables of the same block. Furthermore, this constraint is natural since blocks with the biggest dependencies are wanted. Note that ρ_{kb} seems to be correlated with the Cramer's V since simulations illustrate a monotone relationship between both. Figure 3 presents this result by studying two binary variables. Such a behavior has also been observed in many other situations.

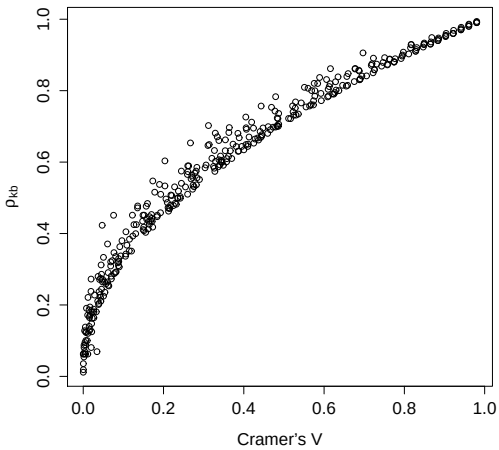


Fig 3: Evolution of ρ_{kb} computed with the identifiability constraint according to the Cramer's V for two binary variables.

4 Parameters estimation

For a fixed model $(g, \boldsymbol{\sigma})$, the parameters have to be estimated. Since the proposed distribution CCM has two latent variables (the classes membership and the intra-block distributions membership), two algorithms derived from the EM algorithm are performed for the associated continuous parameters estimation. The combinatorial problems arising from the discrete parameters estimation are avoided by an algorithm of Metropolis-Hastings.

4.1 Global GEM algorithm

The whole data set of size n composed of independent and identically distributed individuals is denoted as $\mathbf{x} = (\boldsymbol{x}_1, \dots \boldsymbol{x}_n)$ where $\boldsymbol{x}_i \in \mathcal{X}$. The objective is to obtain the maximum log-

likelihood estimator $\hat{\boldsymbol{\theta}}$ defined as (g is now implicit in each expression):

$$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathbf{x}, \boldsymbol{\sigma}) \quad \text{with} \quad L(\boldsymbol{\theta}; \mathbf{x}, \boldsymbol{\sigma}) = \sum_{i=1}^{n} \ln \Big(p(\boldsymbol{x}_i; \boldsymbol{\sigma}, \boldsymbol{\theta}) \Big). \tag{7}$$

The search of maximum likelihood estimates for mixture models leads to solve equations having no analytical solutions. For the mixture models, the assignments of the individuals to the classes can be considered as missing data. This is why the tool generally used is the Expectation-Maximization algorithm (denoted EM algorithm) and its extensions (Dempster *et al.* [9]; McLachlan & Krishnan [22]). Denoting the unknown indicator vectors of the g clusters by $\mathbf{z} = (z_{ik}; i = 1, \ldots, n; k = 1, \ldots, g)$ where $z_{ik} = 1$ if $\boldsymbol{x}_i$ arises from cluster k, $z_{ik} = 0$ otherwise, the mixture model distribution corresponds to the marginal distribution of the random variable $\mathbf{X}$ obtained from the couple distribution of the random variables $(\mathbf{X}, \mathbf{Z})$. In order to maximize the log-likelihood, the EM algorithm uses the complete-data log-likelihood which is defined as:

$$L_{\text{c}}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}, \boldsymbol{\sigma}) = \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \ln \Big(\pi_k p(\boldsymbol{x}_i; \boldsymbol{\sigma}_k, \boldsymbol{\theta}_k) \Big). \tag{8}$$

The EM algorithm is an iterative algorithm which alternates between two steps: the computation of the complete-data log-likelihood conditional expectation (E step) and its maximization (M step). Many algorithms are derived from the EM algorithm and among them the Generalized EM algorithm (GEM) is of interest for us. It works on the same principle as the EM algorithm but the maximization step is replaced by a GM step where the proposed parameters increase the expectation of the complete-data log-likelihood according to its previous value without necessarily maximizing it. We prefer to use the GEM algorithm, since the maximization step in the EM algorithm requires to estimate the continuous parameters for too many possible values of the discrete parameters in order to assure the maximization of the complete-data log-likelihood expectation. Indeed, an exhaustive approach for estimating the discrete parameters is generally impossible when a block contains variables with many modalities and/or many variables, as detailed now. If $S(a, b)$ is the number of possible surjections from a set of cardinal a into a set of cardinal b, then $\boldsymbol{\delta}_{kb}$ is defined in the discrete space of dimension $\prod_{j=1}^{d^{\{kb\}}-1} S(m_j^{\{kb\}}, m_{j+1}^{\{kb\}})$. For example, a block with three variables and $m^{\{kb\}} = (5, 4, 3)$ implies 51 840 possibilities for $\boldsymbol{\delta}_{kb}$. Thus, a stochastic approach is proposed in Section 4.2 to overcome the discrete parameters estimation. Then, the estimation of the continuous parameters conditionally on the discrete parameters is performed by the classical EM algorithm presented in Section 4.3 since their estimation cannot be obtained in closed form.

At the iteration (r), the steps of the global GEM can be written as:

- **$\mathbf{E}_{\text{global}}$ step:** $z_{ik}^{(r)} = \frac{\pi_k^{(r)} p(\boldsymbol{x}_i; \boldsymbol{\sigma}_k, \boldsymbol{\theta}_k^{(r)})}{\sum_{k'=1}^{g} \pi_{k'}^{(r)} p(\boldsymbol{x}_i; \boldsymbol{\sigma}_{k'}, \boldsymbol{\theta}_{k'}^{(r)})}$,
- **$\mathbf{GM}_{\text{global}}$ step:** $\pi_k^{(r+1)} = \frac{n_k^{(r)}}{n}$ where $n_k^{(r)} = \sum_{i=1}^{n} z_{ik}^{(r)}$ and $\forall (k, b)$ $\boldsymbol{\theta}_{kb}^{(r+1)}$ is updated under the constraint that the conditional expectation of complete-data log-likelihood increases (see Sections 4.2 and 4.3).

Initialization of the algorithm: since this algorithm is performed in an stochastic algorithm used for the model selection (see Section 5) and since this latter has an influence on the GEM initialization, this point will be detailed in Section 5.2.

Stopping criterion: the GEM algorithm is stopped after $r_{\max}$ iterations.

Output of the algorithm: we fix $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(r_{\max})}$.

4.2 Detail of the $\mathbf{GM_{global}}$ step of the GEM

The maximization of the expected complete-data log-likelihood is done by optimizing its terms for each (k, b). Thus, the determination of $\boldsymbol{\theta}_{kb}^{(r+1)}$ is performed independently to the parameters of the other blocks. A Metropolis-Hastings algorithm (Robert & Casella [27]) is also performed, for each (k, b), to avoid the combinatorial problems induced by the detection of the discrete parameters $\boldsymbol{\delta}_{kb}$. It performs a random walk over the discrete parameters space and computes the maximum likelihood estimators of continuous parameters $(\rho_{kb}, \boldsymbol{\alpha}_{kb}, \boldsymbol{\tau}_{kb})$ associated with them. This stochastic algorithm allows to find the estimator maximizing the expected complete-data log-likelihood of the block b for the component k:

$$\underset{\boldsymbol{\theta}_{kb}}{\operatorname{argmax}} \sum_{i=1}^{n} z_{ik}^{(r)} \ln p(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\theta}_{kb}). \tag{9}$$

At each iteration (s) of this Metropolis-Hastings algorithm, a discrete parameter noted $\boldsymbol{\delta}_{kb}^{(r,s+\frac{1}{2})}$ is sampled with a uniform distribution in a neighborhood of $\boldsymbol{\delta}_{kb}^{(r,s)}$ noted $\Delta(\boldsymbol{\delta}_{kb}^{(r,s)})$. Then the continuous parameters $(\rho_{kb}^{(r,s+\frac{1}{2})}, \boldsymbol{\alpha}_{kb}^{(r,s+\frac{1}{2})}, \boldsymbol{\tau}_{kb}^{(r,s+\frac{1}{2})})$ are computed, conditionally on the value of $\boldsymbol{\delta}_{kb}^{(r,s+\frac{1}{2})}$, in order to maximize the expected complete-data log-likelihood of the block b for the component k:

$$\sum_{i=1}^{n} z_{ik}^{(r)} \ln p(\boldsymbol{x}_i^{\{kb\}}; \rho_{kb}, \boldsymbol{\alpha}_{kb}, \boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb}^{(r,s+\frac{1}{2})}). \tag{10}$$

The candidate parameters are now denoted as $\boldsymbol{\theta}_{kb}^{(r,s+\frac{1}{2})} = (\rho_{kb}^{(r,s+\frac{1}{2})}, \boldsymbol{\alpha}_{kb}^{(r,s+\frac{1}{2})}, \boldsymbol{\tau}_{kb}^{(r,s+\frac{1}{2})}, \boldsymbol{\delta}_{kb}^{(r,s+\frac{1}{2})})$. The whole block parameters $\boldsymbol{\theta}_{kb}^{(r,s+1)}$ of the next step are then defined as $\boldsymbol{\theta}_{kb}^{(r,s+\frac{1}{2})}$ with the acceptance probability $\mu^{(r,s+1)}$ and $\boldsymbol{\theta}_{kb}^{(r,s)}$ otherwise, where:

$$\mu^{(r,s+1)} = \min \left\{ \frac{\prod_{i=1}^{n} p(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\theta}_{kb}^{(r,s+\frac{1}{2})})^{z_{ik}^{(r)}} |\Delta(\boldsymbol{\delta}_{kb}^{(r,s+\frac{1}{2})})|}{\prod_{i=1}^{n} p(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\theta}_{kb}^{(r,s)})^{z_{ik}^{(r)}} |\Delta(\boldsymbol{\delta}_{kb}^{(r,s)})|}, 1 \right\}, \tag{11}$$

$|\Delta(\boldsymbol{\delta}_{kb}^{(r,s)})|$ denoting the cardinal of $\Delta(\boldsymbol{\delta}_{kb}^{(r,s)})$. Thus, at the iteration (s), the algorithm performs the three following steps:

- **Stochastic step on $\boldsymbol{\delta}_{kb}$:** generate $\boldsymbol{\delta}_{kb}^{(r,s+\frac{1}{2})}$ with a uniform distribution among the elements of $\Delta(\boldsymbol{\delta}_{kb}^{(r,s)})$,
- **Maximization step on the continuous parameters ($\mathbf{M}_{\boldsymbol{\theta}}$ step):** compute the continuous parameters of $\boldsymbol{\theta}_{kb}^{(r,s+\frac{1}{2})}$ (see Section 4.3),
- **Stochastic step on $\boldsymbol{\theta}_{kb}$:** sample $\boldsymbol{\theta}_{kb}^{(r,s+1)} = \begin{cases} \boldsymbol{\theta}_{kb}^{(r,s+\frac{1}{2})} & \text{with probability } \mu^{(r,s+1)} \\ \boldsymbol{\theta}_{kb}^{(r,s)} & \text{otherwise.} \end{cases}$

The neighborhood $\Delta(\boldsymbol{\delta}_{kb}^{(r,s)})$ is defined as the set of the parameters where at most two surjections are different from that of $\boldsymbol{\delta}_{kb}^{(r,s)}$. Figure 4 illustrates this definition.

Initialization of the algorithm: the initialization of the algorithm is done by $\boldsymbol{\theta}_{kb}^{(r+1,0)} = \boldsymbol{\theta}_{kb}^{(r)}$.

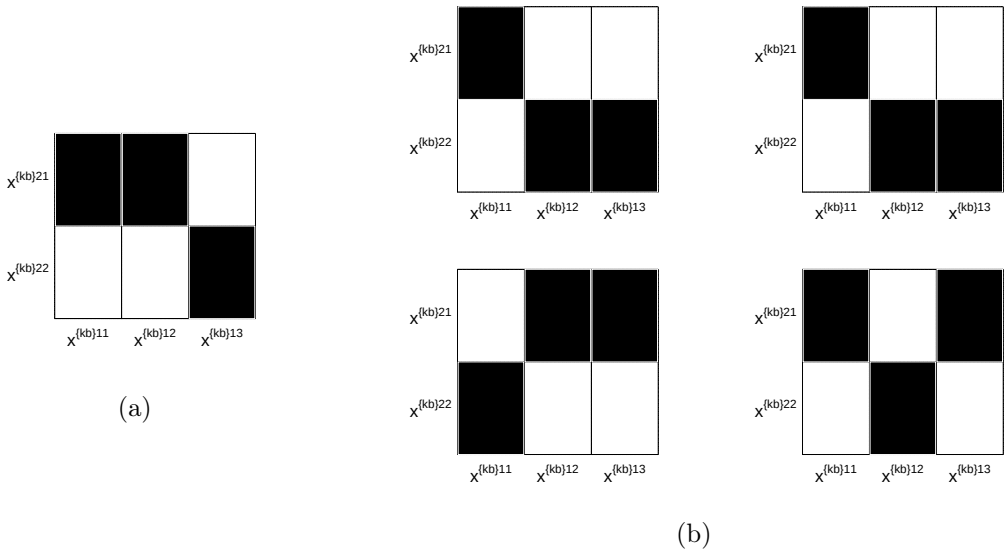


Fig 4: Example of $\Delta(\boldsymbol{\delta}_{kb})$ with $d^{\{kb\}} = 2$ and $\boldsymbol{m}^{\{kb\}} = (3, 2)$. For the row h' and the column h, a black cell indicates that $\delta_{kb}^{h2h'} = 1$ and a white cell that $\delta_{kb}^{h2h'} = 0$: (a) $\boldsymbol{\delta}_{kb}$; (b) $\Delta(\boldsymbol{\delta}_{kb})$.

Stopping criterion: this algorithm is stopped after a number of iterations $s_{\max}$ specified by the user.

Output of the algorithm: at the end of the Metropolis-Hastings algorithm, the best iteration is kept: the estimator $\boldsymbol{\theta}_{kb}^{(r+1)} = \boldsymbol{\theta}_{kb}^{(r+1,\tilde{s})}$ is returned with $\tilde{s} = \underset{s}{\operatorname{argmax}} \sum_{i=1}^{n} z_{ik}^{(r)} \ln p(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\theta}_{kb}^{(r,s)})$. Thus, the proposed initialization ensures the growing of the likelihood at each iteration of the GEM algorithm.

Remark: when the space of possible $\boldsymbol{\delta}_{kb}$ is small (for example when the block groups a small number of binary variables), an exhaustive approach obtains the same results as the proposed algorithm with less computation time. Thus, the retained approach (exhaustive or stochastic) depends on the number of variables and modalities.

4.3 Detail of $\mathrm{M}_{\boldsymbol{\theta}}$ step of the $\mathrm{GM}_{\mathbf{global}}$ step

As there is a second level of mixing, another EM algorithm can be performed for the continuous parameters $(\rho_{kb}, \boldsymbol{\alpha}_{kb}, \boldsymbol{\tau}_{kb})$ estimation by introducing other unknown vectors corresponding to the indicator of the blocks distributions conditionally on $\mathbf{z}$. These vectors are written as $\mathbf{y} = (\mathbf{y}^{\{kb\}}; k = 1, \ldots, g; b = 1, \ldots, B_k)$ with $\mathbf{y}^{\{kb\}} = (y_1^{\{kb\}}, ..., y_n^{\{kb\}})$ where $y_i^{\{kb\}} = 1$ if $\boldsymbol{x}_i^{\{kb\}}$ arises from the *maximum dependence* distribution for the block b of the cluster k and $y_i^{\{kb\}} = 0$ if $\boldsymbol{x}_i^{\{kb\}}$ arises from the *independence* distribution for the block b of the cluster k. The whole mixture model distribution corresponds to the marginal distribution of the random variable $\mathbf{X}$ obtained from the triplet distribution of the random variables $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$. Since the blocks are independent conditionally on $\mathbf{Z}$, the *full* complete-data log-likelihood (both in $\mathbf{Y}$ and $\mathbf{Z}$) is defined

as:

$$L_{\mathrm{c}}^{\mathrm{full}}(\boldsymbol{\theta};\mathbf{x},\mathbf{y},\mathbf{z},\boldsymbol{\sigma}) = \sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik}\Big(\ln\pi_k + \sum_{b=1}^{B_k}\big((1-y_i^{\{kb\}})\ln(1-\rho_{kb}) + (1-y_i^{\{kb\}})\ln \mathring{p}(\boldsymbol{x}_i^{\{kb\}};\boldsymbol{\alpha}_{kb}) + y_i^{\{kb\}}\ln\rho_{kb} + y_i^{\{kb\}}\ln \acute{p}(\boldsymbol{x}_i^{\{kb\}};\boldsymbol{\tau}_{kb},\boldsymbol{\delta}_{kb})\big)\Big). \tag{12}$$

At the iteration (t), the local EM algorithm estimates the continuous parameters of the block b, with fixed values of $\mathbf{z}^{(r)}$ and $\boldsymbol{\delta}_{kb}^{(r,s+\frac{1}{2})}$, by the following two steps:

- **$\mathbf{E}_{\mathbf{local}}$ step:** $y_i^{\{kb\}(r,s+\frac{1}{2},t)} = \frac{\rho_{kb}^{(r,s+\frac{1}{2},t)}\acute{p}(\boldsymbol{x}_i^{\{kb\}};\boldsymbol{\tau}_{kb}^{(r,s+\frac{1}{2},t)},\boldsymbol{\delta}_{kb}^{(r,s+\frac{1}{2})})}{p(\boldsymbol{x}_i^{\{kb\}};\rho_{kb}^{(r,s+\frac{1}{2},t)},\boldsymbol{\alpha}_{kb}^{(r,s+\frac{1}{2},t)},\boldsymbol{\tau}_{kb}^{(r,s+\frac{1}{2},t)},\boldsymbol{\delta}_{kb}^{(r,s+\frac{1}{2})})}$,
- **$\mathbf{M}_{\mathbf{local}}$ step:** $\rho_{kb}^{(r,s+\frac{1}{2},t+1)} = \frac{n_{kb}^{(r,s+\frac{1}{2},t)}}{n_k^{(r)}}$, $\tau_{kb}^{(r,s+\frac{1}{2},t+1)} = \frac{\sum_{i=1}^{n} z_{ik}^{(r)} y_i^{\{kb\}(r,s+\frac{1}{2},t)} x_i^{\{kb\}1h}}{n_{kb}^{(r,s+\frac{1}{2},t)}}$,
  $\alpha_{kb}^{(r,s+\frac{1}{2},t+1)} = \frac{\sum_{i=1}^{n} z_{ik}^{(r)}(1-y_i^{\{kb\}(r,s+\frac{1}{2},t)})x_i^{\{kb\}jh}}{n_k^{(r)}-n_{kb}^{(r,s+\frac{1}{2},t)}}$, where $n_{kb}^{(r,s+\frac{1}{2},t)} = \sum_{i=1}^{n} z_{ik}^{(r)} y_i^{\{kb\}(r,s+\frac{1}{2},t)}$.

Conjecture: during our numerous experiments, we empirically noticed that the log-likelihood function of the mixture between the independence and the maximum dependence distributions had a unique optimum. We conjecture now that this function has an unique maximum.

Initialization of the algorithm: the previous conjecture allows to perform only one initialization of the EM algorithm. This one is fixed to: $(\rho_{kb}^{(r,s+\frac{1}{2},0)}, \boldsymbol{\alpha}_{kb}^{(r,s+\frac{1}{2},0)}, \boldsymbol{\tau}_{kb}^{(r,s+\frac{1}{2},0)}) = (\rho_{kb}^{(r,s)}, \boldsymbol{\alpha}_{kb}^{(r,s)}, \boldsymbol{\tau}_{kb}^{(r,s)})$.

Stopping criterion: this algorithm is stopped after a number of iterations denoted $t_{\max}$.

Output of the algorithm: the algorithm returns the value of block parameters $\boldsymbol{\theta}_{kb}^{(r,s+\frac{1}{2})}$ defined as $\boldsymbol{\theta}_{kb}^{(r,s+\frac{1}{2})} = \boldsymbol{\theta}_{kb}^{(r,s+\frac{1}{2},t_{\max})}$.

Remark: in the specific case where $\boldsymbol{\delta}_{kb}$ are known for each (k,b), the estimation of all the continuous parameters could be performed by a unique EM algorithm where, in the iteration (r), the E step would compute both $\mathbf{z}^{(r)}$ and $\mathbf{y}^{(r)}$ while the M step would estimate all the parameters maximizing the expectation of the *full* complete-data log-likelihood.

5 Model selection

5.1 Gibbs algorithm for exploring the space of models

Since the number of components g determines the dimension of $\boldsymbol{\sigma}$, the model construction is done in two steps: first, the number of components selection and then the determination of the variable repartition per blocks for each component. In a Bayesian context, the best model $(\hat{g},\hat{\boldsymbol{\sigma}})$ is defined as (Robert [26]):

$$(\hat{g},\hat{\boldsymbol{\sigma}}) = \underset{g,\boldsymbol{\sigma}}{\operatorname{argmax}}\ p(g,\boldsymbol{\sigma}|\mathbf{x}). \tag{13}$$

Thus, by considering that $p(g) = \frac{1}{g_{\max}}$ if $g \leq g_{\max}$ and 0 otherwise, where $g_{\max}$ is the maximum number of classes allowed by the user, and by assuming that $p(\boldsymbol{\sigma}|g)$ follows a uniform distribution, the best model is also defined as:

$$(\hat{g}, \hat{\boldsymbol{\sigma}}) = \underset{g}{\operatorname{argmax}} \left[\underset{\boldsymbol{\sigma}}{\operatorname{argmax}}\ p(\mathbf{x}|g, \boldsymbol{\sigma}) \right]. \tag{14}$$

In order to determine $(\hat{g}, \hat{\boldsymbol{\sigma}})$, a Gibbs algorithm is used for estimating $\operatorname{argmax}_{\boldsymbol{\sigma}}\ p(\mathbf{x}|g, \boldsymbol{\sigma})$, for each value of $g \in \{1, \ldots, g_{max}\}$, to avoid the combinatorial problem involved by the detection of the block structure of variables. A reversible jump method could be used (Richardson & Green [25]), however this approach is rarely performed with mixed parameters (continuous and discrete). Indeed, in such a case, it is difficult to define a mapping between the parameters space of two models. This is the reason why we propose to use an easier Gibbs algorithm having $p(\boldsymbol{\sigma}|\mathbf{x}, g)$ as stationary distribution. It alternates between two steps: the generation of a stochastic neighborhood $\Sigma^{[q]}$ conditionally on the current model $\boldsymbol{\sigma}^{[q]}$ by a proposal distribution and the generation of a new pattern $\boldsymbol{\sigma}^{[q+1]}$ included in $\Sigma^{[q]}$ with a probability proportional to its posterior probability. At the iteration $[q]$, it is written as:

- **Neighborhood step:** generate a stochastic neighborhood $\Sigma^{[q]}$ by a proposal distribution given below conditionally on the current model $\boldsymbol{\sigma}^{[q]}$,
- **Pattern step:** $\boldsymbol{\sigma}^{[q+1]} \sim p(\boldsymbol{\sigma}|\mathbf{x}, g, \Sigma^{[q]})$ with $p(\boldsymbol{\sigma}|\mathbf{x}, g, \Sigma^{[q]}) = \begin{cases} \frac{p(\mathbf{x}|g,\boldsymbol{\sigma})}{\sum_{\boldsymbol{\sigma}' \in \Sigma^{[q]}} p(\mathbf{x}|g,\boldsymbol{\sigma}')} & \text{if } \boldsymbol{\sigma} \in \Sigma^{[q]} \\ 0 & \text{otherwise.} \end{cases}$

A possible deterministic neighborhood of $\boldsymbol{\sigma}^{[q]}$ could be defined as the set of models where, at most, one variable is affected, for one component, in another block (possibly creating a new block): $\left\{\boldsymbol{\sigma} : \exists!(k, b, j)\ j \in \boldsymbol{\sigma}^{[q]}_{kb} \text{ and } j \notin \boldsymbol{\sigma}_{kb}\right\} \cup \left\{\boldsymbol{\sigma}^{[q]}\right\}$. However, this deterministic neighborhood can be very large, this is why a proposal distribution allows to reduce it to a stochastic neighborhood $\Sigma^{[q]}$ by reducing the number of (k, b) where $\boldsymbol{\sigma}_{kb}$ could be different to $\boldsymbol{\sigma}^{[q]}_{kb}$. Thus, one component $k^{[q]}$ is randomly sampled in $\{1, \ldots, g\}$ then one block $b^{[q]}_{from}$ is randomly sampled in $\{1, \ldots, B^{[q]}_{k^{[q]}}\}$. Another block $b^{[q]}$ is randomly sampled in $\{1, \ldots, B^{[q]}_{k^{[q]}}\} \setminus b^{[q]}_{from}$ and the set $b^{[q]}_{to} = \{b^{[q]}, B^{[q]}_{k^{[q]}} + 1\}$ is built. The stochastic neighborhood $\Sigma^{[q]}$ is then defined as:

$$\Sigma^{[q]} = \left\{\boldsymbol{\sigma} : \exists!(k, b, j)\ j \in \boldsymbol{\sigma}^{[q]}_{kb},\ j \notin \boldsymbol{\sigma}_{kb} \text{ and } j \in \boldsymbol{\sigma}_{kb'} \text{ with } k = k^{[q]},\ b = b^{[q]}_{\text{from}},\ b' \in b^{[q]}_{\text{to}}\right\} \cup \left\{\boldsymbol{\sigma}^{[q]}\right\}. \tag{15}$$

We denote the elements of $\Sigma^{[q]}$ as $\boldsymbol{\sigma}^{[q+\varepsilon(e)]}$ where $\varepsilon(e) = \frac{e}{|\Sigma^{[q]}|+1}$ and $e = 1, \ldots, |\Sigma^{[q]}|$. Figure 5 shows an illustration of this definition.

At the generation pattern step, the previous algorithm needs the value of $p(\mathbf{x}|g, \boldsymbol{\sigma})\ \forall \boldsymbol{\sigma} \in \Sigma^{[q]}$. By using the BIC approximation (Schwarz [29]; Lebarbier & Mary-Huard [20]), this probability is approximated by:

$$\ln p(\mathbf{x}|g, \boldsymbol{\sigma}) \simeq L(\hat{\boldsymbol{\theta}}; \mathbf{x}, g, \boldsymbol{\sigma}) - \frac{\nu_{\text{CCM}}}{2} \log(n), \tag{16}$$

$\hat{\boldsymbol{\theta}}$ being the maximum likelihood estimator obtained by the GEM algorithm previously described in Section 4. Thus, at the iteration $[q]$, for each $e = 1, \ldots, |\Sigma^{[q]}|$, the estimator $\hat{\boldsymbol{\theta}}^{[q+\varepsilon(e)]}$ associated to the element $\boldsymbol{\sigma}^{[q+\varepsilon(e)]}$ is computed by the GEM algorithm.

Initialization: whatever the initial value selected for $\boldsymbol{\sigma}^{[0]}$, the algorithm converges at the same value of $\boldsymbol{\sigma}$. However, this convergence can be very slow when the initialization is poor. Since blocks are constituted by the most linked variables, a Hierarchical Ascendant Classification

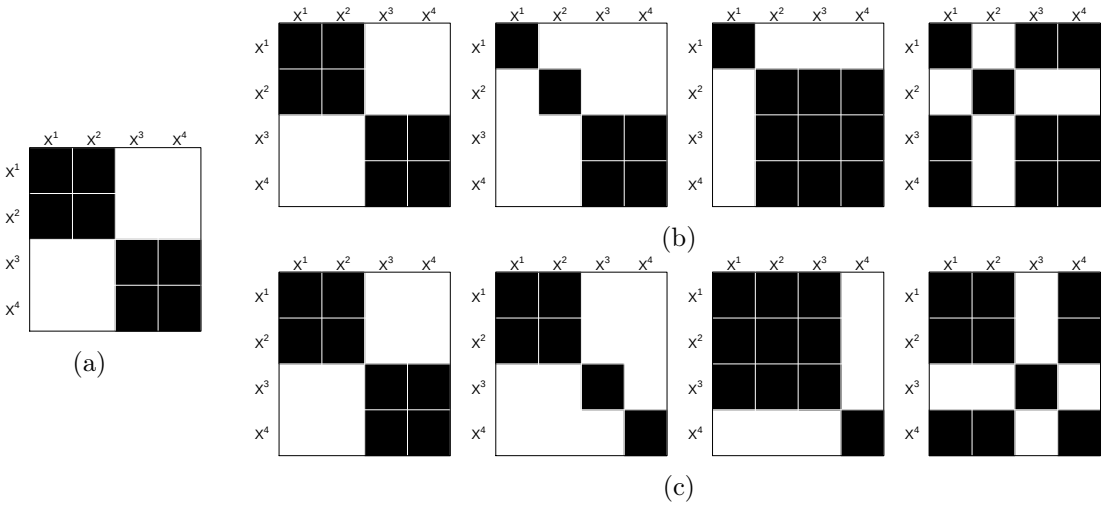


Fig 5: Example of the support of $\Sigma^{[q]}$ in a case of four variables. If the variable of the row j and the variable of the column j' are in the same block then the cell (j, j') is painted in black. This cell is painted in white otherwise. (a) Graphical representation of $\boldsymbol{\sigma}_k^{[q]} = (\{1,2\},\{3,4\})$; (b) Elements of $\Sigma^{[q]}$ if $b_{from}^{[q]} = 1$; (c) Elements of $\Sigma^{[q]}$ if $b_{from}^{[q]} = 2$.

(HAC) is used on the matrix of Cramer's V distances on the couples of variables. The partition produced by the HAC minimizing the block number without blocks holding more than four variables are chosen for each $\boldsymbol{\sigma}_k^{[0]}$. The variables number of a block is limited to four, for the initialization, because very few blocks having more than four variables were exhibited during our experiments. Obviously, the Gibbs algorithm can then modify this setting.

Stopping criterion: the algorithm is stopped when $q_{\max}$ successive iterations have not discovered a better model.

5.2 Consequences of the Gibbs algorithm on the GEM algorithm

Initialization of the GEM algorithm: at the iteration $[q]$ of the Gibbs algorithm, the GEM algorithm estimates $\hat{\boldsymbol{\theta}}^{[q+\varepsilon(e)]}$ associated to the model $\boldsymbol{\sigma}^{[q+\varepsilon(e)]}$ for $e = 1, \ldots, |\Sigma^{[q]}|$. Since these models are closed to $\boldsymbol{\sigma}^{[q]}$, their maximum likelihood estimators should be closed to $\hat{\boldsymbol{\theta}}^{[q]}$. The GEM algorithm initialization is also done by the value of $\hat{\boldsymbol{\theta}}^{[q]}$ for the not modified blocks. Thus, $\boldsymbol{\theta}_{kb}^{[q+\varepsilon(e)](0)} = \hat{\boldsymbol{\theta}}_{kb}^{[q]}$ if the blocks are not modified ($\boldsymbol{\sigma}_{kb}^{[q+\varepsilon(e)]} = \boldsymbol{\sigma}_{kb}^{[q]}$). For the other blocks, the continuous parameters are randomly sampled. For those blocks, in order to avoid the combinatorial problems, we use a sequential method to initialize $\boldsymbol{\delta}_{kb}^{[q+\varepsilon(e)](0)}$: the surjections from $\boldsymbol{x}^{\{kb\}1}$ to $\boldsymbol{x}^{\{kb\}j}$ are sampled, according to $\mathbf{x}$ and to the continuous parameters previously sampled $(\rho_{kb}^{[q+\varepsilon(e)](0)}, \boldsymbol{\alpha}_{kb}^{[q+\varepsilon(e)](0)}, \boldsymbol{\tau}_{kb}^{[q+\varepsilon(e)](0)})$, for each $j = 2, \ldots, d^{\{kb\}}$ as follows:

$$\boldsymbol{\delta}_{kb}^{\cdot j[q+\varepsilon(e)](0)} \propto \prod_{i=1}^{n} p(x_i^{\{kb\}1}, x_i^{\{kb\}j}; \rho_{kb}^{[q+\varepsilon(e)](0)}, \boldsymbol{\alpha}_{kb}^{1[q+\varepsilon(e)](0)}, \boldsymbol{\alpha}_{kb}^{j[q+\varepsilon(e)](0)}, \boldsymbol{\tau}_{kb}^{[q+\varepsilon(e)](0)}, \boldsymbol{\delta}_{kb}^{\cdot j})^{z_{ik}^{[q]}}, \quad (17)$$

where $\boldsymbol{\delta}_{kb}^{\cdot j[q+\varepsilon(e)]} = (\boldsymbol{\delta}_{kb}^{hj[q+\varepsilon(e)]}; h = 1, \ldots, m_1^{\{kb\}})$ and where $z_{ik}^{[q]} = \mathbb{E}\big[\mathbf{Z}\,|\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^{[q]}\big]$.

Remark about $r_{\max}$: as said in Section 4.1, the algorithm is stopped after a fixed number of iterations $r_{\max}$. If the algorithm is stopped before its convergence, the proposed initialization limits the problems. Indeed, if the model has a high *a posteriori* probability, it will stay in the neighborhood $\Sigma^{[q]}$ during some successive iterations, so its log-likelihood will increase.

6 Simulations

Table 2 presents the adjustment parameters values used for all the simulations and applications.

Algorithm	Stopping criterion	Value
Gibbs	successive iterations number without finding a best model	$q_{\max} = 20 \times d$
GEM	iterations number	$r_{\max} = 10$
Metropolis-Hastings	iterations number	$s_{\max} = 1$
EM	iterations number	$t_{\max} = 5$

Table 2: Values of the different stopping criteria.

As these algorithms are interlocked, the iterations number of the most internal algorithms are small. Since the number of possible models increases with d, we propose to fix: $q_{\max} = 20 \times d$. When the best model is selected by the Gibbs algorithm, this latter will stay in this model during lots of iterations so the Metropolis-Hastings and the EM algorithm are performed lots of times. Thus, it is not necessary to have a large iterations number as stopping criterion.

6.1 Study of the algorithm for the $\boldsymbol{\delta}_{kb}$ estimation

In this section, we illustrate the performance of the Metropolis-Hastings algorithm used for the $\boldsymbol{\delta}_{kb}$ estimation (see Section 4.2) and the relevance of its initialization (see Equation (17)). Since this algorithm is interlocked in the Gibbs and in the GEM algorithm, we need that it fastly converges. It is shown in the following simulations that the algorithm stays relevant until six modalities per variables and until six variables per block. These conditions embrace most of the situations.

Samples of size 200 described by variables having the same modalities number are generated by a mixture between an independence distribution and a maximum dependence distribution. The parameters estimation is also performed by the Metropolis-Hastings algorithm, described in Section 4.2, since only one class is generated. The discrete parameters initializations are performed according to Equation (17) with $z_{ik} = 1$ for $i = 1, \ldots, 200$.

Figure 6 shows the box-plots of the iterations number needed by the Metropolis-Hastings algorithm for finding the true links between modalities maximizing the likelihood[1]. Figure 6.(a) illustrates the algorithm behavior according to the modalities number when the samples are described by three variables with the proposed initialization. Figure 6.(b) illustrates the algorithm behavior according to the number of variables, when they have three modalities with the proposed initialization. Figure 6.(c) illustrates the algorithm behavior according to the modalities

[1]In fact, the algorithm is stopped as soon as it finds a discrete estimator involving a likelihood higher than or equal to the likelihood obtained with the true discrete parameters used for the simulation.

number when the samples are described by three variables with a random initialization. Figure 6.(d) illustrates the algorithm behavior according to the number of variables, when they have three modalities with a random initialization.

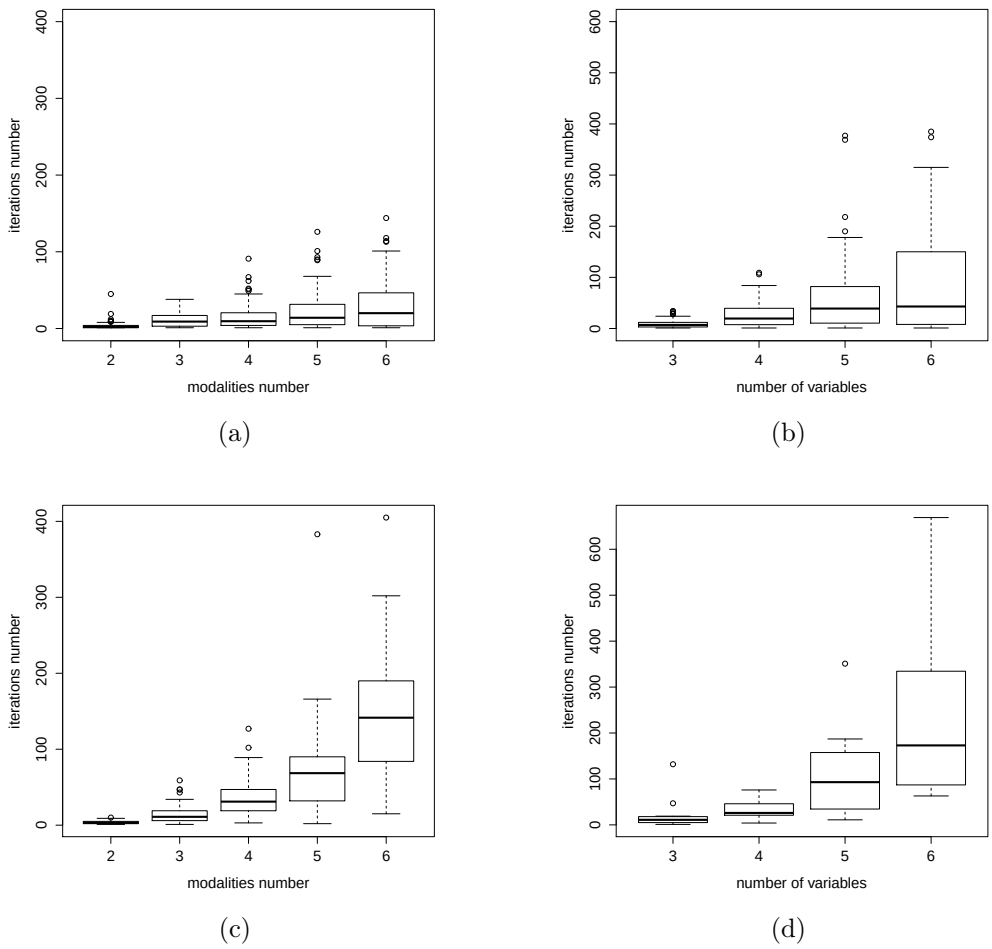


Fig 6: Box-plots of the iterations number needed by the Metropolis-Hastings algorithm for finding the best links between modalities, according to the modalities number when datasets are simulated with a proportion of maximum dependence distribution equal to 0.5. (a) Three variables with the proposed initialization; (b) Three modalities per variables with the proposed initialization; (c) Three variables with a random initialization; (d) Three modalities per variables with a random initialization.

According to these simulations, one can observe that the results of this algorithm are good thanks to its initialization which allows to significantly reduce the iterations number needed to find the maximum likelihood estimators.

6.2 Study of the algorithm for model selection

In order to illustrate the efficiency of the algorithm for the model selection (and also the included estimation process), we want to study the evolution of the Kullback-Leibler divergence

according to the number of variables and to the size of the data set. Thus, 100 samples are generated for many situations according to the CCM with two components. The general model presented in Table 3 is used for the simulations. For each block, in the dependence parameters part, the first line indicates the value of ρ_{kb} and the name of its variables, the other lines present the weight ($\boldsymbol{\tau}_{kb}$) and the kind of the modalities relationships ($\boldsymbol{\delta}_{kb}$). The independence parameters description presents, by column, the value of $\boldsymbol{\alpha}_{kb}^{j}$. Note that the parameter u is introduced for controlling the overlapping of classes: when it is closed to one their overlapping (Bayes error) is closed to one. This parameter fixes the error rate to 0.10 for each studied situations.

Parameters							
Class	Block	Dependence				Independence	
		Strength		Description		Description	
		Variables	Modalities				
		ρ_{kb}	$\boldsymbol{\tau}_{kb}$	$\boldsymbol{\sigma}_{kb}$ and $\boldsymbol{\delta}_{kb}$		$\boldsymbol{\alpha}_{kb}$	
1	b	$0.60(1-u)$		$\boldsymbol{x}_{1b}^{1}$	$\boldsymbol{x}_{1b}^{2}$	$\boldsymbol{\alpha}_{1b}^{1}$	$\boldsymbol{\alpha}_{1b}^{2}$
			0.60	1	1	0.20	0.20
			0.20	2	2	0.20	0.20
			0.20	3	3	0.60	0.60
2	b	$0.60(1-u)$		$\boldsymbol{x}_{2b}^{1}$	$\boldsymbol{x}_{2b}^{2}$	$\boldsymbol{\alpha}_{2b}^{1}$	$\boldsymbol{\alpha}_{2b}^{2}$
			0.60	1	2	$0.075u+0.20(1-u)$	$0.850u+0.20(1-u)$
			0.20	2	3	$0.850u+0.20(1-u)$	$0.075u+0.20(1-u)$
			0.20	3	1	$0.075u+0.60(1-u)$	$0.075u+0.60(1-u)$

Table 3: Parameters used in the simulations.

Table 4 shows the mean and the variance of the Kullback-Leibler divergence between the parameters used for the dataset generation and the estimated parameters according to the number of variables. When n increases, the Kullback-Leibler divergence converges to zero. It confirms the good behavior of the proposed algorithm.

d \ n	100	200	400	800
4	**0.77** (*1.34*)	**0.26** (*0.26*)	**0.15** (*0.05*)	**0.12** (*0.05*)
6	**1.22** (*1.77*)	**0.27** (*0.14*)	**0.09** (*0.07*)	**0.05** (*0.05*)
8	**1.72** (*2.50*)	**0.41** (*0.20*)	**0.09** (*0.05*)	**0.05** (*0.03*)
10	**1.73** (*4.06*)	**0.52** (*0.14*)	**0.10** (*0.03*)	**0.04** (*0.03*)

Table 4: **mean** (*standard deviation*) of the Kullback-Leibler divergence according to the number of variables and to the size of the data set.

7 Applications

In this section, three applications are shown on real data sets. The results obtained by the proposed model are compared to those obtained for the latent class model by the Mixmod software (Biernacki *et al.* [3]). Furthermore, a presentation of the dependence parameters are done for each application in order to show the meaningful aspect of our proposition in

Appendix C. The parameters of the independence distribution are not presented because their analysis is the same as the classical latent class model.

7.1 Classical binary data set: Congressional voting records

This data set presents the votes results of the U.S. House of Representatives Congressmen (Schlimmer [28]) on the 16 key votes identified by the Congressional Quarterly Almanac (handicapped infants *Han*; water project cost sharing *Wat*; adoption of the budget resolution *Bud*; physician fee freeze *Phy*; El Salvador aid *Sal*; religious groups in schools *Rel*; anti satellite test ban *Ant*; aid to Nicaraguan contras *Nic*; mx missile *Mis*; immigration *Imm*; synfuels corporation cutback *Syn*; education spending *Edu*; super-fund right to sue *Sup*; crime *Cri*; duty free exports *Dut*; export administration act South-Africa *Exp*). A clustering analysis is done by hiding the political membership (Democrats and Republicans). The individuals having missing values are removed.

Table 5 compares the results obtained by the classical latent class model (CIM) and by the proposed block model extension (CCM). It indicates the values of the BIC criterion, the continuous parameters numbers and the values of the adjusted Rand criterion (Hubert & Arabie [16]) for both models with different number of classes. This latter criterion allows to compare the quality of the estimated partition with the "true" partition done by the political parties belonging. Furthermore, the error rate is also computed with the true number of classes ($g = 2$).

g	CIM	CCM
1	-2519/16/0	-1960/22/0
2	-1826/33/0.59 Error rate: 0.116	**-1753/43/0.58** Error rate: 0.121
3	**-1789/50/0.50**	-1778/65/0.46
4	-1798/67/0.48	-1798/67/0.48
5	-1821/84/0.42	-1821/84/0.42

Table 5: Results of different models depending on the number of classes for the congressional voting records data set with the BIC criterion. (BIC value / number of parameters / adjusted Rand). Best values according to the BIC criterion are in bold.

The results indicate that the model which is best suited to the data is the CCM because its values of the BIC criterion are always better than them of the CIM. However, when $g \geq 4$, the CCM and the CIM select the same model of conditional independence. It can be explained by the excessive parameters number compared to the individuals number.

When the BIC criterion selects three classes for CIM, it selects only two classes for CCM. It is a bias illustration of the CIM caused by the conditional independence assumption. According to the BIC criterion, the selected parameters better fit the data distribution while it needs less parameters than the selected CIM. Furthermore, the partition obtained by the CCM is closer to the political membership than the partition obtained by the CIM since the adjusted Rand index is better for the first one.

If the number of classes is considered as known ($g = 2$), the BIC criterion is better for CCM than for CIM, however both partitions are roughly the same since the adjusted Rand index and the error rate are very close.

Figure 7 summarizes the results of the best CCM according to the BIC criterion (the dependence parameters are presented in Appendix C.2). In the lines, the estimated classes are

represented with respect to their proportions in decreasing order. Their corresponding area depends on their proportion. The cumulated proportions are indicated on the left side. On the column, three indications are given. The first one is the inter-variables correlations (ρ_{kb}) for all the blocks of the class ordered by their strength of correlation (in decreasing order). The second one is the intra-variables correlations ($\boldsymbol{\tau}_{kb}$) for each block drawn according to their strength dependencies (in decreasing order). The third is the variables repartition per blocks. A black cell indicates that the variable is affected to the block and a white cell indicates that, conditionally on this class, the variable is independent to the variables of this block. For example, this figure shows that the first class has a proportion of 0.53 and it is composed by four blocks. The most correlated block of the first class has $\rho_{kb} \simeq 0.50$ and the strength of the biggest modalities link is closed to 0.80 too. This block is composed of the variables *Wat* and *Sup*.

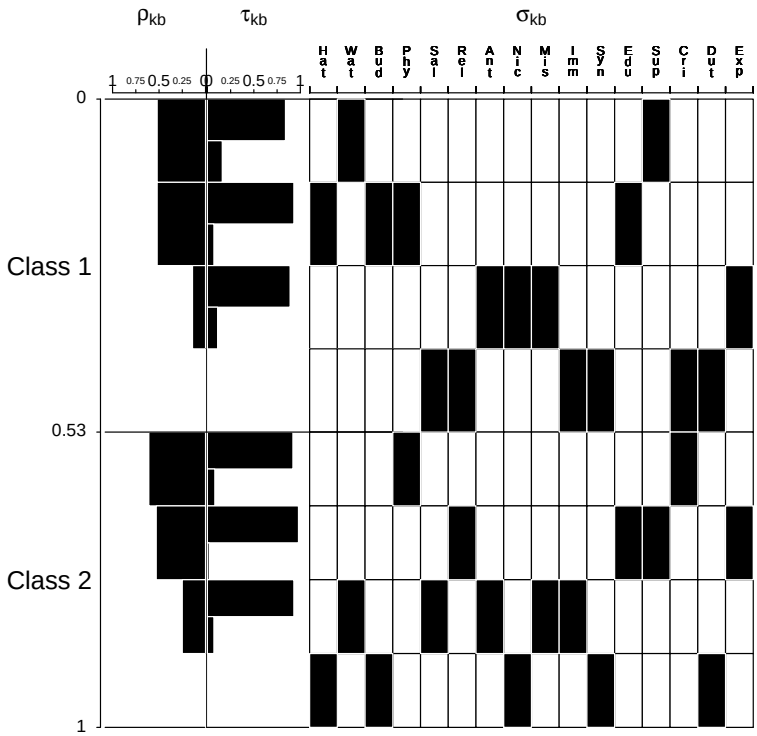


Fig 7: Summary of the best CCM according to BIC for the congressional voting records data set.

In order to put the emphasis on the interpretation possibilities of the proposed model, Table 8 and Table 9 (see Appendix C.1) present the dependence parameters values of the first and second class respectively. However, the description of the variables are not given with this data set, so we cannot analyze the estimated classes. These tables present the partition of variables ordered by the strength of their variables link. For each block, the first line indicates the correlation between the variables and the other lines the links between the modalities ordering by the strength of this link. For example, in Class 1, the most correlated block groups the variables *Water project cost sharing* and *Super-fund right to sue* with a strength of correlation equal to 0.52. The over-represented modalities crossing is the positive answer for both variables with a strength of modalities link equal to 0.84.

7.2 Classical categorical data set: contraceptive method choice

This data set is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey (Lim *et al.* [21]). The samples are 1473 married women who were either not pregnant or do not know if they were at the time of interview. The original problem is to predict the current contraceptive method choice (no use, long-term methods, or short term methods) of a woman based on her demographic and socio-economic characteristics. Each woman is described by nine variables: number of children ever born *Chi* (0, 1, 2, 3, 4, 5 and more), wife's age *WAg* (25 and less, 26-35,36-45, 46 and more), wife's education *WEd* (1=low, 2, 3, 4=high), husband's education *HEd* (1=low, 2, 3, 4=high), husband's occupation *HOc* (1, 2, 3, 4), standard of living index *Liv* (1=low, 2, 3, 4=high), wife's religion *WRe* (Non-Islam or Islam), wife's now working *WWo* (yes or no) and media exposure *Med* (good or not good). For our analysis, the contraceptive method used is blinded, in order to work in a clustering context. Table 6 presents the values of the BIC criterion for the CIM and the CCM. Until four classes, the CCM results are better than them of the CIM. The selection of classes number seems better for the CCM since it selects the "true" number of classes whereas the CIM overestimates it.

g	1	2	3	4	5	6
CIM	-13221	-12566	-12430	-12383	**-12368**	-12410
CCM	-12709	-12378	**-12288**	-12339	-12368	-12410

Table 6: BIC criterion values for both models with different number of classes. Best values according to the BIC criterion are in bold.

Figure 8 summarized the results of the best CCM according to the BIC criterion. It allows to describe the classes by their main features (proportions, intra-class correlations). All the dependence parameters are presented in tables of Appendix C.2. These tables allow a more accurate interpretation of classes since the conditional dependence parameters are presented ($\boldsymbol{\delta}_{kb}$ and $\boldsymbol{\tau}_{kb}$).We present now the classes interpretation for each of the three estimated classes according to the whole dependence parameters.

- **Class 1: young families**
  - **General:** this class proportion is equal to 0.49. There are two dependence blocks and one block of independence.
  - **Block 1:** in this class, the women age and their children number are correlated (ρ_{kb}), with a presence of both extreme situations (young women without child and old women with lots of children explained by both $\boldsymbol{\delta}_{kb}$ and $\boldsymbol{\tau}_{kb}$).
  - **Block 2:** the education level of both members of couple are closed ($\boldsymbol{\delta}_{kb}$) and high education is most present ($\boldsymbol{\tau}_{kb}$).
  - **Block 3:** the practice of Islam is general. The couple members have jobs in category two and three and their living index stays low ($\boldsymbol{\alpha}_{kb}$).
- **Class 2: well-off and not practicing Islam**
  - **General:** this class proportion is equal to 0.37. There are two dependence blocks and one block of independence.

- **Block 1:** there is a strong correlation between the kind of the husband's occupation and the wife's religion (ρ_{kb}). In this class the women practicing Islam have generally a husband with the occupation's level 4 ($\boldsymbol{\delta}_{kb}$ and $\boldsymbol{\tau}_{kb}$).
- **Block 2:** this block shows a link between the number of children and the age of the women. The older are the women, the more they have children ($\boldsymbol{\delta}_{kb}$).
- **Block 3:** in this class both members of the couple have done high studies ($\boldsymbol{\alpha}_{kb}$).

• **Class 3: poor and large families**

- **General:** this class proportion is equal to 0.14. There is one block of independence.
- **Block 1:** this is a class where the number of children is very high (50% of women have at least 5 children). It consists mostly of rather old women with low levels of education, as well as their husbands. They work in groups 2 and 3. The practice of Islam is general. Found in this category all individuals not exposed to the media ($\boldsymbol{\alpha}_{kb}$).

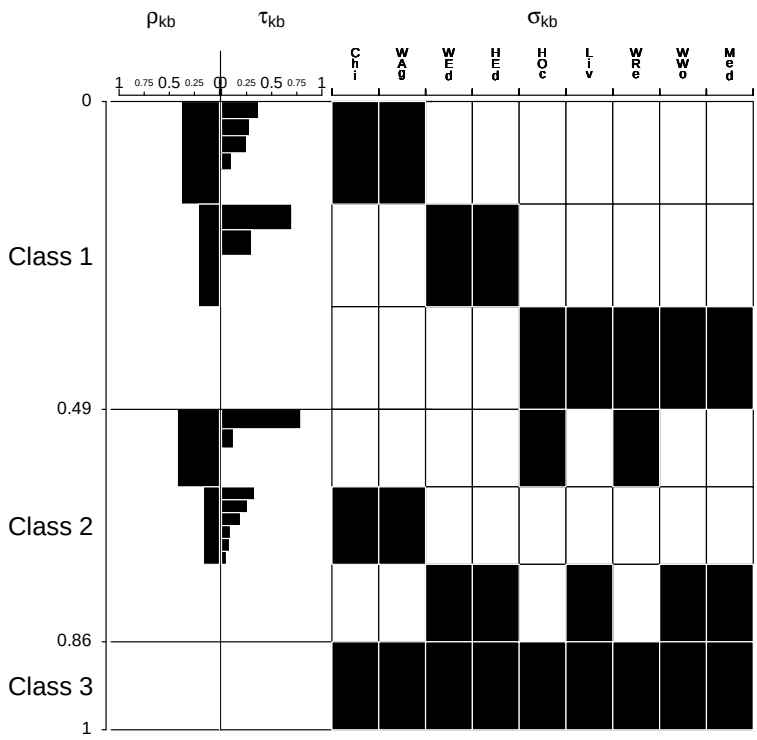


Fig 8: Summary of the best CCM according to BIC for the contraceptive method choice data set.

It is noted that the CCM is more relevant for this data set. Indeed, the number of classes is limited and they are interpretable. In addition, assumption of conditional independence between variables seems too strong for some couples of variable: relationship between age and number of children, relations between the educational level of both members of a couple in a country where caste system is present.

7.3 Real study for improving calves reproduction

The "Genes Diffusion" company has collected informations from the French breeders in order to cluster calves. The 4270 studied calves are described by nine variables of behavior (aptitude for sucking *Apt*, behavior of the mother just before the calving *Iso*) and to the health (treatment against omphalite *TOC*, respiratory disease *TRC* and diarrhea *TDC*, umbilicus disinfection *Dis*, umbilicus emptying *Emp*, mother preventive treatment against respiratory disease *TRM* and diarrhea *TDM*). Table 7 presents the results obtained for the CIM and CCM.

g	1	2	3	4	5	6	7	8
CIM	-28589 17	-26859 35	-26526 53	-26333 71	-26238 89	-26235 107	-26226 125	**-26185** **143**
CCM	-26653 24	-26289 48	-26173 80	-26038 89	**-26025** **112**	-26059 131	-26045 148	-26058 163

Table 7: Results for the CIM and the CMM according to different classes numbers. For both models, first row corresponds to the BIC criterion values and the second row indicates the continuous parameters number. For each model, the best results according to the BIC criterion are in bold.

For the CIM, the BIC criterion selects a high number of classes, since it selected eight classes. The classes interpretation is also difficult and we can assume that the estimators quality is very bad. A description of the five selected classes by CCM ordered by their proportions is done in the following. Figure 9 helps the class interpretation and tables presented in Appendix C.3 detail the dependence parameters allowing the following interpretation.

- **Class 1:**
  - **General:** this class has a proportion equal to 0.29 and it is composed of three blocks of dependence and one block of independence.
  - **Block 1:** there is a strong correlation (ρ_{kb}) between the variables diarrhea treatment of the calve and mother preventive treatment against respiratory disease, especially between the modality no treatment against the calve diarrhea and the absence of preventive treatment against respiratory disease of its mother ($\boldsymbol{\tau_{kb}}$ and $\boldsymbol{\delta_{kb}}$).
  - **Block 2:** there is a strong correlation (ρ_{kb}) between the variables treatment against respiratory illness of the calve and mother preventive treatment against diarrhea, especially between the modality preventive treatment against respiratory illness of the calve and the presence of diarrhea preventive treatment of its mother ($\boldsymbol{\tau_{kb}}$ and $\boldsymbol{\delta_{kb}}$).
  - **Block 3:** there exists another strong link between the behavior of the mother, the emptying of the umbilical and its disinfection ($\boldsymbol{\tau_{kb}}$ and $\boldsymbol{\delta_{kb}}$).
  - **Block 4:** this class is characterized by absence of preventive treatment against omphalite and have 50% of the calves infected by this illness ($\boldsymbol{\alpha}_{kb}$).
- **Class 2:**
  - **General:** this class has a proportion equal to 0.27 and it is composed of two blocks of dependence and one block of independence.
  - **Block 1:** it strongly groups (ρ_{kb}) the clave's treatment against omphalite, the umbilical disinfection and the preventive treatment of the mother against diarrhea.

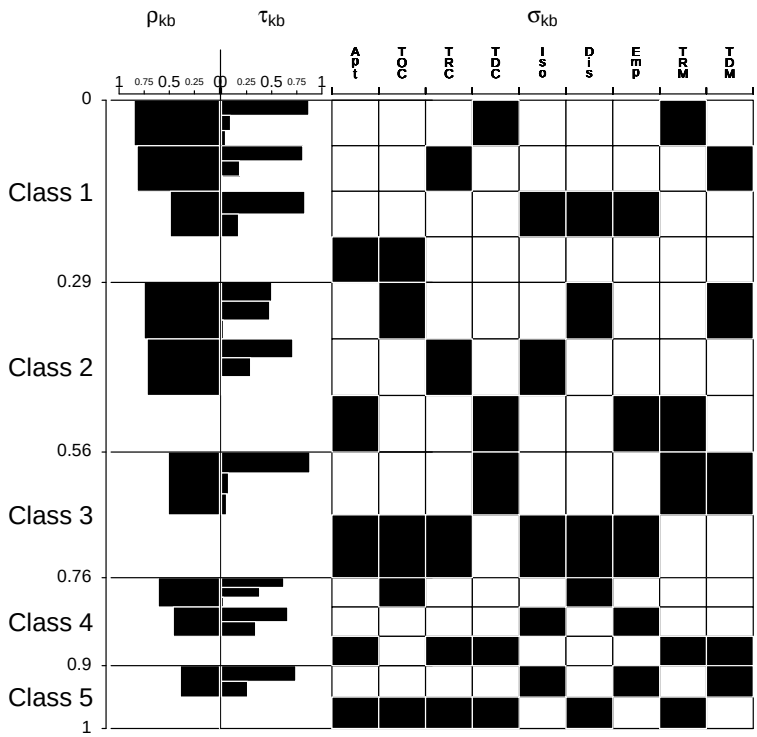


Fig 9: Summary of the best CCM according to BIC for the real study for improving calves reproduction data set.

- **Block 2:** it relates the relationship between the behavior of the mother and the curative treatment of the calve against respiratory illness ($\boldsymbol{\delta}_{kb}$ and $\boldsymbol{\tau}_{kb}$).
- **Block 3:** it is characterized by a strong prevention against the other illness.

- **Class 3:**
  - **General:** this class has a proportion equal to 0.20 and it is composed by one block of dependence and one block of independence.
  - **Block 1:** there exists a correlation between the lack of treatment against diarrhea of the calve and a preventive treatment of its mother against diarrhea but not against respiratory illness (ρ_{kb} and $\boldsymbol{\tau}_{kb}$).
  - **Block 2:** there is a strong prevention against omphalite (preventive treatment, umbilical disinfection and umbilicus emptying). The mother are often isolated and received a preventive treatment against diarrhea ($\boldsymbol{\alpha}_{kb}$).
- **Class 4:**
  - **General:** this class has a proportion equal to 0.14 and it is composed by two blocks of dependence and one block of independence.
  - **Block 1:** there is a link between the absence of preventive treatment against omphalite of the calve and a presence of a preventive treatment against this illness of its mother ($\boldsymbol{\delta}_{kb}$).

- **Block 2:** the isolated mother have calves which need an umbilicus emptying ($\boldsymbol{\delta}_{kb}$ and $\boldsymbol{\alpha}_{kb}$).
- **Block 3:** this class contains the most robust calves since they have no preventive treatment against diarrhea and respiratory illness and it is the same for their mother ($\boldsymbol{\alpha}_{kb}$).

- **Class 5:**
  - **General:** this class has a proportion equal to 0.10 and it is composed by one block of dependence and one block of independence.
  - **Block 1:** there is a correlation between the isolation of the mother, the absence of umbilical emptying and the presence of preventive treatment against diarrhea for the mother (ρ_{kb} and $\boldsymbol{\tau_{kb}}$).
  - **Block 2:** the calves have a preventive treatment only against respiratory illness and their mothers are rarely isolated ($\boldsymbol{\alpha}_{kb}$).

8 Conclusion

By using the block extension of the latent class model, a new mixture model has been proposed for clustering categorical data by taking into account the intra-class correlation. The block distribution is defined as a mixture between an independent distribution and a maximum dependence distribution. This specific distribution stays parsimonious according to the latent class model and allows different levels of interpretation. The blocks of variables enquire about the conditional dependence between variables and its strength is reflected by the proportion of maximum dependence distribution. The parameters of this distribution reflect the links and its strength between modalities.

The parameters estimation and the model selection are simultaneously performed by a Gibbs algorithm. It allows to reduce combinatorial problems of the block structure detection and the links between modalities search for the estimation of the maximum dependence distribution. The results are good when the modalities numbers are small, but when they are more than five, the detection of modalities links involves some persistent difficulties. So the algorithm can be slow in this case.

A code in R/C++ is available on request from the authors. A R package will be available soon. The proposed model can be easily extended to the case of ordinal data. For this, constraints on the dependency structure of each distributions of maximum dependence need to be add.

Acknowledgments: The authors are grateful to Genes Diffusion company for the provision of the data set and especially its members: Amélie Vallée, Julie Hamon and Claude Grenier. We are grateful to Parmeet Bhatia for his precious advices in programming. This work was financed by DGA and Inria.

References

[1] Agresti, A. 2002. *Categorical data analysis.* Vol. 359. John Wiley and Sons.

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

[3] Biernacki, C., Celeux, G., Govaert, G., & Langrognet, F. 2006. Model-based cluster and discriminant analysis with the MIXMOD software. *Computational Statistics & Data Analysis*, **51**(2), 587–600.

[4] Bock, HH. 1986. Loglinear models and entropy clustering methods for qualitative data. *Classification as a tool of research. North Holland, Amsterdam*, 19–26.

[5] Celeux, G., & Govaert, G. 1991. Clustering criteria for discrete data and latent class models. *Journal of classification*, **8**(2), 157–176.

[6] Celeux, G., & Govaert, G. 1995. Gaussian parsimonious clustering models. *Pattern Recognition*, **28**(5), 781–793.

[7] Cheng, J., & Greiner, R. 1999. Comparing Bayesian network classifiers. *Pages 101–108 of: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc.

[8] Chow, C., & Liu, C. 1968. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, **14**(3), 462–467.

[9] Dempster, A.P., Laird, N.M., & Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

[10] Eaves, L.J., Silberg, J.L., Hewitt, J.K., Rutter, M., Meyer, J.M., Neale, M.C., & Pickles, A. 1993. Analyzing twin resemblance in multisymptom data: genetic applications of a latent class model for symptoms of conduct disorder in juvenile boys. *Behavior Genetics*, **23**(1), 5–19.

[11] Friedman, N., Geiger, D., & Goldszmidt, M. 1997. Bayesian network classifiers. *Machine learning*, **29**(2), 131–163.

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

[13] Gordon, A.D. 1981. Classification. Methods for the exploratory analysis of multivariate data. *Monographs on Applied Probability and Statistics. London: New York, Chapman and Hall xii, 193p.*

[14] Govaert, G., & Nadif, M. 2003. Clustering with block mixture models. *Pattern Recognition*, **36**(2), 463–473.

[15] Hand, D.J., & Yu, K. 2001. Idiot's Bayes—Not So Stupid after All? *International Statistical Review*, **69**(3), 385–398.

[16] Hubert, L., & Arabie, P. 1985. Comparing partitions. *Journal of classification*, **2**(1), 193–218.

[17] Jajuga, K., Sokołowski, A., & Bock, H.H. 2002. *Classification, clustering and data analysis: recent advances and applications*. Springer Verlag.

[18] Jorgensen, M., & Hunt, L. 1996. Mixture model clustering of data sets with categorical and continuous variables. *Proceedings of the Conference ISIS*, **96**, 375–384.

[19] Keel, P.K., Fichter, M., Quadflieg, N., Bulik, C.M., Baxter, M.G., Thornton, L., Halmi, K.A., Kaplan, A.S., Strober, M., Woodside, D.B., *et al.* . 2004. Application of a latent class analysis to empirically define eating disorder phenotypes. *Archives of General Psychiatry*, **61**(2), 192.

[20] Lebarbier, E., & Mary-Huard, T. 2006. Une introduction au critère BIC : fondements théoriques et interprétation. *Journal de la SFdS*, **147**(1), 39–57.

[21] Lim, T.S., Loh, W.Y., & Shih, Y.S. 2000. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning*, **40**(3), 203–228.

[22] McLachlan, G.J., & Krishnan, T. 1997. *The EM algorithm.* Wiley-Interscience, New York: Wiley Series in Probability and Statistics: Applied Probability and Statistics.

[23] McLachlan, G.J., & Peel, D. 2000. *Finite mixutre models.* Wiley-Interscience, New York: Wiley Series in Probability and Statistics: Applied Probability and Statistics.

[24] Meila, M., & Jordan, M.I. 2001. Learning with mixtures of trees. *The Journal of Machine Learning Research*, **1**, 1–48.

[25] Richardson, S., & Green, P.J. 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59**(4), 731–792.

[26] Robert, C.P. 2005. *Le choix bayesien: principes et pratique.* Springer France Editions.

[27] Robert, C.P., & Casella, G. 2004. *Monte Carlo statistical methods.* Springer Verlag.

[28] Schlimmer, Jeffrey Curtis. 1987. *Concept acquisition through representational adjustment.* Ph.D. thesis, University of California, Irvine.

[29] Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Appendices

A Identifiability of the mixture model of distributions per independent blocks

Main idea:

In order to prove the generic identifiability (denoted by GI in the following) of the mixture model per independent blocks, we use Theorem 4 of Allman *et al.* [2]. This theorem demonstrates the GI of the conditional independent mixture model. Its demonstration needs to built a tri-partition of the variables (verified since $d \geq 3$). To applied the previous result on the mixture model per independent blocks, we use conditional independence between blocks by added some constraint on the tri-partition construction: if two variables are grouped in the same block for one component, then these two variables have to be in the same subset of the tri-partition. The following example illustrates this constraint by representing a mixture model per independent blocks with two components and five variables: in Case (a) it is possible to create a tri-partition (denoted by S) according to the block repartition of the variables, in Case (b) this is not possible.

$\boldsymbol{\sigma}_1 = (\{1,2\},\{3,4\},\{5\})$	$\boldsymbol{\sigma}_1 = (\{1,2\},\{3,4\},\{5\})$
$\boldsymbol{\sigma}_2 = (\{1\},\{2\},\{3,4\},\{5\})$	$\boldsymbol{\sigma}_2 = (\{1,3\},\{2,4\},\{5\})$
$S = (\{1,2\},\{3,4\},\{5\})$	no tri-partition
Case (a)	Case (b)

Three sufficient conditions:

The three sufficient conditions to GI are now described:

- *C1*: the block distribution is GI,
- *C2*: there exists a tri-partition of $\boldsymbol{\sigma}_k$, identical for each k, into three disjoint non-empty subsets S_1, S_2, S_3:

$$\forall k \in \{1, \dots, g\},\ \forall \boldsymbol{\sigma}_{kb} \in \boldsymbol{\sigma}_k,\ \exists u \in \{1, 2, 3\} \text{ as } \boldsymbol{\sigma}_{kb} \in S_u,$$

- *C3*: by denoting κ_u the dimension of the space generated by the variables associated to the partition S_u with $u \in \{1, 2, 3\}$, $\kappa_u = \prod_{j \in S_u} m_j$, the condition is:

$$\min(g, \kappa_1) + \min(g, \kappa_2) + \min(g, \kappa_3) \geq 2g + 2.$$

Influence of the condition C2:

We have to add the condition *C2* to *C1* and *C3* already present in the Allman's theorem. However this condition is not very restrictive because if there are more than two variables ($d \geq 3$) and that it exists at least two variables which are conditionally independent for each class ($\exists (j_1, j_2),\ j_1 \neq j_2$ and $\forall j \in \{j_1, j_2\}\ \forall (k, b)$ as $j \in \boldsymbol{\sigma}_{kb}$ with $\rho_{kb} = 0$), then *C2* is verified. Furthermore, if $\boldsymbol{\sigma}_k = \boldsymbol{\sigma}_{k'}\ \forall (k, k')$ and $B_k \geq 3$, then the application conditions of the theorem is verified. In case where *C2* is not validate, then the theorem can not be applied. However, the proposed mixture model seems generically identifiable in practice since, during our numerous simulations, we did not encounter cases of non-identifiability when *C1* is validate and that there are at least three conditionally independent variables per class. We also note that the case where $B_k = 1$ for each k is not GI if the block distribution is any mixture model. So, our proposed block distribution involved that the model is not GI in this case.

B Identifiability of the specific block distribution

In this section, we demonstrate that the block distribution (mixture between independence and maximum dependence distributions) is GI when the number of the block variables or when the modalities number of the second variable of the block is at least equal to three. For remind, a distribution $p(.; \boldsymbol{\theta})$ is GI when the set $\left\{\boldsymbol{\theta} :\ \exists \tilde{\boldsymbol{\theta}} \text{ as } \forall \boldsymbol{x}\ p(\boldsymbol{x}; \boldsymbol{\theta}) = p(\boldsymbol{x}; \tilde{\boldsymbol{\theta}})\right\}$ has a Lebesgue measure equal to zero. In Section B.1, we show that this distribution is GI for two tri-modalities variables and for three binary variables. In Section B.2, we show that if a model is GI then it stays GI by addition of modalities or variables. In order to simplify the notations, we consider this distribution:

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = (1 - \rho) \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\alpha^{jh})^{x^{jh}} + \rho \sum_{h=1}^{m_1} \tau^h \mathbb{1}_{\{\boldsymbol{x} = \boldsymbol{\delta}^h\}} \qquad \text{where } \boldsymbol{\theta} = (\rho, \boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{\delta}). \tag{18}$$

Thus, Figure 10 illustrates the proposed reasoning: the cross ($d = 2$ and $m_d = 2$) means that the model is not GI, the two triangles represent the two situations where the GI is demonstrated in Section B.1 ($d = 2$ and $m_d = 3$; $d = 3$ and $m_d = 2$) and the dots mean that the models are GI since they can be built by addition of variables and/or modalities from one of the two models described in Section B.1.

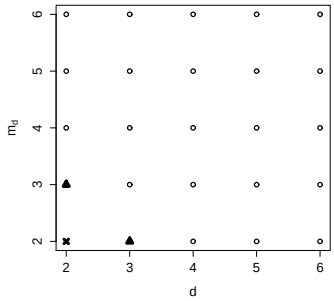


Fig 10: Generic identifiability of the block distribution according to the variables number and to the number of modalities of the last variable. The meanings of symbols are given in the text.

B.1 Generic identifiability of the both basic cases

In order to prove the generic identifiability of the proposed distribution when $d = 2$ and $\boldsymbol{m} = (3, 3)$, two situations are studied: when the dependence relations between modalities involved by the maximum dependence distribution are the same and when they are different. The same reasoning can be used for the case $d = 3$ and $\boldsymbol{m} = (2, 2, 2)$. Two discrete parameters defining a maximal dependence distribution are shown in Figure 11 for reminding the role of these discrete parameters (black cell indicates that the modalities crossing is allowed by the maximal dependence distribution and white cell indicates that it is not allowed).

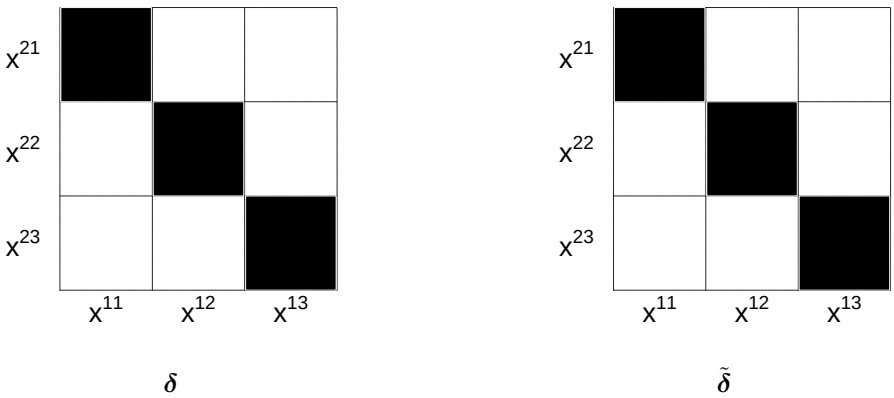


Fig 11: Example of two discrete parameters defining a maximal dependence distribution.

B.1.1 Same discrete parameter $\boldsymbol{\delta}$

Without loss of generality, it is considered that $\delta^{h2h} = 1$ and $\delta^{h2h'} = 0$ if $h \neq h'$ (as previously illustrated by Figure 11) and two parameters $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\tau}, \rho, \boldsymbol{\delta})$ and $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\tau}}, \tilde{\rho}, \boldsymbol{\delta})$ defining the same distribution. The following system is also defined:

$$\begin{cases} (1-\rho)\alpha^{1h}\alpha^{2h} + \rho\tau^{h} & = (1-\tilde{\rho})\tilde{\alpha}^{1h}\tilde{\alpha}^{2h} + \tilde{\rho}\tilde{\tau}^{h} \\ (1-\rho)\alpha^{1h}\alpha^{2h'} & = (1-\tilde{\rho})\tilde{\alpha}^{1h}\tilde{\alpha}^{2h'}. \end{cases} \tag{19}$$

From the previous system, it is easy to show that $\alpha^{jh}(1-\tilde{\alpha}^{jh}) = \tilde{\alpha}^{jh}(1-\alpha^{jh})$ with $j = 1, 2$ and $h = 1, 2, 3$ so $\boldsymbol{\alpha} = \tilde{\boldsymbol{\alpha}}$, $\boldsymbol{\tau} = \tilde{\boldsymbol{\tau}}$ and $\rho = \tilde{\rho}$ so the distribution is GI.

B.1.2 Different discrete parameter $\boldsymbol{\delta}$ and $\tilde{\boldsymbol{\delta}}$

It is considered that the dependence relations between modalities are different for two modalities crossing. The following reasoning can be applied for the case where the dependence relations between modalities are different for the three modalities crossing. Without loss of generality, it is supposed that $\delta^{h2h'} = 1$ if $h = h'$ and $\delta^{h2h'} = 0$ otherwise and $\tilde{\delta}^{h2h'} = 1$ if $h = 4 - h'$ and $\tilde{\delta}^{h2h'} = 0$ otherwise. The following system is obtained:

$$\left\{\begin{array}{rl} (1-\tilde{\rho})\tilde{\alpha}^{12}\tilde{\alpha}^{21} & = (1-\rho)\alpha^{12}\alpha^{21} \\ (1-\tilde{\rho})\tilde{\alpha}^{11}\tilde{\alpha}^{22} & = (1-\rho)\alpha^{11}\alpha^{22} \\ (1-\tilde{\rho})\tilde{\alpha}^{12}\tilde{\alpha}^{23} & = (1-\rho)\alpha^{12}\alpha^{23} \\ (1-\tilde{\rho})\tilde{\alpha}^{13}\tilde{\alpha}^{21} + \tilde{\rho}\tilde{\tau}^{3} & = (1-\rho)\alpha^{13}\alpha^{21} \\ (1-\tilde{\rho})\tilde{\alpha}^{13}\tilde{\alpha}^{22} & = (1-\rho)\alpha^{13}\alpha^{22} \\ (1-\tilde{\rho})\tilde{\alpha}^{13}\tilde{\alpha}^{23} & = (1-\rho)\alpha^{13}\alpha^{23} + \rho\tau^{3}. \end{array}\right. \tag{20}$$

From the previous system, it is deduced that $\tilde{\alpha}^{23} = \frac{\alpha^{23}}{\alpha^{21}}\tilde{\alpha}^{21}$ and $\tilde{\alpha}^{13} = \frac{\alpha^{13}}{\alpha^{11}}\tilde{\alpha}^{11}$ and the following system is studied:

$$\left\{\begin{array}{rl} (1-\tilde{\rho})\frac{\alpha^{13}}{\alpha^{11}}\tilde{\alpha}^{11}\tilde{\alpha}^{21} + \tilde{\rho}\tilde{\tau}^{3} & = (1-\rho)\alpha^{13}\alpha^{21} \\ (1-\tilde{\rho})\frac{\alpha^{13}}{\alpha^{11}}\frac{\alpha^{23}}{\alpha^{21}}\tilde{\alpha}^{11}\tilde{\alpha}^{21} & = (1-\rho)\alpha^{13}\alpha^{23} + \rho\tau^{3}. \end{array}\right. \tag{21}$$

Equation (21) implies a contradiction because from this equation, we can deduce that $\rho\tau^{3}\frac{\alpha^{21}}{\alpha^{23}} + \tilde{\rho}\tilde{\tau}^{3} = 0$ so $\tilde{\rho} < 0$ or $\tilde{\tau} < 0$ but by definition $\tilde{\rho} \geq 0$ or $\tilde{\tau} \geq 0$. So the distribution is GI.

B.2 Generic identifiability extension

It is shown that if the distribution is not GI with the modalities vector $\boldsymbol{m} = (m_1, \ldots, m_d)$ then it is not GI with the modalities vector $\tilde{\boldsymbol{m}} = (\tilde{m}_1, \ldots, \tilde{m}_d)$ with $\tilde{m}_j > 2$ for $j = 1, 2$ and $\tilde{m}_j \geq 2$ for $j = 3, \ldots, d$ defined by the fusion of two modalities h_0 et h_1 of the variable j_0 involving the same realization for the variable $j_0 + 1$ under the maximum dependence distribution, so $\exists h\ \delta^{hj_0h_0} + \delta^{hj_0h_1} = 2$.

The data where the fusion between the modalities h_0 and h_1 of the variable j_0 (with $h_0 < h_1$) is applied is denoted by $\tilde{\boldsymbol{x}}$. By supposing two different parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ defining the same distribution for $\boldsymbol{x}$, two different parameters $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}'$ are defined for parameterize the distribution of $\tilde{\boldsymbol{x}}$ as:

- If $j_0 > 1$ then $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\alpha}}, \boldsymbol{\tau}, \rho, \boldsymbol{\delta})$ with :

$$\tilde{\alpha}^{jh} = \left\{\begin{array}{rl} \alpha^{j_0h_0} + \alpha^{j_0h_1} & \text{if } j = j_0 \text{ and } h = h_0 \\ \alpha^{j_0h-1} & \text{if } j = j_0 \text{ and } h > h_1 \\ \alpha^{jh} + \alpha^{j_0h_1} & \text{otherwise.} \end{array}\right. \tag{22}$$

- If $j_0 = 1$ then $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\alpha}}, \boldsymbol{\tau}, \rho, \boldsymbol{\delta})$ with $\tilde{\boldsymbol{\alpha}}$ defined by (22) and :

$$\tilde{\boldsymbol{\delta}}^{h} = \left\{\begin{array}{rl} \delta^{h-1} & \text{if } h > h_1 \\ \delta^{h} & \text{otherwise.} \end{array}\right. \quad \text{and} \quad \tilde{\tau}^{jh} = \left\{\begin{array}{rl} \tau^{h_0} + \tau^{h_1} & \text{if } h = h_0 \\ \tau^{h-1} & \text{if } h > h_1 \\ \tau^{h} & \text{otherwise.} \end{array}\right. \tag{23}$$

In these two situations $\tilde{\boldsymbol{\theta}}'$ is defined as $\tilde{\boldsymbol{\theta}}$ with its parameters. So $\tilde{\boldsymbol{\theta}} \neq \tilde{\boldsymbol{\theta}}'$ and $\forall \tilde{\boldsymbol{x}}\ p(\tilde{\boldsymbol{x}}; \tilde{\boldsymbol{\theta}}) = p(\tilde{\boldsymbol{x}}; \tilde{\boldsymbol{\theta}}')$. If the distribution is not GI with the modalities vector $\tilde{\boldsymbol{m}}$ then it is not GI with the modalities vector $\boldsymbol{m}$. So the proposed distribution stays GI by addition of modalities. It is easy to show that the proposed distribution stays GI by addition of variable. So the block distribution is GI when the number of variables or the modalities number of the second variable is at least equal to three.

C Estimated dependence parameters of the applications

C.1 Classical binary data set: Congressional voting records

Block	Dependence parameters							
	Strength		Description					
	Variables	Modalities						
	ρ_{kb}	$\boldsymbol{\tau}_{kb}$	$\boldsymbol{\sigma}_{kb}$ and $\boldsymbol{\delta}_{kb}$					
1	0.52		Wat	Sup				
		0.84	yes	no				
		0.16	no	yes				
2	0.52		Han	Bud	Phy	Edu		
		0.93	no	no	yes	yes		
		0.07	yes	yes	no	no		
3	0.14		Ant	Nic	Mis	Exp		
		0.89	no	no	no	no		
		0.11	yes	yes	yes	yes		
4	0		Sal	Rel	Imm	Syn	Cri	Dut

Table 8: Dependence parameters of the class 1 for the Congressional voting records data set.

Block	Dependence parameters						
	Strength		Description				
	Variables	Modalities					
	ρ_{kb}	$\boldsymbol{\tau}_{kb}$	$\boldsymbol{\sigma}_{kb}$ and $\boldsymbol{\delta}_{kb}$				
1	0.61		Phy	Cri			
		0.92	no	no			
		0.08	yes	yes			
2	0.53		Rel	Edu	Sup	Exp	
		0.98	no	no	no	yes	
		0.02	yes	yes	yes	no	
3	0.25		Wat	Sal	Ant	Mis	Imm
		0.93	no	no	no	no	
		0.07	yes	yes	yes	yes	
4	0		Han	Bud	Nic	Syn	Dut

Table 9: Dependence parameters of the class 2 for the Congressional voting records data set.

C.2 Classical categorical data set: contraceptive method choice

Block	Dependence parameters						
	Strength		Description				
	Variables	Modalities					
	ρ_{kb}	$\boldsymbol{\tau}_{kb}$	$\boldsymbol{\sigma}_{kb}$ and $\boldsymbol{\delta}_{kb}$				
1	0.38		Chi	WAg			
		0.37	2	1			
		0.28	6	3			
		0.25	3	1			
		0.10	1	1			
		0.00	5	4			
		0.00	4	2			
2	0.21		WEd	HEd			
		0.70	4	4			
		0.30	2	2			
		0.00	1	1			
		0.00	3	3			
3	0		HOc	WRe	Liv	WWo	Med

Table 10: Dependence parameters of the class 1 for the Contraceptive method choice data set.

Block	Dependence parameters						
	Strength		Description				
	Variables	Modalities					
	ρ_{kb}	$\boldsymbol{\tau}_{kb}$	$\boldsymbol{\sigma}_{kb}$ and $\boldsymbol{\delta}_{kb}$				
1	0.42		HOc	WRe			
		0.79	1	2			
		0.12	2	2			
		0.01	3	2			
		0.00	4	1			
2	0.16		Chi	WAg			
		0.33	3	2			
		0.26	2	4			
		0.19	6	4			
		0.09	1	1			
		0.08	5	4			
		0.05	4	3			
3	0		Wed	Hed	Liv	WWo	Med

Table 11: Dependence parameters of the class 2 for the Contraceptive method choice data set.

C.3 Real study for improving calves reproduction

In order to present the dependence block, the variables follows this coding : aptitude for sucking (Apt), treatment against omphalite of the calve (TOC), treatment against respiratory illness of the calve (TRC), treatment against diarrhea of the calve (TDC), isolated mother (Iso), umbilicus disinfection (Dis), umbilicus emptying (Emp), preventive treatment against respiratory illness of the mother (TRM) and preventive treatment of the mother against diarrhea (TDM).

Block	Dependence parameters				
	Strength		Description		
	Variables	Modalities			
	ρ_{kb}	$\boldsymbol{\tau}_{kb}$	$\boldsymbol{\sigma}_{kb}$ and $\boldsymbol{\delta}_{kb}$		
1	0.85		TDC	TRM	
		0.87	no	no	
		0.09	curative	no	
		0.04	preventive	yes	
2	0.82		TRC	TDM	
		0.81	preventive	yes	
		0.18	no	no	
		0.01	curative	no	
3	0.49		Iso	Dis	Emp
		0.83	no	no	no
		0.17	yes	yes	yes
4	0		Apt	TOC	

Table 12: Dependence parameters of class 1 for the Genes Diffusion data set.

Block	Dependence parameters					
	Strength		Description			
	Variables	Modalities				
	ρ_{kb}	$\boldsymbol{\tau}_{kb}$	$\boldsymbol{\sigma}_{kb}$ and $\boldsymbol{\delta}_{kb}$			
1	0.75		TOC	Dis	TDM	
		0.50	preventive	yes	yes	
		0.48	no	no	no	
		0.02	curative	no	no	
2	0.72		TRC	Iso		
		0.71	preventive	yes		
		0.29	no	no		
		0.00	curative	no		
3	0		Apt	TDC	Emp	TRM

Table 13: Dependence parameters of class 2 for the Genes Diffusion data set.

Block	Dependence parameters				
	Strength		Description		
	Variables	Modalities			
	ρ_{kb}	$\boldsymbol{\tau}_{kb}$	$\boldsymbol{\sigma}_{kb}$ and $\boldsymbol{\delta}_{kb}$		
1	0.51		TDC	TRM	TDM
		0.88	no	no	yes
		0.07	preventive	yes	no
		0.05	curative	no	yes
2	0		Apt TOC TRC Iso Dis Emp		

Table 14: Dependence parameters of class 3 for the Genes Diffusion data set.

Block	Dependence parameters			
	Strength		Description	
	Variables	Modalities		
	ρ_{kb}	$\boldsymbol{\tau}_{kb}$	$\boldsymbol{\sigma}_{kb}$ and $\boldsymbol{\delta}_{kb}$	
1	0.61		TOC	Dis
		0.62	no	no
		0.38	preventive	yes
		0.02	curative	no
2	0.46		Iso	Emp
		0.66	yes	no
		0.34	no	yes
3	0		Apt TRC TDC TRM TDM	

Table 15: Dependence parameters of class 4 for the Genes Diffusion data set.

Block	Dependence parameters				
	Strength		Description		
	Variables	Modalities			
	ρ_{kb}	$\boldsymbol{\tau}_{kb}$	$\boldsymbol{\sigma}_{kb}$ and $\boldsymbol{\delta}_{kb}$		
1	0.39		Iso	Emp	TDM
		0.74	yes	no	yes
		0.26	no	yes	no
2	0		Apt TOC TRC TDC Dis TRM		

Table 16: Dependence parameters of class 5 for the Genes Diffusion data set.

RESEARCH CENTRE
LILLE – NORD EUROPE

Parc scientifique de la Haute-Borne
40 avenue Halley - Bât A - Park Plaza
59650 Villeneuve d'Ascq

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399