Model-based clustering for categorical and mixed data sets

Matthieu Marbac-Lourdelle

NUMÉRO D'ORDRE: 41516

Université de Lille 1 — Laboratoire Paul Painlevé

Thèse

présentée pour l'obtention du

Diplôme de doctorat de l'université de Lille 1

Spécialité : **Mathématiques**

Modèles de mélange pour la classification non supervisée de données qualitatives et mixtes

---

par

Matthieu Marbac-Lourdelle

Soutenue le 23 septembre 2014 devant le jury composé de :

Dimitris	Karlis	Professeur, University of Athens	Rapporteur
Jean-Michel	Marin	Professeur, Université Montpellier 2	Rapporteur
Gilles	Celeux	Directeur de Recherche, Inria Saclay	Examinateur
Nicolas	Wicker	Professeur, Université Lille 1	Examinateur
Christophe	Biernacki	Professeur, Université Lille 1	Directeur
Vincent	Vandewalle	Professeur, Université Lille 2	Directeur

Laboratoire Paul Painlevé — Université Lille1 Sciences et Technologies — Inria informatics mathematics — DGA

Thèse préparée au
Département de Mathématiques
Laboratoire Paul Painlevé (UMR CNRS 8524)
Université Lille 1
59 655 Villeneuve d'Ascq CEDEX

Remerciements

First of all, I sincerely would like to thank Dimitris Karlis and Jean-Michel Marin for accepting to report on this thesis, as well as Gilles Celeux and Nicolas Wicker for their kind participation as examiners in the Ph.D. defense.

J'adresse également de profonds remerciements à mes deux directeurs de thèse Christophe Biernacki et Vincent Vandewalle pour leur encadrement au cours de ces trois années. Vous avez su distiller de nombreux conseils tout en me laissant une liberté quant aux axes de recherches que j'ai pu explorer, ce que j'ai fortement apprécié. Je vous remercie de m'avoir donné l'opportunité de m'initier à la recherche lors de mon stage de master et de vous être démenés pour me trouver un financement de thèse.

Je remercie Chi Tran qui m'a donné goût à la recherche lors de nos réunions de TER que j'avais souvent décalées... Je suis particulièrement reconnaissant envers Bernd Beckermann qui m'a encouragé depuis mon master et qui a toujours su prendre le temps de me prodiguer de précieux conseils.

J'adresse également mes remerciements à l'ensemble de l'équipe MODAL où mon intégration a été facilitée grâce à Sandrine Meilen.

Au cours de ces trois années de thèse, j'ai eu le plaisir de découvrir le monde de l'enseignement avec Karin Sahmer, Cristian Preda et Julien Jacques.

Je remercie le personnel du laboratoire Painlevé, en particulier: Sabine, Carine, Cathy, Hélène, Frédérique, Jean-Jacques (GiGi), François (la tomate) et Omar (le roi de la muscu).

Pendant cette thèse, j'ai eu le plaisir d'échanger avec de nombreuses personnes parmi lesquelles: Serge, Parmeet, Alexandru, Benjamin, Pierre-Louis (dit Pilou), Hubert et Julien. Les midis ont été particulièrement mémorables grâce à Antoine (Mac'), Émilie et Pierre. Merci à Elodie (tu as vu, je n'ai pas mis l'accent) pour ses conseils, en particulier pendant ma rédaction.

Je remercie les membres de ma famille (qu'ils aiment ou non les ananas) pour leur soutien et parce que, j'en suis sur, ils ont tous retenu le titre de ma thèse... J'adresse une mention spéciale pour Cléo qui me supporte au quotidien.

Enfin, je présente mes excuses à ceux que j'ai honteusement oublié de mentionner et qui, je l'espère, ne m'en tiendrons pas rigueur...

Contents

Foreword

On the need to cluster

Data acquisition has become increasingly easy thanks to the increased performance of computing. Therefore, practitioners are facing data sets which are increasingly information-rich but also increasingly abstruse. Indeed, the information contained in the data sets can be directly unattainable for two main reasons: the *quantity* of data and their *complexity*. Thus, statistical methods are mandatory to analyze such data sets.

Clustering is an approach which reduces the problem caused by the large quantity of data. Indeed, its aim is to group the individuals into few specific classes. Thus, it provides a meaningful summary of the data set throughout few characteristic individuals. The idea of the individual (or object) clustering is natural. For instance: the living world is divided into two classes: plants and animals; the animals are split into invertebrate and vertebrate; the vertebrate are classified into five classes (mammals, fishes, birds, amphibians, reptiles)...

The probabilistic methods permit to perform the cluster analysis in a rigorous context. Among these methods, the finite mixtures of parametric distributions summarize the data by the few parameters of each class. Moreover, in this context, the classical probabilistic tools are available to answer the difficult questions of cluster analysis like the choice of the number of classes. If the bibliography is prolific about continuous data sets, we note a shortage when the data are more *complex*. In this context, the aim of this manuscript is to study existing probabilistic methods and to propose new ones to cluster *complex* data sets.

The two objectives of this work

We focus on two situations where the data sets are *complex*: the case where individuals are described by *categorical* variables and the case where they are described by *mixed* variables (different kinds of variables). Thus, two thematics are studied.

— The model-based clustering for categorical data sets.

— The model-based clustering for mixed data sets.

The *categorical variables* are difficult to cluster since they leave the statistician facing with many combinatorial challenges. This difficulty increases when the variables are dependent in the same class. Indeed, the models require a large number of parameters in order to take into account the intra-class dependencies. Thus, the classical approach assumes the conditional independence between variables. However

this approach is biased when the conditional independence assumption is violated. Many alternative approaches have been proposed but their answers stay incomplete. Indeed, the combinatorial problem of the model selection is not always solved. Moreover, these methods can suffer from instability or from a lack of interpretability. In this context, our contribution consists in two parsimonious mixture models which allow to cluster categorical data presenting intra-class dependencies. The main idea of these models is to group the variables into conditionally independent blocks. By setting specific distributions for these blocks, both models consider the intra-class dependencies between the variables. Both models provide few parameters to summarize the data, propose a rigorous approach to perform the model selection and give a user-friendly visualization of the parameters. The challenge of the categorical data clustering is motivated by the fact that categorical data are easily accessible. As they are numerous, the risk to observe intra-class correlated data increases.

The study of the cluster analysis of *mixed data sets* is the second objective of this work. This problem is motivated by the fact that the current data sets are often composed with different kinds of data. A classical approach uses factor analyzers methods. The interpretation of such a method is often complex since the parameters are not relative to the variables in their native space. Other classical approaches consist in applying a specific mixture model on these data. The challenge is due to the lack of classical distributions for mixed variables. We propose two mixture models to fill this gap. The first one is classical since it is an extension of a well-known method to cluster categorical data sets. Indeed, the model combines Gaussian distributions and linear logistic regressions. Thus, this model analyzes data sets with continuous and categorical variables. The second model is the main contribution of this thesis since it allows to analyze data sets with any kind of variables admitting a cumulative distribution function. This model is defined as a mixture of Gaussian copulas, thus preserving classical one-dimensional margin distribution for each variables of each component. Furthermore, this approach modelizes the intra-class dependencies. Finally, note that a visualization tool is available as a by-product of this model.

Organization of the manuscript

The manuscript is divided into two main parts corresponding to the two kinds of data of interest. More precisely, it is organized as follows.

— **Chapter 1** is a brief overview of the main clustering methods in a general framework. It focuses on the different aspects related to finite mixture models. It introduces the general notions and algorithms used in the following chapters.

— **Part I: Model-based clustering for categorical data**

  — **Chapter 2** consists in the state of the art of the methods performing the cluster analysis of categorical data sets.

— **Chapter 3** presents our first contribution to the categorical data analysis framework. The proposed model groups the variables into conditionally independent blocks. The specific distribution of the blocks modelizes the intra-class dependencies and provides a specific coefficient summarizing the strength of these dependencies. All these results are part of the article *Model-based clustering for conditionally correlated categorical data* [MBV13a].

— **Chapter 4** presents our second contribution to the categorical data analysis framework. This model consists in a mixture model which groups the variables into conditionally independent blocks. Each block follows a parsimonious multinomial distribution where the few free parameters correspond to its modes. All these results are part of the article *Finite mixture model of conditional dependencies modes to cluster categorical data* [MBV14a].

— **Chapter 5** illustrates both R packages which perform the inference of both proposed models. This chapter can also be used as a tutorial of both packages since it provides a presentation of their main functions and many scripts allowing to perform the cluster analysis.

— **Part II: Model-based clustering for mixed data**

— **Chapter 6** consists in the state of the art of the methods performing the cluster analysis of mixed data sets.

— **Chapter 7** presents our first contribution to the cluster analysis framework of mixed data sets with continuous and categorical variables. The model is derived from the multilevel latent class model developed to cluster categorical data sets. For this model, the component distributions of the continuous variables are Gaussian and those of the categorical variables conditionally on the continuous ones are linear logistic regressions. The model selection and the parameter estimation are simultaneously performed by a GEM algorithm.

— **Chapter 8** presents the main contribution of this thesis. It consists in a mixture model of Gaussian copulas. Thus, the model performs the cluster analysis of data sets composed of any kind of variables admitting a cumulative distribution function. All these results are part of the article *Model-based clustering of Gaussian copulas for mixed data* [MBV14b].

Main abbreviations and notations

Main abbreviations

General

MAP	maximum *a posteriori*
MAPE	maximum *a posteriori* estimate
MLE	maximum likelihood estimate
cdf	cumulative distribution function
pdf	probability distribution function

Algorithms

EM	Expectation-Maximization algorithm
GEM	Generalized Expectation-Maximization algorithm
MCMC	Markov chain Monte Carlo
SEM	Stochastic Expectation-Maximization algorithm

Information criteria

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
ICL	Integrate Complete-Likelihood

Main notations

The notations have been defined with the following rules:

— The variables are denoted with Arabic letters while the parameters are denoted with Greek letters.

— The multidimensional objects are denoted by bold symbols while the unidimensional objects are denoted by thin symbols.

Variables and observations

$\boldsymbol{X}_i$	set of the e random variables related to individual i
$\boldsymbol{x}_i$	observed values of $\boldsymbol{X}_i$
$\boldsymbol{x}_i'$	transpose of $\boldsymbol{x}_i$
$\boldsymbol{x}_i^{\mathrm{c}}$	subset of $\boldsymbol{x}_i$ composed of the c continuous variables
$\boldsymbol{x}_i^{\mathrm{d}}$	subset of $\boldsymbol{x}_i$ composed of the d discrete variables
m_j	number of modalities of variable j
$\boldsymbol{Z}_i$	random variable of the class membership of the individual i
$\boldsymbol{z}_i$	observed values of $\boldsymbol{Z}_i$
$\boldsymbol{y}_i$	second latent variable related to individual i (if required)
$\mathbf{x}$	n sample $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$
$\mathbf{z}$	n sample $\mathbf{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$
$\mathbf{y}$	n sample $\mathbf{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$

Parameters

$\boldsymbol{\theta}$	whole parameters of the mixture
$\boldsymbol{\pi}$	vector of proportions
$\boldsymbol{\alpha}_k$	parameters related to component k
$\boldsymbol{\Gamma}_k$	matrix of the correlation related to component k
ν	number of parameters

Important integers

c	number of continuous variables
d	number of categorical variables
e	number of variables, *i.e.* $c + d = e$
g	number of classes
n	size of the sample
n_k	size of class k computed on the fuzzy partition
n_k	size of class k computed on the hard partition

Classical distributions

$\mathcal{D}_g(.)$	Dirichlet distribution of size g
$\mathcal{G}(.)$	Gamma distribution
$\mathcal{N}_c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	c-variate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathcal{M}_g(.)$	multinomial distribution of size g
$\mathcal{P}(.)$	Poisson distribution

Classical tools

$p(.;.)$	probability distribution function
$P(.;.)$	cumulative distribution function
$\phi_c(.;\boldsymbol{\mu},\boldsymbol{\Sigma})$	pdf of $\mathcal{N}_c(\boldsymbol{\mu},\boldsymbol{\Sigma})$
$\Phi_c(.;\boldsymbol{\mu},\boldsymbol{\Sigma})$	cdf of $\mathcal{N}_c(\boldsymbol{\mu},\boldsymbol{\Sigma})$
$\Phi_1(.)$	pdf of $\mathcal{N}_1(0,1)$
$KL(f_1,f_2)$	Kullback-Leibler divergence from f_1 to f_2 (f_2: reference)
$p(\mathbf{x};\boldsymbol{\theta})$	observed-data likelihood
$L(\boldsymbol{\theta};\mathbf{x})$	observed-data log-likelihood
$p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})$	complete-data likelihood
$L(\boldsymbol{\theta};\mathbf{x},\mathbf{z})$	complete-data log-likelihood
$t_{ik}(\boldsymbol{\theta})$	probability that $\boldsymbol{x}_i$ is drawn by component k

Chapter 1

Cluster analysis: state of the art

The main purpose of this chapter is to review the literature about cluster analysis. Note that our aim is not to be exhaustive, thus we principally focus on the model-based approaches in order to define the different notions developed in this manuscript.
Firstly, we present different approaches (geometric and probabilistic) to cluster the data. Secondly, we review the frequentist and the Bayesian approaches used to infer the finite mixture models. Finally, we present some criteria performing the model selection in a probabilistic context.
Two toy examples illustrate the different notions and the algorithms through this chapter, in a continuous case since it is the easiest one.

The good story lay in half-told things
which must be filled in out of the
hearer's own experience.
John Steinbeck — Tortilla Flat

1.1 Overview of the clustering approaches

1.1.1 Clustering challenge

Nowadays, practitioners are often facing complex data sets, that we denote in this manuscript by $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, describing n individuals $\boldsymbol{x}_i = (x_i^1, \ldots, x_i^e)$ by e variables. This complexity is generally involved by the large number of individuals overwhelming the practitioners under embedded informations. Furthermore, this complexity can be increased by the descriptors due to the number of variables or to their nature (for instance, categorical or mixed variables).

Clustering is a general answer to this problem which increasingly emerges with the computer development. Indeed, this technique summarizes the data by grouping

the individuals into g classes according to both following principles: the *class homogeneity* (grouping similar individuals into the same class) and the *class separability* (two individuals arisen from two different classes are strongly different). Even if the exact definition of a class is specific to the clustering method selected by the practitioner, it is always based on these two principles.

According to these principles, the clustering methods try to determine the latent vector $\mathbf{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ where the vector $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ig})$ indicates the class membership of the individual $\boldsymbol{x}_i$ by using a complete disjunctive coding (*i.e.* $z_{ik} = 1$ if $\boldsymbol{x}_i$ is affiliated into class k and $z_{ik} = 0$ otherwise). Note that classes have to be interpretable for the specialist of the domain where the data come from. Indeed, a clustering method provides an efficient summary of the data only if its resulting classes are meaningful.

The class memberships of the individuals have to be estimated but the number of classes is generally unknown. Thus, an efficient clustering method provides tools to help the practitioner for the selection of the number of classes. The clustering has three main goals: *to estimate the partition*, *to provide meaningful classes* and *to select automatically the number of classes*.

For the large data sets (n and e large), practitioners can simultaneously cluster the individuals and the variables. Two partitions can thus be searched: one among the individuals and one among the variables. This approach is named *co-clustering* [GN08, GN10] but it is not developed in this thesis where we only study the clustering problem.

The methods performing the cluster analysis are divided into two large families: the *geometric* methods based on some distances between individuals and the *probabilistic* methods modelizing the data generation. In this section, both approaches are detailed in a general framework. Nevertheless, they are illustrated on a bivariate continuous data set presented below since it allows to easily visualize the results. Two specific states of the art relative to more complex situations (categorical and mixed data sets) are given later in the introductions of Part 1 and Part 2.

Faithful data set 1.1 (Data presentation).

The `Faithful` data set [AB90] is available on the R package `mass`. This data set contains the waiting time between 272 eruptions and the duration of the eruptions for the Old Faithful geyser in Yellowstone National Park (Wyoming, USA) displayed by Figure 1.1. The aim is to provide a meaningful summary of the data set $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ where each individual $\boldsymbol{x}_i \in \mathbb{R}^2$. Thus, in this example $n = 272$ and $e = 2$.

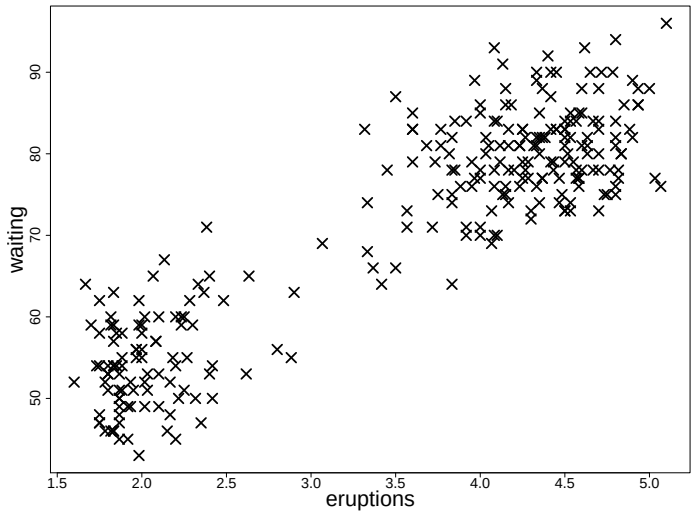


Figure 1.1 – The Faithful data set.

1.1.2 Geometric approaches

Generalities

Main idea The geometric approaches group the set of the clustering methods based on the distance measurement between individuals, in order to assign the most similar individuals into the same class. By exacerbating the class *homogeneity* concept, an (intractable) objective could be the following: *two individuals arising from the same class have to be more similar than two individuals having a different class membership*. This aim can be expressed by the following mathematical relation: for a distance $D(.,.)$ and for all (i_0, i_1, i_2, i_3) such that $\boldsymbol{z}_{i_0} = \boldsymbol{z}_{i_1}$ and $\boldsymbol{z}_{i_2} \neq \boldsymbol{z}_{i_3}$:

$$D(\boldsymbol{x}_{i_0}, \boldsymbol{x}_{i_1}) \leq D(\boldsymbol{x}_{i_2}, \boldsymbol{x}_{i_3}).$$

A NP-hard problem Since such a relation has to be satisfied by all the quadruplets, this objective generally leads to empty set solution (huge number of constraints). Furthermore, it is not realizable to perform an exhaustive approach which computes the objective criterion for all the possible partitions. Indeed, for a sample of size $n = 40$ that we cluster in $g = 3$ classes, the number of the possible partitions is roughly equal to 2.10^{18}. Thus, a computer performing 10^9 partitions per second needs 64000 years to evaluate all the possibilities.

Global criterion In practice, it is advised to optimize a global criterion relating the class homogeneity. Different criteria, often based on heuristic ideas, have been also proposed (see the running example). These criteria are easily optimized by an algorithm avoiding an exhaustive approach which is intractable.

Structure of this section Firstly, we present the three most common criteria of interest when the variables are continuous. Secondly, we detail the *K-means*

algorithm which is the most classical geometric approach to cluster individuals described by continuous variables [Ber06]. Finally, this algorithm is illustrated on the `Faithful` data set.

Faithful data set 1.2 (Optimized criteria for continuous variables)**.**

Let us introduce the matrix of the whole sample covariance, denoted by $\boldsymbol{T}$ and defined as

$$\boldsymbol{T} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})', \tag{1.1}$$

where $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$ is the mean vector of the whole sample. This matrix can be written as a sum of two matrices

$$\boldsymbol{T} = \boldsymbol{W} + \boldsymbol{B}, \tag{1.2}$$

where the intra-class covariance matrix $\boldsymbol{W}$ and where the inter-class covariance matrix $\boldsymbol{B}$ are defined by

$$\boldsymbol{W} = \frac{1}{n}\sum_{k=1}^{g}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)' \text{ and } \boldsymbol{B} = \frac{1}{n}\sum_{k=1}^{g}\mathrm{n}_k(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}})', \tag{1.3}$$

$\bar{\boldsymbol{x}}_k = \frac{1}{\mathrm{n}_k}\sum_{i=1}^{n} z_{ik}\boldsymbol{x}_i$ is the mean vector and $\mathrm{n}_k = \sum_{i=1}^{n} z_{ik}$ the size of class k. These matrices characterize both construction's principles of the classes. Indeed, if classes are *homogeneous*, then the distances between the individuals assigned to a class and its center are small, so $\boldsymbol{W}$ is "small", while if classes are well *separated*, then the centers of the classes are mutually taken away, so $\boldsymbol{B}$ is "large".
Thus, the practitioner can cluster the `Faithful` data set by optimizing different criteria related to these matrices. Among them, the most classical are the following: $\min \mathrm{trace}(\boldsymbol{W})$, $\min \det(\boldsymbol{W})$ or $\max \mathrm{trace}(\boldsymbol{B}\boldsymbol{W}^{-1})$. We refer to the book *Cluster analysis* by B.S. Everitt, S. Landau, M. Leese and D. Stahl [ELLS11] for more details.

K-means algorithm

Main idea S. Lloyd proposed the *K-means* algorithm around 1950 and waited 1982 to publish it [Llo82]. Associated to a distance $D(.,.)$, this algorithm aims at minimizing the following inertia

$$I(\mathbf{z}, \boldsymbol{\theta}; \mathbf{x}) = \sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik}D^2(\boldsymbol{x}_i, \boldsymbol{\mu}_k), \tag{1.4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)$ and where $\boldsymbol{\mu}_k$ is the center of the class k. Starting from an initial value of the class centers, the *K-means* algorithm alternates between two steps: the assignment of each individual to the class minimizing the distance between him and the class center, and the computation of the class centers.

Algorithm 1.3 (The K-means algorithm)**.**

Starting from an initial value $\boldsymbol{\theta}^{[0]}$, iteration $[r]$ is written as follows

— **Class membership** $\mathbf{z}^{[r]} = \underset{\mathbf{z}}{\operatorname{argmin}}\, I(\mathbf{z}, \boldsymbol{\theta}^{[r]}; \mathbf{x})$:

$$z_{ik}^{[r]} = \begin{cases} 1 & \text{if } k = \underset{k'}{\operatorname{argmin}}\, D^2(\boldsymbol{x}_i, \boldsymbol{\mu}_{k'}^{[r]}) \\ 0 & \text{otherwise.} \end{cases} \tag{1.5}$$

— **Centroid estimation** $\boldsymbol{\theta}^{[r+1]} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, I(\mathbf{z}^{[r]}, \boldsymbol{\theta}; \mathbf{x})$:

$$\boldsymbol{\mu}_k^{[r+1]} = \frac{1}{\mathrm{n}_k^{[r]}} \sum_{i=1}^{n} z_{ik}^{[r]} \boldsymbol{x}_i, \tag{1.6}$$

where $\mathrm{n}_k^{[r]} = \sum_{i=1}^{n} z_{ik}^{[r]}$.

As this algorithm converges to a local optimum of $I(\mathbf{z}, \boldsymbol{\theta}; \mathbf{x})$, it is mandatory to perform different initializations and to keep the couple $(\mathbf{z}, \boldsymbol{\theta})$ minimizing the objective inertia.

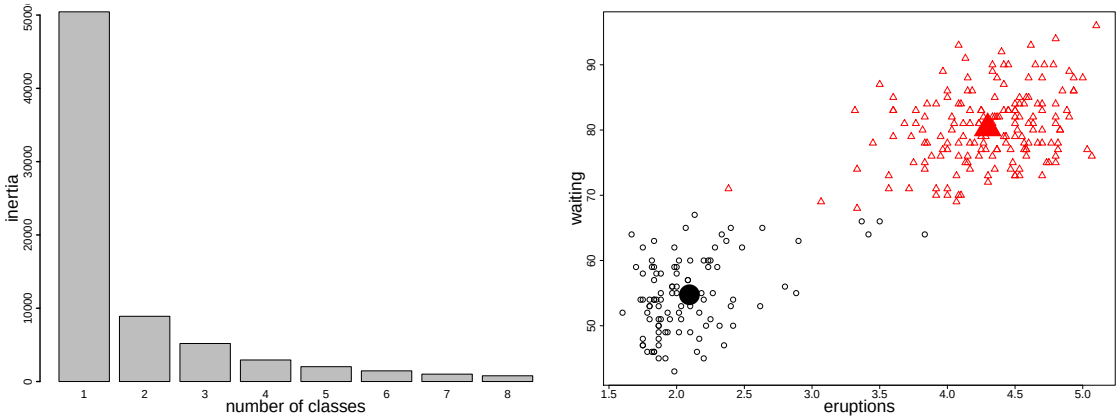
Remark 1.4 (K-means algorithm and optimized criterion)**.** If the *K-means* algorithm clusters continuous data by using the Euclidean distance, then it optimizes the criterion $\min \operatorname{trace}(\boldsymbol{W})$.

Extensions of the K-means algorithm Some approaches attempt to reduce the drawbacks of the *K-means* algorithm. For instance, the *K-means++* algorithm [AV07] extends the classical one by a randomized seeding technique improving the speed and the accuracy of the *K-means*.

How many classes? The selection of the number of classes can not be directly performed by the inertia criterion defined in (1.4) since this latter is decreasing with the number of classes g. However, the objective criterion reaches a plateau when g increases. Indeed, when this plateau is reached, the added classes are no more *homogeneous* while the class overlapping increases. A heuristic criterion consists in selecting the first number of classes of this plateau. But, it is clear that this kind of criterion is not very rigorous. In practice, the criterion can be unhelpful when some plateaus are observed. Other criteria are available, see for instance [Ber06], but they are based on a heuristic approach.

Faithful data set 1.5 (K-means algorithm approach).

We use the *K-means* algorithm to cluster the `Faithful` data set. According to Figure 1.2a drawing the evolution of the inertia for different numbers of classes (from one to eight), we can select two classes. The partition and the class centers are displayed by Figure 1.2b. The `Faithful` data set is also summarized by two profiles of eruptions: the eruptions with short waiting time and duration (center at $(2.09, 54.75)$) are displayed with black circles and the eruptions with larger waiting time and duration (center at $(4.30, 80.28)$) are displayed with red triangles.



(a) Inertia (b) Scatter plot and class centers

Figure 1.2 – Outputs of the `Faithful` cluster analysis performed by a *K-means* algorithm. The scatter plot indicates the partition by the colors and the thin symbols, while the class centers are represented by the bold symbols in the color of their class.

Limits of the geometric approaches

Different authors point out the drawbacks inherent to the geometric approaches (see for instance the book *Data analysis*, chapter 9 by G. Govaert [Gov10]) whose the three main drawbacks are presented here.

Model selection performed by heuristic approaches Even if the geometric approaches are a possible answer to the clustering challenge, many theoretical problems may arise [Gov10]. They generally select the number of classes by using heuristic approaches like the slope of the criterion values. Furthermore, both choices of the metric and the criterion used are important aspects of these methods, since they involve many hidden assumptions which are generally ignored by the practitioners. However, their impact is crucial as they involve different partitions.

Extension of the conclusions to the whole population If the conclusions of a cluster analysis based on a sample have to be extended to the whole population, it is mandatory to understand (so, to modelize) the data generation. The extension of the conclusions obtained by a geometric method are not allowed. In such a case, the probabilistic framework is also mandatory.

Dealing with missing values The geometric approaches cannot directly manage data sets with missing values. Indeed, they have to perform an arbitrary imputation or they have to ignore individuals with missing values while the probabilistic approaches are able to rigorously manage such data.

Links between geometric and generative approaches

Many geometric approaches can be interpreted as probabilistic ones revealing their probabilistic hidden assumptions (some examples are given in this thesis). In order to have probabilistic tools and to reveal the assumptions made by the clustering methods, we develop this thesis in a probabilistic framework.

1.1.3 Generative approaches

Generalities

Main idea The mixture models are natural tools to cluster the data by approaching their distributions, because their probabilistic framework explains the data generation. In this context, the notion of class *homogeneity* is defined by the following idea: *the individuals of the same class arise from the same probability distribution.* If this distribution is generally assumed to be uni-modal, this assumption can be relaxed (for instance, the component distribution can be itself a mixture of parametric distributions to increase the model flexibility [BRC+10]).

Latent variable and class membership The class membership of the individual i is a qualitative random variable having g modalities and denoted by $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ig})$ by using a disjunctive coding. Thus, the class membership follows a multinomial distribution

$$\boldsymbol{Z}_i \sim \mathcal{M}_g(\pi_1, \ldots, \pi_g), \tag{1.7}$$

where π_k denotes the proportion of class k also interpreted as the probability *a priori* that an individual arises from class k. The class proportion π_k respects both following constraints $0 < \pi_k \leq 1$ and $\sum_{k=1}^{g} \pi_k = 1$. Note that the clustering challenge is to estimate the value of the realization $\boldsymbol{z}_i$ of the latent variable $\boldsymbol{Z}_i$ conditionally on the observed data $\boldsymbol{x}_i$.

Observed variables Class k is characterized by the distribution of the e-variate random variable $\boldsymbol{X}_i = (X_i^1, \ldots, X_i^e)$ defined on the space $\mathcal{X}$ conditionally on the realization $\boldsymbol{z}_i$ of the random variable $\boldsymbol{Z}_i$. This distribution is denoted by $p_k(\boldsymbol{x}_i)$ where k is such that $z_{ik} = 1$ and

$$\boldsymbol{X}_i | \boldsymbol{Z}_i = \boldsymbol{z}_i \sim p_{\{k:z_{ik}=1\}}(\boldsymbol{x}_i). \tag{1.8}$$

Distribution of both observed and latent variables By using this probability decomposition $P(\boldsymbol{X}_i, \boldsymbol{Z}_i) = P(\boldsymbol{Z}_i)P(\boldsymbol{X}_i|\boldsymbol{Z}_i)$, the probability distribution function (pdf) of $(\boldsymbol{x}_i, \boldsymbol{z}_i)$, denoted by the generic notation $p(.)$, is defined as follows

$$p(\boldsymbol{x}_i, \boldsymbol{z}_i) = \prod_{k=1}^{g} (\pi_k p_k(\boldsymbol{x}_i))^{z_{ik}} . \tag{1.9}$$

Since this model is used to cluster, the labels $\boldsymbol{z}_i$ are considered as missing values. Thus, we obtain both definitions of the finite mixture model and its generative model, by summing the previous equation over all the possible values of $\boldsymbol{Z}_i$.

Definition 1.6 (Finite mixture model)**.** The finite mixture model with g components defines the margin distribution of the random variable $\boldsymbol{X}_i$. Its pdf is written as

$$p(\boldsymbol{x}_i) = \sum_{k=1}^{g} \pi_k p_k(\boldsymbol{x}_i). \tag{1.10}$$

Generative model The sampling from the mixture model defined by (1.10) is performed by the following generative model divided into two steps:

— Step 1: the class membership sampling $\boldsymbol{Z}_i \sim \mathcal{M}_g(\pi_1, \ldots, \pi_g)$.

— Step 2: the conditional data sampling $\boldsymbol{X}_i|\boldsymbol{Z}_i = \boldsymbol{z}_i \sim p_{\{k:\boldsymbol{z}_{ik}=1\}}(\boldsymbol{x}_i)$.

Classification rule

Fuzzy and hard partition When the data distribution $p(\boldsymbol{x}_i)$ is known, the definition of $\boldsymbol{Z}_i|\boldsymbol{X}_i = \boldsymbol{x}_i$ is straightforward

$$\boldsymbol{Z}_i|\boldsymbol{X}_i = \boldsymbol{x}_i \sim \mathcal{M}_g(t_{i1}, \ldots, t_{ig}), \tag{1.11}$$

where t_{ik} is the conditional probability that $\boldsymbol{x}_i$ is drawn from component k which is defined by

$$t_{ik} = \frac{P(Z_{ik} = 1, \boldsymbol{X}_i = \boldsymbol{x}_i)}{P(\boldsymbol{X}_i = \boldsymbol{x}_i)} = \frac{\pi_k p_k(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)}. \tag{1.12}$$

Vector $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{ig})$ also defines a *fuzzy* partition which can be used to compute the risk associated to the *hard* partition $\boldsymbol{z}_i$.

Error risk and classification rule From this fuzzy partition, we can define the classification error $e(.)$ associated to $(\boldsymbol{z}_i, \boldsymbol{t}_i)$ by

$$e(\boldsymbol{z}_i, \boldsymbol{t}_i) = 1 - \sum_{k=1}^{g} (t_{ik})^{z_{ik}}. \tag{1.13}$$

The *maximum a posteriori* rule (MAP) minimizes the classification error by assigning an individual into the class having the largest probability. Thus, it defines the classification rule $r(.)$ as follows

$$\forall \boldsymbol{x}_i \in \mathcal{X},\ r(\boldsymbol{x}_i) = k \text{ if } \forall k'\ t_{ik} \geq t_{ik'}. \tag{1.14}$$

The evaluation of the risk of the classification error is a great advantage of the probabilistic methods, since the geometric ones cannot quantify the error risk associated to their classification rule.

1.2 Generalities on finite mixture models

Main idea These models assume that the observed individuals are independently drawn from the same distribution. We now quickly describe the semi-parametric mixture models which make few assumptions on the component distributions. Then, we describe the full parametric mixture models which assume that the component distributions are parametric ones. Note that the model description, presented here, is based on *Finite mixture models* by G.J. McLachlan and D. Peel [MP00]. In this thesis, we focus on the full parametric mixture models since they permit an easier interpretation of the classes.

1.2.1 Semi-parametric mixture models

Few constraints on the component distributions The semi-parametric approaches do not assume that the components follow parametric distributions. However, for reasons of identifiability, some constraints have to be imposed for the components. For instance, distributions have to belong to the family of uni-modal distributions or symmetric distributions [HWH07].

Inference The estimation of the component distributions can be performed by algorithms inspired from the EM algorithm [BCH09] which uses Kernel approaches [CHL10]. The R package `mixtools` [BCH09] allows us to cluster the data by using a semi-parametric mixture model.

Non-identifiability risk The semi-parametric mixture models are very flexible, so they can easily fit the data distribution. However, this flexibility involves an important risk of non-identifiability and a large variance of the estimated model. Furthermore, the class interpretation can be difficult since the components can not be summarized by few parameters as in the full parametric mixture models. So, in this thesis, we only study the parametric mixture models for which the general properties are now developed.

1.2.2 Full parametric mixture models

Generalities

Main idea These models make the supplementary assumption that each component follows a parametric distribution, so $p_k(\boldsymbol{x}_i) = p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k)$ where $\boldsymbol{\alpha}_k$ groups the parameters of component k. The individuals are also drawn by a parametric distribution $p(\boldsymbol{x}_i) = p(\boldsymbol{x}_i; \boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha})$ denotes the whole parameter, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ is the vector of the class proportions and where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_g)$ groups the parameters of the components.

Definition 1.7 (Finite parametric mixture model)**.** The pdf of the finite parametric mixture models with g components is defined by

$$p(\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k). \tag{1.15}$$

Interpretation via the parameters These models are more meaningful than the semi-parametric approaches since each class can be summarized by its proportion π_k and the parameters of its distribution $\boldsymbol{\alpha}_k$. The probabilities of the class memberships t_{ik} are also parametrized by $\boldsymbol{\theta}$, so they are now denoted by $t_{ik}(\boldsymbol{\theta})$ with

$$t_{ik}(\boldsymbol{\theta}) = \frac{\pi_k p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k)}{\sum_{k'=1}^{g} \pi_{k'} p(\boldsymbol{x}_i; \boldsymbol{\alpha}_{k'})}. \tag{1.16}$$

Components: trade off between relevance and number This approach is not really restrictive since a mixture of parametric distributions can approach any distribution with any precision, if all the distributions (component distributions and approached distribution) have the same support, just by increasing the number of components and the sample size (see the following example). Thus, a mixture of standard distributions can modelize highly complex distributions. However, the mixture is more meaningful when the number of classes stays small. Furthermore, as the mixture model is estimated on a *finite* sample, the practitioner searches the model performing the best *trade off* between its *bias* (with the "true" model) and its *variance* (caused by the fluctuations of the sampling). Therefore, it is important that the components follow standard distributions which are adapted to the data.

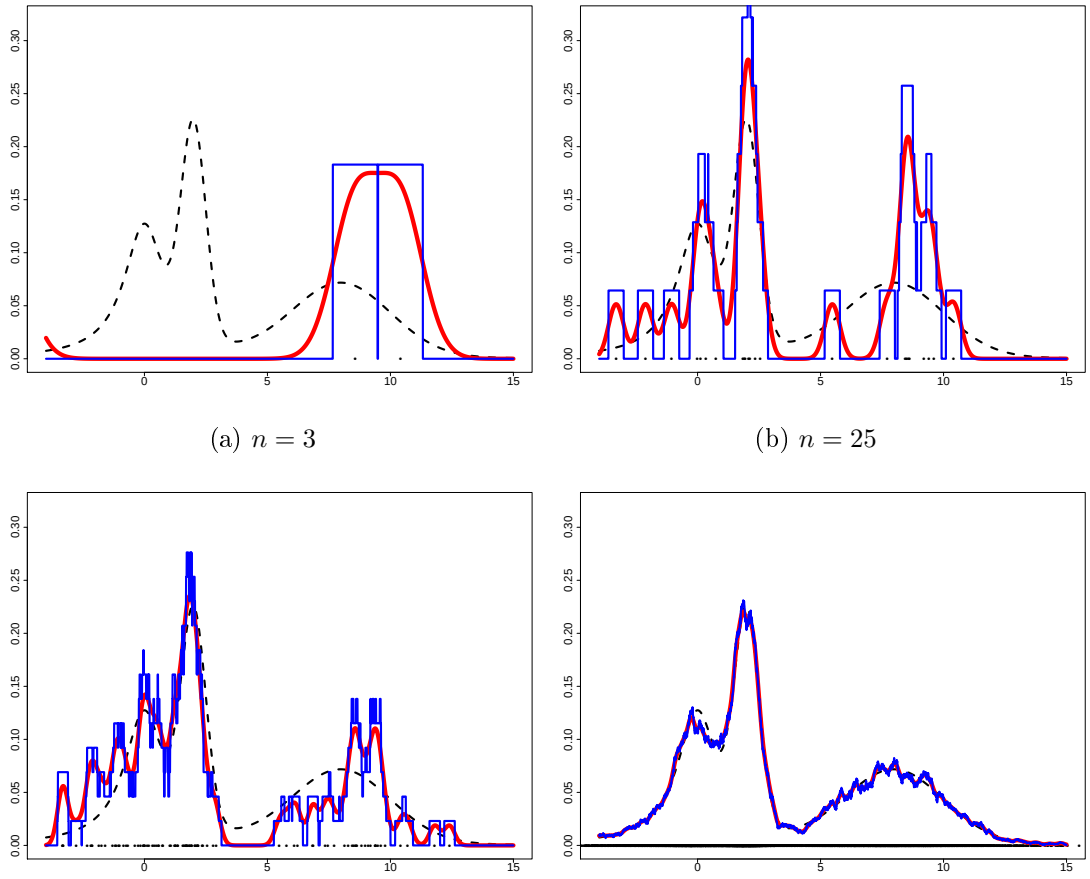
Example 1.8 (Density estimation and Parzen-Rosenblatt estimator)**.** *Let* $\mathbf{x}$ *to be the sample of size* n *where each individual* $x_i \in \mathbb{R}$ *is independently drawn by the distribution characterized by its pdf* $f(x)$*. The Parzen-Rosenblatt estimator [Par62] is defined as*

$$p(y; \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{1}{nh} K\left(\frac{y - x_i}{h}\right). \tag{1.17}$$

Note that (1.17) *defines a finite mixture model with* n *components where each proportion is equal to* $1/(nh)$ *and where the distribution of component* i *is parameterized by one* x_i*, for* $i = 1, \ldots, n$*. Under some regularity conditions on the function* $K(.)$ *and under some relations between the sample size* n *and the bandwidth* h*, the pdf* $p(y; \boldsymbol{\theta})$ *converges to the true pdf* $f(x)$ *(see for instance [ZD12] for more details). Figure 1.3 illustrates the Parzen-Rosenblatt estimator approaching a pdf by using the uniform kernel or the Gaussian one*

$$K_{uniform}(y) = \begin{cases} 1/2 & \text{if } |y| < 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{and } K_{Gaussian}(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}. \tag{1.18}$$

Mixture and kind of variables The parametric mixture models can cluster individuals by approaching the distribution of the variables in their native space. Obviously, the distributions of the components have to respect the nature of the variables. Thus, the mixture models are used to analyze different kinds of data sets. For instance, the mixture of Poisson distributions [KM07] can cluster integer data while the mixtures of Student distributions [PM00] can cluster the continuous ones, but the mixture models are also used to cluster networks [LBA10, FRW13], rank data [JB12] or functional data [JP14]. The case of categorical data is developed in Part I while Part II is focused on the clustering of mixed data. However, the most common mixture model is the Gaussian one that we detail now.



(a) $n = 3$ (b) $n = 25$

(c) $n = 100$ (d) $n = 1000$

Figure 1.3 – The pdf of the true distribution (black dotted curve), and its estimates obtained by the uniform kernel (thin blue curve) and by the Gaussian kernel (bold red curve) where the bandwidth is $h = \ln n$.

Gaussian mixture model

Main idea The *Gaussian mixture model* was introduced simultaneously to the parametric mixture model in order to cluster the Pearson's crab data set [Pea94]. It is a powerful tool to cluster continuous data whose success is due to two main reasons. On the one hand, its elliptical definition of a class is in accordance with the natural definition of a class. On the other hand, its computational tractability permits an easy inference. The Gaussian mixture model assumes that each random variable $\boldsymbol{X}_i|Z_{ik} = 1$ is an e-variate Gaussian variable whose the mean vector is denoted by $\boldsymbol{\mu}_k$ and whose the covariance matrix is denoted by $\boldsymbol{\Sigma}_k$, so

$$\boldsymbol{X}_i|Z_{ik} = 1 \sim \mathcal{N}_e(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{1.19}$$

Thus, we obtain the following definition of the Gaussian mixture model.

Definition 1.9 (Gaussian mixture model)**.** Let $\boldsymbol{x}_i \in \mathbb{R}^e$ be the continuous variable arisen from a Gaussian mixture model with g components. Its pdf is written as

follows

$$p(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) \text{ with } p(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) = \phi_e(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k), \tag{1.20}$$

where $\phi_e(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{e/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\right)$ and where $\boldsymbol{\alpha}_k = (\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$.

Class interpretation The Gaussian mixture model provides a summary of each class throughout its central position $\boldsymbol{\mu}_k$ and its dispersion matrix $\boldsymbol{\Sigma}_k$ relating its dependencies between the pairs of variables.

Parsimonious models When the samples are small, the information about the intra-class dependencies between variables is not present in the data. In general, the bias/variance trade off may be better if constraints on the parameter space are added. The resulting models are called parsimonious models. For instance, based on the spectral decomposition of $\boldsymbol{\Sigma}_k$ proposed in [BR93], fourteen parsimonious models were built [CG95]. As these models are sensitive to the unit of measurement of the variables, a new family of Gaussian mixture model, named RTV, was proposed by C. Biernacki and A. Lourme [BL13]. Finally, note that the spectral decomposition of $\boldsymbol{\Sigma}_k$ can be used to cluster high-dimensional data [BB14].

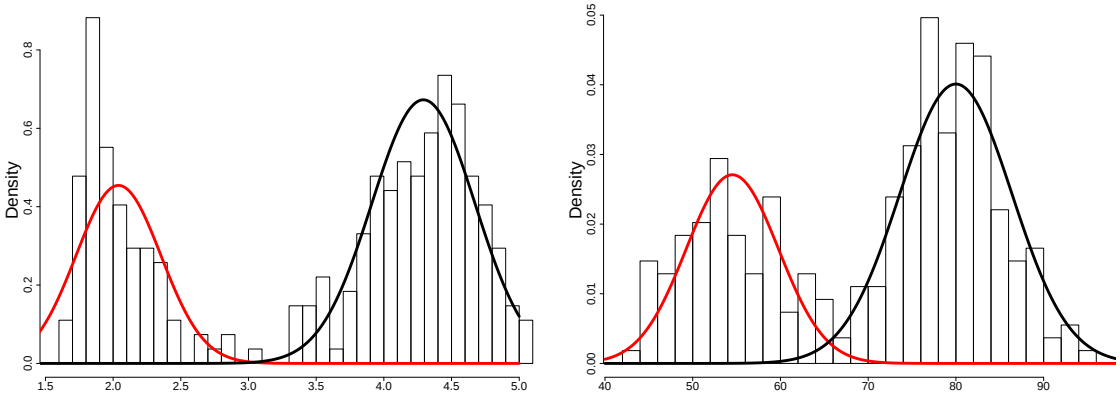
Softwares Many softwares performing the estimation of the Gaussian mixture models are available. Their impact in the diffusion of these models was decisive. Among them, one can cite the three followers: Mclust [FR06], Mixmod [LIL+12] and Mixtool [BCHY09].

Faithful data set 1.10 (Gaussian mixture model clustering)**.**

As both variables of `Faithful` data set are continuous, we estimate a bi-component Gaussian mixture model.
The histograms of both variables are displayed in Figure 1.4 and the estimated marginal pdf of both components are superimposed.
The model summarizes the data set as follows.

- The majority class ($\pi_1 = 0.64$) is the class of the strong eruptions since ($\boldsymbol{\mu}_k = (4.29, 80.00)$).
- The minority class ($\pi_2 = 0.36$) is the class of the weak eruptions since ($\boldsymbol{\mu}_k = (2.04, 54.51)$).

The class of the strong eruptions is more dispersed than the class of the weak eruptions since the variance of both variables are larger (respectively $(0.15, 40.90)$ and $(0.10, 57.68)$). Even if the variables are positively correlated in both classes, their dependency strength is larger in the minority class than in the majority one. Indeed, the coefficient of correlation is equal to 0.50 in the class of the weak eruptions while it is equal to 0.23 in the class of the strong eruptions.



(a) Eruptions (b) Waiting

Figure 1.4 – Histograms and marginal densities of the bi-component mixture model for `Faithful` data set.

Faithful data set 1.11 (Comparison between both clustering results)**.**

We remark that the summary provided by the Gaussian mixture model is more precise than the suumary obtained by the *K-means* algorithm. Indeed, the *K-means* algorithm does not consider the class proportions, so it implicitly assumes that both kinds of eruptions are equiprobable. Furthermore, the Gaussian mixture model provides an analysis of the intra-class dependencies as displayed by Figure 1.5 which draws the scatter plot of the partition and the ellipses of equiprobability.

1.2.3 Mixture model with conditional independence assumption

Main idea The conditional independence model (CIM), also known as *naive Bayes* or *latent class model*, is a mixture model assuming the conditional independence between variables. So, the conditional probability of the e-variate random variable $\boldsymbol{X}_i = (X_i^1, \ldots, X_i^e)$ is written as

$$P(\boldsymbol{X}_i|Z_{ik} = 1) = \prod_{j=1}^{d} P(X_i^j|Z_{ik} = 1). \tag{1.21}$$

Obviously, this assumption is weaker than the global independence assumption. Indeed, the dependency between the variables is modelized by the structure in classes of the distribution.

Definition 1.12 (CIM)**.** The CIM model is a mixture model assuming the conditional independence between the variables. Thus, the pdf of the individual $\boldsymbol{x}_i$ arisen from

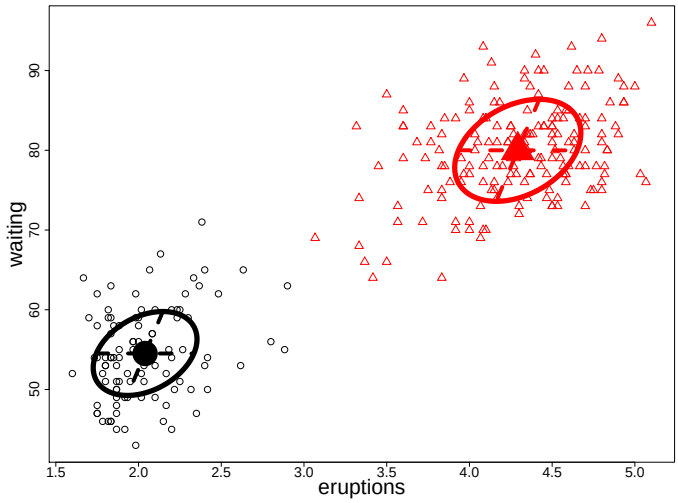


Figure 1.5 – Outputs of the `Faithful` cluster analysis performed by a bi-component Gaussian mixture model. The scatter plot indicates the partition by the colors and the thin symbols while the intra-class dependencies are depicted by the ellipses of equiprobability of the Gaussian components: the individuals belonging to the class of the strong eruptions are displayed with red triangles and those belonging to the class of the weak eruptions are displayed with black circles.

the CIM model is written as

$$p(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \prod_{j=1}^{d} p(x_i^j;\boldsymbol{\alpha}_{kj}), \tag{1.22}$$

where $\boldsymbol{\alpha}_{kj}$ denotes the margin parameters related to variable j for component k.

Example 1.13 (Correlated data drawn by CIM)**.** *Figure 1.6 displays a sample drawn by the* CIM *model with four bivariate Gaussian components. In this example, it is straightforward that both variables are not independent.*

Meaningful results in practice The CIM model obtains good results in practice since it requires few parameters, as discussed by D.J. Hand and K. Yu [HY01]. Thus, it can realize a good trade off between the bias and the variance. Its sparsity is a great advantage for the small data sets since the information of the intra-class dependency is generally not present. The success of the CIM model is also explained by its meaningful aspect. Indeed, when the marginal distributions of the components are classical (for instance when they belong to the exponential family), each class can be summarized by the parameters of its margins.

Bias of the intra-class correlated data When the conditional independence assumption is violated, the CIM model suffers from severe biases. In such a case, if the class number is known then the partition becomes biased, while if the class number is unknown then the CIM model overestimates it to better fit the data (see for instance the application presented in [VHH09]).

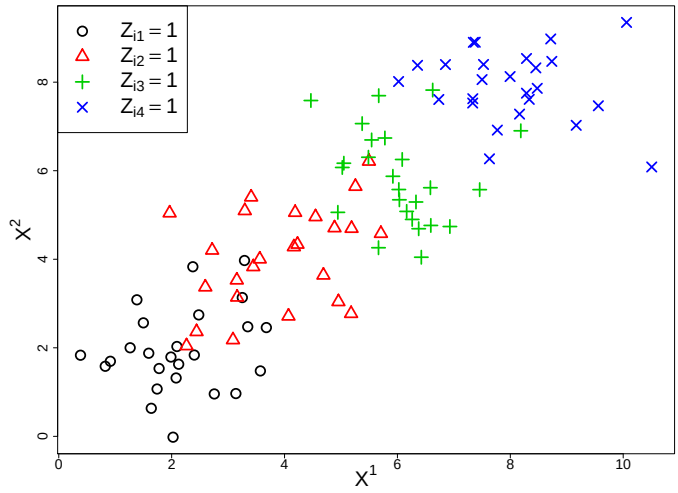


Figure 1.6 – Bivariate sample arisen from the CIM model with a global dependency between variables: $p(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{4} \frac{1}{4}\mathcal{N}_2(2\boldsymbol{k}, \boldsymbol{I})$.

Example 1.14 (Biased partition)**.** *Let the bi-components homoscedastic Gaussian mixture model, whose the pdf of* $\boldsymbol{x}_i \in \mathbb{R}^2$ *is denoted by* $f(\boldsymbol{x}_i) = \frac{1}{3}\phi_2(\boldsymbol{x}_i;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}) + \frac{2}{3}\phi_2(\boldsymbol{x}_i;\boldsymbol{\mu}_2,\boldsymbol{\Sigma})$ *where* $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ *and* $\rho \neq 0$*. The optimal classification rules (minimizing the Bayes' error) is*

$$r_{Bayes}(\boldsymbol{x}_i) : \; z_{ik} = \begin{cases} 1 & \text{if } (\boldsymbol{x}_i - \boldsymbol{\mu}_k)'\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k) < (\boldsymbol{x}_i - \boldsymbol{\mu}_\ell)'\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_\ell) \text{ with } \ell = 2 - k \\ 0 & \text{otherwise.} \end{cases}$$

Let the CIM *model having the same one-dimensional margin distributions than the model defined by* $f(\boldsymbol{x}_i)$*. The pdf of this* CIM *model is written as*

$$p(\boldsymbol{x}_i;\boldsymbol{\theta}) = \frac{1}{3}\phi_2(\boldsymbol{x}_i;\boldsymbol{\mu}_1,\boldsymbol{\Gamma}) + \frac{2}{3}\phi_2(\boldsymbol{x}_i;\boldsymbol{\mu}_2,\boldsymbol{\Gamma}) \text{ with } \boldsymbol{\Gamma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The classification rules associated to this model is

$$r_{\text{CIM}}(\boldsymbol{x}_i) : \; z_{ik} = \begin{cases} 1 & \text{if } (\boldsymbol{x}_i - \boldsymbol{\mu}_k)'(\boldsymbol{x}_i - \boldsymbol{\mu}_k) < (\boldsymbol{x}_i - \boldsymbol{\mu}_\ell)'(\boldsymbol{x}_i - \boldsymbol{\mu}_\ell) \text{ with } \ell = 2 - k \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the partition estimated by the CIM *model is biased since the set* Ω *which groups the individuals where the classification rules disagree has not a measure equals to zero*

$$\Omega = \{\boldsymbol{x}_i \in \mathbb{R}^2 : r_{Bayes}(\boldsymbol{x}_i) \neq r_{\text{CIM}}(\boldsymbol{x}_i)\}.$$

1.2.4 The identifiability of the mixture models

Essential condition The classes provided by a cluster analysis are interpreted throughout the parameters of the components. It is also crucial that the parameters are unique for a fix distribution. Thus, two models having the same distribution must have the same parameters, we also refer to the identifiability of the model.

Definition 1.15 (Identifiability)**.** Let two mixture models having the same nature for all the components and respectively parametrized by $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, then the model is identifiable if

$$\forall \boldsymbol{x}_i \in \mathcal{X}\ p(\boldsymbol{x}_i;\boldsymbol{\theta}) = p(\boldsymbol{x}_i;\boldsymbol{\theta}') \Leftrightarrow \boldsymbol{\theta} = \boldsymbol{\theta}'. \tag{1.23}$$

Problem due to the relabeling of the components Obviously, the mixture models are not strictly identifiable but only up to label swapping, since the classes can be relabeled as illustrated by the following example. However, this case of non identifiability is not a severe drawback since the interpretation of the partition stays identical.

Example 1.16 (Relabeling of the components)**.** *Let the bi-component mixture models parametrized by* $\boldsymbol{\theta} = (\pi_1, \pi_2, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ *and* $\boldsymbol{\theta}' = (\pi_2, \pi_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_1)$*, then both parameter sets define the same distribution:*

$$\begin{aligned}\forall \boldsymbol{x}_i \in \mathcal{X},\ p(\boldsymbol{x}_i;\boldsymbol{\theta}) &= \pi_1 p(\boldsymbol{x}_i;\boldsymbol{\alpha}_1) + \pi_2 p(\boldsymbol{x}_i;\boldsymbol{\alpha}_2)\\ &= \pi_2 p(\boldsymbol{x}_i;\boldsymbol{\alpha}_2) + \pi_1 p(\boldsymbol{x}_i;\boldsymbol{\alpha}_1)\\ &= p(\boldsymbol{x}_i;\boldsymbol{\theta}').\end{aligned}$$

Relabeling of the components and inference Note that this condition of non identifiability does not disturb the EM algorithm (see next section) but, in a Bayesian framework, it can involve the label switching phenomenon [Ste00a].

Weakly identifiable mixture models In order to avoid the problem due to the component relabeling, the notion of weak identifiability was introduced for the mixture models [Tei63].

Definition 1.17 (Weak identifiability)**.** The mixture model having $p(\boldsymbol{x}_i;\boldsymbol{\theta})$ as pdf is weakly identifiable when

$$\forall \boldsymbol{x}_i \in \mathcal{X}\ p(\boldsymbol{x}_i;\boldsymbol{\theta}) = p(\boldsymbol{x}_i;\boldsymbol{\theta}') \Leftrightarrow \boldsymbol{\theta} \text{ and } \boldsymbol{\theta}' \text{ are equivalent.} \tag{1.24}$$

Many mixture models are weakly identifiable, among them one can cite the finite mixtures of Gaussian distributions, the finite mixtures of Gamma distributions and the finite mixtures of Poisson distributions. Three main articles study the identifiability of the mixture models [Tei63, Tei67, YS68]. In order to give an idea of their reasoning, we present the theorem of H. Teicher [Tei63] which demonstrates the weak identifiability of some univariate mixture models (for instance the univariate Gaussian mixture model).

Theorem 1.18 (Conditions of weak identifiability [Tei63])**.** *Let* $\mathcal{P} = \{P\}$ *be a family of one-dimensional cumulative distribution functions with transforms* $\psi(t)$ *defined for* $t \in S_\psi$ *(the domain of definition of* ψ*) such that the mapping* $M : P \mapsto \psi$ *is linear and one-to-one. Suppose that there exists a total ordering* $(\preceq)$ *of* $\mathcal{P}$ *such that* $F_1 \prec F_2$ *implies (i)* $S_{\psi_1} \subseteq S_{\psi_2}$*, (ii) the existence of some* $t_1 \in \bar{S}_{\psi_1}$ *(*t_1 *being independent of* ψ_2*) such that* $\lim_{t \to t_1} \psi_2(t)/\psi_1(t) = 0$*. Then, the class of all finite mixtures of* $\mathcal{P}$ *is weakly identifiable.*

Proposition 1.19 (Identifiability of the univariate Gaussian mixture model [Tei63])**.** *The class of all finite mixtures of univariate Gaussian distributions is weakly identifiable.*

Proof. Let $\Phi_1(.;\mu,\sigma^2)$ denote the Gaussian cumulative distribution function with mean μ and variance $\sigma^2 > 0$. Its bilateral Laplace transform is given by $\psi(t) = \exp(\sigma^2 t^2/2 - \mu t)$. Order the family lexicographically by $\Phi_1(x_i;\mu_1,\sigma_1^2) \prec \Phi_1(x_i;\mu_2,\sigma_2^2)$ if $\sigma_1 > \sigma_2$ or if $\sigma_1 = \sigma_2$ but $\mu_1 < \mu_2$. Then, Theorem 1.18 applies with $S_\psi = (-\infty,\infty)$ and $t_1 = +\infty$. □

However, some mixture models are not identifiable but are of interest since they provide meaningful results in practice and since their parameters seem identifiable.

Generic identifiability As the identifiability condition could be too stringent, a less restrictive condition, named generic identifiability, was introduced.

Definition 1.20 (Generic identifiability)**.** A model is generically identifiable when the parameter space, where the model is not identifiable up to the component relabeling, has a measure equal to zero.

Based on the conditional independence assumption, E.S Allman, C. Matias and J.A. Rhodes [AMR09] find a sufficient condition of the generic identifiability for the mixture of multinomial distributions. This model is studied in Part I, where details on the proof of its generic identifiability are given. However, some mixture models are not generically identifiable, so their parameters cannot be interpreted as illustrated by the following example.

Example 1.21 (Non-generic identifiability of the mixture of uniform distributions)**.** *Let $x_i \in [a_1;b_2]$ drawn by the bi-component mixture model of uniform distributions whose the pdf is*

$$p(x_i;\boldsymbol{\theta}) = \pi U[a_1,b_1] + (1-\pi)U[a_2,b_2], \tag{1.25}$$

where $\boldsymbol{\theta} = (\pi,a_1,b_1,a_2,b_2)$, $U[.,.]$ denotes the pdf of a uniform distribution and where $a_1 < b_1$, $a_2 < b_2$, $a_1 < a_2$, $b_1 < b_2$ and $a_2 < b_1$. This model is equivalent to the following tri-component mixture model of uniform distributions whose the pdf is

$$p(x_i;\boldsymbol{\theta}') = \varepsilon_1 U[a_1,a_2] + \varepsilon_2 U[a_2,b_1] + (1-\varepsilon_1-\varepsilon_2)U[b_1,b_2], \tag{1.26}$$

where $\boldsymbol{\theta}' = (\varepsilon_1,\varepsilon_2,a_1,a_2,b_1,b_2)$, $\varepsilon_1 = \pi\frac{a_2-a_1}{b_1-a_1}$ and $\varepsilon_2 = \pi\frac{b_1-a_2}{b_1-a_1} + (1-\pi)\frac{b_1-a_2}{b_2-a_2}$. Thus, the distribution of x_i can be modelized with two different parametrizations $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$.

1.3 Parameter estimation

Structure of this section The clustering of a data set by a finite mixture model requires the estimation of the model parameters. In this section, we present the two most popular estimates in the mixture model context: the maximum likelihood estimate (further denoted by MLE) and the maximum *a posteriori* estimate (further denoted by MAPE). For both estimates, we present the estimation algorithms and their features specific to the mixture models.

Running example In this section, definitions and algorithms are given in a generalize case and they are illustrated by the following running example extracted from the article *Bayesian Modelling and Inference on Mixtures of Distributions* of J.M. Marin, K. Mengersen and C.P. Robert [MMR05].

Running example 1.22 (Bi-component univariate Gaussian mixture)**.**

Let $x_i \in \mathbb{R}$ and let the bi-component univariate Gaussian mixture model whose the pdf is written as

$$p(x_i;\boldsymbol{\theta}) = \sum_{k=1}^{2} \pi_k \phi_1(x_i;\mu_k,\sigma_k^2), \tag{1.27}$$

where $\boldsymbol{\theta} = (\pi_1,\mu_1,\sigma_1^2,\pi_2,\mu_2,\sigma_2^2)$ and where $\phi_1(.;\mu_k,\sigma_k^2)$ denotes the pdf of the univariate Gaussian variable $\mathcal{N}_1(\mu_k,\sigma_k^2)$.

1.3.1 Maximum likelihood estimation

In this section, we define the likelihood function and the estimate of the maximum likelihood whose we describe the main properties. Then, we introduce the notion of complete-data and the likelihood function associated to it, named complete-data likelihood function.

Maximum likelihood estimate

Main idea The likelihood function holds the whole information contained in the data set. However, as it is more comfortable to work with the logarithm of this function, we present the definitions of both functions.

Definition 1.23 (Likelihood function)**.** For an *i.i.d.* sample $\mathbf{x}$, this function computed at the point $\boldsymbol{\theta}$ is defined by $p(\mathbf{x};\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i;\boldsymbol{\theta})$.

Definition 1.24 (Log-likelihood function)**.** The log-likelihood function computed from the *i.i.d.* sample $\mathbf{x}$ and evaluated on the point $\boldsymbol{\theta}$ is defined by

$$L(\boldsymbol{\theta};\mathbf{x}) = \sum_{i=1}^{n} \ln p(\boldsymbol{x}_i;\boldsymbol{\theta}). \tag{1.28}$$

Running example 1.25 (Likelihood and log-likelihood functions)**.**

For an i.i.d. sample $\mathbf{x}$ composed with n individuals $x_i \in \mathbb{R}$, then the likelihood and the log-likelihood functions are defined as

$$p(\mathbf{x};\boldsymbol{\theta}) = \prod_{i=1}^{n}\sum_{k=1}^{2} \pi_k \phi_1(x_i;\mu_k,\sigma_k^2) \text{ and } L(\boldsymbol{\theta};\mathbf{x}) = \sum_{i=1}^{n} \ln \sum_{k=1}^{2} \pi_k \phi_1(x_i;\mu_k,\sigma_k^2). \tag{1.29}$$

In a frequentist framework, we want to infer according to the information given by the data. So, a natural approach is to search the maximum likelihood estimate (MLE), denoted by $\hat{\boldsymbol{\theta}}$.

Definition 1.26 (MLE)**.** The maximum likelihood estimate is defined by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}; \mathbf{x}). \tag{1.30}$$

Thus, if the log-likelihood function (or similarly the likelihood function) is twice differentiable — this condition is generally verified for the mixture models —, the MLE is obtained by solving the equations which annul of the gradient and which give a non positive definite Hessian matrix

$$\nabla L(\boldsymbol{\theta}; \mathbf{x}) = \mathbf{0}. \tag{1.31}$$

Properties of the MLE The MLE is a popular estimate since it has — under few restrictive conditions — good properties:

— It is unique with a probability tending to 1 as the sample size grows to infinity.
— It is consistent.
— It is asymptotically unbiased.
— It is asymptotically Gaussian.
— It asymptotically minimizes the Kullback-Leibler divergence.

Details on the conditions involving these properties are given in *Theory of Point Estimation*, chapter 6 by E.L. Lehmann and G. Casella [LC98]. As the MLE has good properties, it is natural to study its existence and, if it exists, the methods performing its estimation.

Degeneracy The existence and the uniqueness of the MLE is not guarantee for the mixture models. Indeed, the log-likelihood function can be not upper bounded (see the running example). In such a case, the likelihood function can tend to the infinity. This situation, named *degeneracy* [Bie07], involves inconsistent estimators. In such a case, the estimator verifying (1.31) and involving a finite log-likelihood is also searched.

Running example 1.27 (Degeneracy)**.**

We want to fit a bi-component univariate Gaussian mixture model on the sample $\mathbf{x}$. If we assume that $\mu_1 = x_1$, we observe a model degeneracy since

$$\lim_{\substack{\sigma_1^2 \to 0 \\ \sigma_2^2 > 0}} L(\boldsymbol{\theta}; \mathbf{x}) = \infty. \tag{1.32}$$

Definition 1.28 (MLE and unbounded log-likelihood)**.** When the log-likelihood function is not upper bounded, the MLE is defined as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \{\boldsymbol{\theta}: L(\boldsymbol{\theta};\mathbf{x}) < +\infty \text{ and } \nabla L(\boldsymbol{\theta};\mathbf{x}) = \mathbf{0}\}}{\operatorname{argmax}} L(\boldsymbol{\theta}; \mathbf{x}). \tag{1.33}$$

Note that the log-likelihood function has generally several local optima which increase the difficulty in finding the MLE. This phenomenon is now illustrated on the running example.

Running example 1.29 (Log-likelihood optima with unknown means)**.**

We generate a sample of size 150 from the bi-component univariate mixture model whose the parameters are defined as follows

$$\pi_1 = 1/3,\ \pi_2 = 2/3,\ \mu_1 = -1,\ \mu_2 = 3.5 \text{ and } \sigma_1^2 = \sigma_2^2 = 1. \tag{1.34}$$

We assume that the proportions and the variances are known and we want to estimate the means by maximum likelihood. Figure 1.7 displays the log-likelihood function according to the values on both parameters. One can observe that this function has two optima. The global one is located around $(-1, 3.5)$ while the local one is located around $(3.5, -1)$.

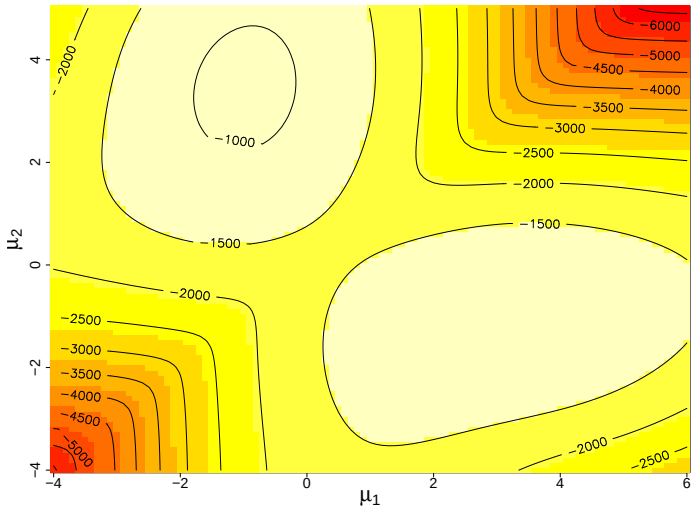


Figure 1.7 – Log-likelihood values for the sample of size 150 according to the values of (μ_1, μ_2).

No explicit solution For the mixture models, the search of the MLE involves to solve equations having no analytical solution. The direct computation of the MLE is not easy because of the specific form of the log-likelihood function (sum of logarithms of sums of pdf). So, some iterative procedures have also be used, like the Newton-Raphson algorithm detailed, for instance, in *Numerical optimization: theoretical and practical aspects* by J.F. Bonnans, J.C. Gilbert, C. Lemarechal and C.A. Sagastizabal [BGLS06]. However, its implementation is often complex since it involves the computation of the derivatives of the likelihood.

The mixture models have been disseminated because of the invention of the EM algorithm whose the implementation is simple since no derivative of the likelihood is involved. As this algorithm is specialized for the *missing data*, we firstly define the notion of *complete-data* for the mixture model, and we secondly detail it.

Observed data and complete-data For a sample $\mathbf{x}$, the generative models assume that the drawing of each individual $\boldsymbol{x}_i$ involves the preliminary sampling of $\boldsymbol{z}_i$ (see Paragraph Generative model in Section 1.1.3). As vector $\mathbf{z}$ is unobserved, it is considered as a missing value. Thus, $\mathbf{x}$ is named the observed data while the couple $(\mathbf{x}, \mathbf{z})$ is named the complete-data. In the same way, the log-likelihood function computed on $(\mathbf{x}, \mathbf{z})$ is named the *complete-data log-likelihood* function and it is written as follows

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) &= \sum_{i=1}^{n} \ln p(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \ln \left(\prod_{k=1}^{g} (\pi_k p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k))^{z_{ik}} \right) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \ln \left(\pi_k p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) \right), \end{aligned} \tag{1.35}$$

By starting from the relation between the pdf $p(\boldsymbol{x}_i; \boldsymbol{\theta}) p(\boldsymbol{z}_i | \boldsymbol{x}_i; \boldsymbol{\theta}) = p(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta})$, one can deduce the following relation

$$L(\boldsymbol{\theta}; \mathbf{x}) = L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) + e(\mathbf{z}, \mathbf{x}; \boldsymbol{\theta}). \tag{1.36}$$

where $e(\mathbf{z}, \mathbf{x}; \boldsymbol{\theta}) = -\sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \ln t_{ik}(\boldsymbol{\theta})$. Thus, $L(\boldsymbol{\theta}; \mathbf{x}) \geq L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z})$ since $t_{ik}(\boldsymbol{\theta})$ is a probability.

Running example 1.30 (Complete-data log-likelihood).

For the bi-component univariate Gaussian mixture model, the complete-data log-likelihood function computed on $\boldsymbol{\theta}$ for the sample $\mathbf{x}$ and the partition $\mathbf{z}$ is defined as

$$L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \sum_{k=1}^{2} n_k \ln \frac{\pi_k}{\sigma_k} - \frac{1}{2} \sum_{k=1}^{2} \sum_{i=1}^{n} z_{ik} \frac{(x_i - \mu_k)^2}{\sigma_k^2} - n \ln \sqrt{2\pi}. \tag{1.37}$$

1.3.2 Algorithms for a maximum likelihood estimation

The MLE of the mixture models is generally obtained via an EM algorithm. This section is devoted to the presentation of this algorithm and of its extensions.

The EM algorithm

Presentation of the EM algorithm The *Expectation-Maximization* algorithm (further denoted by EM) was proposed by Dempster, Laird, and Rubin in 1977 [DLR77]. Its domains of application are vaster than the mixture models since it is specialized in the case of missing values. In the context of the mixture models, the class memberships of the individuals are interpreted as missing values. Note that this algorithm allows to cluster a data set with missing values by making weakly restrictive assumptions. The main advantage of this algorithm is its simplicity since it optimizes the likelihood function without computing its derivatives. Furthermore, since its implementation can be parallelized, it stays efficient when its is confronted with large data sets. This section is just an overview of this algorithm, the reader needing more details could refer to *The EM algorithm and Extensions* by G.J. McLachlan and T. Krishnan [MK97].

Definition of the EM algorithm The EM algorithm is an iterative one, starting from an initial value of the parameter which alternates between the two following steps: the computation of the expectation of the complete-data log-likelihood (E step) and its maximization (M step).

Algorithm 1.31 (The EM algorithm)**.**

Starting from an initial value $\boldsymbol{\theta}^{[0]}$, iteration $[r]$ is written as

— **E step**: calculate $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{[r]})$ where

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{[r]}) = \mathbb{E}_{\boldsymbol{\theta}^{[r]}} \left[L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) \right], \tag{1.38}$$

— **M step**: select $\boldsymbol{\theta}^{[r+1]}$ such as

$$\boldsymbol{\theta}^{[r+1]} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{[r]}). \tag{1.39}$$

Stopping criteria Two criteria are generally used. The most common one consists in stopping the algorithm when the increase of the log-likelihood is lower than the threshold ε chosen by the user, so when

$$L(\boldsymbol{\theta}^{[r+1]}; \mathbf{x}) - L(\boldsymbol{\theta}^{[r]}; \mathbf{x}) < \varepsilon. \tag{1.40}$$

The second one fixes in advance the number of iterations performed by the algorithm.

EM avoids the difficulties inherent to the mixture structure The estimation of the parameters is doable for a mixture model (without constraint between classes) if the inference of such a model can be made when the class memberships of the individuals are known. Thus, a mixture model whose the estimate is tractable in the discriminant analysis —so, when the labels are known— can always be inferred in a clustering problem (for instance all the mixture models whose the component

distributions belong to the exponential family and having no constraint together can be explicitly computed).

Running example 1.32 (EM algorithm)**.**

We consider the classical Gaussian mixture model whose the components are defined by (1.20). Iteration $[r]$ of the EM algorithm is written as

— **E step**: calculate conditional probabilities

$$t_{ik}(\boldsymbol{\theta}^{[r]}) = \frac{\pi_k^{[r]} p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k^{[r]})}{p(\boldsymbol{x}_i; \boldsymbol{\theta}^{[r]})}. \tag{1.41}$$

— **M step**: maximization of the expectation of the complete-data log-likelihood

$$\pi_k^{[r+1]} = \frac{n_k^{[r]}}{n}, \quad \mu_k^{[r+1]} = \frac{1}{n_k^{[r]}} \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]}) \boldsymbol{x}_i,$$

$$\sigma_k^{2[r+1]} = \frac{1}{n_k^{[r]}} \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]})(\boldsymbol{x}_i - \mu_k^{[r+1]})^2, \tag{1.42}$$

where $n_k^{[r]} = \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]})$.

Properties of the EM algorithm Under few restrictive assumptions [Wu83], the EM algorithm provides a sequence of estimates $\boldsymbol{\theta}^{[r]}$ which converges to a local optimum of the log-likelihood function. This optimum only depends on the initialization $\boldsymbol{\theta}^{[0]}$. Indeed, the likelihood function increases at each iteration of the EM algorithm

$$\forall [r],\ L(\boldsymbol{\theta}^{[r+1]}; \mathbf{x}) \geq L(\boldsymbol{\theta}^{[r]}; \mathbf{x}). \tag{1.43}$$

This algorithm converges to a local optimum. So, it is mandatory to perform several different initializations in order to hope to get the MLE. An other drawback of the EM algorithm is its speed of convergence. Indeed, this algorithm can converge slowly, especially when the classes are overlapped. Thus, many authors have been interested in the acceleration of the EM algorithm (see for instance [VR08, BR12]). After the description of the EM algorithm applied on the Gaussian mixture models, we present three of its extensions reducing its drawbacks.

Running example 1.33 (Naive example and EM algorithm).

We consider the bi-component univariate Gaussian mixture model with known proportions and variances. In such a case, the M step only consists in computing $\mu_k^{[r+1]}$ since the other parameters are known. Figure 1.8 displays the values of the likelihood computed at each iterations of two runs of the EM algorithm. The run printed with triangles is initialized at $(-1, -1/2)$ and converges to the MLE while the run printed with squares is initialized at $(3.5, 3)$ and converges to a local maximum of the likelihood function.

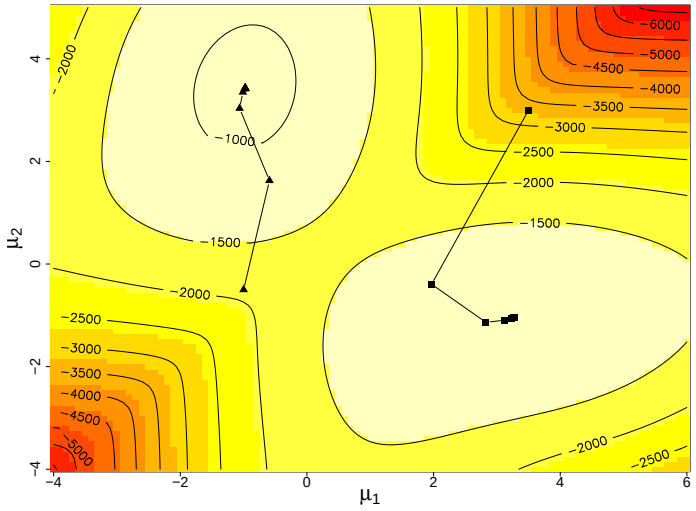


Figure 1.8 – Log-likelihood values associated with two sequences of parameters providing by an EM algorithm.

Extensions of the EM algorithm

GEM algorithm Sometimes, the solution of the M step is not explicit. In such a case, the Generalized-EM (GEM) algorithm can be used. The M step is also replaced by a GM one which only requires the increase of the expectation of the complete-data log-likelihood. Thus, at iteration $[r]$, the E step is unchanged while the GM step determines $\boldsymbol{\theta}^{[r+1]}$ such as

$$Q(\boldsymbol{\theta}^{[r+1]}; \boldsymbol{\theta}^{[r]}) \geq Q(\boldsymbol{\theta}^{[r]}; \boldsymbol{\theta}^{[r]}). \tag{1.44}$$

The GEM algorithm keeps the monotonic property of the increase of the likelihood function for each iteration which is inherited from the EM algorithm. However, this algorithm requires more iterations than the EM since its convergence is slower.

SEM and SAEM algorithms In order to overcome the three main drawbacks of the EM algorithm (*i.e.* strong dependency with the initialization point, local optimum convergence and slow convergence), the Stochastic-EM (SEM) algorithm was proposed [CD+87]. The algorithm incorporates a stochastic step (S step) between the E step and the M step directed by the *random imputation principle*. The sequence generated by the SEM algorithm converges to a unique stationary distribution close to $p(\boldsymbol{\theta}|\mathbf{x})$.

Algorithm 1.34 (The SEM algorithm for the mixture models)**.**

Starting from an initial value $\boldsymbol{\theta}^{[0]}$, iteration $[r]$ is written as

— **E step**: calculate the conditional probabilities

$$t_{ik}(\boldsymbol{\theta}^{[r]}) = \frac{\pi_k^{[r]} p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k^{[r]})}{p(\boldsymbol{x}_i; \boldsymbol{\theta}^{[r]})}. \tag{1.45}$$

— **S step**: sample the class membership such as

$$\boldsymbol{z}_i^{[r]} \sim \mathcal{M}(t_{i1}(\boldsymbol{\theta}^{[r]}), \ldots, t_{ig}(\boldsymbol{\theta}^{[r]})). \tag{1.46}$$

— **M step**: select $\boldsymbol{\theta}^{[r+1]}$ such as

$$\boldsymbol{\theta}^{[r+1]} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}^{[r]}). \tag{1.47}$$

The algorithm is stopped after a number of iterations chosen by the user. Note that, another version of this algorithm, named SAEM, provides an almost surely convergence to the unique stationary distribution [CD92]. It is a trade off between a version *simulated annealing*-like EM algorithm and the SEM algorithm. Indeed, the A step (annealing) is introduced after the S step in order to reduce the impact of the random perturbations performed by the S step. This reduction increases with the number of iterations. Thus, when the SAEM starts, it works like the SEM algorithm then it tends to the EM algorithm when the number of iterations increases.

CEM algorithm The Classification-EM (CEM) algorithm [CG92] is a general algorithm to compute the estimate and to find the partition under the classification approach. Thus, it provides the couple $(\boldsymbol{\theta}, \mathbf{z})$ maximizing the complete-data log-likelihood

$$\underset{(\boldsymbol{\theta}, \mathbf{z})}{\operatorname{argmax}}\ L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}). \tag{1.48}$$

Even if the estimate of the maximum complete-data is biased and inconsistent [Gov10], its results can be better than those of the MLE when the sample size is small and when the classes are well separated. Furthermore, the CEM algorithm converges faster than the EM algorithm. The CEM algorithm incorporates a classification step between the E step and the M step according to the MAP principle. Thus, its convergence speed is expected to be faster than the convergence speed of the EM algorithm.

Algorithm 1.35 (The CEM algorithm for the mixture models)**.**

Starting from an initial value $\boldsymbol{\theta}^{[0]}$, its iteration $[r]$ is written as

— **E step**: calculate $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{[r]})$ where

$$t_{ik}(\boldsymbol{\theta}^{[r]}) = \frac{\pi_k^{[r]} p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k^{[r]})}{p(\boldsymbol{x}_i; \boldsymbol{\theta}^{[r]})}. \tag{1.49}$$

— **C step**: minimize $e(\mathbf{z}, \mathbf{x}; \boldsymbol{\theta}^{[r]})$ so

$$z_{ik}^{[r]} = \begin{cases} 1 & \text{if } t_{ik}(\boldsymbol{\theta}^{[r]}) \geq t_{i\ell}(\boldsymbol{\theta}^{[r]}) \ \forall \ell = 1, \ldots, g \\ 0 & \text{otherwise.} \end{cases} \tag{1.50}$$

— **M step**: select $\boldsymbol{\theta}^{[r+1]}$ such as

$$\boldsymbol{\theta}^{[r+1]} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}^{[r]}). \tag{1.51}$$

Remark 1.36 (CEM for the spherical Gaussian mixture model and K-means algorithm)**.** The *K-means* algorithm is equivalent to the CEM one when the model at hand is the spherical Gaussian mixture model with equal proportions [CG91].

Drawbacks of the maximum likelihood approaches Even if the extensions of the EM algorithm reduce its main drawbacks, three problems stay inherent to the maximum likelihood approach applied on mixture models.

— The first one is the difficulty to find the global maximum of the likelihood function. Note that this problem is more present when the samples are small since the likelihood function can be very lumpy.

— The second one is due to the upper-bound of the likelihood function. Indeed, if this function is upper-bounded for categorical data, it can be upper-unbounded in other situations (see for instance the heteroscedastic Gaussian mixture model). Thus, the estimate returned by the algorithm can be on the degeneracy way. In such a case, this estimate is also biased and inconsistent.

— The third one is about the regularity conditions which are often violated for the small data sets. Thus, the estimation can involve an over-fitting.

1.3.3 Maximum *a posteriori* estimation

Bayesian framework In the Bayesian framework, the parameter $\boldsymbol{\theta}$ is assumed to be itself a random variable whose the *prior* distribution is denoted by $p(\boldsymbol{\theta})$. This distribution contains the information on $\boldsymbol{\theta}$ given by an expert. Thus, the term *prior* can be interpreted as *before to observe the data*. There are different prior distributions and their impact on the inference can be not negligible, especially for the small data sets. These distributions are presented in *The Bayesian choice* by C.P. Robert [Rob07] giving large details on the Bayesian framework. If the *prior* distribution contains the information given by an expert, the distribution from which

inferences are made is the posterior one. The *posterior* distribution contains the prior information given by the expert ($p(\boldsymbol{\theta})$) and by the data ($\mathbf{x}$). Thus, the term *posterior* can be interpreted as *after to observe the data.*

Posterior distribution and likelihood function The Bayes' rule involves that the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$ is defined as

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})}. \tag{1.52}$$

Note that the information given by the data is related to the likelihood function $p(\mathbf{x}|\boldsymbol{\theta})$. Thus, the definition of the likelihood function is crucial since this function contains all the informations given by the data, in both frequentist and Bayesian frameworks. This function is also a common base for both communities. Since $p(\mathbf{x})$ is a constant according to $\boldsymbol{\theta}$, the following relation is used when $p(\mathbf{x})$ is not computable

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{1.53}$$

Main advantages of the Bayesian approach The Bayesian approaches for the mixture model are detailed in *Finite Mixture and Markov Switching Models* by S. Frühwirth-Schnatter [FS08] from where one can extract the four following main qualities:

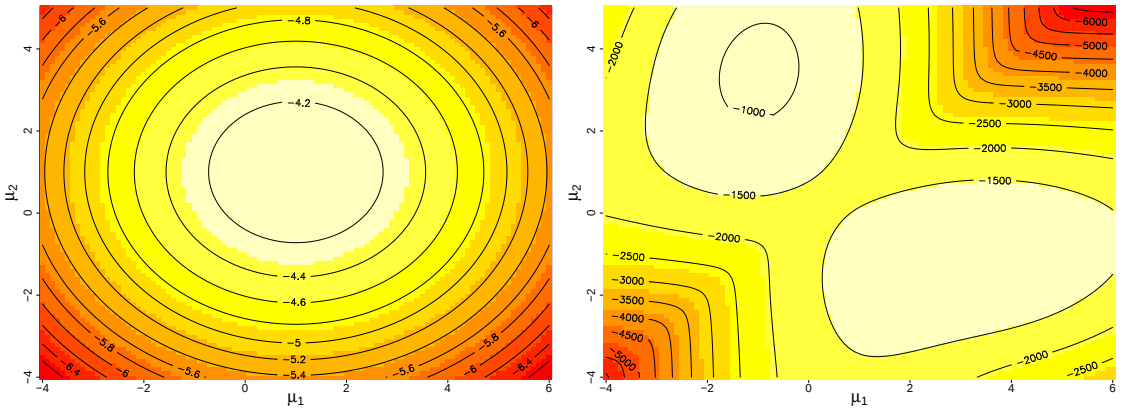
- The prior gives a smooth effect avoiding the problems of the degenerate solution.
- These methods take into account the parameter uncertainty.
- They stay valid in case where regularity conditions are violated (small data set, mixture with small component proportions) since they do not rely an asymptotic normality.
- Their implementation is not complex when the component distributions belong to the exponential family. Indeed, in such case, the conjugate prior distributions provide explicit posterior distributions [Rob07].

Running example 1.37 (Bayesian framework)**.**

We consider the bi-component univariate Gaussian mixture model with known proportions and variances. In this case $\boldsymbol{\theta} = (\mu_1, \mu_2)$, we assume independence between the prior distributions and we use the Jeffreys non informative ones, so

$$p(\boldsymbol{\theta}) = p(\mu_1)p(\mu_2) \text{ with } \mu_1 \sim \mathcal{N}(\xi, \kappa) \text{ and } \mu_2 \sim \mathcal{N}(\xi, \kappa), \tag{1.54}$$

where ξ and κ are hyper-parameters. Figure 1.9 displays the values of the *prior* and of the *posterior* distributions computed on a sample of size 150.



(a) prior distribution (b) posterior distribution

Figure 1.9 – Prior and posterior distributions for a bi-component univariate mixture model with $\xi = 1$ and $\kappa = 9$.

Smooth effect and degenerate solution The Bayesian framework can avoid some degeneracy problems by using prior distributions which provides a smooth effect as illustrated below.

Running example 1.38 (Smooth effect and degenerate solution)**.**

From the sample $\mathbf{x}$ arisen from the bi-component univariate Gaussian mixture model with known proportions with $\mu_1 = 0$ and $\mu_2 = x_1$, the aim is to infer the parameter $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2)$. The frequentist way can suffer from degeneracy solution as illustrated in Running example 1.27. In a Bayesian way, classical independence assumption between prior distributions associated with conjugate prior distributions, involves that

$$p(\boldsymbol{\theta}) = p(\sigma_1^2)p(\sigma_2^2) \text{ where } 1/\sigma_1^2 \sim \mathcal{G}\left(c_0, C_0\right) \text{ and } 1/\sigma_2^2 \sim \mathcal{G}\left(c_0, C_0\right). \quad (1.55)$$

Since the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x}) = \sum_{\mathbf{z}\in\mathcal{Z}} p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z})p(\mathbf{z}|\mathbf{x})$, where $\mathcal{Z} = \{1, 2\}^n$, then this distribution is upper-bounded by

$$p(\boldsymbol{\theta}|\mathbf{x}) \leq \sum_{\mathbf{z}\in\mathcal{Z}} p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z}) \text{ and } p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z}) = p(\sigma_1^2|\mathbf{x}, \mathbf{z})p(\sigma_2^2|\mathbf{x}, \mathbf{z}). \quad (1.56)$$

Since $1/\sigma_k^2|\mathbf{x}, \mathbf{z} \sim \mathcal{G}\left(c_0 + \frac{\mathrm{n}_k}{2}, C_0 + \sum_{\{i:z_{ik}=1\}} \frac{(x_i-\mu_k)^2}{2}\right)$ and since the mode of a $\mathcal{G}(\alpha, \beta)$ is $\frac{\alpha-1}{\beta}$ when $\alpha \leq 1$, then

$$p(\boldsymbol{\theta}|\mathbf{x}) < +\infty \quad \forall \boldsymbol{\theta}. \quad (1.57)$$

Note that the conjugate prior distributions are often used since they significantly simplify the inference. However, the choice of the hyper-parameters (parameters of the prior distribution) can be delicate when the Jeffery's non informative prior is

not available. A solution is also to fix the hyper-parameters by using an empirical Bayesian approach which determines the hyper-parameters according to the data (see for instance [Raf96] for the Gaussian mixture model).

Bayesian inference The Bayesian approaches often need simulation methods like Markov chain Monte Carlo (MCMC) to be inferred. So, their development was recent because it is related to the computational power. Since the *posterior* distribution contains the whole information about $\boldsymbol{\theta}$ (information given by the expert and by the data), any inference on $\boldsymbol{\theta}$ are based on this distribution. It is also natural to adopt the approach similar to the one used in the frequentist framework. So, we want to obtain the estimate of the maximum *a posteriori* estimate (MAPE) denoted by $\tilde{\boldsymbol{\theta}}$.

Definition 1.39 (Maximum *a posteriori* estimate)**.** The maximum *a posteriori* estimate is defined by

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta}|\mathbf{x}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{1.58}$$

Remark 1.40 (Link between the MLE and MAPE)**.** Note that if the prior follows a uniform distribution, then $p(\mathbf{x}|\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\mathbf{x})$. In such a case, the MLE is also equal to the MAPE.

Thus, the MAPE is a reasonable estimate but it can be difficult to obtain. It can be also replaced by the mean or the median of the *posterior* distribution. Both latter estimates are more easily obtained via MCMC algorithms. We now detail the main algorithms performing the Bayesian inference on $\boldsymbol{\theta}$.

1.3.4 Algorithms for a maximum *a posteriori* estimation

EM algorithm for Bayesian estimation

Main idea The EM algorithm can be modified to provide the MAPE or the estimate of the maximum penalized likelihood [Gre90]. This objective is achieved by using the following relation obtained by applying the Bayes' rule and by using the logarithm function

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta}|\mathbf{x}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta};\mathbf{x}) + \ln p(\boldsymbol{\theta}). \tag{1.59}$$

To obtain the MAPE, the M step of the EM algorithm consists in the maximization of the expectation of the complete data posterior distribution $p(\boldsymbol{\theta}|\mathbf{x},\mathbf{z})$. At iteration $[r]$, the M step determines the parameter $\boldsymbol{\theta}^{[r+1]}$ as such that

$$\boldsymbol{\theta}^{[r+1]} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta};\boldsymbol{\theta}^{[r]}) + \ln p(\boldsymbol{\theta}). \tag{1.60}$$

Running example 1.41 (Bayesian estimation of Gaussian mixture)**.**

Let bi-component univariate Gaussian mixture model whose only the means are unknown. The prior distribution of the parameters is defined in (1.54), so the M step is written as

$$\mu_k^{[r+1]} = \frac{\sigma_k^2 \xi + \kappa \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]}) x_i}{\sigma_k^2 + \kappa n_k^{[r]}}. \tag{1.61}$$

MCMC algorithms and Bayesian estimation

Structure of this section We now present a short overview of three main algorithms used to infer the parameters of a mixture model: the Gibbs sampler, the Metropolis-Hastings algorithm and the Metropolis-within-Gibbs sampler. The reader wanting more details can report on *Monte Carlo Statistical Methods* by C.P. Robert [RC04]. These algorithms are MCMC ones whose the Markov chain has the posterior distribution $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x})$ as the stationary distribution. Thus, they sample a sequence of parameters according to their posterior distribution since this approach allows us to perform the Bayesian inference.

Remark 1.42 (Almost-absorbing states involving different initializations of the algorithms)**.** Even if the MCMC algorithms having an irreducible and ergodic Markov chain are not, theoretically, sensitive to the local optima, their behavior is not so perfect in practice (see for instance [MMR05]). Indeed, there are trapping states which are almost-absorbing states requiring a so large number of iterations to escape from them that the algorithm is generally stopped before.

The Gibbs sampler

Main idea The Gibbs sampler is the most popular approach to perform the Bayesian inference of a mixture model since it uses the latent structure of the data. This algorithm is built on full conditional distributions from which it is easy to sample.

Gibbs sampler and mixture models The Gibbs sampler is an iterative algorithm whose one iteration is split in two main steps for the mixture model framework. Indeed, this algorithm alternatively samples the class memberships conditionally on the parameters and on the data, and the parameters conditionally on the class memberships and on the data. Thus, its stationary distribution is $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x})$, therefore the sequences of the generated parameters are sampled from their posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$.

Algorithm 1.43 (The Gibbs sampler for the mixture models).

This algorithm, having $p(\boldsymbol{\theta}|\mathbf{x})$ as marginal stationary distribution, starts from an initial value $\boldsymbol{\theta}^{[0]}$ then alternates between two steps. At iteration $[r]$, it performs the two following steps

$$\mathbf{z}^{[r]} \sim \mathbf{z}|\boldsymbol{\theta}^{[r]}, \mathbf{x} \tag{1.62}$$

$$\boldsymbol{\theta}^{[r+1]} \sim \boldsymbol{\theta}|\mathbf{z}^{[r]}, \mathbf{x}. \tag{1.63}$$

Sampling of the class membership Independence between individuals allows to easily sample the vector $\mathbf{z}$ since $p(\mathbf{z}|\boldsymbol{\theta}^{[r]}, \mathbf{x}) = \prod_{i=1}^{n} p(\boldsymbol{z}_i|\boldsymbol{\theta}^{[r]}, \boldsymbol{x}_i)$. Indeed, each $\boldsymbol{z}_i^{[r]}$ is sampled from the following multinomial distribution

$$\boldsymbol{z}_i^{[r]}|\boldsymbol{\theta}^{[r]}, \boldsymbol{x}_i \sim \mathcal{M}(t_{i1}(\boldsymbol{\theta}^{[r]}), \ldots, t_{ig}(\boldsymbol{\theta}^{[r]})). \tag{1.64}$$

Sampling of the parameters When there is no constraint between the parameters of different classes, the following decomposition is used to sample $\boldsymbol{\theta}^{[r+1]}$

$$p(\boldsymbol{\theta}^{[r+1]}|\mathbf{z}^{[r]}, \mathbf{x}) = p(\boldsymbol{\pi}^{[r+1]}|\mathbf{z}^{[r]}) \prod_{k=1}^{g} p(\boldsymbol{\alpha}_k^{[r+1]}|\mathbf{z}^{[r]}, \mathbf{x}). \tag{1.65}$$

Note that $\boldsymbol{\pi}$ is independent of the data conditionally on the class memberships. The usual prior of $\boldsymbol{\pi}$ is the conjugate Jeffrey's non informative prior. In such a case, the prior and the posterior distributions of the class proportions are respectively defined by

$$\boldsymbol{\pi} \sim \mathcal{D}_g\left(\frac{1}{2}, \ldots, \frac{1}{2}\right) \text{ and } \boldsymbol{\pi}|\mathbf{z}^{[r]} \sim \mathcal{D}_g\left(\frac{1}{2} + \mathrm{n}_1^{[r]}, \ldots, \frac{1}{2} + \mathrm{n}_g^{[r]}\right), \tag{1.66}$$

where we remind that $\mathrm{n}_k^{[r]} = \sum_{i=1}^{n} z_{ik}^{[r]}$. We now illustrate this algorithm with the running example.

Running example 1.44 (Gibbs sampler)**.**

We assume that the prior of σ_k^2 is $\mathcal{G}^{-1}(c_0, C_0)$ and that the prior of μ_k conditionally on σ_k^2 is $\mathcal{N}_1(b_0, B_0^{-1}\sigma_k^2)$. Iteration $[r]$ of the Gibbs sampler having $p(\boldsymbol{\theta}|\mathbf{x})$ as stationary distribution is written as follows

$$\forall i = 1, \ldots, n \quad \boldsymbol{z}_i^{[r]}|\boldsymbol{x}_i, \boldsymbol{\theta}^{[r]} \sim \mathcal{M}_2(t_{i1}(\boldsymbol{\theta}^{[r]}), t_{i2}(\boldsymbol{\theta}^{[r]})) \tag{1.67}$$

$$\boldsymbol{\pi}^{[r+1]}|\mathbf{z}^{[r]} \sim \mathcal{D}_2\left(\frac{1}{2} + \mathrm{n}_1^{[r]}, \frac{1}{2} + \mathrm{n}_2^{[r]}\right) \tag{1.68}$$

$$\forall k = 1, 2 \quad \mu_k^{[r+1]}|\mathbf{x}, \mathbf{z}^{[r]}, \sigma_k^{2[r]} \sim \mathcal{N}_1(b_k^{[r]}, B_k^{[r]}) \tag{1.69}$$

$$\forall k = 1, 2 \quad \sigma_k^{2[r+1]}|\mathbf{x}, \mathbf{z}^{[r]}, \mu_k^{[r+1]} \sim \mathcal{G}^{-1}(c_k^{[r]}, C_k^{[r]}), \tag{1.70}$$

where $b_k^{[r]} = \frac{B_0 b_0 + \sum_{i=1}^n z_{ik}^{[r]} x_i}{B_0 + \mathrm{n}_k^{[r]}}$, $B_k^{[r]} = \frac{\sigma_k^{2[r]}}{B_0 + \mathrm{n}_k^{[r]}}$, $c_k^{[r]} = c_0 + \frac{\mathrm{n}_k^{[r]}+1}{2}$ and $C_k^{[r]} = C_0 + \frac{1}{2}(\sum_{i=1}^n z_{ik}^{[r]}(x_i - \mu_k^{[r+1]}) + B_0(\mu_k^{[r+1]} - b_0)^2)$.

Simple sampling condition As the Gibbs sampler has to perform a huge number of iterations, it is absolutely necessary that each step involves a small sampling time. If the full conditional distributions of $\boldsymbol{z}_i^{[r]}$ and $\boldsymbol{\pi}^{[r+1]}$ are explicit, the sampling of $\boldsymbol{\alpha}^{[r+1]}$ can be more complicated. The conjugate prior distributions are also generally used since they provide classical posterior distribution. Thus, the sampling of $\boldsymbol{\alpha}^{[r+1]}$ is easy when there is no constraint between the parameters. In the case where the simulation of $p(\boldsymbol{\alpha}_k^{[r+1]}|\mathbf{z}^{[r]}, \mathbf{x})$ is too much time consuming, another approach than the Gibbs sampler has to be used.

The Metropolis-Hastings algorithm

Main idea The aim of the Metropolis-Hastings algorithm is to sample a sequence of $\boldsymbol{\theta}$ according to its posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$. This algorithm requires an *instrumental* distribution, denoted by $q(.;\boldsymbol{\theta})$, defined with respect to the dominating measure of the model. At iteration $[r]$, the instrumental distribution generates a candidate $\boldsymbol{\theta}^\star$ conditionally on the current value of $\boldsymbol{\theta}$. Then, the candidate is accepted with a probability $\lambda^{[r]}$ defined by

$$\lambda^{[r]} = \min\left\{\frac{p(\boldsymbol{\theta}^\star|\mathbf{x})q(\boldsymbol{\theta}^{[r]};\boldsymbol{\theta}^\star)}{p(\boldsymbol{\theta}^{[r]}|\mathbf{x})q(\boldsymbol{\theta}^\star;\boldsymbol{\theta}^{[r]})}; 1\right\}. \tag{1.71}$$

Algorithm 1.45 (The Metropolis-Hastings algorithm).

This algorithm has $p(\boldsymbol{\theta}|\mathbf{x})$ as stationary distribution. Starting from an initial value $\boldsymbol{\theta}^{[0]}$, its iteration $[r]$ is written as

$$\boldsymbol{\theta}^{\star} \sim q(\boldsymbol{\theta}; \boldsymbol{\theta}^{[r]}) \tag{1.72}$$

$$\boldsymbol{\theta}^{[r+1]} = \begin{cases} \boldsymbol{\theta}^{\star} & \text{with probability } \lambda^{[r]} \\ \boldsymbol{\theta}^{[r]} & \text{with probability } 1-\lambda^{[r]}. \end{cases} \tag{1.73}$$

The hybrid MCMC

Main idea When a step of a Gibbs sampler is difficult to perform, the hybrid MCMC algorithms are often used. The most popular approach is to sample a sequence of $\boldsymbol{\theta}$ according to a Metropolis-within-Gibbs sampler. In this approach, the difficult steps of the Gibbs sampler are replaced by *one* iteration of a Metropolis-Hastings algorithm. However, the stationary distribution of the Markov chain stays equal to $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x})$.

Algorithm 1.46 (The Metropolis-within-Gibbs sampler).

This algorithm, performing the inference for the mixture models, has $p(\boldsymbol{\theta}|\mathbf{x})$ as marginal stationary distribution. Starting from an initial value $\boldsymbol{\theta}^{[0]}$, its iteration $[r]$ is written as

$$\mathbf{z}^{[r]} \sim \mathbf{z}|\boldsymbol{\theta}^{[r]}, \mathbf{x} \tag{1.74}$$

$$\boldsymbol{\theta}^{\star} \sim q(\boldsymbol{\theta}; \boldsymbol{\theta}^{[r]}) \tag{1.75}$$

$$\boldsymbol{\theta}^{[r+1]} = \begin{cases} \boldsymbol{\theta}^{\star} & \text{with probability } \lambda^{[r]} \\ \boldsymbol{\theta}^{[r]} & \text{with probability } 1-\lambda^{[r]}, \end{cases} \tag{1.76}$$

where $q(.;\boldsymbol{\theta})$ is the instrumental distribution of the Metropolis-Hastings step and where $\lambda^{[r]}$ is its acceptance probability defined by

$$\lambda^{[r]} = \min\left\{ \frac{p(\boldsymbol{\theta}^{\star}|\mathbf{z}^{[r]}, \mathbf{x}) q(\boldsymbol{\theta}^{[r]}; \boldsymbol{\theta}^{\star})}{p(\boldsymbol{\theta}^{[r]}|\mathbf{z}^{[r]}, \mathbf{x}) q(\boldsymbol{\theta}^{\star}; \boldsymbol{\theta}^{[r]})}; 1 \right\}. \tag{1.77}$$

We now illustrate this algorithm with the running example.

Running example 1.47 (The Metropolis-within-Gibbs sampler)**.**

Iteration $[r]$ of this algorithm is written as

$$\mathbf{z}^{[r]} \sim \mathbf{z}|\boldsymbol{\theta}^{[r]}, \mathbf{x} \tag{1.78}$$

$$\boldsymbol{\pi}^{[r+1]} \sim \boldsymbol{\pi}|\mathbf{z}^{[r]} \tag{1.79}$$

$$\boldsymbol{\alpha}^{\star} \sim q(\boldsymbol{\alpha}; \boldsymbol{\alpha}^{[r]}) \tag{1.80}$$

$$\boldsymbol{\alpha}^{[r+1]} = \begin{cases} \boldsymbol{\alpha}^{\star} & \text{with probability } \lambda^{[r]} \\ \boldsymbol{\alpha}^{[r]} & \text{with probability } 1-\lambda^{[r]}, \end{cases} \tag{1.81}$$

where $q(.;.)$ is the instrumental distribution of the Metropolis-Hastings step and where $\lambda^{[r]}$ is its acceptance probability defined by

$$\lambda^{[r]} = \min\left\{\frac{p(\boldsymbol{\alpha}^{\star}|\mathbf{z}^{[r]}, \mathbf{x})q(\boldsymbol{\alpha}^{[r]}; \boldsymbol{\alpha}^{\star})}{p(\boldsymbol{\alpha}^{[r]}|\mathbf{z}^{[r]}, \mathbf{x})q(\boldsymbol{\alpha}^{\star}; \boldsymbol{\alpha}^{[r]})}; 1\right\}. \tag{1.82}$$

1.4 Model selection

1.4.1 On the model selection challenge

Definition 1.48 (Model)**.** Let us consider the general finite mixture model

$$p(\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k), \tag{1.83}$$

$\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where the parameter space $\boldsymbol{\Theta}$ is defined by the number of components and the nature of each component. The model $\boldsymbol{m}$ groups the set of the distributions defined by (1.83), so

$$\boldsymbol{m} = \{p(\boldsymbol{x}_i; \boldsymbol{\theta}) : \ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}. \tag{1.84}$$

Aim The model $\boldsymbol{m}$ defines the nature of the component distributions and the number of components. As it is generally unknown, the model has to be inferred according to the data. Thus, we define $\boldsymbol{\Delta}$ as the set of the models considered by the practitioner and the aim is to find the "best" model among $\boldsymbol{\Delta}$.

Log-likelihood function and embedded models The likelihood function generally allows to estimate the "best" model according to the data. However, this approach can not be directly applied in the mixture model context. Indeed, in such a case, a lot of models are embedded (for instance a Gaussian mixture model with three components always obtains a best likelihood values than the Gaussian models with two components). Thus, the "best" model is the model which makes the best trade off between its quality of adjustment to the data (given by its likelihood value) and its complexity (number of parameters).

Information criteria We saw, in Section 1.1.2, that heuristic criteria (like the slope of the likelihood function) can be used to select the number of classes especially for the geometric clustering methods. If these approaches can be used to select the model of probabilistic method, it is more convenient to use *information criteria* (IC) proposed by the probabilistic framework. These criteria, further detailed, rigorously perform the model selection according to an objective of a data adjustment (AIC, BIC criteria) or an objective of classification (ICL criterion). Generally, these criteria require the MLE related to each model in $\boldsymbol{\Delta}$, since they can often be written as a penalization of the log-likelihood function

$$\text{IC}_{\boldsymbol{m}} = L(\hat{\boldsymbol{\theta}}; \mathbf{x}) - h(\nu_{\boldsymbol{m}}), \tag{1.85}$$

where $\hat{\boldsymbol{\theta}}$ is the MLE of the model $\boldsymbol{m}$, where $\nu_{\boldsymbol{m}}$ is the parameters number of the model $\boldsymbol{m}$ and where $h(.)$ is a function defined by the criterion. Note that a quality which could be wanted for the information criterion is the consistency in dimension assuring a good asymptotic behavior.

Definition 1.49 (Consistency in dimension for a criterion)**.** A criterion is consistent in dimension if it selects the simplest true model with a probability one when the sample size tends to the infinity.

1.4.2 Information criteria for the data adjustment

This section is devoted to the model selection which stays a difficult problem (see [FS08] Chapter 4) for the mixture models (especially the selection of the class number) principally since the models are embedded and since the information criteria are only asymptotically true.

Frequentist criterion

In a frequentist framework, the aim is to find the model minimizing the Kullback-Leibler (KL) divergence [KL51] of the "true" distribution relative to the estimated one.

Definition 1.50 (Kullback-Leibler divergence)**.** Let $\boldsymbol{x}_i \in \mathbb{R}^e$, the Kullback-Leibler (KL) divergence of the pdf $f(\boldsymbol{x}_i)$ relative to the pdf $g(\boldsymbol{x}_i)$ is

$$\text{KL}(f, g) = \int_{\boldsymbol{x}_i \in \mathcal{X}} f(\boldsymbol{x}_i) \ln f(\boldsymbol{x}_i) d\boldsymbol{x}_i - \int_{\boldsymbol{x}_i \in \mathcal{X}} f(\boldsymbol{x}_i) \ln g(\boldsymbol{x}_i) d\boldsymbol{x}_i. \tag{1.86}$$

Let $f(\boldsymbol{x}_i)$ be the pdf of the true model, find the model minimizing KL is equivalent to find the model minimizing the term on left-hand side of the previous equation. Thus, for a model $\boldsymbol{m}$, the aim is to compute

$$\eta(\boldsymbol{x}_i; f, \boldsymbol{m}, \hat{\boldsymbol{\theta}}) = \int_{\boldsymbol{x}_i \in \mathcal{X}} f(\boldsymbol{x}_i) \ln p(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}) d\boldsymbol{x}_i, \tag{1.87}$$

where $p(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}})$ is the pdf of the model $\boldsymbol{m}$ parametrized in its MLE $\hat{\boldsymbol{\theta}}$. As the distribution f is unknown, we use the natural estimator of $\eta(\boldsymbol{x}_i; f, \boldsymbol{m}, \hat{\boldsymbol{\theta}})$ defined by

$$\hat{\eta}(\boldsymbol{x}_i; \hat{f}, \boldsymbol{m}, \hat{\boldsymbol{\theta}}) = \frac{1}{n} L(\hat{\boldsymbol{\theta}}; \boldsymbol{x}_i). \tag{1.88}$$

However, this estimator suffers from the following bias

$$b = \mathbb{E}_f \left[\hat{\eta}(\boldsymbol{x}_i; \hat{f}, \boldsymbol{m}, \hat{\boldsymbol{\theta}}) - \eta(\boldsymbol{x}_i; f, \boldsymbol{m}, \hat{\boldsymbol{\theta}}) \right]. \tag{1.89}$$

Thus, the best model among $\boldsymbol{\Delta}$ maximizes the correct log-likelihood

$$\underset{\boldsymbol{m} \in \boldsymbol{\Delta}}{\operatorname{argmax}} \, L(\hat{\boldsymbol{\theta}}; \mathbf{x}) - b. \tag{1.90}$$

Akaike [Aka73] showed that the corrected term is asymptotically equal to the number of parameters.

Definition 1.51 (The AIC criterion)**.** The Akaike Information Criterion (AIC) is defined as

$$\text{AIC}(\boldsymbol{m}) = L(\hat{\boldsymbol{\theta}}; \mathbf{x}) - \nu, \tag{1.91}$$

where $\hat{\boldsymbol{\theta}}$ is the MLE of the model $\boldsymbol{m}$ and ν its number of parameters.

Thus, the AIC criterion is an estimator of the expectation of the mean of the log-likelihood. A study of the AIC criterion properties (and of its extension) is available in [Boz87]. However, the behavior of the AIC criterion can be inconsistent.

Proposition 1.52 (AIC is not consistent)**.** AIC *is not consistent in dimension when models with the same number of components are embedded.*

Proof. Let a model $\boldsymbol{m}_0$ whose the MLE of dimension ν_0 is denoted by $\hat{\boldsymbol{\theta}}_0$ and let $\boldsymbol{m}_1$ whose the MLE of dimension ν_1 is denoted by $\hat{\boldsymbol{\theta}}_1$ such as $\boldsymbol{m}_0$ is the true number, $\boldsymbol{m}_0$ and $\boldsymbol{m}_1$ are embedded models with the same number of components and $\nu_0 < \nu_1$. Then,

$$\begin{aligned} 2\left(\text{AIC}(\boldsymbol{m}_1) - \text{AIC}(\boldsymbol{m}_0)\right) &= 2\left(L(\mathbf{x}; \hat{\boldsymbol{\theta}}_1) - L(\mathbf{x}; \hat{\boldsymbol{\theta}}_0)\right) - 2(\nu_1 - \nu_0) \\ &\xrightarrow{D} \chi^2_{\nu_1 - \nu_0} - 2(\nu_1 - \nu_0). \end{aligned} \tag{1.92}$$

Thus, the AIC criterion is not consistent (*i.e.* $\lim\limits_{n \to \infty} P\left(\text{AIC}(\boldsymbol{m}_1) > \text{AIC}(\boldsymbol{m}_0)\right) > 0$) since $P(\chi^2_{\nu_1 - \nu_0} > 2(\nu_1 - \nu_0)) > 0$. Note that the demonstration can not be performed to select the number of classes. Indeed, in such case, the convergence to the likelihood ratio is unknown since it involves a Taylor's development on the border of the parameters space. However, it has often been observed that the AIC criterion selects more complicated models, even if the true one is in the list of the models [BCG00]. □

Bayesian criterion

Let $p(\boldsymbol{m})$ the prior distribution of $\boldsymbol{m} \in \boldsymbol{\Delta}$, the posterior distribution of interest is defined by using the Bayes' rule as follows

$$p(\boldsymbol{m}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{m})p(\boldsymbol{m})}{p(\mathbf{x})}. \tag{1.93}$$

Indeed, in a Bayesian point of view, the "best" model $\boldsymbol{m}^\star$ maximizes the posterior distribution

$$\boldsymbol{m}^\star = \underset{\boldsymbol{m}\in\boldsymbol{\Delta}}{\operatorname{argmax}}\, p(\boldsymbol{m}|\mathbf{x}) = \underset{\boldsymbol{m}\in\boldsymbol{\Delta}}{\operatorname{argmax}}\, p(\mathbf{x}|\boldsymbol{m})p(\boldsymbol{m}). \tag{1.94}$$

Thus, the quantity performing the model selection is the integrated likelihood also named marginal likelihood or evidence defined by

$$p(\mathbf{x}|\boldsymbol{m}) = \int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} p(\mathbf{x},\boldsymbol{\theta}|\boldsymbol{m})p(\boldsymbol{\theta}|\boldsymbol{m})d\boldsymbol{\theta}, \tag{1.95}$$

where the parameter space $\boldsymbol{\Theta}$ depends on $\boldsymbol{m}$. If this quantity can be approached via many methods (see the review of [FW12]), the most classical one is to use the BIC approximation [Sch78] which approximates $\ln p(\mathbf{x}|\boldsymbol{m})$ by using a Laplace approximation and by replacing the MAPE by the MLE.

Definition 1.53 (The BIC criterion)**.** The Bayesian Information Criterion (BIC) is defined as

$$\text{BIC}(\boldsymbol{m}) = L(\hat{\boldsymbol{\theta}};\mathbf{x}) - \frac{\nu}{2}\ln n, \tag{1.96}$$

where $\hat{\boldsymbol{\theta}}$ is the MLE of the model $\boldsymbol{m}$ and where ν denotes its number of parameters.

This criterion assumes regularity conditions on the pdf which may be not verified by the mixture models. Furthermore, this approximation is only asymptotically true.

Proposition 1.54 (BIC is consistent)**.** *The* BIC *criterion is consistent in dimension when models with the same number of components are embedded.*

Proof. Let a model $\boldsymbol{m}_0$ whose the MLE of dimension ν_0 is denoted by $\hat{\boldsymbol{\theta}}_0$ and let $\boldsymbol{m}_1$ whose the MLE of dimension ν_1 is denoted by $\hat{\boldsymbol{\theta}}_1$ such as $\boldsymbol{m}_0$ is the true number, $\boldsymbol{m}_0$ and $\boldsymbol{m}_1$ are embedded models with the same number of components and $\nu_0 < \nu_1$. Then,

$$\begin{aligned} 2\left(\text{BIC}(\boldsymbol{m}_1) - \text{BIC}(\boldsymbol{m}_0)\right) &= 2\left(L(\mathbf{x};\hat{\boldsymbol{\theta}}_1) - L(\mathbf{x};\hat{\boldsymbol{\theta}}_0)\right) - 2(\nu_1-\nu_0)\ln n \\ &\xrightarrow{D} \chi^2_{\nu_1-\nu_0} - 2(\nu_1-\nu_0)\ln n. \end{aligned} \tag{1.97}$$

By using the notation $\Delta_\nu = \nu_1 - \nu_0$, the following result is obtained by applying the Tchebychev's inequality when n is large

$$P(\text{BIC}(\boldsymbol{m}_1) > \text{BIC}(\boldsymbol{m}_0)) \leq P(|\chi^2_{\Delta_\nu} - \Delta_\nu| > \Delta_\nu(-1+\ln n)) \tag{1.98}$$

$$\leq \frac{2\Delta_\nu}{(\Delta_\nu(-1+\ln n))^2} \overset{n\to\infty}{\to} 0, \tag{1.99}$$

since $\mathbb{E}[\chi^2_{\Delta_\nu}] = \Delta_\nu$ and $\text{Var}(\chi^2_{\Delta_\nu}) = 2\Delta_\nu$. Thus, the BIC criterion is consistent in dimension when models with the same number of components are embedded. Note that the demonstration can not be performed to select the number of classes. Indeed, in such case, the convergence to the likelihood ratio is unknown since it involves a Taylor's development on the border of the parameters space. However, if the BIC criterion is more robust than the AIC one, it can overestimates the number of components when the "true" model is not in $\boldsymbol{\Delta}$ [BCG00]. □

Note that, by using a locally conic parametrization, [Ker00] shows that the BIC criterion is a consistent estimator of the correct number of components in the distribution.

Reversible jump If $\boldsymbol{\Delta}$ is large or if the estimation of the parameters is complex, then the exhaustive approach is not doable. Indeed, this approach consists in a computation of an information criterion for all the models, so it is time consuming. Furthermore, the practitioner only uses the estimate associated to the best model. Thus, all the other estimates are not used for the data analysis. This drawback is avoided by the approach of the *reversible jump* [Gre95, RG97] where the model and the parameters are simultaneously estimated. Unfortunately, this approach involves the computation of the probabilities of the model transition which can be complex. However, this objective consists in avoiding the estimation of the parameters of all the models in $\boldsymbol{\Delta}$ what can become mandatory when the model space becomes huge (see Part I).

1.4.3 Information criterion for the partition adjustment

Main idea Paradoxically, the consistency of the information criterion can be a drawback. Indeed, as all the models are wrong, the BIC criterion asymptotically overestimates the number of components according to the class separation. So, C. Biernacki, G. Celeux and G. Govaert [BCG00] propose to include a classification objective in the information criterion. In such case, the best model maximizes the integrated complete-data likelihood. As the vector $\mathbf{z}$ is unknown, it is replaced by its MAPE, denoted by $\hat{\mathbf{z}}$, evaluated with the MLE.

Definition 1.55 (ICL exact)**.** The Integrated Complete-data Likelihood (ICL) asses a model with a classification aim. It is defined as

$$\begin{aligned} \text{ICL}_{\text{ex}}(\boldsymbol{m}) &= \ln p(\mathbf{x}, \hat{\mathbf{z}}|\boldsymbol{m}) && (1.100)\\ &= \ln \int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} p(\mathbf{x}, \hat{\mathbf{z}}|\boldsymbol{m}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{m}) d\boldsymbol{\theta}. && (1.101) \end{aligned}$$

However, even if the integral can be explicit (see for instance the study in [BCG10] for the CIM of multinomial distributions), it is not generally the case. Thus, an approximated version of this criterion is available

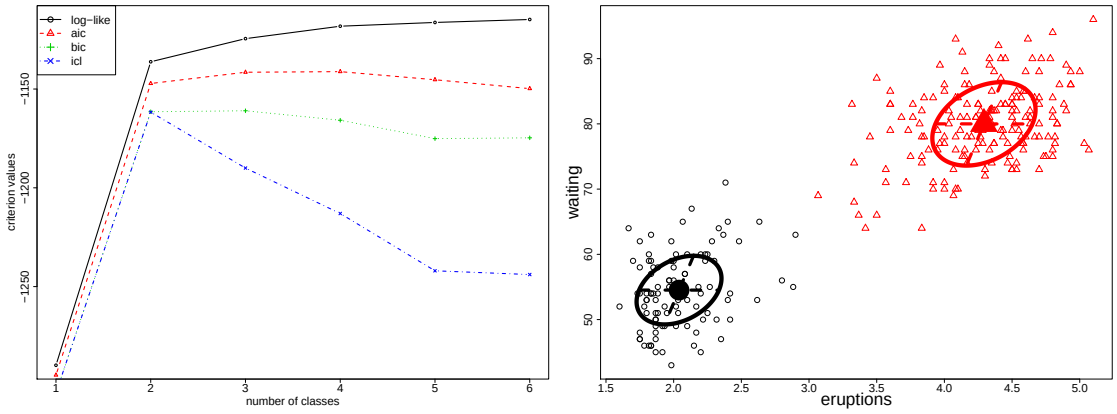
Definition 1.56 (ICL-BIC)**.** The Integrated Complete-data Likelihood (ICL) can be approximated by

$$\text{ICL}_{\text{BIC}}(\boldsymbol{m}) = \ln p(\mathbf{x}, \hat{\mathbf{z}}|\boldsymbol{m}, \hat{\boldsymbol{\theta}}) - \frac{\nu}{2} \ln n. \quad (1.102)$$

1.4.4 Application on real data set of the information criteria

Faithful data set 1.57 (Model selection by information criteria).

We give an example of the using of the information criteria on the Faithful data set. For different numbers of classes, the fourteen parsimonious Gaussian mixture models [CG95] compose the set of the considered models. Figure 1.10 displays the values of the information criteria for different numbers of classes.
The log-likelihood function is increasing with the class number. The AIC criterion selects four classes while the BIC criterion selects three classes. However, both criteria hesitate to select the number of classes. Thus, the practitioner using AIC (respectively BIC) can also analyse the partition in three classes (respectively two classes). Finally, the ICL criterion strongly selects the partition in two classes. This partition seems realistic according to the scatter plot.



(a) Information criteria

(b) Scatter plot with the partition in two classes selected by ICL

Figure 1.10 – Criterion values for the best of the fourteen Gaussian mixture models for different number of classes to cluster the Faithful data set.

1.5 Conclusion

This bibliographical chapter has presented the general framework of the mixture models and has been illustrated by the bi-component Gaussian mixture model. Thus, it has pointed-out that the general CIM model allows to easily cluster complex data (categorical or mixed data), but with a risk of bias when the data are intra-class correlated.

Both estimation methods (frequentist and Bayesian) have been presented. In this thesis, we favor the frequentist approach to infer the parameters since it does not

require some information *a priori*. However, if the MLE is intractable, we perform the inference in the Bayesian framework. In such a case, we favor the conjugate prior distributions in order to easily estimate the parameters. The hyper-parameters of the prior distributions are selected to be weakly informative.

The model selection is performed by using the BIC criterion is the most common information criterion for the mixture models.

The purpose of the following chapters is to study and to propose new mixture models allowing to cluster complex data without assuming the conditional independence between the variables. We now focus on the categorical data set clustering.

Part I

Model-based clustering for categorical data

This part, devoted to the cluster analysis of categorical data, is split into four chapters.
The first one presents an overview of the clustering approaches devoted to the categorical data sets. We mainly focus on three main model-based approaches: the log-linear mixture models, the mixtures of trees and the multilevel latent class models. These models are illustrated on a small real data set.
The second and the third chapters present our contributions to this framework. We present two new mixture models which are extensions of the classical latent class model. For such models, the variables are grouped into conditionally independent blocks. The specific distributions of the blocks modelize the intra-class dependencies. These results are part of two submitted articles.
The last chapter is devoted to proposed model comparison illustrated on the example of the overview chapter. The second purpose of this chapter is to present our R packages performing the inference of both proposed models.

Those lonely fishermen who believed that the fish bite at high tide left their rocks, and their places were taken by others, who were convinced that the fish bite at low tide.
John Steinbeck — Tortilla flat

Table of Contents

Chapter 2

Cluster analysis of categorical data sets: state of the art

The purpose of this chapter is to present the main approaches to cluster categorical data sets.
The first section is devoted to the geometric methods and, more precisely, to the K-means-like ones. The other sections describe the three main model-based methods. More precisely, the second section focuses on the log-linear mixture models. It also details the classical latent class model which is a specific model of the log-linear mixture one assuming the conditional independence between the variables. This model is of interest for us since the proposed models introduced in the two following chapters consist in two extensions of this model. The third section is devoted to the presentation of the tree mixture models. The last section presents the multilevel latent class models.
These models are illustrated on a real data set throughout this chapter.

Always do sober what you said you'd do drunk. That will teach you to keep your mouth shut.
Ernest Hemingway — The Short Happy Life of Francis Macomber

2.1 Challenge of cluster analysis for categorical data

Introduction The categorical variables are often present in the data sets since they are easily accessible. The difficulty involved by such variables is double. Firstly, it is not convenient to *visualize* categorical data in their native space. So, this lack has to be counterbalanced by an easy interpretation of the partition. Secondly, the

combinatorial problems are ubiquitous when some intra-class dependencies have to be modelized (huge number of parameters and huge number of models in competition). Thus, it appears important for us that the models used to cluster respect the two following objectives.

Two crucial objectives Based on the models presented in this bibliographic chapter, we put the light on the following crucial objectives:

1. Models have to provide few meaningful parameters to counterbalance the lack of visualization.
2. Models have to take into account the intra-class dependencies by limiting the combinatorial problems related to both of the number of parameters and of the model selection.

The data Throughout this part, we consider the d-variate vector of categorical variables denoted by $\boldsymbol{x}_i = (\boldsymbol{x}_i^1, \ldots, \boldsymbol{x}_i^d)$ and defined in space $\mathcal{X}$. Each categorical variable $\boldsymbol{x}_i^j = (x_i^{jh}; h = 1, \ldots, m_j)$ has m_j modalities and uses a complete disjunctive coding as such $x_i^{jh} = 1$ if individual i takes modality h for variable j and $x_i^{jh} = 0$ otherwise.

Structure of this chapter Section 2.2 focuses on the geometric methods permitting to cluster categorical data sets. The other sections are devoted to probabilistic methods. Indeed, Section 2.3 presents the log-linear mixture model which is the reference to cluster categorical data in a probabilistic framework. Section 2.4 presents the mixture of dependency trees while Section 2.5 presents the multilevel latent models.

Running example During this chapter, the different methods allowing to cluster categorical data sets are illustrated on the Handelman's Dentistry data [EH89] which is a classical categorical data set.

Running example 2.1 (The dentistry data set)**.**

This binary data set, presented in Table 2.1, consists in the diagnoses given by five dentists for 3869 premolars and molars. The aim is to characterize the behavior of the dentists according to their diagnoses.

2.2 Geometric approaches

2.2.1 Methods on the native space

K-means algorithm for categorical data

Main idea The K-means algorithm for categorical data, proposed by H. Ralambondrainy [Ral95], uses the complete disjunctive coding of the categorical variables.

Dentist						Dentist					
1	2	3	4	5	Frequency	1	2	3	4	5	Frequency
S	S	S	S	S	1880	C	S	S	S	S	22
S	S	S	S	C	789	C	S	S	S	C	26
S	S	S	C	S	43	C	S	S	C	S	6
S	S	S	C	C	75	C	S	S	C	C	14
S	S	C	S	S	23	C	S	C	S	S	1
S	S	C	S	C	63	C	S	C	S	C	20
S	S	C	C	S	8	C	S	C	C	S	2
S	S	C	C	C	22	C	S	C	C	C	17
S	C	S	S	S	188	C	C	S	S	S	2
S	C	S	S	C	191	C	C	S	S	C	20
S	C	S	C	S	17	C	C	S	C	S	6
S	C	S	C	C	67	C	C	S	C	C	27
S	C	C	S	S	15	C	C	C	S	S	3
S	C	C	S	C	85	C	C	C	S	C	72
S	C	C	C	S	8	C	C	C	C	S	1
S	C	C	C	C	56	C	C	C	C	C	100

Table 2.1 – Radiographic cross-diagnosis of 3869 molars and premolars by five dentists [EH89]. Teeth are diagnosed as sound (S) or carious (C).

Indeed, for such approach each x_i^{jh} is considered as a binary variable. The distance used to cluster is the chi-square one that we below detail.

Chi-square distance The chi-square distance takes into account the weight of each modality in the distance computed by giving more importance to the rare modalities than to the most common ones.

Definition 2.2 (Chi-square distance)**.** Let $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ to be the sample composed by n individuals $\boldsymbol{x}_i$ described by d categorical variables. The chi-square distance between $\boldsymbol{x}_{i_1}$ and $\boldsymbol{x}_{i_2}$, with $1 \leq i_1, i_2 \leq n$, is defined by

$$D_{\chi^2}(\boldsymbol{x}_{i_1}; \boldsymbol{x}_{i_2}) = \sum_{j=1}^{d} \sum_{h=1}^{m_j} \frac{(x_{i_1}^{jh} - x_{i_2}^{jh})^2}{n^{jh}}, \tag{2.1}$$

where $n^{jh} = \sum_{i=1}^{n} x_i^{jh}$.

Comments The main drawback of this approach is that the vector of the class means—real values between zero and one—does not indicate the characteristics of the classes. The obtained partition is also weakly interpretable.

K-modes algorithm for categorical data

Main idea The K-modes algorithm has been proposed by Z. Huang [Hua98] to avoid the problem related to the cluster analysis of large categorical data sets. This

algorithm extends the K-means one by using a simple matching dissimilarity measure for categorical variables. At each iteration, it updates the modes by a frequency based method in order to minimize a cost function.

Dissimilarity and modes

Definition 2.3 (Matching dissimilarity)**.** Let two d-variate categorical variables $\boldsymbol{x}_{i_1}$ and $\boldsymbol{x}_{i_2}$. The matching dissimilarity counts the mismatches between both $\boldsymbol{x}_{i_1}$ and $\boldsymbol{x}_{i_2}$. This dissimilarity is defined by

$$D_1(\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}) = \sum_{j=1}^{d} \delta(\boldsymbol{x}_{i_1}^j, \boldsymbol{x}_{i_2}^j) \text{ with } \delta(\boldsymbol{x}_{i_1}^j, \boldsymbol{x}_{i_2}^j) = \begin{cases} 1 & \text{if } \boldsymbol{x}_{i_1}^j \neq \boldsymbol{x}_{i_2}^j \\ 0 & \text{if } \boldsymbol{x}_{i_1}^j = \boldsymbol{x}_{i_2}^j. \end{cases} \tag{2.2}$$

Definition 2.4 (Mode [Hua98])**.** A mode of the sample $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is a vector $\boldsymbol{\mu} \in \mathcal{X}$, as such $\boldsymbol{\mu} = (\mu^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$, which minimizes

$$D(\mathbf{x}; \boldsymbol{\mu}) = \sum_{i=1}^{n} D_1(\boldsymbol{x}_i; \boldsymbol{\mu}). \tag{2.3}$$

Note that $\boldsymbol{\mu}$ is not necessarily an element of $\mathbf{x}$ and that it is not necessarily unique.

Optimized criterion When the dissimilarity defined by Definition 2.3 is used, then the K-modes algorithm optimizes the following criterion

$$I(\mathbf{z}, \boldsymbol{\theta}; \mathbf{x}) = \sum_{i=1}^{n} \sum_{k=1}^{g} \sum_{j=1}^{d} \delta(\boldsymbol{x}_i^j, \boldsymbol{\mu}_k^j), \tag{2.4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g)$ and where $\boldsymbol{\mu}_k$ is the mode of class k.

The algorithm

Algorithm 2.5 (The K-modes algorithm)**.**

Starting from an initial value $\boldsymbol{\theta}^{[0]}$, its iteration $[r]$ is written

— **Class membership** $\mathbf{z}^{[r]} = \underset{\mathbf{z}}{\operatorname{argmin}}\, I(\mathbf{z}, \boldsymbol{\theta}^{[r]}; \mathbf{x})$:

$$z_{ik}^{[r]} = \begin{cases} 1 & \text{if } k = \underset{k'}{\operatorname{argmin}}\, D_1(\boldsymbol{x}_i, \boldsymbol{\mu}_{k'}^{[r]}) \\ 0 & \text{otherwise.} \end{cases}$$

— **Centroid estimation** $\boldsymbol{\theta}^{[r+1]} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, I(\mathbf{z}^{[r]}, \boldsymbol{\theta}; \mathbf{x})$:

$$\boldsymbol{\mu}_k^{[r+1]} = \underset{\boldsymbol{\mu}_k}{\operatorname{argmin}} \sum_{i=1}^{n} z_{ik}^{[r]} D_1(\boldsymbol{x}_i; \boldsymbol{\mu}_k).$$

Comments The K-modes algorithm, like the K-means one, converges to a local minimum of the function $I(\mathbf{z}, \boldsymbol{\theta}; \mathbf{x})$. It is also mandatory to perform different initializations in order to hope to get the global minimum of this function. Finally, note that the centroid estimation step is facilitated by the definition of $D_1(.,.)$. Indeed, this optimization is performed coordinates by coordinates.

Running example 2.6 (K-modes clustering).

Approach Based on the criterion values computed for different numbers of classes, two partitions could be of interest.

Interpretation The first one splits the data into two classes whose the modes are defined by both most present diagnoses (all dentists claim that the tooth is sound and all dentists claim that the tooth is sound except the last dentist). The second one splits the data into three classes. It adds, at the two previous modes, the diagnosis where all dentists claim that the tooth is carious except the first dentist. Basically, the K-modes approach does not allow to really understand this data set.

2.2.2 Methods on the factorial space

Main idea When many variables are correlated, they provide some redundant information. Thus, it can be efficient to perform a selection of variables or a reduction of the space dimension. When the variables are categorical, a Multiple Correspondence Analysis (MCA) can be used in order to reduce the space dimension. Indeed, this method provides numerical coordinates for each individual. Therefore, it is possible to use classical geometric approach to cluster numerical data like the K-means algorithm. We present the method of H. Hwang, W.R. Dillon and Y. Takane [HDT06] which combines MCA and K-means algorithm in a unified framework.

Notations We remind that the sample $\mathbf{x} = (\boldsymbol{x}_i; i = 1, \ldots, n)$ is composed with individuals described by d categorical variables which use a disjunctive coding. Let $\boldsymbol{f}$ denoting the matrix of size $n \times \mathfrak{d}$ where $\mathfrak{d} \leq m_j$ corresponds to the $\mathfrak{d}$-dimensional representation of the d categorical variables. Let $\boldsymbol{w}^j$ the matrix of weights of size $m_j \times \mathfrak{d}$. We consider $\mathbf{z}$ as the matrix of size $n \times g$ where the rows correspond to the individuals and the column to the class. We denote by $\boldsymbol{\theta}$ the matrix of the centroid values of the cluster in the factorial space.

Optimized criterion The aim is to combine MCA and K-means algorithm, so the problem is equivalent to the minimization of the following criterion

$$I_{\alpha_1,\alpha_2}(\mathbf{z}, \boldsymbol{\theta}; \mathbf{x}) = \alpha_1 \sum_{j=1}^{d} SS(\boldsymbol{f} - \mathbf{x}^j \boldsymbol{w}^j) + \alpha_2 SS(\boldsymbol{f} - \mathbf{z}\boldsymbol{\theta}), \tag{2.5}$$

where $\alpha_1 > 0$, $\alpha_2 > 0$, $\alpha_1 + \alpha_2 = 1$, where $\mathbf{x}^j$ is the matrix where the element (i, h) is equal to one if individual i takes modality h for variable j and is equal to zero otherwise, and where $SS(\boldsymbol{f}) = \text{trace}(\boldsymbol{f}'\boldsymbol{f})$.

Remark 2.7 (On the couple (α_1, α_2))**.** When $\alpha_1 = 1$, the criterion defined by (2.5) is the standard one for MCA. When $\alpha_2 = 1$, the criterion defined by (2.5) is equivalent to the standard one for the K-means algorithm. Thus, for others values of (α_1, α_2), this criterion performs a trade-off between the MCA and the K-means objectives.

The algorithm The estimation of $(\boldsymbol{f}, \boldsymbol{w}^j, \mathbf{z}, \boldsymbol{\theta})$ is performed by the alternating least squares algorithm proposed by [HDT06]. This algorithm converges to a local minimum of the function $I_{\alpha_1,\alpha_2}(\mathbf{z}, \boldsymbol{\theta}; \mathbf{x})$. So, several different initializations of this algorithm have to be done in order to obtain the estimators minimizing this criterion.

Algorithm 2.8 (Algorithm minimizing $I_{\alpha_1,\alpha_2}(\mathbf{z}, \boldsymbol{\theta}; \mathbf{x})$ [HDT06])**.**

Starting from an initial value $\boldsymbol{\theta}^{[0]}$, its iteration $[r]$ is written

— **Weight matrix and centroids optimization**

$$\boldsymbol{w}^{j[r]} = (\mathbf{x}^{j\prime}\mathbf{x}^j)^{-1}\mathbf{x}^{j\prime}\boldsymbol{f} \text{ and } \boldsymbol{\theta}^{[r]} = (\mathbf{z}^{[r]\prime}\mathbf{z}^{[r]})^{-1}\mathbf{z}^{[r]\prime}\boldsymbol{f}. \tag{2.6}$$

— **Factorial space optimization**

$$\boldsymbol{f}^{[r+1]} = \underset{\boldsymbol{f}}{\text{argmax}}\ \text{trace}\left(\boldsymbol{f}'\left[\alpha_1 \sum_{j=1}^{d} \mathbf{x}^j(\mathbf{x}^{j\prime}\mathbf{x}^j)^{-1}\mathbf{x}^{j\prime} + \alpha_2\mathbf{z}^{[r]}(\mathbf{z}^{[r]\prime}\mathbf{z}^{[r]})^{-1}\mathbf{z}^{[r]\prime}\right]\boldsymbol{f}\right). \tag{2.7}$$

— **Partition optimization**

$$\mathbf{z}^{[r+1]} = \underset{\mathbf{z}}{\text{argmin}}\ SS(\boldsymbol{f}^{[r+1]} - \mathbf{z}\boldsymbol{\theta}^{[r]}). \tag{2.8}$$

Comments In addition to the classical limits of the geometric approaches, three problems are rised by this method.

— The first problem is about the parameters (α_1, α_2) which are fixed by the user. Indeed, there is no rule which efficiently determines them while their impacts on the partition are significant.

— The second problem is about the size of the factorial space. Indeed, the space dimension $\mathfrak{d}$ is arbitrary fixed by the user with a risk of loosing information.

— The last problem is about the class interpretation which is complex. Indeed, the classes are summarized by the centroids which are not defined in the native space but in the factorial space created by combinations of the original variables.

2.3 Log-linear mixture models

Main idea We have seen in Chapter 1 that the Gaussian model can be used as component distribution when the variables are numeric. In the same way, the log-linear model (see *Categorical Data Analysis* by A. Agresti [Agr02]) is naturally used as component distribution when the variables are categorical. However, the complete log-linear model estimates the probability of all the modality crossings. It is also mandatory to impose constraints on this model to build the log-linear mixture model.

Structure of this section A classical approach assumes conditional independence between the variables. This model is named *Latent class model* or *naive Bayes* [Goo74] and has been detailed in Section 2.3.1. Other log-linear mixture models, which relax the conditional independence assumption, are presented in Section 2.3.2.

2.3.1 Latent class model

Model presentation

Main idea This mixture model assumes that the variables are independent conditionally on class, thus its components follow a product of multinomial distributions.

Definition 2.9 (Latent class model)**.** Let $\boldsymbol{x}_i$ be the d-variate categorical variable using a disjunctive coding. If $\boldsymbol{x}_i$ arise from the latent class model with g components, then its pdf is written as follows

$$p(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \mathring{p}(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) \text{ with } \mathring{p}(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) = \prod_{j=1}^{d}\prod_{h=1}^{m_j} (\alpha_k^{jh})^{x_i^{jh}}, \tag{2.9}$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha})$, where $\boldsymbol{\pi}$ is defined on the simplex of size g, where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_g)$ and where $\boldsymbol{\alpha}_k = (\boldsymbol{\alpha}_k^1, \ldots, \boldsymbol{\alpha}_k^d)$ as such that $\boldsymbol{\alpha}_k^j = (\alpha_k^{jh}; h = 1, \ldots, m_j)$ is defined on the simplex of size m_j. Note that α_k^{jh} denotes the probability that an individual arisen from component k takes modality h for variable j.

Despite its simplicity, the latent class model leads to good results in practice [HY01] for different areas like the behavioral sciences [RSS+06] or in medicine [SRAT+06].

Model identifiability The generic identifiability of the latent class model was proved by E.S. Allman, C. Matias and J.A. Rhodes [AMR09]. We now present their theorem. The reader interested by its proof can refer to the article [AMR09].

Theorem 2.10 (Generic identifiability of the latent class model [AMR09])**.** *Let the model defined by Definition 2.9 with $d \geq 3$. Suppose there exists a tripartition of the set $S = \{1, \ldots, d\}$ into three disjoint nonempty subsets S_1, S_2, S_3, such that if $\kappa_b = \prod_{j \in S_b} m_j$ then*

$$\min(g, \kappa_1) + \min(g, \kappa_2) + \min(g, \kappa_3) \geq 2g + 2. \tag{2.10}$$

Then model parameters are generically identifiable, up to a label swapping. Moreover, the statement remains valid when the mixing proportions $\boldsymbol{\pi}$ are held fixed and positive.

Links with the geometric approaches As shown in [Gov10], the geometric approach looking for the partition into g classes maximizing the information criterion or the χ^2 criterion is approximately equivalent to assume that individuals are drawn by a latent class model.

Parameter estimation

The inference of the latent class model can be performed in a frequentist or in a Bayesian framework.

In a frequentist point of view, the estimation of the MLE can be performed via an EM algorithm, presented below, or by its extensions. Note that the likelihood function is upper-bounded, so there is no degeneracy problem.

In a Bayesian framework, the estimation can be performed by a Gibbs sampler. Note that by choosing the Jeffreys non informative conjugate priors, the posterior distributions are explicit and an exact information criterion can be computed.

We now detail both frequentist and Bayesian approaches.

Frequentist framework The MLE can be easily obtained by the following EM algorithm.

Algorithm 2.11 (EM algorithm for the latent class model)**.**

Starting from the initial value of $\boldsymbol{\theta}^{[0]}$, iteration $[r]$ of the EM algorithm is written as

— **E step**: calculate conditional probabilities

$$t_{ik}(\boldsymbol{\theta}^{[r]}) = \frac{\pi_k^{[r]} \mathring{p}(\boldsymbol{x}_i; \boldsymbol{\alpha}_k^{[r]})}{p(\boldsymbol{x}_i; \boldsymbol{\theta}^{[r]})}. \tag{2.11}$$

— **M step**: maximization of the expectation of the complete-data log-likelihood

$$\pi_k^{[r+1]} = \frac{n_k^{[r]}}{n} \text{ and } \alpha_k^{jh[r+1]} = \frac{\sum_{i=1}^n t_{ik}(\boldsymbol{\theta}^{[r]}) x_i^{jh}}{n_k^{[r]}}, \tag{2.12}$$

where $n_k^{[r]} = \sum_{i=1}^n t_{ik}(\boldsymbol{\theta}^{[r]})$.

Bayesian framework The classical assumption of the independence between the prior distributions of the class proportions $\boldsymbol{\pi}$ and of the class parameters $\boldsymbol{\alpha}_k^j$ involves

that

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}) \prod_{k=1}^{g} \prod_{j=1}^{d} p(\boldsymbol{\alpha}_k^j). \tag{2.13}$$

As the Jeffreys non informative prior for a multinomial distribution is a conjugate Dirichlet one, the prior distribution is written as follows

$$p(\boldsymbol{\pi}) = \mathcal{D}_g\left(\frac{1}{2}, \ldots, \frac{1}{2}\right) \text{ and } p(\boldsymbol{\alpha}_k^j) = \mathcal{D}_{m_j}\left(\frac{1}{2}, \ldots, \frac{1}{2}\right). \tag{2.14}$$

The inference is also made by the following Gibbs sampler which generates a sequence of parameters from their posterior distributions. Note that this algorithm is easily performed since conjugate prior distributions involve explicit posterior distributions.

Algorithm 2.12 (Gibbs sampler for the latent class model)**.**

Starting from the initial value of $\boldsymbol{\theta}^{[0]}$, iteration $[r]$ of the Gibbs sampler having $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x})$ as stationary distribution is written as

$$\forall i = 1, \ldots, n \quad \boldsymbol{z}_i^{[r]} | \boldsymbol{x}_i, \boldsymbol{\theta}^{[r]} \sim \mathcal{M}_g\left(t_{i1}(\boldsymbol{\theta}^{[r]}), \ldots, t_{ig}(\boldsymbol{\theta}^{[r]})\right) \tag{2.15}$$

$$\boldsymbol{\pi}^{[r+1]} | \mathbf{z}^{[r]} \sim \mathcal{D}_g\left(\frac{1}{2} + \mathrm{n}_1^{[r]}, \ldots, \frac{1}{2} + \mathrm{n}_g^{[r]}\right) \tag{2.16}$$

$$\forall (k, j) \quad \boldsymbol{\alpha}_k^{j[r+1]} | \mathbf{x}, \mathbf{z}^{[r]} \sim \mathcal{D}_{m_j}\left(\frac{1}{2} + \mathrm{n}_k^{j1[r]}, \ldots, \frac{1}{2} + \mathrm{n}_k^{jm_j[r]}\right), \tag{2.17}$$

where $\mathrm{n}_k^{[r]} = \sum_{i=1}^{n} z_{ik}^{[r]}$ and $\mathrm{n}_k^{jh[r]} = \sum_{i=1}^{n} z_{ik}^{[r]} x_i^{jh}$.

Exact criterion By using the properties of the conjugate prior distributions, [BCG10] proposed an exact version of the ICL criterion for the latent class model. Indeed, the integrated complete-data likelihood of this model, defined by

$$p(\mathbf{x}, \mathbf{z}) = \int_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{2.18}$$

is explicit by using the prior distributions defined in (2.14). For any couple $(\mathbf{x}, \mathbf{z})$, the integrated complete-data likelihood is equal to

$$p(\mathbf{x}, \mathbf{z}) = \frac{\Gamma(\frac{g}{2})}{\Gamma(\frac{1}{2})^g} \frac{\prod_{k=1}^{g} \Gamma(\mathrm{n}_k + \frac{1}{2})}{\Gamma(n + \frac{g}{2})} \prod_{k=1}^{g} \prod_{j=1}^{d} \frac{\Gamma(\frac{m_j}{2})}{\Gamma(\frac{1}{2})^{m_j}} \frac{\prod_{h=1}^{m_j} \Gamma(\mathrm{n}_k^{jh} + \frac{1}{2})}{\Gamma(\mathrm{n}_k + \frac{m_j}{2})}. \tag{2.19}$$

Vector $\mathbf{z}$ is replaced, in the above equation, by its maximum likelihood estimate $\hat{\mathbf{z}}$ by using the MAP rule. Then, the exact ICL criterion is defined as follows

$$\mathrm{ICL}_{\mathrm{ex}} = \ln p(\mathbf{x}, \hat{\mathbf{z}}). \tag{2.20}$$

In [BCG10], the authors propose to use (2.19) in order to compute the complete-data likelihood by using importance sampling approach. They underline, by their numerical experiments, that the exact criterion outperforms the classical asymptotic information criteria (BIC and ICL). Thus, when the exact criteria are available, they have to be favored.

Running example 2.13 (Latent class model clustering).

Approach The MLE are estimated for different numbers of classes and the BIC criterion is used to select the best number of classes.

Interpretation The best model is the latent class model with three components. The estimated classes can be interpreted as follows.

— The majority class ($\pi_1 = 0.72$) groups the teeth diagnosed as sound with a strong probability by all the dentists. This probability is upper than 0.90 for the first four dentists and equal to 0.74 for the last one.

— The second class ($\pi_2 = 0.20$) groups the teeth claimed as sound by the first four dentists with more incertitude than in the previous class (probability between 0.50 and 0.90) while they are claimed as carious by the last dentist with probability 0.76.

— The third class ($\pi_3 = 0.08$) groups the teeth mainly declared as carious especially by the fifth dentist.

Parsimonious versions of the latent class model

The number of parameters required by the *latent class model* is equal to

$$(g-1) + g\sum_{j=1}^{d}(m_j - 1). \tag{2.21}$$

Thus, this number is generally strongly smaller than the number of parameters required by the *full log-linear model* which is equal to $\prod_{j=1}^{d} m_j$.

However, a better bias/variance trade off can be obtained by reducing the number of parameters for the latent class model. Thus, five parsimonious versions of the latent class model was introduced by G. Celeux and G. Govaert [CG91] for binary variables then these models was extended to the categorical variables [Gov10]. The constraints added on the parameter space require the introduction of a new model parametrization. With this new parameterization, the multinomial distribution of variable j for component k is determined by its center $\boldsymbol{a}_k^j$ denoting the majority modality and its a dispersion parameter $\boldsymbol{\varepsilon}_k^j$.

Definition 2.14 (Alternative parametrization of the parsimonious latent class model)**.** The latent class model can be parametrized as follows

$$p(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{n}\pi_k \prod_{j=1}^{d}\left((1-\varepsilon_k^j)^{a_k^{jh}}(\varepsilon_k^j/(m_j-1))^{1-a_k^{jh}}\right)^{x_i^{jh}}, \tag{2.22}$$

Thus, with $0 < \varepsilon_k^j < 1$, the parsimonious model assumes that one mode corresponding to the most likely modality is characteristic for each multinomial while the remaining probability mass is uniformly spread among the other modalities. This model requires $(g-1)+gd$ parameters. The other parsimonious models are obtained by assuming the equality of ε_k^j between the class or between the variables or between the class and the variables.

Limits of the latent class model

The latent class model may suffer from severe biases when the data are intra-class correlated. For instance, an application presented in [VHH09] shows that latent class model dramatically over-estimates the number of classes when the conditional independence assumption is violated. We now present three alternative mixture models relaxing the conditional independence assumption. Note that the larger is the number of variables, the higher is the risk to observe conditionally correlated variables in a data set, and consequently the higher is the risk to involve such biases by using the latent class model.

2.3.2 Log-linear mixture models with intra-class dependencies

Main idea The log-linear models [Agr02] purpose is to modelize the individual log-probability by selecting interactions between variables. Thus, the log-linear mixture model has been used for a long time [Har72, Hag88] to cluster categorical data set with intra-class dependencies. Note that some constraints have to be imposed on each log-linear model in order to obtain the model identifiability.

Running example 2.15 (Log-linear mixture model clustering [EH89]).

Approach M.A. Espeland and S.L. Handelman [EH89] apply a log-linear mixture model to fit the data. Note that authors estimate several models fixed by advance whose the best one considers a mixture with four components.

Interpretation The first two components take into account the interactions between the dentists 3 and 4. The last two components are specific since their allow only one modality interaction, when all the diagnoses are respectively carious and sound.

Comments Note that these assumptions are required by the authors due to their realistic nature. Indeed, this model fits the data better than the CIM model. On the other hand, its interpretation needs the analysis of four classes, so the data summary is more complex. Finally, we could criticize the building of the two specific classes modeling only one modality crossing. Indeed, these classes appears as artificially added in order to modelize the conditional dependencies.

Model selection By considering intra-class dependency of order one, the authors of [EH89] obtain good results for the clustering of radiographic cross-diagnostics. These authors perform the model selection by using a *forward* method which determines the intra-class interaction. However, note that this approach is sub-optimal and converges to a local optimum of the information criterion used by the practitioner. The model presented in [VHH09] considers the interactions of order two but during in the application there are only interactions of order one which are estimated. As for the previous article, authors have to determine by advance the intra-class interactions. The model selection for the log-linear mixture models is a complex problem since the number of models becomes huge with the number of variables.

Too many parameters The number of parameters required by the log-linear mixture model increases with the number of modalities and with the considered order of interactions. Thus, this model can fit well the data but it may need too many parameters. So, there is an over-fitting risk and the interpretation becomes harder. Furthermore, the parameters can be poorly meaningful if there are too numerous.

Conclusion The log-linear mixture model is a powerful tool to cluster categorical data. However, it is important to impose constraints on the parameters space in order to provide a meaningful model. Both mixture models presented in Chapter 3 and Chapter 4 can be interpreted as log-linear mixture models with specific

constraints which control the number of parameters and which provide meaningful classes. Both models are given with an efficient approach to perform the model selection in a Bayesian framework.

2.4 Mixtures of trees

2.4.1 Dependence trees

Main idea This approach, proposed by C. Chow and C. Liu [CL68], consists in approximating discrete multivariate probability distribution with dependence trees, *i.e.* with a product of second-order distributions.

Definition 2.16 (Pdf of a dependence tree distribution)**.** Let the tree $T = \{E, V\}$ where $E = \{1, \ldots, d\}$ and $V = \{(j, j') : \ j \in E \text{ and } j' \in E \setminus j\}$. If variable $\boldsymbol{x}_i$ is sampled from a dependence tree distribution defined by T, then its pdf is written as follows

$$p(\boldsymbol{x}_i; \boldsymbol{\alpha}) = \frac{\prod_{(j,j') \in V} p(\boldsymbol{x}_i^j, \boldsymbol{x}_i^{j'}; \boldsymbol{\beta}^{jj'})}{\prod_{j=1}^{d} p(\boldsymbol{x}_i^j; \boldsymbol{\alpha}^j)^{v_j - 1}}, \tag{2.23}$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^j, \boldsymbol{\beta}^{jj'}; j = 1, \ldots, d; j' \text{ as such } (j, j') \in V)$, where v_j denotes the cardinal of the neighbor of edge j. The pdf of component k is defined by

$$p(\boldsymbol{x}_i^j; \boldsymbol{\alpha}^j) = \prod_{h=1}^{m_j} (\alpha^{jh})^{x_i^{jh}} \text{ and } p(\boldsymbol{x}_i^j, \boldsymbol{x}_i^{j'}; \boldsymbol{\beta}^{jj'}) = \prod_{h=1}^{m_j} \prod_{h'=1}^{m_{j'}} (\beta^{jj'hh'})^{x_i^{jh} x_i^{j'h'}}, \tag{2.24}$$

with $\boldsymbol{\alpha}^j = (\alpha^{jh}; h = 1, \ldots, m_j)$ and $\boldsymbol{\beta}^{jj'} = (\boldsymbol{\beta}^{jj'hh'}; h = 1, \ldots, m_j; h' = 1, \ldots, m_{j'})$. The parameter $\boldsymbol{\alpha}^{jh}$ denotes the probability that variable j takes modality h and the parameter $\boldsymbol{\beta}^{jj'hh'}$ denotes the probability that the couple of variables (j, j') takes the couple of modalities (h, h').

Estimation As shown in [CL68], the maximum likelihood estimate can be directly obtained by using the Kruskal algorithm which estimates the tree of minimal length [Kru56]. The value of the branch weight between the two random variables $\boldsymbol{X}^j$ and $\boldsymbol{X}^{j'}$ is given by the mutual information defined as

$$I(\boldsymbol{X}^j, \boldsymbol{X}^{j'}) = \sum_{h=1}^{m_j} \sum_{h'=1}^{m_{j'}} p(X^{jh} = 1, X^{j'h'} = 1) \ln \frac{p(X^{jh} = 1, X^{j'h'} = 1)}{p(X^{jh} = 1) p(X^{j'h'} = 1)}. \tag{2.25}$$

From this definition, the empirical mutual information is deduced for a sample $\mathbf{x}$.

Definition 2.17 (Empirical mutual information)**.** The empirical mutual information between $\mathbf{x}^j = (\boldsymbol{x}_i^j; i = 1, \ldots, n)$ and $\mathbf{x}^{j'} = (\boldsymbol{x}_i^{j'}; i = 1, \ldots, n)$ computed from the sample $\mathbf{x}$ is defined as

$$\hat{I}(\mathbf{x}^j, \mathbf{x}^{j'}) = \sum_{h=1}^{m_j} \sum_{h'=1}^{m_{j'}} f(x^{jh}, x^{j'h'}) \ln \frac{f(x^{jh}, x^{j'h'})}{f(x^{jh}) f(x^{j'h'})}, \tag{2.26}$$

where $f(x^{jh}, x^{j'h'}) = \frac{1}{n} \sum_{i=1}^{n} x_i^{jh} x_i^{j'h'}$ and $f(x^{jh}) = \frac{1}{n} \sum_{i=1}^{n} x_i^{jh}$.

Algorithm 2.18 (Estimation of the dependence tree).

1. **Compute** $\hat{I}(\mathbf{x}^j, \mathbf{x}_i^{j'}), \forall (j, j')$
2. **Index** the $d(d-1)/2$ branches according to their weight . So the weight b_ℓ is greater than or equal to the weight $b_{\ell'}$ whenever $j < j'$.
3. **Select** b_1 and b_2
4. For $\ell = 3$ to $d(d-1)/2$: **add** the branch b_ℓ if it does not form a cycle with the set previously selected.

Remark 2.19 (Unique solution). If the weights are all different, then the solution of Algorithm 2.18 is unique.

2.4.2 Tree mixture model

Main idea This approach, proposed by M. Meila and M.I. Jordan [MJ01], generalizes the probabilistic trees to the mixture model framework. The authors assume that each component follows a distribution per dependence tree defined in (2.23).

Estimation In a frequentist framework, the MLE is easily obtained by an EM algorithm. The M step maximizes the expectation of the complete likelihood by using Algorithm 2.18 where the empirical mutual information is computed according to the conditional probabilities of the class memberships. In a Bayesian framework, the MAPE is also obtained by a specific EM algorithm maximizing the posterior distribution.

Running example 2.20 (Tree mixture model clustering).

Approach We cluster the data set with mixture models of dependency trees with different numbers of classes and we use the BIC criterion to select the best one.

Interpretation The best model is the bi-component one. Note that this model requires the estimation of 19 parameters while the latent class model requires only 11 parameters. If its BIC criterion value is better than the bi-component latent class model (respectively -7490 and -7511), its global result is not better since the tri-component latent class model obtains a BIC criterion value of -7481. If the BIC criterion values are relatively close, the partitions are different as shown by the confusion matrices presented in Table 2.2.

	c1-tree	c2-tree
c1-bi-LCM	3037	191
c2-bi-LCM	55	586

(a)

	c1-tree	c2-tree
c1-tri-LCM	2922	0
c2-tri-LCM	170	484
c3-tri-LCM	7	293

(b)

Table 2.2 – Confusion matrices between the partition obtained by the bi-component mixture of trees and the partition obtained by: (a) the bi-component latent class model; (b) the tri-component latent class model.

Conclusion The main problem of these models is that they require too often an intractable number of parameters. Furthermore, the tree structure is often unstable. Indeed, if the data set is a little bit changed, then the tree structure can be very different. Thus, the interpretation based on this structure can be irrelevant. Finally, note that the mixtures of trees are meaningful principally when the tree structure explains some causal relationships.

2.5 Multilevel latent class model

Main idea The idea is to consider two latent variables. The first one is categorical and is relative to the class membership. The second one is continuous (univariate or multivariate) and modelizes the intra-class dependencies.

Overview of these methods When covariates are available, the conditional dependencies between the categorical ones can be modeled by a logistic function [For92, RIW08]. By assuming that these covariates are unobserved, the *multilevel latent class model* [Ver03, Ver07] naturally incorporates the intra-class dependencies. This model has connections with the approach of Y. Qu, M. Tan and M.H. Kutner [QTK96] where the intra-class dependencies are modeled by a latent continuous variable with a probit function. The *hybrid model* [Mut08] in which, for each class, a factor analysis model is fitted to either all categorical variables or to those categorical variables having dependencies is a more general approach. Recently, I. Gollini and T.B. Murphy [GM13] have proposed the *mixture model of latent traits analyzers* which assumes that the distribution of the categorical variables depends on both a categorical latent variable (the class) and many continuous latent traits variables. The inference is also a difficult point which is solved via a variational approach. If all these models consider the intra-class dependencies, their main drawback is that these dependencies have to be interpreted among relations with a latent variable. Thus, pertinent interpretation can be difficult.

Focus on the [QTK96] approach The multilevel latent class model proposed in [QTK96] is introduced to analyze binary data sets. It explains the intra-class dependencies by a logit function.

Definition 2.21 ([QTK96] multilevel latent class model)**.** The pdf of component k

is written as

$$p(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) = \int_{\mathbb{R}} \prod_{j=1}^{d} \Phi(a_{kj} + b_{kj}t)^{x_i^{j1}} (1 - \Phi(a_{kj} + b_{kj}t))^{1-x_i^{j1}} d\Phi(t), \tag{2.27}$$

where $\Phi(.)$ denotes the cumulative distribution function of a standard normal variable.

In practice, this pdf is approximated by using the Gauss-Hermit quadrature. The MLE is obtained by using an EM algorithm. We now present the result of this model for the running example.

Running example 2.22 (Clustering with the random effects model)**.**

Approach As the last specific model with four classes proposed in [EH89] seems artificial, the authors of [QTK96] prefer to use the random effects models in a latent class analysis with two classes. They assume that conditional dependencies can be modelized by a single continuous latent variable which varies among the individuals. According to the authors, the latent continuous variable can reflect the influence of the condition of images.

Interpretation According to the authors, one class represents the sound teeth and the other represents the carious ones. The random effect represents all the patient specific unrecorded characteristics of the x-ray images. Their model does not require the two additional artificial classes. Thus, their interpretation is easier even if it is not easy to evaluate the strength of the intra-class dependencies

Conclusion By adding two levels of latent variables, the multilevel latent class models permits to consider the intra-class dependencies. However, it is not easy to characterize these dependencies since there is none parameter reflecting the strength of these dependencies.

2.6 Conclusion

The classical latent class model is often biased when the sample size is large because its conditional independence assumption is violated. Different methods allow to cluster the data by taking into account the intra-class dependencies. However, there is not any model which provides one coefficient to characterize the strength of these dependencies.

In this overview, we have not spoken about the mixture of factor analyzers since we want to favor the models which are easily interpretable. So, we have focused on models which cluster the individuals by modeling the distribution of the variables in their native space.

The log-linear mixture models seems to be the most general one. We also propose, in the two following chapters, two mixture models which add specific constraints to this general model. Thus, both proposed models allow to summarize the intra-class dependencies with few parameters.

Chapter 3

Model-based clustering with blocks of extreme distributions

This chapter introduces a new extension of the latent class model. This model groups the variables into conditionally independent blocks. The specific distribution of the blocks modelizes the intra-class dependencies and provides one coefficient summarizing the strength of these dependencies.
A maximum likelihood inference is performed by a GEM *algorithm while the combinatorial problems of the model selection are avoided by a* MCMC *algorithm.*
Numerical experiments, on simulated and real data sets, underline the main characteristics of this new mixture model.

Science never solves a problem without creating ten more.
George Bernard Shaw

3.1 Introduction

We propose to extend the classical latent class model for categorical data, by a new mixture model which relaxes the conditional independence assumption between the variables. We refer to the proposed model as the *mixture of extreme dependency distributions per blocks* (denoted by MEDD).

The MEDD model groups the variables into *conditionally independent blocks* given the class. The main intra-class dependencies are thus underlined by the repartition of the variables into these blocks. This approach, allowing modeling of the main conditional interactions, was first proposed by M. Jorgensen and L. Hunt [JH96, HJ99] in order to cluster data sets with continuous and categorical variables. For the MEDD model, each block follows a particular dependency distribution which consists in a bi-component mixture of the *independence* and the *maximal dependency* distribution according to the Cramer's V criterion. This specific distribution of the blocks

provides one parameter summarizing the strength of the conditional dependencies of the variables. This crucial parameter is the proportion of the maximum dependency distribution. Furthermore, the nature of the conditional dependencies is bring out by the relation defined by the maximum dependency distribution. Thus, the model puts the light on the main conditional dependencies and their strengths.

The proposed model can be interpreted as a two-level parsimonious version of a log-linear mixture model and thus benefits from its interpretative power. The first level defines the considered interactions by grouping in the same block the variables which are conditionally dependent. The strength of this dependency is reflected by the proportion of the distribution of maximum dependency compared to that of the independence distribution. The second level of sparsity is induced by the small fraction of the parameters of the maximum dependency distribution of the block. As for all log-linear mixture models, the selection of the pertinent interactions is a combinatorial problem. Therefore, we propose to perform the model selection via a MCMC algorithm in order to avoid the enumeration of all the models. Thus, this general approach could also select the interactions of a more general log-linear mixture model.

Structure of this chapter This chapter is organized as follows. Section 3.2 presents the mixture model of conditionally independent blocks of variables. Section 3.3 presents the new mixture model taking into account the intra-class dependencies. Section 3.4 is devoted to the estimation of the parameters by maximization of the likelihood in the case where the class number and the blocks of variables are supposed to be known. Section 3.5 presents a MCMC algorithm avoiding combinatorial difficulties inherent to block selection. Section 3.6 presents results on simulated data. Section 3.7 illustrates the MEDD model on two real clustering challenges. A conclusion is given in Section 3.8. Note that a tutorial of the R package `Clustericat` [1] performing the model selection and the estimation of the parameters of MEDD is given in Chapter 5. All these results are part of the article *Model-based clustering for conditionally correlated categorical data* [MBV13a].

3.2 Mixture of intra-class independent blocks

Main idea The mixture model of intra-class independent blocks considers that, *conditionally* on class k, variables are grouped into B_k *independent blocks* and each block follows a specific distribution.

A partition of the variables per class The repartition of the variables into blocks determines a partition $\boldsymbol{\sigma}_k = (\boldsymbol{\sigma}_{k1}, \ldots, \boldsymbol{\sigma}_{k\mathrm{B}_k})$ 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 the vector of categorical variables $\boldsymbol{x}_i^{\{kb\}} = \boldsymbol{x}_i^{\boldsymbol{\sigma}_{kb}} = (\boldsymbol{x}_i^{\{kb\}j}; j = 1, \ldots, d^{\{kb\}})$ which is the subset of $\boldsymbol{x}_i$ associated to $\boldsymbol{\sigma}_{kb}$. The integer $d^{\{kb\}} = \mathrm{card}(\boldsymbol{\sigma}_{kb})$

1. The R package `Clustericat` is available on Rforge website at the following url: *https://r-forge.r-project.org/R/?group_id=1803*

is the number of variables affiliated to block b of component k. The vector $\boldsymbol{x}_i^{\{kb\}j} = (x_i^{\{kb\}jh}; h = 1, \ldots, m_j^{\{kb\}})$ corresponds to variable j of block b for component k and uses a complete disjunctive coding where $m_j^{\{kb\}}$ is the number of modalities for the variable $\boldsymbol{x}_i^{\{kb\}j}$. Thus, $x_i^{\{kb\}jh} = 1$ if individual i takes modality h for variable $\boldsymbol{x}_i^{\{kb\}j}$ and $x_i^{\{kb\}jh} = 0$ otherwise.

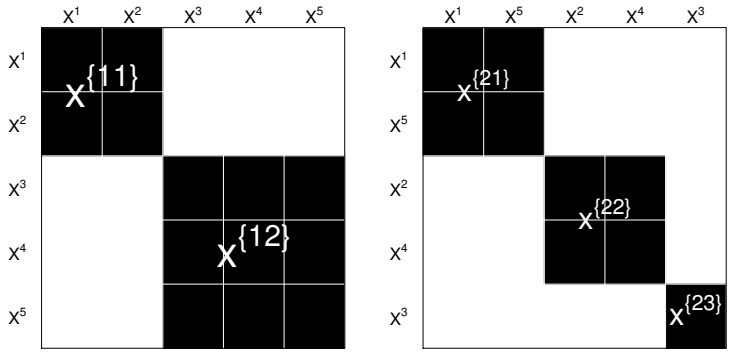
Remark 3.1 (Different intra-class dependencies)**.** Different variables repartitions in blocks are allowed for each component and they are grouped into $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_g)$.

Definition 3.2 (Mixture model of conditionally independent blocks of variables)**.** Let $\boldsymbol{x}_i$ to be the d-variate categorical variable arisen from a mixture model of conditionally independent blocks of variables whose the partition is denoted by $\boldsymbol{\sigma}$ and the parameters by $\boldsymbol{\theta}$. Then, its pdf is written as follows

$$p(\boldsymbol{x}_i; \boldsymbol{\sigma}, \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}_i; \boldsymbol{\sigma}_k, \boldsymbol{\alpha}_k) \text{ with } p(\boldsymbol{x}_i; \boldsymbol{\sigma}_k, \boldsymbol{\alpha}_k) = \prod_{b=1}^{\mathrm{B}_k} p(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\alpha}_{kb}), \qquad (3.1)$$

where $\boldsymbol{\alpha}_k = (\boldsymbol{\alpha}_{k1}, \ldots, \boldsymbol{\alpha}_{k\mathrm{B}_k})$ and where $p(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\alpha}_{kb})$ is the pdf of the block b of the component k parametrized by $\boldsymbol{\alpha}_{kb}$.

Example 3.3 (Bi-component mixture model of conditionally independent blocks of variables)**.** *Let $\boldsymbol{x}_i = (\boldsymbol{x}_i^1, \ldots, \boldsymbol{x}_i^5)$ be the vector of five categorical variables following the bi-component mixture model of conditionally independent blocks. The partition of the variables of this model is $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$ with $\boldsymbol{\sigma}_1 = (\{1, 2\}, \{3, 4, 5\})$ and $\boldsymbol{\sigma}_2 = (\{1, 5\}, \{2, 4\}, \{3\})$. Figure 3.1 illustrates the intra-class dependencies taken into account by the model: blank cell indicates that the intra-class correlation is neglected and black cell indicates that this correlation is taken into account.*



(a) Class 1 (b) Class 2

Figure 3.1 – Intra-class dependencies taken into account by the bi-component mixture model of conditionally independent blocks of variables with $\boldsymbol{\sigma}_1 = (\{1, 2\}, \{3, 4, 5\})$ and $\boldsymbol{\sigma}_2 = (\{1, 5\}, \{2, 4\}, \{3\})$.

Note that the classical latent class model with conditional independence would be represented by white cells off the diagonal and black on the latter.

Related models The approach per conditionally independent blocks is very general, since any distribution can be chosen for each block distribution $p(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\alpha}_{kb})$. The mixture model by conditional independent blocks is a parsimonious version of the log-linear mixture model. Indeed, the distribution of each block determines which interactions are considered. Note that the order of these interactions is determined by the number of the variables into the block. Finally, the interactions between variables of different blocks will be zero and those between variables of the same block can be modelized by the specific distribution of the block. The limiting case of this model where $\mathrm{B}_k = d$ for each class is equivalent to the latent class model.

Generic identifiability The generic identifiability of the mixture models for categorical data can be difficultly proved. However, by adding some constraints on the repartition of the variables into blocks, we can use Theorem 2.10 [AMR09]. Thus, the generic identifiability of the model is obtained by using its conditional independence assumption between blocks under two sufficient conditions.

Corollary 3.4 (Generic identifiability of the mixture model of conditionally independent blocks of variables equal between classes)**.** *If* $\boldsymbol{\sigma}_1 =, \ldots, = \boldsymbol{\sigma}_g$ *with* $\mathrm{B}_1 \geq 3$, *and if that the block distributions* $p(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\alpha}_{kb})$ *are identifiable and have* υ_b *degrees of freedom, then suppose there exists a tripartition of the set* $S = \{\boldsymbol{\sigma}_{11}, \ldots, \boldsymbol{\sigma}_{1\mathrm{B}_1}\}$ *into three disjoint nonempty subsets* S_1, S_2 *and* S_3, *such that if* $\kappa_u = \prod_{\{j \in S_u\}} \upsilon_j$ *then*

$$\min(g, \kappa_1) + \min(g, \kappa_2) + \min(g, \kappa_3) \geq 2g + 2. \tag{3.2}$$

Then model parameters are generically identifiable, up to label swapping.

Proof. There is a bijection from $\boldsymbol{x}_i^{\{kb\}}$ to $\tilde{\boldsymbol{x}}_i^{\{kb\}}$ where $\tilde{\boldsymbol{x}}_i^{\{kb\}}$ is a categorical variable having υ_b modalities. The variable $\tilde{\boldsymbol{x}}_i^{\{kb\}}$ follows the latent class model, so its identifiability is defined by Theorem 2.10. Thus, we conclude to the generic identifiability of the model drawing $\boldsymbol{x}_i^{\{kb\}}$. □

Corollary 3.5 (Generic identifiability of the mixture model of conditionally independent blocks of variables)**.** *If there exists a tri-partition of* $\boldsymbol{\sigma}_k$, *equals for each* $k = 1, \ldots, g$, *into three disjoint non-empty subsets* S_1, S_2, S_3*:*

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

and if that the block distributions $p(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\alpha}_{kb})$ *are identifiable and have* υ_b *degrees of freedom such that if* $\kappa_u = \prod_{\{j \in S_u\}} \upsilon_j$ *then*

$$\min(g, \kappa_1) + \min(g, \kappa_2) + \min(g, \kappa_3) \geq 2g + 2. \tag{3.3}$$

Then model parameters are generically identifiable, up to label swapping.

Proof. The proof is similar than the proof of Corollary 3.4. Note that the existence of $\tilde{\boldsymbol{x}}_i^{\{kb\}}$ is assured by the equality of the tri-partition of the $\boldsymbol{\sigma}_k$ between classes. □

3.3 Parsimonious block distribution

Main idea The aim is to define a parsimonious distribution for each block that takes into account the dependency between variables. Furthermore, the parameters of the distribution inside block must be meaningful for the practitioner. In this context, we propose to modelize the distribution of each block by a mixture of the two extreme distributions according to the Cramer's V criterion computed on all the couples of variables. The model results in a bi-component mixture between an independence distribution and a maximum dependency distribution which can be easily interpreted by the user.

The maximum dependency distribution is introduced first, then the mixture model of extreme dependency distributions per blocks (MEDD) is secondly detailed.

Remark 3.6 (Ordered variables)**.** Without loss of generality, the variables are considered as ordered by decreasing number of modalities in each block

$$\forall (k,b)\ m_j^{\{kb\}} \geq m_{j+1}^{\{kb\}} \text{ where } j = 1, \ldots, d^{\{kb\}} - 1.$$

3.3.1 Maximum dependency distribution

Main idea The maximum dependency distribution is defined as the "opposite" distribution of independence according to the Cramer's V criterion computed on all the couples of variables. Indeed, the independence distribution minimizes this criterion while the maximum dependency distribution maximizes it. Under this distribution, the modality knowledge of one variable provides the maximum information on all the subsequent variables.

Remark 3.7 (Non-reciprocal functional link)**.** Note that it is a non-reciprocal functional link between variables. Indeed, if $\boldsymbol{x}_i^{\{kb\}}$ arises from this distribution, the knowledge of the variable having the largest number of modalities determines exactly the others but the reverse does not necessarily apply.

Remark 3.8 (Successive surjections)**.** This distribution defines successive surjections from the space of $x_i^{\{kb\}j}$ to the space of $x_i^{\{kb\}j+1}$ with $j = 1, \ldots, d^{\{kb\}} - 1$ (recall that the variables are ordered by decreasing number of modalities in each block). In fact, it is a reciprocal functional link only when $m_j^{\{kb\}} = m_{j+1}^{\{kb\}}$.

Parametrization 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 the continuous vector

$$\boldsymbol{\tau}_{kb} = (\tau_{kb}^h; h = 1, \ldots, m_1^{\{kb\}}) \text{ with } \tau_{kb}^h \geq 0 \text{ and } \sum_{h=1}^{m_1^{\{kb\}}} \tau_{kb}^h = 1, \tag{3.4}$$

and the product between the conditional distributions defined as specific multinomial distributions. So, conditionally on $x_i^{\{kb\}1h} = 1$, then for $j = 2, \ldots, d^{\{kb\}}$, $\boldsymbol{x}_i^{\{kb\}j}$

follows a multinomial distribution parametrized by the discrete vector

$$\boldsymbol{\delta}_{kb}^{hj} = (\delta_{kb}^{hjh'}; h' = 1, \ldots, m_j^{\{kb\}}) \text{ with } \delta_{kb}^{hjh'} \in \{0,1\}, \sum_{h'=1}^{m_j^{\{kb\}}} \delta_{kb}^{hjh'} = 1 \text{ and } \sum_{h=1}^{m_1^{\{kb\}}} \delta_{kb}^{hjh'} \geq 1. \tag{3.5}$$

Note that the above constraints define the successive surjections. By denoting $\boldsymbol{\delta}_{kb} = (\boldsymbol{\delta}_{kb}^{hj}; h = 1, \ldots, m_1^{\{kb\}}; j = 2, \ldots, d^{\{kb\}})$, the distribution of maximum dependency can be now defined.

Definition 3.9 (Maximum dependency distribution)**.** Let $\boldsymbol{x}_i^{\{kb\}}$ be the $d^{\{kb\}}$-variate categorical variable following the maximum dependency distribution whose the discrete parameters are denoted by $\boldsymbol{\delta}_{kb}$ and whose the continuous ones are denoted by $\boldsymbol{\tau}_{kb}$. Then, its pdf is written as follows

$$\begin{aligned} \acute{p}(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb}) &= p(\boldsymbol{x}_i^{\{kb\}1}; \boldsymbol{\tau}_{kb}) \prod_{j=2}^{d^{\{kb\}}} p(\boldsymbol{x}_i^{\{kb\}j} | \boldsymbol{x}_i^{\{kb\}1}; \{\boldsymbol{\delta}_{kb}^{hj}\}_{h=1,\ldots,m_1^{\{kb\}}}) \\ &= \prod_{h=1}^{m_1^{\{kb\}}} \left(\tau_{kb}^h \prod_{j=2}^{d^{\{kb\}}} \prod_{h'=1}^{m_j^{\{kb\}}} (\delta_{kb}^{hjh'})^{x_i^{\{kb\}jh'}} \right)^{x_i^{\{kb\}1h}}. \end{aligned} \tag{3.6}$$

Example 3.10 (Bivariate and tri-variate maximum dependency distributions)**.** *Let the mixture model whose the blocks of variables for the first component are defined by* $\boldsymbol{\sigma}_1 = (\{1,2\}, \{3,4,5\})$*. The distributions of the blocks are maximum dependency ones whose the parameters are the following*

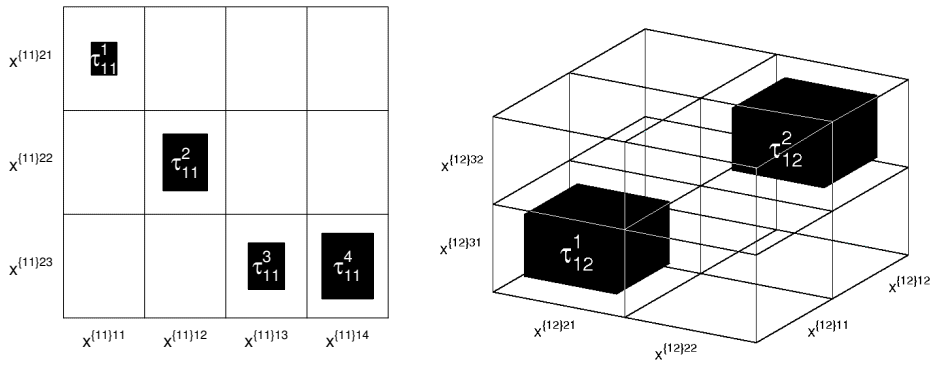
$$\begin{aligned} \delta_{11}^{111} = \delta_{11}^{212} = \delta_{11}^{313} = \delta_{11}^{413} = \delta_{12}^{1j1} = \delta_{12}^{2j2} = 1, \\ \boldsymbol{\tau}_{11} = (0.1, 0.3, 0.2, 0.4) \text{ and } \boldsymbol{\tau}_{12} = (0.5, 0.5). \end{aligned}$$

Figure 3.2 displays the probabilities of the joint distributions by the area of dark boxes. Note that $\boldsymbol{\delta}_{kb}$ *defines the* locations *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).*

Identifiability A sufficient condition of identifiability is to impose $\tau_{kb}^h > 0$, for all $h = 1, \ldots, m_1^{\{kb\}}$. This distribution has very limited interest because it is so unrealistic that it can almost never be used alone, we now present how to use it in a more efficient way.

3.3.2 Block distribution: mixture of two extreme distributions

Main idea We assume that the blocks composed by at least two variables follow a bi-components mixture between an *independence* distribution and a *maximum dependency* distribution while the block composed by one variable follow a multinomial distribution.



(a) First block of class 1 (b) Second block of class 1

Figure 3.2 – Two examples of block distributions following a maximum dependency distribution where $m_1^{\{11\}} = 4$, $m_2^{\{11\}} = 3$ and $m_1^{\{12\}} = m_2^{\{12\}} = m_3^{\{12\}} = 2$.

Definition 3.11 (Model-based clustering of blocks of extreme distributions)**.** A d-variate categorical variable $\boldsymbol{x}_i$ is generated by a MEDD model if it is drawn by a mixture model of conditionally independent blocks whose the pdf is written as follows

$$p(\boldsymbol{x}_i; \boldsymbol{\sigma}, \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \prod_{b=1}^{\mathrm{B}_k} p(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\alpha}_{kb}). \tag{3.7}$$

Moreover, the pdf of block b for component k is written as

$$p(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\alpha}_{kb}) = \begin{cases} (1-\rho_{kb})\mathring{p}(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\xi}_{kb}) + \rho_{kb}\acute{p}(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb}) & \text{if } d^{\{kb\}} > 1 \\ \mathring{p}(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\xi}_{kb}) & \text{otherwise,} \end{cases} \tag{3.8}$$

where $\acute{p}(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb})$ is the pdf of the maximum dependency distribution defined by (3.6) and where $\mathring{p}(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\xi}_{kb})$ is the pdf of the independence distribution defined by $\mathring{p}(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\xi}_{kb}) = \prod_{j=1}^{d^{\{kb\}}} \prod_{h=1}^{m_j^{\{kb\}}} (\xi_{kb}^{jh})^{x_i^{\{kb\}jh}}$. The parameter $\boldsymbol{\alpha}_{kb} = (\rho_{kb}, \boldsymbol{\xi}_{kb}, \boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb})$ groups the parameters of block b for component k. Finally, the real $\rho_{kb} \in [0, 1]$ is the proportion of the maximum dependency distribution.

Number of parameters The MEDD model requires little additional parameters compared with the CIM model. Indeed, for each block with at least two variables, the number of additional parameters depends only on the number of modalities of the first variable of the block and not on the number of variables into the block. The number of parameters of MEDD, denoted by ν_{MEDD}, is also defined by

$$\nu_{\text{MEDD}} = (g-1) + g\sum_{j=1}^{d}(m_j - 1) + \sum_{\{(k,b)|d^{\{kb\}}>1\}} m_1^{\{kb\}}. \tag{3.9}$$

In addition, the MEDD model 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 dependency distribution are no longer defined.

Meaningful block distribution Under this distribution, the proportion of the maximum dependency distribution reflects the deviation from independence under the assumption that the alternative distribution is the maximum dependency distribution. The parameter ρ_{kb} gives an indicator of the *inter-variable dependency* of the block. It is not here a pairwise dependency among variables but a dependency between 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-variable dependencies* are defined by $\boldsymbol{\delta}_{kb}$. The strength of these dependencies is explained by $\boldsymbol{\tau}_{kb}$. Indeed, this vector gives the *weight of the over-represented modality crossings* compared with the independence distribution.

Parsiominous log-linear mixture model We interpreted the MEDD model as a two-level parsimonious version of the log-linear mixture model. The first one is defined by the repartition of the variables into blocks determining the conditional interactions to be modeled. The specific block distribution adds a second level of parsimony since among the interactions allowed by each component, only those corresponding to the maximum dependency distribution are modeled while the other ones are considered as null.

Identifiability The proposed distribution is identifiable under the condition that the block is composed by at least three variables ($d^{\{kb\}} > 2$) or that the modality number of the last variable of the block is greater than two ($m_2^{\{kb\}} > 2$). This result is demonstrated in Appendix A.1. We remind that the parameter ρ_{kb} is a new indicator allowing to measure the dependency between variables, not limited to dependency between couples of variables. However, if $d^{\{kb\}} = 2$ and $m_2^{\{kb\}} = 2$ then the block distribution is not identifiable so 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 dependency 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 as illustrated by the following example.

Example 3.12 (Cramer's V and ρ_{kb}: two measures of the dependency)**.** *Figure 3.3 presents the link between the Cramer's V and ρ_{kb} on simulated bivariate binary variables. Such a behavior has also been observed in many other situations.*

3.4 Maximum likelihood estimation via a GEM algorithm

Aim Let $\mathbf{x} = (\boldsymbol{x}_1, \dots \boldsymbol{x}_n)$ be the sample composed with n independent and identically distributed individuals assumed to arise from the MEDD model. From this sample, the aim is to estimate the MLE for a fixed model $\boldsymbol{m}$ defined by $(g, \boldsymbol{\sigma})$.

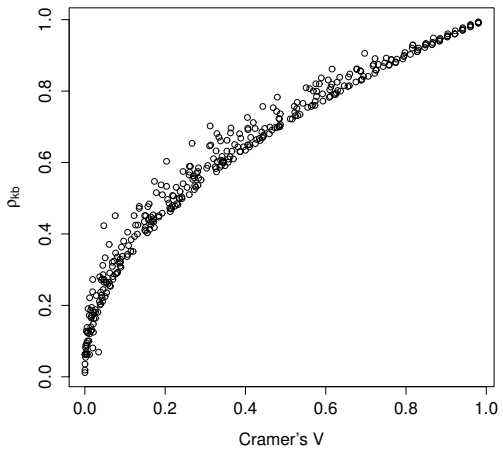


Figure 3.3 – Evolution of ρ_{kb} computed with the identifiability constraint according to the Cramer's V for two binary variables.

Combinatorial problem We have seen in Section 1.3.2 that the inference for a mixture model can be performed via an EM algorithm or one of its extensions if the maximization of the complete-data log-likelihood is easy. However, it is not the case for the MEDD model since the estimation of the discrete parameters of the maximum dependency distribution is a combinatorial problem. Indeed, 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 $\prod_{j=1}^{d^{\{kb\}}-1} S(m_j^{\{kb\}}, m_{j+1}^{\{kb\}})$. So, an exhaustive enumeration for estimating the discrete parameters is generally impossible when a block contains variables with many modalities and/or many variables.

Example 3.13 (Combinatorial problem involved by the discrete parameters)**.** *A block with three variables and* $m^{\{kb\}} = (5, 4, 3)$ *implies* 51 840 *possibilities for* $\boldsymbol{\delta}_{kb}$*.*

Estimation map The parameters are estimated via a GEM algorithm avoiding the classical problem involved by the unknown class membership. At its GM step, the maximization of the expectation of the complete-data likelihood is independently performed on the parameters of each block. Thus, at the GM step of iteration $[r]$, the combinatorial problem of the discrete parameter estimation for block b of component k is overcame by a Metropolis-Hastings algorithm whose the stationary distribution is close to $p(\boldsymbol{\delta}_{kb}|\mathbf{x}^{\{kb\}}, \mathbf{z}^{[r]})$. The proposal distribution of this algorithm randomly samples the candidate $\boldsymbol{\delta}_{kb}^{\star}$ while the candidate $(\rho_{kb}^{\star}, \boldsymbol{\xi}_{kb}^{\star}, \boldsymbol{\tau}_{kb}^{\star})$ is deterministically computed in order to maximize $p(\rho_{kb}^{\star}, \boldsymbol{\xi}_{kb}^{\star}, \boldsymbol{\tau}_{kb}^{\star}, \boldsymbol{\delta}_{kb}^{\star}|\mathbf{x}^{\{kb\}}, \mathbf{z}^{[r]})$. Note that the continuous parameters $(\rho_{kb}^{\star}, \boldsymbol{\xi}_{kb}^{\star}, \boldsymbol{\tau}_{kb}^{\star})$ are conditionally obtained by an EM algorithm by introducing a second latent variable denoting the membership of the dependency distributions of the block (independence or maximum dependency distribution).

3.4.1 Global GEM algorithm

Main idea The inference could be performed via an EM algorithm overcoming the problem of the class membership. However, as the optimization of the expectation of the complete-data log-likelihood on the discrete parameters is performed via a stochastic algorithm, we can only assure the increase of the expectation of the complete-data log-likelihood and not its maximization. So, the inference is performed via the following GEM algorithm.

Algorithm 3.14 (The GEM algorithm to obtain the MEDD model MLE)**.**

Starting from an initial value $\boldsymbol{\theta}^{[0]}$, its iteration $[r]$ is written as

— **E step**: calculate conditional probabilities

$$t_{ik}(\boldsymbol{\theta}^{[r]}) = \frac{\pi_k^{[r]} p(\boldsymbol{x}_i; \boldsymbol{\sigma}_k, \boldsymbol{\alpha}_k^{[r]})}{p(\boldsymbol{x}_i; \boldsymbol{\sigma}, \boldsymbol{\theta}^{[r]})}.$$

— **GM step**: increase of the expectation of the complete-data log-likelihood

$$\pi_k^{[r+1]} = \frac{n_k^{[r]}}{n} \text{ and } \boldsymbol{\alpha}_{kb}^{[r+1]} \text{ as such } L_{kb}(\boldsymbol{\alpha}_{kb}^{[r+1]}; \mathbf{x}, \mathbf{t}^{[r]}) \geq L_{kb}(\boldsymbol{\alpha}_{kb}^{[r]}; \mathbf{x}, \mathbf{t}^{[r]}),$$

where $L_{kb}(\boldsymbol{\alpha}_{kb}; \mathbf{x}, \mathbf{t}^{[r]}) = \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]}) \ln p(\boldsymbol{x}_i^{kb}; \boldsymbol{\alpha}_{kb})$ with $\mathbf{t}^{[r]} = (t_{ik}(\boldsymbol{\theta}^{[r]}); i = 1, \ldots, n; k = 1, \ldots, g)$ and $n_k^{[r]} = \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]})$.

This algorithm is stopped after $r_{\max}$ iterations. The optimization on each $\boldsymbol{\alpha}_{kb}$ is independently performed for each (k, b) at the GM step by the following Metropolis-Hastings algorithm.

3.4.2 Focus on the GM step of the GEM algorithm

Main idea A Metropolis-Hastings algorithm is independently executed for each (k, b) in order to perform the GM step of Algorithm 3.14. For a fix (k, b), this algorithm has a stationary distribution close to $p(\boldsymbol{\alpha}_{kb}|\mathbf{x}, \mathbf{t}^{[r]})$ when it is performed at iteration $[r]$ of the global GEM algorithm. It samples a sequence of the block parameters $(\boldsymbol{\alpha}_{kb}^{[r,0]}, \ldots, \boldsymbol{\alpha}_{kb}^{[r,s_{\max}]})$ where $s_{\max}$ is the number of iterations fixed by the user. As the algorithm aims at finding the value maximizing the expectation of the complete-data log-likelihood, we put

$$\boldsymbol{\alpha}_{kb}^{[r+1]} = \underset{s=1,\ldots,s_{\max}}{\operatorname{argmax}} L_{kb}(\boldsymbol{\alpha}_{kb}^{[r,s]}; \mathbf{x}, \mathbf{t}^{[r]}). \tag{3.10}$$

We now detail the Metropolis-Hastings algorithm then we detail its instrument distribution $q(.; \boldsymbol{\alpha}_{kb}^{[r,s]})$.

Algorithm 3.15 (The Metropolis-Hastings algorithm)**.**

Starting from the initial value $\boldsymbol{\alpha}_{kb}^{[r,0]} = \boldsymbol{\alpha}_{kb}^{[r]}$, its iteration $[s]$ is written as

$$\boldsymbol{\alpha}_{kb}^{\star} \sim q(\boldsymbol{\alpha}_{kb}; \boldsymbol{\alpha}_{kb}^{[r,s]}) \tag{3.11}$$

$$\boldsymbol{\alpha}_{kb}^{[r,s+1]} = \begin{cases} \boldsymbol{\alpha}_{kb}^{\star} & \text{with probability } \lambda^{[r,s]} \\ \boldsymbol{\alpha}_{kb}^{[r,s]} & \text{with probability } 1 - \lambda^{[r,s]}. \end{cases} \tag{3.12}$$

Focus on the proposal distribution The instrumental distribution $q(\boldsymbol{\alpha}_{kb}; \boldsymbol{\alpha}_{kb}^{[r,s]})$ samples the candidate $\boldsymbol{\alpha}_{kb}^{\star}$ in two steps. Firstly, it uniformly samples the candidate $\boldsymbol{\delta}_{kb}^{\star}$ among the neighborhood of $\boldsymbol{\delta}_{kb}^{[r,s]}$ denoted by $\Delta(\boldsymbol{\delta}_{kb}^{[r,s]})$. This neighborhood is defined as the set of the parameters where at most two surjections are different from those of $\boldsymbol{\delta}_{kb}^{[r,s]}$. Secondly, it computes the continuous parameters conditionally on $\Delta(\boldsymbol{\delta}_{kb}^{[r,s]})$ as such

$$(\rho_{kb}^{\star}, \boldsymbol{\xi}_{kb}^{\star}, \boldsymbol{\tau}_{kb}^{\star}) = \underset{\rho_{kb}, \boldsymbol{\xi}_{kb}, \boldsymbol{\tau}_{kb}}{\operatorname{argmax}} L_{kb}(\rho_{kb}, \boldsymbol{\xi}_{kb}, \boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb}^{\star}; \mathbf{x}, \mathbf{t}^{[r]}). \tag{3.13}$$

Note that the maximization of the expectation of the complete-data log-likelihood stays not straightforward, even when the discrete parameters are known. However, by remarking that the block distribution is itself a mixture, we introduce a second latent variable indicating the block distribution membership (independence or maximum dependency distribution). Thus, the continuous parameters defined by the previous equation are obtained by an EM algorithm detailed in the next section.

Example 3.16 (Neighborhood of the discrete parameter)**.** *Figure 3.4 states the elements of* $\Delta(\boldsymbol{\delta}_{kb})$ *with* $d^{\{kb\}} = 2$, $m^{\{kb\}1} = 3$, $m^{\{kb\}2} = 2$, *and with* $\delta_{kb}^{121} = \delta_{kb}^{221} = \delta_{kb}^{322} = 1$ *and* $\delta_{kb}^{h2h'} = 0$ *otherwise.*

Focus on the acceptance probability In order to complete the definition of the Metropolis-Hastings algorithm, we precise the acceptance probability which is defined by

$$\lambda^{[r,s]} = \min \left\{ \frac{p(\mathbf{x}^{\{kb\}}, \mathbf{t}^{[r]}; \boldsymbol{\alpha}_{kb}^{\star})}{p(\mathbf{x}^{\{kb\}}, \mathbf{t}^{[r]}; \boldsymbol{\alpha}_{kb}^{[r,s]})} \frac{|\Delta(\boldsymbol{\delta}_{kb}^{\star})|}{|\Delta(\boldsymbol{\delta}_{kb}^{[r,s]})|}; 1 \right\}, \tag{3.14}$$

$|\Delta(\boldsymbol{\delta}_{kb}^{[r,s]})|$ denoting the cardinal of $\Delta(\boldsymbol{\delta}_{kb}^{[r,s]})$.

Remark 3.17 (Exhaustive approach *vs.* stochastic one)**.** 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.

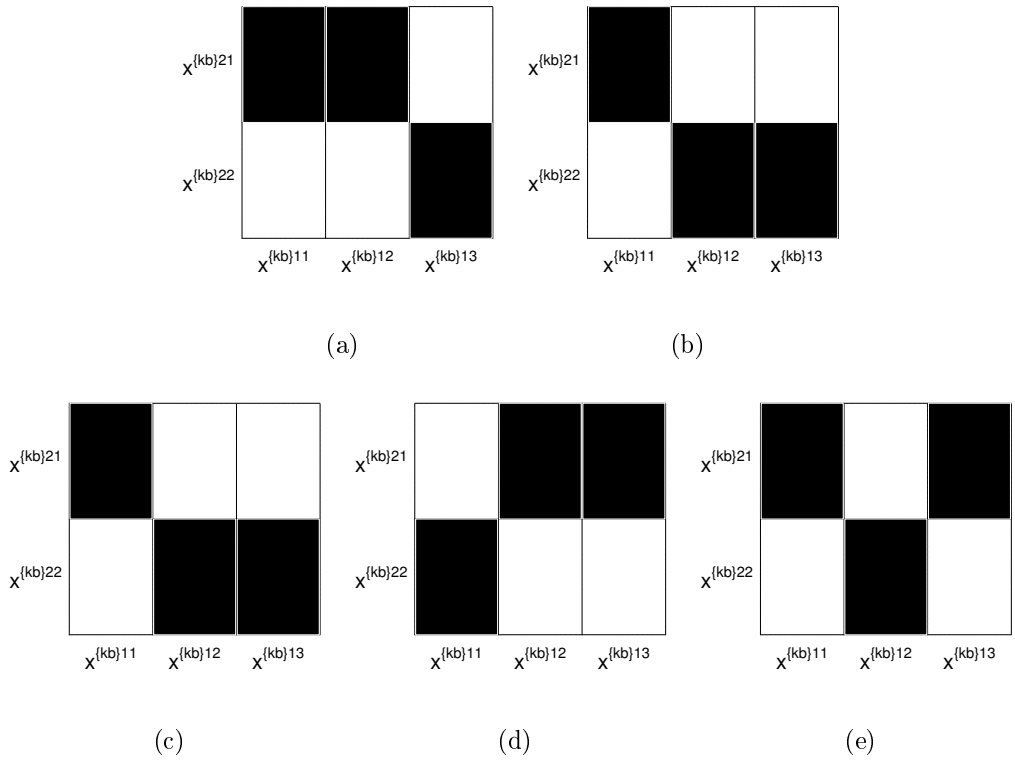


Figure 3.4 – For the row h' and the column h, a black cell indicates that $\delta_{kb}^{h2h'} = 1$ and a white cell that $\delta_{kb}^{h2h'} = 0$: (a) $\boldsymbol{\delta}_{kb}$; (b), (c), (d), (e) are the elements of $\Delta(\boldsymbol{\delta}_{kb})$.

3.4.3 Determination of $(\rho_{kb}^{\star}, \boldsymbol{\xi}_{kb}^{\star}, \boldsymbol{\tau}_{kb}^{\star})$ by the proposal distribution

A second latent variable If the first latent vector $\mathbf{z}$ indicates the class membership, a second latent vector denotes the block distribution membership. It is denoted by $\mathbf{y} = (y_i^{\{kb\}}; i = 1, \ldots, n; k = 1, \ldots, g; b = 1, \ldots, \mathrm{B}_k)$ where $y_i^{\{kb\}} = 1$ if $\boldsymbol{x}_i^{\{kb\}}$ arises from the *maximum dependency* distribution for block b of class k and $y_i^{\{kb\}} = 0$ if $\boldsymbol{x}_i^{\{kb\}}$ arises from the *independence* distribution for block b of class k.

A full complete-data log-likelihood 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

$$L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y}, \mathbf{z}) = \sum_{k=1}^{g} \mathrm{n}_k \ln \pi_k + \sum_{k=1}^{g} \sum_{b=1}^{\mathrm{B}_k} L_{kb}(\boldsymbol{\alpha}_{kb}; \mathbf{x}, \mathbf{y}, \mathbf{z}), \tag{3.15}$$

where $L_{kb}(\boldsymbol{\alpha}_{kb}; \mathbf{x}, \mathbf{y}, \mathbf{z})$ denotes the full complete-data log-likelihood of block b for component k defined by

$$L_{kb}(\boldsymbol{\alpha}_{kb}; \mathbf{x}, \mathbf{y}, \mathbf{z}) = \sum_{i=1}^{n} z_{ik} \Big((1-y_i^{\{kb\}}) \ln \big((1-\rho_{kb}) \mathring{p}(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\xi}_{kb}) \big) + y_i^{\{kb\}} \ln \big(\rho_{kb} \acute{p}(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb}) \big) \Big).$$

Conditional estimation of the continuous parameters At iteration $[s]$ of Algorithm 3.15 performed at iteration $[r]$ of Algorithm 3.14, the discrete candidate parameter $\boldsymbol{\delta}_{kb}^{\star}$ is sampled. Then, the continuous candidate parameters are defined as follows:

$$(\rho_{kb}^{\star}, \boldsymbol{\xi}_{kb}^{\star}, \boldsymbol{\tau}_{kb}^{\star}) = \underset{\rho_{kb}, \boldsymbol{\xi}_{kb}, \boldsymbol{\tau}_{kb}}{\operatorname{argmax}} L_{kb}(\rho_{kb}, \boldsymbol{\xi}_{kb}, \boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb}^{\star}; \mathbf{x}, \mathbf{t}^{[r]}).$$

So, conditionally on $(\boldsymbol{\delta}_{kb}^{\star}, \mathbf{x}, \mathbf{t}^{[r]})$, the continuous parameters are obtained by the following EM algorithm.

Algorithm 3.18 (The EM algorithm to obtain $(\rho_{kb}^{\star}, \boldsymbol{\xi}_{kb}^{\star}, \boldsymbol{\tau}_{kb}^{\star})$)**.**

Starting from an initial value $(\rho_{kb}^{[0]}, \boldsymbol{\alpha}_{kb}^{[0]}, \boldsymbol{\tau}_{kb}^{[0]})$, iteration $[\ell]$ is written as

— **E step**: calculate the conditional expectation of $y_i^{\{kb\}}$

$$u_i(\boldsymbol{\alpha}_{kb}^{[\ell]}) = \frac{\rho_{kb}^{[\ell]} \acute{p}(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\tau}_{kb}^{[\ell]}, \boldsymbol{\delta}_{kb}^{\star})}{(1-\rho_{kb}^{[\ell]}) \mathring{p}(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\xi}_{kb}^{[\ell]}) + \rho_{kb}^{[\ell]} \acute{p}(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\tau}_{kb}^{[\ell]}, \boldsymbol{\delta}_{kb}^{\star})},$$

— **M step**: maximization of the expectation of the complete-data log-likelihood

$$\rho_{kb}^{[\ell+1]} = \frac{n_{kb}^{[\ell]}}{n_k^{[r]}}, \ \xi_{kb}^{jh[\ell+1]} = \frac{\mathring{n}_{kb}^{jh[\ell]}}{n_k^{[r]} - n_{kb}^{[\ell]}} \text{ and } \tau_{kb}^{h[\ell+1]} = \frac{\acute{n}_{kb}^{h[\ell]}}{n_{kb}^{[\ell]}},$$

with $n_{kb}^{[\ell]} = \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]}) u_i(\boldsymbol{\alpha}_{kb}^{[\ell]})$, $\acute{n}_{kb}^{h[\ell]} = \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]}) u_i(\boldsymbol{\alpha}_{kb}^{[\ell]}) x_i^{\{kb\}1h}$ and $\mathring{n}_{kb}^{jh[\ell]} = \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]})(1-u_i(\boldsymbol{\alpha}_{kb}^{[\ell]})) x_i^{\{kb\}jh}$.

Conjecture 3.19 (One optimum)**.** *During our experiments, we empirically noticed that the log-likelihood function of the mixture between the independence and the maximum dependency distributions had a unique optimum. We conjecture that this function has indeed a unique maximum.*

Remark 3.20 (One EM algorithm if the discrete parameters are known)**.** In the specific case where $\boldsymbol{\delta}_{kb}$ are known for each (k, b), all the continuous parameters could be estimated via a unique EM algorithm. At iteration $[r]$ of this algorithm, the E step would compute both expectations of $\mathbf{z}^{[r]}$ and $\mathbf{y}^{[r]}$ while the M step would estimate all the continuous parameters maximizing the expectation of the *full* complete-data log-likelihood.

3.5 Model selection via a MCMC algorithm

Aim The aim is to select the model defined by $(\hat{g}, \hat{\boldsymbol{\sigma}})$ which better fit the data. Thus, in a Bayesian framework, the aim is to find the model having the largest posterior probability.

Prior distributions We consider 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 we assume that $p(\boldsymbol{\sigma}|g)$ follows a uniform distribution.

Posterior distributions The best model maximizes its posterior distribution. According to the prior distributions, it is defined as

$$(\hat{g}, \hat{\boldsymbol{\sigma}}) = \underset{g}{\operatorname{argmax}} \left[\underset{\boldsymbol{\sigma}}{\operatorname{argmax}}\, p(\mathbf{x}|g, \boldsymbol{\sigma})\right] \text{ where } p(\mathbf{x}|g, \boldsymbol{\sigma}) \propto \int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} p(\mathbf{x}|\boldsymbol{\theta}, g, \boldsymbol{\sigma}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{3.16}$$

To find $(\hat{g}, \hat{\boldsymbol{\sigma}})$, a MCMC algorithm is used for estimating $\operatorname{argmax}_{\boldsymbol{\sigma}}\, p(\mathbf{x}|g, \boldsymbol{\sigma})$, for each value of $g \in \{1, \ldots, g_{max}\}$. This method limits the combinatorial problem involved by the detection of the block structure of variables since it provides a random walk among the $\boldsymbol{\sigma}$ of interest.

Remark 3.21 (On the reversible jump)**.** A reversible jump method could be used [RG97], 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. So, we propose to use an easier MCMC algorithm having $p(\boldsymbol{\sigma}|\mathbf{x}, g)$ as stationary distribution.

3.5.1 Exploration of the space of the models by a MCMC algorithm

Main idea This algorithm alternates between two steps: the generation of a neighborhood conditionally on the current model by a proposal distribution and the generation of a new model belonging to this neighborhood according to its posterior probability.

Algorithm 3.22 (MCMC algorithm to explore the models)**.**

This MCMC algorithm has $p(\boldsymbol{\sigma}|\mathbf{x}, g)$ as stationary distribution. Starting from an initial value of the repartition of the variables into blocks $\boldsymbol{\sigma}^{[0]}$, its iteration $[q]$ is written as

— **Neighborhood step**: sampling of a stochastic neighborhood $\Sigma^{[q]}$

$$\Sigma^{[q]} \sim q(\Sigma; \boldsymbol{\sigma}^{[q]}). \tag{3.17}$$

— **Model step**: sampling of the repartition of the variables into blocks $\boldsymbol{\sigma}^{[q+1]}$

$$\boldsymbol{\sigma}^{[q+1]} = p(\boldsymbol{\sigma}|\mathbf{x}, g, \Sigma^{[q]}) \text{ where } p(\boldsymbol{\sigma}|\mathbf{x}, g, \Sigma^{[q]}) \propto \begin{cases} p(\mathbf{x}|g, \boldsymbol{\sigma}) & \text{if } \boldsymbol{\sigma} \in \Sigma^{[q]} \\ 0 & \text{otherwise.} \end{cases} \tag{3.18}$$

We now detail both steps of the above MCMC algorithm.

Details of the Neighborhood step A 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 (possibility to build a new block):

$$\left\{\boldsymbol{\sigma} : \exists!(k, b, j)\ j \in \boldsymbol{\sigma}_{kb}^{[q]} \text{ and } j \notin \boldsymbol{\sigma}_{kb}\right\} \cup \left\{\boldsymbol{\sigma}^{[q]}\right\}.$$

However, as this deterministic neighborhood can be very large, our proposal distribution allows reducing it to a stochastic neighborhood $\Sigma^{[q]}$ by limiting the number of (k, b) where $\boldsymbol{\sigma}_{kb}$ could be different to $\boldsymbol{\sigma}_{kb}^{[q]}$. Thus, the sampling according to $q(.; \boldsymbol{\sigma}^{[q]})$ is performed by the three following steps:

— **Component sampling**

$$k^{[q]} \sim \mathcal{U}[\{1, \ldots, g\}].$$

— **Leaving block sampling**

$$b_{from}^{[q]} \sim \mathcal{U}[\{1, \ldots, B_{k^{[q]}}^{[q]}\}].$$

— **Arriving block sampling**

$$b_{to}^{[q]} = \{b^{[q]}, B_{k^{[q]}}^{[q]} + 1\} \text{ where } b^{[q]} \sim \mathcal{U}[\{1, \ldots, B_{k^{[q]}}^{[q]}\} \setminus b_{from}^{[q]}].$$

The stochastic neighborhood $\Sigma^{[q]}$ is then defined as:

$$\Sigma^{[q]} = \left\{\boldsymbol{\sigma} : \exists!(k, b, j)\ j \in \boldsymbol{\sigma}_{kb}^{[q]},\ j \notin \boldsymbol{\sigma}_{kb} \text{ and } j \in \boldsymbol{\sigma}_{kb'} \text{ with } k = k^{[q]},\ b = b_{\text{from}}^{[q]},\ b' \in b_{\text{to}}^{[q]}\right\} \cup \left\{\boldsymbol{\sigma}^{[q]}\right\}. \tag{3.19}$$

We denote by $\boldsymbol{\sigma}^{[q+\varepsilon(\mathfrak{e})]}$ the elements of $\Sigma^{[q]}$ where $\varepsilon(\mathfrak{e}) = \frac{\mathfrak{e}}{|\Sigma^{[q]}|+1}$ and $\mathfrak{e} = 1, \ldots, |\Sigma^{[q]}|$.

Example 3.23 (neighborhood $\Sigma^{[q]}$)**.** *Figure 3.5 shows an illustration of this definition of the neighborhood* $\Sigma^{[q]}$ *when* $\boldsymbol{\sigma}_k^{[q]} = (\{1, 2\}, \{3, 4\})$.

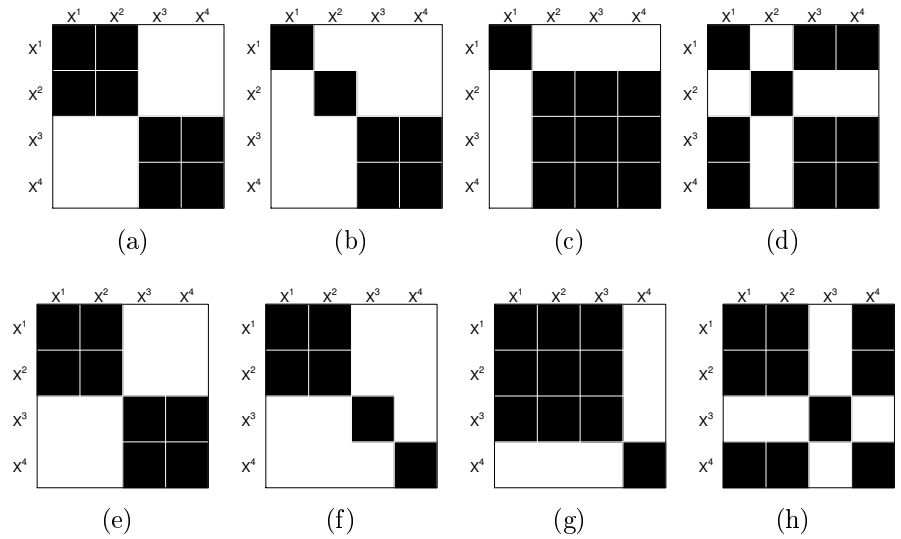


Figure 3.5 – Example of the support of $\Sigma^{[q]}$ in the case of four variables. If the variables of the j^{th} row and of the j'^{th} column are in the same block then the cell (j, j') is painted in black. This cell is painted in white otherwise. (a—d) Elements of $\Sigma^{[q]}$ if $b^{[q]}_{from} = 1$; (e—h) Elements of $\Sigma^{[q]}$ if $b^{[q]}_{from} = 2$.

Details of the Pattern step At the generation pattern step, the algorithm needs the value of $p(\mathbf{x}|g, \boldsymbol{\sigma})\ \forall \boldsymbol{\sigma} \in \Sigma^{[q]}$ to implement Algorithm 3.22. By using the BIC approximation, this probability is approximated by

$$\ln p(\mathbf{x}|g, \boldsymbol{\sigma}) \simeq L(\hat{\boldsymbol{\theta}}; \mathbf{x}, g, \boldsymbol{\sigma}) - \frac{\nu_{\text{MEDD}}}{2} \log(n), \tag{3.20}$$

$\hat{\boldsymbol{\theta}}$ being the maximum likelihood estimator obtained by the GEM algorithm previously described in Section 3.4. Thus, at iteration $[q]$, for each $\mathfrak{e} = 1, \ldots, |\Sigma^{[q]}|$, the estimator $\hat{\boldsymbol{\theta}}^{[q+\varepsilon(\mathfrak{e})]}$ associated to the element $\boldsymbol{\sigma}^{[q+\varepsilon(\mathfrak{e})]}$ is computed by Algorithm 3.14.

Initialization Whatever the initial value selected for $\boldsymbol{\sigma}^{[0]}$, the algorithm converges to the same stationary distribution. However, this convergence can be very slow when the initialization is poor. Since blocks consist in the most correlated variables, a Hierarchical Ascendant Classification (HAC) is applied on the matrix of Cramer's V distances on the couples of variables. We select, for $\boldsymbol{\sigma}^{[0]}_k$, the partition provided by the HAC which minimizes the number of blocks and which excludes the blocks consisting of more than four variables. Note that the number of the variables affected into a block is limited to four, for the initialization, because very few blocks having more than four variables were observed during our experiments. Obviously, the MCMC algorithm can then violate this initial constraint if necessary.

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

3.5.2 Consequences of the model selection on the GEM algorithm

Main idea At iteration $[q]$ of the MCMC algorithm performing the model selection (*i.e.* Algorithm 3.22), the GEM algorithm (*i.e.* Algorithm 3.14) estimates $\hat{\boldsymbol{\theta}}^{[q+\varepsilon(\mathfrak{e})]}$ associated to the model $\boldsymbol{\sigma}^{[q+\varepsilon(\mathfrak{e})]}$ for $\mathfrak{e} = 1, \ldots, |\Sigma^{[q]}|$. Since these models are close to $\boldsymbol{\sigma}^{[q]}$, their maximum likelihood estimates should be closed to $\hat{\boldsymbol{\theta}}^{[q]}$.

Parameters of the non-modified blocks The GEM algorithm initialization is also done by the value of $\hat{\boldsymbol{\theta}}^{[q]}$ for the non modified blocks. Thus, in such a case $(\boldsymbol{\sigma}_{kb}^{[q+\varepsilon(\mathfrak{e})]} = \boldsymbol{\sigma}_{kb}^{[q]})$, $\boldsymbol{\theta}_{kb}^{[q+\varepsilon(\mathfrak{e})][0]} = \hat{\boldsymbol{\theta}}_{kb}^{[q]}$.

Parameters of the modified blocks For the other blocks, the continuous parameters are randomly sampled. In order to avoid the combinatorial problems, we use a sequential method to initialize $\boldsymbol{\delta}_{kb}^{[q+\varepsilon(\mathfrak{e})][0]}$. The surjections from $\boldsymbol{x}_i^{\{kb\}1}$ to $\boldsymbol{x}_i^{\{kb\}j}$ are sampled, according to $\mathbf{x}$ and the continuous parameters previously sampled $(\rho_{kb}^{[q+\varepsilon(\mathfrak{e})][0]}, \boldsymbol{\alpha}_{kb}^{[q+\varepsilon(\mathfrak{e})][0]}, \boldsymbol{\tau}_{kb}^{[q+\varepsilon(\mathfrak{e})][0]})$, for each $j = 2, \ldots, d^{\{kb\}}$ as follows:

$$\boldsymbol{\delta}_{kb}^{\cdot j[q+\varepsilon(\mathfrak{e})][0]} \propto \prod_{i=1}^{n} p(x_i^{\{kb\}1}, x_i^{\{kb\}j}; \rho_{kb}^{[q+\varepsilon(\mathfrak{e})][0]}, \boldsymbol{\alpha}_{kb}^{1[q+\varepsilon(\mathfrak{e})][0]}, \boldsymbol{\alpha}_{kb}^{j[q+\varepsilon(\mathfrak{e})][0]}, \boldsymbol{\tau}_{kb}^{[q+\varepsilon(\mathfrak{e})][0]}, \boldsymbol{\delta}_{kb}^{\cdot j})^{z_{ik}^{[q]}}, \tag{3.21}$$

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

Remark 3.24 (About the number of iterations of the GEM algorithm $r_{\max}$)**.** As said in Section 3.4.1, 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. As these algorithms are interlocked, the number of iterations of Algorithm 3.18 (the most internal algorithm) is small. When the best model is selected by Algorithm 3.22, this latter will stay in this model during many iterations so the Metropolis-Hastings (Algorithm 3.15) and the EM algorithm (Algorithm 3.18) are performed lots of times. Thus, it is not necessary to have a large number of iterations as stopping criterion.

3.6 Numerical experiments on simulated data sets

Table 3.1 presents the adjustment parameters values used for all the simulations.

Algorithms	MCMC	GEM	Metropolis-Hastings	EM
Criteria	$q_{\max} = 20 \times d$	$r_{\max} = 10$	$s_{\max} = 1$	$t_{\max} = 5$

Table 3.1 – Values of the different stopping criteria.

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

Aim In this section, we illustrate the performance of the Metropolis-Hastings algorithm estimating $\boldsymbol{\delta}_{kb}$ (see Section 3.4.2) and the relevance of its initialization defined by (3.21). Since this algorithm is interlocked in the MCMC algorithm and in the GEM algorithm which respectively estimate the model and the parameters, we need it to converge quickly. It is shown in the following simulations that the algorithm stays relevant up to six modalities per variable and up to six variables per block. These conditions hold in most situations.

Experimental conditions Samples of size 200 described by variables having the same number of modalities are generated by a mixture between an independence distribution and a maximum dependency distribution. The parameters are also estimated by the Metropolis-Hastings algorithm, described in Section 3.4.2, since only one class is generated. The initializations of the discrete parameters are performed according to Equation (3.21) with $z_{i1} = 1$ for all $i = 1, \ldots, 200$.

Results Figure 3.6 shows the box-plots of the numbers of iterations required by the Metropolis-Hastings algorithm in order to find the true links between modalities maximizing the likelihood[2]. According to these simulations, one observes that the results of this algorithm are good thanks to its initialization which allows significantly reducing the number of iterations needed in order to find the maximum likelihood estimators.

3.6.2 Study of the algorithm for model selection

Aim 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.

Experimental conditions In many situations, 100 samples are generated according to the MEDD model with two components. Note that the parameter u is introduced for controlling the overlapping of classes: when it is close to one then the classes are absolutely overlapped. This parameter fix the error rate to 0.10 for each studied situation:

$$\boldsymbol{\sigma}_{kb} = (d/b, 1 + d/b) \quad \rho_{kb} = 0.6(1 - u) \quad \boldsymbol{\tau}_{kb} = (0.60, 0.20, 0.20),$$

$$\delta_{1b}^{h2h'} = 1 \text{ iff } h = h' \quad \delta_{1b}^{122} = \delta_{1b}^{223} = \delta_{1b}^{321} = 1 \quad \boldsymbol{\alpha}_{1b}^{j} = (0.20, 0.20, 0.60),$$

$$\boldsymbol{\alpha}_{2b}^{1} = \boldsymbol{\alpha}_{1b}^{1}(1 - u) + (0.075, 0.850, 0.075)u \quad \text{and} \quad \boldsymbol{\alpha}_{2b}^{2} = \boldsymbol{\alpha}_{1b}^{2}(1 - u) + (0.850, 0.075, 0.075)u.$$

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

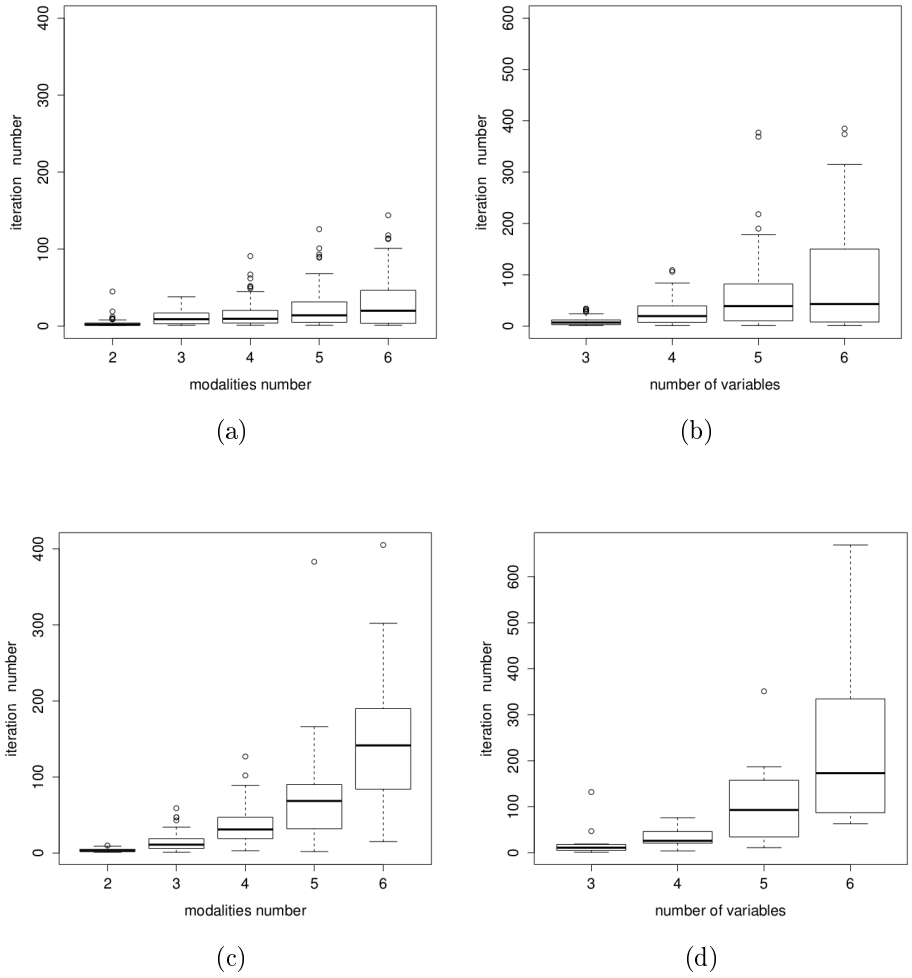


Figure 3.6 – Box-plots of the number of iterations required by the Metropolis-Hastings algorithm in order to find the best links between modalities, according to the number of modalities when datasets are simulated with a proportion of maximum dependency distribution equal to 0.5. (a) Three variables 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.

Results Table 3.2 shows the mean and the standard deviation of the Kullback-Leibler divergence between the parameters used for the data set 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.

$d \setminus n$	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 3.2 – **mean** (*standard deviation*) of the Kullback-Leibler divergence.

3.7 Analysis of two real data sets

3.7.1 Contraceptive method choice

The data This data set is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey ([LLS00]). It is composed with 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 the analysis, the contraceptive method used is blinded, in order to work in a clustering context.

Model selection Table 3.3 presents the values of the BIC criterion for the CIM and the MEDD models. Until four classes, the results of the MEDD model are better than them of the CIM model. The selection of class number is better for the MEDD model since it selects the "true" number of classes while the CIM model overestimates it.

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

Table 3.3 – Values of the BIC criterion obtained by both models with different numbers of classes. Best values according to the BIC criterion are in bold.

Model interpretation Figure 3.7 summarized the results of the best MEDD model according to the BIC criterion. It allows to describe the classes by their main features (proportions, intra-class correlations). On ordinates, the estimated classes are represented with respect to their proportion in decreasing order. Note that their corresponding area depends on their proportion. The cumulated proportions are

indicated on the left side. On abscissa, 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 the strength of their dependencies (in decreasing order). The third is the variables repartition per blocks. A black cell indicates that the variable is assigned to the block and a white cell indicates that, conditionally on this class, the variable is independent of the variables of this block. For example, this figure shows that the first class has a proportion of 0.49 and that all the variables are split into three blocks.

- **Class 1: young families**
  - **General:** this class proportion is equal to 0.49. There are two dependency 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 dependency 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 children they have ($\boldsymbol{\delta}_{kb}$).
  - **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 ($\boldsymbol{\alpha}_{kb}$).

Conclusion It is noted that the MEDD model is more relevant for this data set. Indeed, the number of classes is limited and they are interpretable. In addition, the assumption of conditional independence between variables seems too stringent 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.

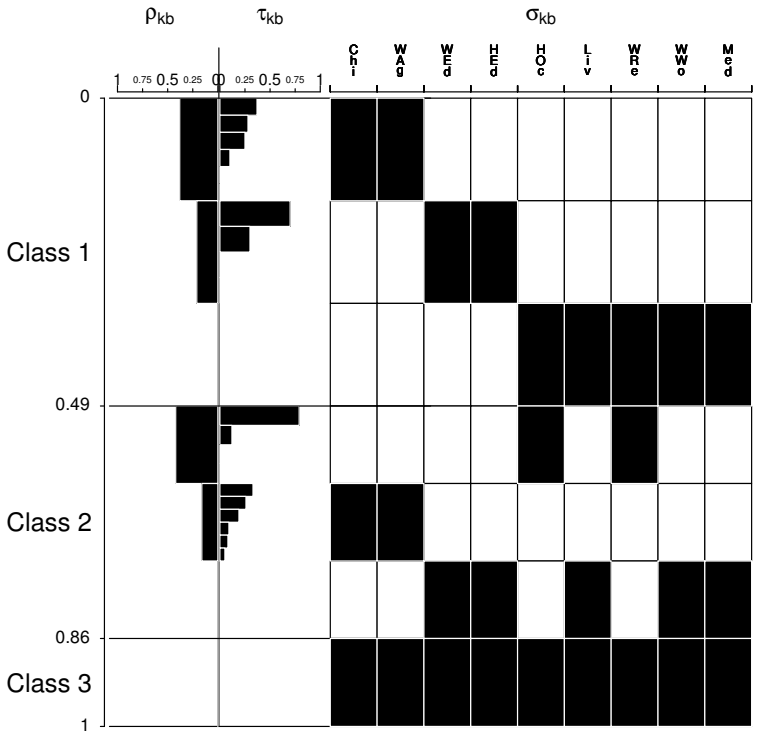


Figure 3.7 – Summary of the best MEDD model according to the BIC criterion for the contraceptive method choice data set.

3.7.2 Calves clustering

The data The "Genes Diffusion" company has collected information 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 health related (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*).

Information criteria Table 3.4 displays the BIC criterion values and the number of parameters required by the CIM and the MEDD models. Furthermore, the computing time in minutes (obtained with a processor Intel Core i5-3320M) to estimate the MEDD model by starting 20 MCMC chains with a stopping criterion of $q_{\max} = 180$ while the CIM model needs 3 sec with the R package RMixmod [LIL+12].

For the CIM model, the BIC criterion selects a high number of classes, since it selected eight classes. The interpretation of the clusters is also difficult and we can assume that the quality of the estimate is very poor. Figure 3.8 helps the interpretation for the MEDD model with five components (best model according to the BIC criterion). Its interpretation is the same as the interpretation of Figure 5.1. For example, this figure shows that the first class has a proportion of 0.29 and it is composed of four blocks. The most correlated block of the first class has $\rho_{kb} \simeq 0.80$

	g	1	2	3	4	5	6	7	8
CIM	BIC	-28589	-26859	-26526	-26333	-26238	-26235	-26226	**-26185**
	ν_{CIM}	17	35	53	71	89	107	125	**143**
MEDD	BIC	-26653	-26289	-26173	-26038	**-26025**	-26059	-26045	-26058
	ν_{MEDD}	24	48	80	89	**112**	131	148	163
	time (min)	0.97	3.32	6.16	6.56	**10.03**	11.76	12.31	14.92

Table 3.4 – Results for the CIM and the MEDD models according to different numbers of classes. For both models, first row corresponds to the BIC criterion values and the second row indicates the number of continuous parameters. Best results according to the BIC criterion are in bold. Computing time for the MEDD model estimation is given in minutes.

and the strength of the biggest modalities link is close to 0.85. This block consists in the variables *TDC* and *TRM*.

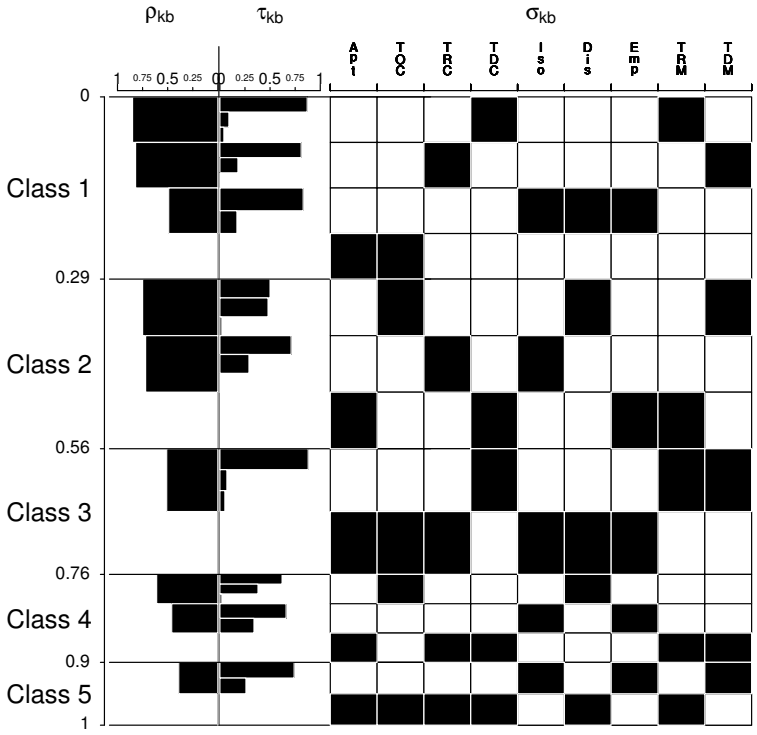


Figure 3.8 – Summary of the best MEDD model according to BIC criterion for the calves data set.

Interpretation of Class 1 Here is now a possible interpretation of Class 1 (note that the others classes are also meaningful; see details in [MBV13b]):

— **General:** this class has a proportion equal to 0.29 and consists of three blocks of dependency and one block of independence.

— **Block 1:** there is a strong correlation (ρ_{11}) 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 ($\boldsymbol{\tau}_{11}$ and $\boldsymbol{\delta}_{11}$).

— **Block 2:** there is a strong correlation (ρ_{12}) 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 ($\boldsymbol{\tau}_{12}$ and $\boldsymbol{\delta}_{12}$).

— **Block 3:** there exists another strong link between the behavior of the mother, the emptying of the umbilical and its disinfection ($\boldsymbol{\tau}_{13}$ and $\boldsymbol{\delta}_{13}$).

— **Block 4:** this block is characterized by an absence of preventive treatment against omphalite and contains 50% of the calves infected by this illness ($\boldsymbol{\alpha}_{14}$).

3.8 Conclusion

By using the block extension of the CIM model, a new mixture model called the MEDD model has been proposed to cluster categorical data by taking into account the intra-class dependency. The block distribution of the MEDD model is defined as a mixture between an independence distribution and a maximum dependency distribution. This specific distribution stays parsimonious and allows different levels of interpretation. The first level is given by the blocks of variables which bring out the conditional dependencies between variables, and by the proportions of the maximum dependency distributions which characterize the strength of these dependencies. The second level is more precise since the parameters of the block distribution reflect the links between modalities and their strengths. The MEDD model has been compared to the full latent class model on two real data sets.

The parameter and the model are simultaneously estimated via a MCMC algorithm. This algorithm allows to reduce the combinatorial problems of the block structure detection and the links between modalities search for the estimation of the maximum dependency distribution. The results are good when the number of modalities is small for each variable. For more than six modalities, the detection of other links meets some persistent difficulties. So, the algorithm can be slow in this case. The proposed approach to estimate the block structure is not adapted for data sets with lots of variables.

The main drawback of this algorithm is its need to compute the MLE associated to each candidate model. This estimation is time consuming and only the MLE associated to the best model is interpreted. Thus, we propose in the next chapter a new mixture model avoiding this drawback since its integrated complete-data likelihood is explicit. This properties avoids the needs to use the MLE to perform the model selection.

Finally, the proposed model can be easily extended to the case of ordinal data. For this, some additional constraints on the dependency structure of each distribution of maximum dependency need to be added. Note that these constraints also limit the combinatorial research of the dependency structures.

Chapter 4

Model-based clustering with conditional dependency modes

This chapter presents our second contribution to the model-based framework permitting to cluster categorical data. This contribution consists in a mixture model which groups the variables into conditionally independent blocks. Each block follows a parsimonious multinomial distribution where the few free parameters correspond to its modes.
The inference is easily performed via an EM *algorithm while the challenge of the model selection is facilitated by an efficient approximation of the integrated complete-data likelihood.*
Numerical experiments, on simulated and real data sets, underline the main characteristics of this new mixture model.

Karma police, arrest this man
He talks in maths
He buzzes like a fridge
He's like a detuned radio
Radiohead — Karma Police

4.1 Introduction

In this chapter, we present a sparse mixture model which relaxes the conditional independence assumption in order to overcome the biases caused by the latent class model and which overcomes the main drawback of the MEDD model. Indeed, the model selection can be easily and efficiently performed by avoiding the combinatorial problems. This step does not require the MLE since the integrated complete-data likelihood can be precisely approached. Firstly, the model selection can be performed by a MCMC algorithm where the MLE is not required. Secondly, the parameters are only estimated for the best model which significantly limits the computation time.

This new model, named *Conditional Modes Model* (referred in this article by CMM), groups the variables into *conditionally independent blocks*, for considering the main conditional dependencies. Moreover, the specific distribution of the block is a multinomial distribution per modes. This distribution assumes that few modality crossings, named *modes*, are characteristic and that the other ones follow a uniform distribution. Thus, the associated multinomial distribution is parsimonious, since its free parameters are limited to the few parameters of the modes.

This simple mixture model (CMM) is a good challenger. On the one hand, the CMM model challenges the mixture model with conditional independence assumption (CIM), since it avoids many biases through modelizing of the main conditional dependencies. On the other hand, it challenges the mixture models relaxing this assumption since its flexible distribution of the block requires few parameters. Note that, as the MEDD model, the CMM model can be interpreted as a parsimonious version of the log-linear mixture model. Indeed, the repartition of the variables into blocks defines the considered interactions while the distribution per modes into blocks defines a specific distribution for each interaction. Furthermore, resulting classes are meaningful since the intra-class dependencies are brought out by two complementary levels: the block variable interaction level and the associated mode interaction level (through locations and probabilities). Note that the CMM model is a comprehensive approach since it includes the CIM model and a part of its parsimonious versions presented in Section 2.3.1.

For a fixed model (number of classes, repartition of the variables into blocks and numbers of modes), the maximum likelihood estimate is obtained via an EM algorithm. The model is selected via a Metropolis-within-Gibbs algorithm. Indeed, this algorithm is a Gibbs sampler which generates a new repartition of the variables into blocks and a new number of modes by one Metropolis-Hastings step. It is performed for a fixed number of classes and avoids combinatorial problems involved by the selection of the blocks of variables and by the estimation of the numbers of modes. This algorithm is based on the fact that the integrated complete-data likelihood, required for the acceptance probability computation of the Metropolis-Hastings inside the Gibbs sampler, is accessible and non ambiguous through weakly informative conjugate prior. Finally, this approach has two main advantages. It permits to reduce the bias of the BIC-like approach. Let us mention that the overestimation of the number of modes by this approach is illustrated during our numerical experiments. Furthermore, it allows us to perform an efficient model selection in a reasonable computational time since the parameters are only estimated for the unique selected model. Thus, this approach is a possible answer to the combinatorial model selection problem which is known to be a real challenge for a log-linear mixture model.

Structure of this chapter This paper is organized as follows. Section 4.2 presents the Conditional Modes Model. Section 4.3 is devoted to maximum likelihood estimation via an EM algorithm. Section 4.4 presents the Metropolis-within-Gibbs sampler performing the model selection through the integrated complete-data likelihood. In Section 4.5, we show that the proposed approach for computing the integrated complete-data likelihood reduces the biases of the BIC-like approach. More-

over, we numerically emphasize the good behavior of the Metropolis-within-Gibbs sampler and the flexibility of the CMM model on simulated data. Section 4.6 presents two cluster analysis of biological data sets performed by the R package `CoModes`[1]. A conclusion is drawn and future extensions are discussed in Section 4.7. All these results are part of the article *Finite mixture model of conditional dependencies modes to cluster categorical data* [MBV14a].

4.2 Mixture model of multinomial distributions per modes

Main idea The proposed model, referred as Conditional Modes Model (CMM), assumes that data arise independently from a mixture of g components of *conditionally independent blocks*, where the repartition of the variables into *blocks* is *equal between classes*. Each block follows a *multinomial distribution per modes* which is a multinomial distribution having few free parameters corresponding to the *modes* of the distribution. More precisely, the modes are defined as the locations of the largest probabilities, while the other parameters are equal.

4.2.1 Conditionally independent blocks equal between classes

The repartition of the variables is assumed to be equal between classes, so we use the notations of the mixture model of conditionally independent blocks defined in Section 3.2 by omitting the indexation on k.

A partition of the variables equals between classes The repartition of the d categorical variables $\boldsymbol{x}_i = (\boldsymbol{x}_i^1, \ldots, \boldsymbol{x}_i^d)$ into B blocks determines a partition $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_B)$ of $\{1, \ldots, d\}$ in B disjoint non-empty subsets. This partition defines new univariate categorical variables $\boldsymbol{x}_i^{\{b\}} = \boldsymbol{x}_i^{\boldsymbol{\sigma}_b} = (x_i^{\{b\}h}; h = 1, \ldots, m^{\{b\}})$ obtained by the concatenation of the subset of $\boldsymbol{x}_i$ associated to $\boldsymbol{\sigma}_b$ where $m^{\{b\}} = \prod_{j \in \boldsymbol{\sigma}_b} m_j$ is the number of the modality crossings into block b. The variable $\boldsymbol{x}_i^{\{b\}}$ uses a disjunctive coding since $x_i^{\{b\}h} = 1$ if individual i takes modality h for the new categorical variable (*i.e.* the modality crossing h of the initial variables affected to the block b) and $x_i^{\{b\}h} = 0$ otherwise.

Triplet building a model A specific model is defined by the number of components, the repartition of the variables into blocks and the number of modes for each multinomial distribution. So, it is defined by the triplet $\boldsymbol{\omega} = (g, \boldsymbol{\sigma}, \boldsymbol{\ell})$ where $\boldsymbol{\ell} = (\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_g)$ groups all the numbers of modes with $\boldsymbol{\ell}_k = (\ell_{k1}, \ldots, \ell_{kB})$ and where ℓ_{kb} is the number of modes of $\boldsymbol{x}_i^{\{b\}}$ for class k (with $0 < \ell_{kb} < m^{\{b\}}$).

Definition 4.1 (Mixture of conditionally independent blocks equal between classes)**.** The categorical variable $\boldsymbol{x}_i$ is drawn by a CMM model defined by $\boldsymbol{\omega}$ and parametrized

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

by $\boldsymbol{\theta}$ if its pdf is written as

$$p(\boldsymbol{x}_i;\boldsymbol{\theta},\boldsymbol{\omega}) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}_i;\boldsymbol{\alpha}_k,\boldsymbol{\sigma},\boldsymbol{\ell}_k) \text{ with } p(\boldsymbol{x}_i;\boldsymbol{\alpha}_k,\boldsymbol{\sigma},\boldsymbol{\ell}_k) = \prod_{b=1}^{\text{B}} p(\boldsymbol{x}_i^{\{b\}};\boldsymbol{\alpha}_{kb},\ell_{kb}), \tag{4.1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi},\boldsymbol{\alpha})$ denotes the whole mixture parameters, where $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_g)$ is the vector of class proportions with $0 < \pi_k \leq 1$ and $\sum_{k=1}^{g}\pi_k = 1$, and where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_g)$ is the vector which groups the parameters of the multinomial distributions per modes with $\boldsymbol{\alpha}_k = (\boldsymbol{\alpha}_{k1},\ldots,\boldsymbol{\alpha}_{k\text{B}})$.

4.2.2 Multinomial distribution per modes

We now specify the multinomial distribution per modes. So, let us introduce its parameter space before to define its pdf.

Definition 4.2 (Parameter space of a multinomial distribution per modes)**.** Let $\boldsymbol{\alpha}_{kb} = (\alpha_{kb}^h; h = 1,\ldots,m^{\{b\}})$ be the vector of size $m^{\{b\}}$ and let τ_{kb} be the mapping from $\{1,\ldots,m^{\{b\}}\}$ to $\{1,\ldots,m^{\{b\}}\}$ ordering the elements of $\boldsymbol{\alpha}_{kb}$ by decreasing values. If $\boldsymbol{\alpha}_{kb}$ denotes the parameters of the multinomial distribution per ℓ_{kb} modes, then it is defined in the constrained simplex $S(\ell_{kb}, m^{\{b\}})$ defined as follows

$$S(\ell_{kb}, m^{\{b\}}) = \left\{ \boldsymbol{\alpha}_{kb} :\ 0 \leq \alpha_{kb}^h \leq 1,\ \sum_{h=1}^{m^{\{b\}}} \alpha_{kb}^h = 1,\ \alpha_{kb}^{(\ell_{kj}+1)} = \ldots = \alpha_{kb}^{(m^{\{b\}})} \right\}. \tag{4.2}$$

where we use the shorter notation $\alpha_{kb}^{(h)} = \alpha_{kb}^{\tau_{kj}(h)}$, so $\alpha_{kb}^{(h)} \geq \alpha_{kb}^{(h+1)}$ $(1 \leq h < m^{\{b\}})$.

Definition 4.3 (Multinomial distribution per modes)**.** The univariate categorical data $\boldsymbol{x}_i^{\{b\}}$ has $m^{\{b\}}$ modalities and follows a multivariate distribution per ℓ_{kb} modes. Its pdf is also written as

$$p(\boldsymbol{x}_i^{\{b\}};\boldsymbol{\alpha}_{kb},\ell_{kb}) = \prod_{h=1}^{m^{\{b\}}} \left(\alpha_{kb}^h\right)^{x_i^{\{b\}h}}, \tag{4.3}$$

where $\boldsymbol{\alpha}_{kb} = (\alpha_{kb}^h; h = 1,\ldots,m^{\{b\}}) \in S(\ell_{kb}, m^{\{b\}})$ and α_{kb}^h is the probability that individual i takes modality h of the concatenated categorical variable $\boldsymbol{x}_i^{\{b\}}$.

4.2.3 Mixture model of conditional modes

Definition 4.4 (Mixture model of conditional modes)**.** The categorical variable $\boldsymbol{x}_i$ is drawn by a CMM model defined by $\boldsymbol{\omega}$ and parametrized by $\boldsymbol{\theta}$ if its pdf is given by

$$p(\boldsymbol{x}_i;\boldsymbol{\theta},\boldsymbol{\omega}) = \sum_{k=1}^{g} \pi_k \prod_{b=1}^{\text{B}} \prod_{h=1}^{m^{\{b\}}} \left(\alpha_{kb}^h\right)^{x_i^{\{b\}h}}. \tag{4.4}$$

Remark 4.5. The CIM model is included in the CCM model, since the conditional independence assumption between the initial variables is defined by putting one variable per block (so $d = \text{B}$ and $\boldsymbol{\sigma} = (\{1\},\ldots,\{\text{B}\})$) and by fixing the number of modes as the number of modalities of the variables minus one ($\ell_{kj} = m_j - 1$).

4.2.4 Properties of the mixture model per conditional modes

Two levels of interpretation The CMM model has two levels of interpretation. Firstly, the intra-class dependencies of variables (equal between classes) are emphasized by the repartition of the variables into blocks given by $\boldsymbol{\sigma}$. Secondly, the intra-class and intra-block dependencies of modalities (possibly different between classes) are summarized by the modes (locations and probabilities).

Two compact terms A shorter summary for each distribution is also available by using the following compact terms κ_{kb} and ρ_{kb} defined on $[0, 1]$ by

$$\kappa_{kb} = \frac{\ell_{kb}}{m^{\{b\}} - 1} \text{ and } \rho_{kb} = \sum_{h=1}^{\ell_{kb}} \alpha_{kb}^{(h)}. \tag{4.5}$$

They reflect respectively the *complexity* and the *strength* of the intra-class and intra-block dependencies. For instance, the smaller is κ_{kb} and the larger is ρ_{kb}, the more massed in few characteristic modality crossings is the distribution. Indeed, the modes are interpreted as an over-contribution at the uniform distribution among all the modality crossings.

Identifibailiy Note that the repartition of the variables guarantees the model generic identifiability since it is equal between classes. Indeed, with this constraint, the results of [AMR09] can be applied to prove the generic identifiability of the CMM model (details are given in Appendix A.2). Despite the constraint to have of the same repartition of the variables into blocks for all the classes, the model stays flexible because of the specific block distribution.

Number of parameters The main idea of the former parsimonious versions of the CIM model (Conditional Independence Model) proposed in [CG91] is to consider only one mode for each multinomial distribution of the initial variable (see Section 2.3.1). Different constraints of equality are then added between the variables and/or classes. In fact, many of these models are included into the model family of CMM by putting $\text{B} = d$ and $\ell_{kb} = 1$. In addition, the CMM models need ν_{CMM} parameters defined by

$$\nu_{\text{CMM}} = (g - 1) + \sum_{k=1}^{g} \sum_{b=1}^{\text{B}} \ell_{kb}. \tag{4.6}$$

Thus, a model of the CMM family can require less parameters than a CIM model—with $\nu_{\text{CIM}} = (g-1)+g\times\sum_{b=1}^{\text{B}}(m^{\{b\}}-1)$ parameters—although it takes into account the conditional dependencies.

4.2.5 New parametrization of the block distribution

Main idea The parsimonious versions of the CIM model introduced in [CG91] are meaningful since each multinomial distribution is expressed with two types of parameters: a discrete one determines the location of the mode of the distribution and a continuous one gives its probability (see Section 2.3.1). By using the same idea,

we propose a new parametrization of the block distribution denoted by $(\boldsymbol{\delta}_{kb}, \boldsymbol{a}_{kb})$. This parametrization facilitates the interpretation and the writing of the prior and posterior distributions related to the block parameters (see Section 4.4).

New parametrization The discrete parameter $\boldsymbol{\delta}_{kb} = \{\delta_{kb}^h; h = 1, \ldots, \ell_{kb}\}$ determines the mode locations, since δ_{kb}^h indicates the modality crossing where the mode h is located, with $\delta_{kb}^h \in \{1, \ldots, m^{\{b\}}\}$ and $\delta_{kb}^h \neq \delta_{kb}^{h'}$ if $h \neq h'$. The continuous parameter $\boldsymbol{a}_{kb} = (a_{kb}^h; h = 1, \ldots, \ell_{kb} + 1)$ determines the probability mass of the ℓ_{kb} modes by its first ℓ_{kb} elements (a_{kb}^h with $h = 1, \ldots, \ell_{kb}$) and the probability mass of the non-mode by its last element ($a_{kb}^{\ell_{kb}+1}$). The parameter $\boldsymbol{a}_{kb}$ is defined on the following truncated simplex

$$S^t(m^{\{b\}}) = \left\{ \boldsymbol{a}_{kb} : \ 0 \leq a_{kb}^h \leq 1, \forall h \leq \ell_{kb} + 1 \text{ and } a_{kb}^h \geq \frac{a_{kb}^{\ell_{kb}+1}}{m^{\{b\}} - \ell_{kb}}, \ \forall h \leq \ell_{kb} \right\}. \tag{4.7}$$

The parameter $\boldsymbol{\alpha}_{kb}$ and the couple $(\boldsymbol{\delta}_{kb}, \boldsymbol{a}_{kb})$ are related by

$$\alpha_{kb}^h = \begin{cases} a_{kb}^{h'} & \text{if } \exists h' \text{ such that } \delta_{kb}^{h'} = h \\ \frac{a_{kb}^{\ell_{kb}+1}}{m^{\{b\}} - \ell_{kb}} & \text{otherwise.} \end{cases} \tag{4.8}$$

4.3 Maximum likelihood estimation via an EM algorithm

Aim Let $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be the sample composed with n independent and identically distributed individuals assuming to be drawn by the CMM model. From this sample, the aim is to estimate the MLE for a fixed model defined by $\boldsymbol{\omega}$.

When $\boldsymbol{\omega}$ is known, the CMM model can be interpreted as a CIM model applied on the concatenated variables $\boldsymbol{x}_i^{\{b\}}$, where constraints are added between parameters. Thus, the MLE can be easily obtained by the following EM algorithm.

Algorithm 4.6 (The EM algorithm to obtain the MLE of a CMM model).

Starting from an initial value $\boldsymbol{\theta}^{[0]}$, iteration $[r]$ is written as

— **E step:** conditional probabilities computation

$$t_{ik}(\boldsymbol{\theta}^{[r]}) = \frac{\pi_k^{[r]} p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k^{[r]}, \boldsymbol{\sigma}, \boldsymbol{\ell}_k)}{\sum_{k'=1}^{g} \pi_{k'}^{[r]} p(\boldsymbol{x}_i; \boldsymbol{\alpha}_{k'}^{[r]}, \boldsymbol{\sigma}, \boldsymbol{\ell}_{k'})}. \tag{4.9}$$

— **M step:** maximization of the expectation of the complete-data log-likelihood

$$\pi_k^{[r+1]} = \frac{n_k^{[r]}}{n} \text{ and } \alpha_{kb}^{(h)[r+1]} = \begin{cases} \frac{n_{kb}^{(h)[r]}}{n_k^{[r]}} & \text{if } (1 \leq h \leq \ell_{kb}) \\ \frac{1-\sum_{h'=1}^{\ell_{kj}} \alpha_{kb}^{(h')[r+1]}}{m^{\{b\}}-\ell_{kb}} & \text{otherwise,} \end{cases} \tag{4.10}$$

by using the notations $n_k^{[r]} = \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]})$ and $n_{kb}^{h[r+1]} = \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]}) \boldsymbol{x}_i^{\{b\}h}$.

Remark 4.7 (On the function τ_{kb}). Note that, at the M step of iteration $[r]$, the function τ_{kb} is redefined as the decreasing ordering function of the $n_{kb}^{h[r+1]}$ and allows us to define $n_{kb}^{(h)[r+1]}$ with $n_{kb}^{(h)[r+1]} \geq n_{kb}^{(h+1)[r+1]}$.

4.4 Model selection via a Metropolis-within-Gibbs sampler

Prior distributions We assume that $p(g) = \frac{1}{g_{\max}}$ for $g = 1, \ldots, g_{\max}$ and that $p(\boldsymbol{\sigma})$ (remind that g and $\boldsymbol{\sigma}$ are independent) and $p(\boldsymbol{\ell}|g, \boldsymbol{\sigma})$ follow uniform distributions.

Aim The aim is to obtain the model $\hat{\boldsymbol{\omega}} = (\hat{g}, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\ell}})$ which has the largest posterior probability

$$\hat{\boldsymbol{\omega}} = \underset{g,\boldsymbol{\sigma},\boldsymbol{\ell}}{\operatorname{argmax}}\, p(\mathbf{x}|g, \boldsymbol{\sigma}, \boldsymbol{\ell}) = \underset{g,\boldsymbol{\sigma},\boldsymbol{\ell}}{\operatorname{argmax}}\, p(g, \boldsymbol{\sigma}, \boldsymbol{\ell}|\mathbf{x}). \tag{4.11}$$

Let $g_{\max}$ models denoted by $\boldsymbol{\omega}^{(g)} = (g, \boldsymbol{\sigma}^{(g)}, \boldsymbol{\ell}^{(g)})$, for $g = 1, \ldots, g_{\max}$, where

$$(\boldsymbol{\sigma}^{(g)}, \boldsymbol{\ell}^{(g)}) = \underset{\boldsymbol{\sigma},\boldsymbol{\ell}}{\operatorname{argmax}}\, p(\mathbf{x}|g, \boldsymbol{\sigma}, \boldsymbol{\ell}) = \underset{\boldsymbol{\sigma},\boldsymbol{\ell}}{\operatorname{argmax}}\, p(\boldsymbol{\sigma}, \boldsymbol{\ell}|\mathbf{x}, g). \tag{4.12}$$

The best model is also defined as

$$\hat{\boldsymbol{\omega}} = \underset{g}{\operatorname{argmax}}\, p(\boldsymbol{\omega}^{(g)}|\mathbf{x}), \tag{4.13}$$

and is found by applying the BIC approximation among those $g_{\max}$ selected models.

Main idea An exhaustive search strategy is not doable for two correlated reasons. Firstly, the number of couples $(\boldsymbol{\sigma}, \boldsymbol{\ell})$ can be excessively huge, and, secondly, the estimation of the MLE for each of them is an unnecessary waste of computation time. A Metropolis-within-Gibbs sampler strategy overcomes these two drawbacks at the same time, as we now describe.

For a fix value of g, the couple $(\boldsymbol{\sigma}^{(g)}, \boldsymbol{\ell}^{(g)})$ is estimated by the following Metropolis-within-Gibbs sampler [RC04] having $p(\boldsymbol{\sigma}, \boldsymbol{\ell}|g, \mathbf{x})$ as stationary distribution.

Algorithm 4.8 (The Metropolis-within-Gibbs sampler to obtain $\boldsymbol{\omega}^{(g)}$)**.**

This algorithm has $p(\boldsymbol{\sigma}, \boldsymbol{\ell}|g, \mathbf{x})$ as marginal stationary distribution. Starting from an initial value $(\boldsymbol{\sigma}^{[0]}, \boldsymbol{\ell}^{[0]})$, iteration $[s]$ is written as

$$\boldsymbol{\theta}^{[s+1]} \sim \boldsymbol{\theta}|\boldsymbol{\omega}^{[s]}, \mathbf{x}, \mathbf{z}^{[s]} \tag{4.14}$$

$$\mathbf{z}^{[s+1]} \sim \mathbf{z}|\boldsymbol{\omega}^{[s]}, \mathbf{x}, \boldsymbol{\theta}^{[s+1]} \tag{4.15}$$

$$(\boldsymbol{\sigma}^{[s+1]}, \boldsymbol{\ell}^{[s+1]}) \sim \boldsymbol{\sigma}, \boldsymbol{\ell}|\boldsymbol{\omega}^{[s]}, \mathbf{x}, \mathbf{z}^{[s+1]}, \tag{4.16}$$

where $\boldsymbol{\omega}^{[s]} = (g, \boldsymbol{\sigma}^{[s]}, \boldsymbol{\ell}^{[s]})$.

Remark 4.9 (On the model sampling)**.** A direct sampling from (4.16) is difficult. This step is also performed by one iteration of a Metropolis-Hastings algorithm whose the stationary distribution is $p(\boldsymbol{\sigma}, \boldsymbol{\ell}|g, \mathbf{x}, \mathbf{z}^{[r+1]})$. More details are given in Section 4.4.2.

Sampling of the class memberships As the observed data are independent, the full conditional distribution of $\mathbf{z}$ is classical and is written as

$$p(\mathbf{z}|\boldsymbol{\omega}, \mathbf{x}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{z}_i|\boldsymbol{\omega}, \boldsymbol{x}_i, \boldsymbol{\theta}) \text{ with } p(\boldsymbol{z}_i|\boldsymbol{\omega}, \boldsymbol{x}_i, \boldsymbol{\theta}) = \prod_{k=1}^{g} (t_{ik}(\boldsymbol{\theta}))^{z_{ik}}. \tag{4.17}$$

In this section, we firstly detail the full conditional distributions sampling the parameters (denoted by *instrumental elements*) by using the block parametrization given in Section 4.2.5, and we secondly detail the sampling of $(\boldsymbol{\sigma}, \boldsymbol{\ell})$ (considered as the *interest elements*).

4.4.1 Sampling of the instrumental elements

We now detail the sampling from (4.14) defined by $p(\boldsymbol{\theta}|\boldsymbol{\omega}^{[s]}, \mathbf{x}, \mathbf{z}^{[s]})$.

Prior assumption We assume the *a priori* independence between the class proportions and the parameters of the block distributions. So, the prior of the whole parameter is written as follows

$$p(\boldsymbol{\theta}|\boldsymbol{\omega}) = p(\boldsymbol{\pi}|\boldsymbol{\omega}) \prod_{k=1}^{g} \prod_{b=1}^{m^{\{b\}}} p(\boldsymbol{\alpha}_{kb}|\boldsymbol{\omega}). \tag{4.18}$$

Note that this property of conditional independence is kept by the distribution of $\boldsymbol{\theta}$ conditionally on $(\boldsymbol{\omega}, \mathbf{x}, \mathbf{z})$, since we have

$$p(\boldsymbol{\theta}|\boldsymbol{\omega}, \mathbf{x}, \mathbf{z}) = p(\boldsymbol{\pi}|\boldsymbol{\omega}, \mathbf{x}, \mathbf{z}) \prod_{k=1}^{g} \prod_{b=1}^{m^{\{b\}}} p(\boldsymbol{\alpha}_{kb}|\boldsymbol{\omega}, \mathbf{x}, \mathbf{z}). \tag{4.19}$$

Prior and posterior distributions of $\boldsymbol{\pi}$ The Jeffreys non informative prior distribution, for a multinomial, is a conjugate Dirichlet distribution [Rob07]. So, the prior and the posterior distributions of $\boldsymbol{\pi}$ [BCG10] are respectively defined by

$$\boldsymbol{\pi}|\boldsymbol{\omega} \sim \mathcal{D}_g\Big(\frac{1}{2}, \ldots, \frac{1}{2}\Big) \text{ and } \boldsymbol{\pi}|\boldsymbol{\omega}, \mathbf{x}, \mathbf{z} \sim \mathcal{D}_g\Big(\frac{1}{2} + \mathrm{n}_1, \ldots, \frac{1}{2} + \mathrm{n}_g\Big), \tag{4.20}$$

where $\mathrm{n}_k = \sum_{i=1}^{n} z_{ik}$.

Prior distribution of $\boldsymbol{\alpha}_{kb}$ We now use the parametrization of the block distribution $(\boldsymbol{\delta}_{kb}, \boldsymbol{a}_{kb})$ (defined in Section 4.2.5). We assume the independence between the prior of $\boldsymbol{\delta}_{kb}$ and of $\boldsymbol{a}_{kb}$, so

$$p(\boldsymbol{\alpha}_{kb}|\boldsymbol{\omega}) = p(\boldsymbol{\delta}_{kb}|\boldsymbol{\omega}) p(\boldsymbol{a}_{kb}|\boldsymbol{\omega}). \tag{4.21}$$

We use a uniform distribution among all the mode locations and a conjugate truncated Dirichlet distribution [2] as prior of $\boldsymbol{a}_{kb}$, so

$$p(\boldsymbol{\delta}_{kb}|\boldsymbol{\omega}) = \binom{m^{\{b\}}}{\ell_{kb}}^{-1} \text{ and } \boldsymbol{a}_{kb}|\boldsymbol{\omega} \sim D^t_{\ell_{kb}+1}\Big(\gamma^1_{kb}, \ldots, \gamma^{\ell_{kb}+1}_{kb}; m^{\{b\}}\Big), \tag{4.22}$$

where the γ^h_{kb} are the parameters of the truncated Dirichlet distribution so that $\boldsymbol{a}_{kb}|\boldsymbol{\omega} \in S^t(m^{\{b\}})$. We now fix $\gamma^h_{kb} = 1$, a justification is given in Appendix A.3. The proposed prior is also weakly informative since it is an uniform distribution.

Posterior distribution of $\boldsymbol{\alpha}_{kb}$ The posterior distribution of $\boldsymbol{\alpha}_{kb}$ is written as

$$p(\boldsymbol{\alpha}_{kb}|\boldsymbol{\omega}, \mathbf{x}, \mathbf{z}) = p(\boldsymbol{\delta}_{kb}|\boldsymbol{\omega}, \mathbf{x}, \mathbf{z}) p(\boldsymbol{a}_{kb}|\boldsymbol{\omega}, \boldsymbol{\delta}_{kb}, \mathbf{x}, \mathbf{z}). \tag{4.23}$$

The distribution of $\boldsymbol{\delta}_{kb}|\boldsymbol{\omega}, \mathbf{x}, \mathbf{z}$ is a multinomial one with too many values to be computable. Let $\tilde{\boldsymbol{\delta}}_{kb} = \{\tilde{\delta}^h_{kb}; h = 1, \ldots, \ell_{kb}\}$ be the set containing the indices of the ℓ_{kb} largest values of $\mathrm{n}^h_{kb} = \sum_{i=1}^{n} z_{ik} x_i^{\{b\}h}$ ordered

$$\forall h \in \{1, \ldots, \ell_{kb} - 1\}, \quad \mathrm{n}^{\tilde{\delta}^h_{kb}}_{kb} \geq \mathrm{n}^{\tilde{\delta}^{h+1}_{kb}}_{kb}. \tag{4.24}$$

We assume that the difference between the mode probabilities and the non-mode probabilities are significant. So, we can approximate the full conditional distribution

2. $p(\boldsymbol{a}_{kb}|\boldsymbol{\omega}) \propto \prod_{h=1}^{\ell_{kb}+1} (a^h_{kb})^{\gamma^h_{kb}-1} \mathbb{1}_{\left\{a^h_{kb} \geq \frac{a^{\ell_{kb}+1}_{kb}}{m^{\{b\}} - \ell_{kb}}\right\}}$.

of $\boldsymbol{\delta}_{kb}$ by a Dirac in $\tilde{\boldsymbol{\delta}}_{kb}$. This approximation is strengthened by the fast convergence speed of the discrete parameters [CS12]. Concerning now $\boldsymbol{a}_{kb}$, as its prior is conjugated, its conditional distribution is explicitly defined as

$$\boldsymbol{a}_{kb}|\boldsymbol{\omega}, \boldsymbol{\delta}_{kb}, \mathbf{x}, \mathbf{z} \sim \mathcal{D}^{t}_{\ell_{kb}+1}\left(1 + \mathrm{n}^{(1)}_{kb}, \ldots, 1 + \mathrm{n}^{(\ell_{kj})}_{kb}, 1 + \bar{\mathrm{n}}^{\ell_{kb}}_{kb}; m^{\{b\}}\right), \tag{4.25}$$

where $\mathrm{n}^{(h)}_{kb}$ is the hth larger value of the set $\{\mathrm{n}^{h}_{kb}; h = 1, \ldots, m^{\{b\}}\}$ and $\bar{\mathrm{n}}^{\ell_{kb}}_{kb} = \mathrm{n}_k - \sum_{h=1}^{\ell_{kb}} \mathrm{n}^{(h)}_{kb}$.

4.4.2 Sampling of a new model $(\boldsymbol{\sigma}^{[s+1]}, \boldsymbol{\ell}^{[s+1]})$

Main idea The sampling of $\boldsymbol{\omega}^{[s+1]} = (g, \boldsymbol{\sigma}^{[s+1]}, \boldsymbol{\ell}^{[s+1]})$ according to (4.16) is performed by one iteration of the following MCMC algorithm whose the stationary distribution is $p(\boldsymbol{\sigma}, \boldsymbol{\ell}|g, \mathbf{x}, \mathbf{z}^{[r+1]})$. This algorithm is divided in two steps. Firstly, it samples, by one iteration of a Metropolis-Hastings algorithm, a new repartition of the variables into blocks and the mode number of the modified blocks denoted respectively by $\boldsymbol{\sigma}^{[s+1]}$ and $\boldsymbol{\ell}^{[s+1/2]}$. Secondly, it samples the mode number of each block by one MCMC iteration. Thus, the sampling of $\boldsymbol{\omega}^{[s+1]}$ is decomposed into the two following steps.

Algorithm 4.10 (The MCMC algorithm)**.**

This algorithm has $p(\boldsymbol{\sigma}, \boldsymbol{\ell}|g, \mathbf{x}, \mathbf{z}^{[s+1]})$ as stationary distribution. At the iteration $[s]$ of Algorithm 4.8, the sampling of $\boldsymbol{\omega}^{[s+1]}$ is performed according to both following steps

$$(\boldsymbol{\sigma}^{[s+1]}, \boldsymbol{\ell}^{[s+1/2]}) \sim \boldsymbol{\sigma}, \boldsymbol{\ell}|\boldsymbol{\omega}^{[s]}, \mathbf{x}, \mathbf{z}^{[s+1]} \tag{4.26}$$

$$\boldsymbol{\ell}^{[s+1]} \sim \boldsymbol{\ell}|\boldsymbol{\omega}^{[s+1/2]}, \mathbf{x}, \mathbf{z}^{[s+1]}, \tag{4.27}$$

where $\boldsymbol{\omega}^{[s+1/2]} = (g, \boldsymbol{\sigma}^{[s+1]}, \boldsymbol{\ell}^{[s+1/2]})$.

Metropolis-Hastings algorithm to sample $\boldsymbol{\omega}^{[s+1/2]}$

The sampling of $\boldsymbol{\omega}^{[s+1/2]}$ is performed by one iteration of the Metropolis-Hastings algorithm divided into two steps. Firstly, the instrumental distribution $q(.; \boldsymbol{\omega}^{[s]})$ generates a candidate $\boldsymbol{\omega}^{\star} = (g, \boldsymbol{\sigma}^{\star}, \boldsymbol{\ell}^{\star})$. Secondly, $\boldsymbol{\omega}^{[s+1]}$ is sampled according to the acceptance probability.

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

Acceptance probability The acceptance probability $\lambda^{[s]}$ is defined by

$$\lambda^{[s]} = \min\left\{\frac{p(\mathbf{x}, \mathbf{z}^{[s+1]}|\boldsymbol{\omega}^\star)}{p(\mathbf{x}, \mathbf{z}^{[s+1]}|\boldsymbol{\omega}^{[s]})}\frac{q(\boldsymbol{\omega}^{[s]};\boldsymbol{\omega}^\star)}{q(\boldsymbol{\omega}^\star;\boldsymbol{\omega}^{[s]})}; 1\right\}. \tag{4.28}$$

The computation of $\lambda^{[s]}$ involves to compute the integrated complete-data likelihood. We now describe how to solve this problem without using the biased BIC approximation or using too much time consuming MCMC methods. The sampling of $\boldsymbol{\omega}^{[s+1/2]}$ is also performed by the following Metropolis-Hastings algorithm.

Algorithm 4.11 (The Metropolis-Hastings algorithm)**.**

This algorithm has $p(\boldsymbol{\sigma}, \boldsymbol{\ell}|g, \mathbf{x}, \mathbf{z}^{[s+1]})$ as stationary distribution. Starting from an initial value $\boldsymbol{\theta}^{[0]}$, iteration $[r]$ is written as

$$\boldsymbol{\omega}^\star \sim q(\boldsymbol{\omega}; \boldsymbol{\omega}^{[s]}) \tag{4.29}$$

$$\boldsymbol{\omega}^{[s+1/2]} = \begin{cases} \boldsymbol{\omega}^\star & \text{with a probability } \lambda^{[s]} \\ \boldsymbol{\omega}^{[s]} & \text{with a probability } 1-\lambda^{[s]}. \end{cases} \tag{4.30}$$

MCMC algorithm to sample $\boldsymbol{\ell}^{[s+1]}$

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

$$p(\ell_{kb}|\boldsymbol{\omega}^{[s+1/2]}, \mathbf{x}, \mathbf{z}^{[s+1]}) \propto \begin{cases} p(\mathbf{x}^{\{b\}}|\mathbf{z}^{[s+1]}, \ell_{kb}) & \text{if } |\ell_{kb} - \ell_{kb}^{[s+1/2]}| < 2 \\ & \text{and } \ell_{kb} \notin \{0, m^{\{b\}}\}. \\ 0 & \text{otherwise,} \end{cases} \tag{4.31}$$

where $\mathbf{x}^{\{b\}} = (\boldsymbol{x}_i^{\{b\}}; i = 1, \ldots, n)$. Thus, this algorithm requires the computation of $p(\mathbf{x}^{\{b\}}|\mathbf{z}, \ell_{kb})$ defined by

$$p(\mathbf{x}^{\{b\}}|\mathbf{z}, \ell_{kb}) = \int_{S(\ell_{kb}, m^{\{b\}})} \prod_{h=1}^{m^{\{b\}}} (\alpha_{kb}^h)^{\mathrm{n}_{kb}^h} d\boldsymbol{\alpha}_{kb} \tag{4.32}$$

and that we detail now.

The integrated complete-data likelihood

The integrated complete-data likelihood is defined as

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\omega}) = p(\mathbf{z}|\boldsymbol{\omega}) \prod_{k=1}^{g} \prod_{b=1}^{\mathrm{B}} p(\mathbf{x}^{\{b\}}|\mathbf{z}, \ell_{kb}). \tag{4.33}$$

Note that the quantities $p(\mathbf{x}, \mathbf{z}|\boldsymbol{\omega})$ and $p(\mathbf{x}^j|\mathbf{z}, \ell_{kb})$ are respectively required to compute the acceptance probability of the Metropolis-Hastings algorithm defined by (4.28) and to sample the number of modes from (4.31). It can be evaluated by

BIC-like approximations. For instance, the integrated complete-data likelihood is approximated by

$$\ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{\omega}) = \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}^\star, \boldsymbol{\omega}) - \frac{\nu}{2} \ln n + \mathcal{O}(1), \tag{4.34}$$

where $\boldsymbol{\theta}^\star$ is the maximum complete-data likelihood estimate. However, this kind of approximation is only asymptotically true and can over-estimate the mode numbers (see Section 4.5.1). As $\mathbf{z}|\boldsymbol{\omega}$ follows a uniform distribution among all the possible partitions, we propose to compute each $p(\mathbf{x}^{\{b\}}|\mathbf{z}, \ell_{kb})$ to obtain $p(\mathbf{x}, \mathbf{z}|\boldsymbol{\omega})$. This computation is not easy since $\boldsymbol{\alpha}_{kb}$ is defined on $S(\ell_{kb}; m^{\{b\}})$ and not on the whole simplex of size ℓ_{kb} (except when $\ell_{kb} = m^{\{b\}} - 1$; in such case we can use the approach of the CIM model [BCG10]). An explicit formula is given in the following proposition by performing an exact computation of the integral over the continuous parameters and an approximation on the discrete ones (for the proof see in Appendix A.3).

Proposition 4.12. *The integrated complete-data likelihood is approximated, by neglecting the sum over the discrete parameters of the modes locations and by performing the exact computation on the continuous parameters, so*

$$p(\mathbf{x}^{\{b\}}|\mathbf{z}, \ell_{kb}) \approx \left(\frac{1}{m^{\{b\}} - \ell_{kb}}\right)^{\bar{\mathrm{n}}_{kb}^{\ell_{kb}}} \prod_{h=1}^{\ell_{kb}} \frac{Bi\left(\frac{1}{m^{\{b\}}-h+1}; \mathrm{n}_{kb}^{(h)} + 1; \bar{\mathrm{n}}_{kb}^{h} + 1\right)}{m^{\{b\}} - h}, \tag{4.35}$$

where $Bi(x; a, b) = B(1; a, b) - B(x; a, b)$ and where $B(x; a, b)$ is the incomplete beta function defined by $B(x; a, b) = \int_0^x w^a (1-w)^b dw$.

From the previous expression, it is straightforward to obtain $p(\mathbf{x}, \mathbf{z}|\boldsymbol{\omega})$.

4.5 Numerical experiments on simulated data sets

4.5.1 Integrated complete-data likelihood: comparison of both approaches

Aim During this experiment, we highlight the biases of the BIC criterion for the selection of the number of modes and the gain given by the proposed computation of the integrated complete-data likelihood.

Data generation We want to compare both approaches for the selection of the number of modes. So, we simulate samples composed with n i.i.d individuals arisen from a multinomial distribution per modes $\mathcal{M}_s(r, r, r, \frac{1-3r}{s-3}, \ldots, \frac{1-3r}{s-3})$ with s modalities and three modes having a probability r. For different sizes of sample, 10^5 samples are generated with different values of (r, s).

Results Figure 4.1 gives a comparison between the proposed approach and the BIC-like approximation for the selection of the number of modes. The proposed criterion obtains better results than the BIC criterion in the four studied situations for the large sample sizes. Furthermore, it allows to never overestimates the number

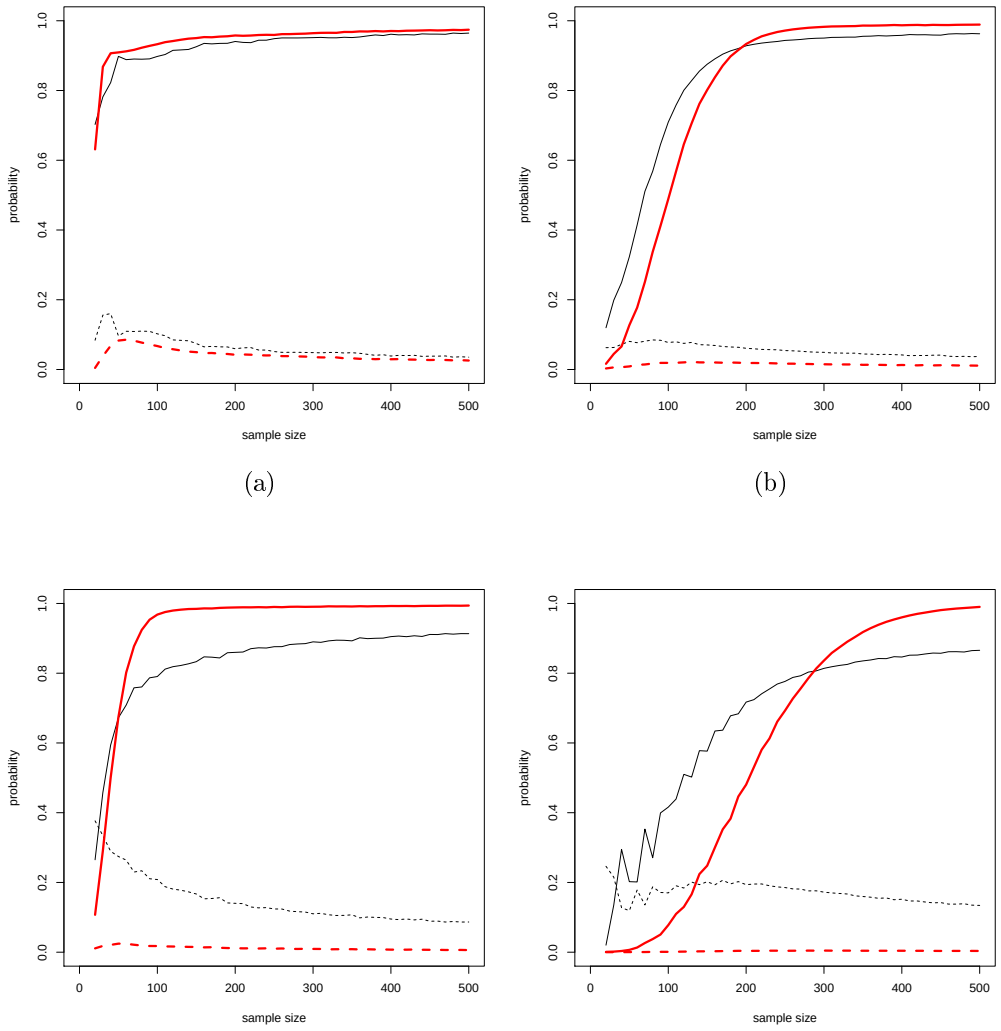


Figure 4.1 – Probability that the BIC criterion (represented in fine black lines) and the proposed approach (represented in bold red lines) select the true number of modes (represented in plain line) and overestimate it (represented in dotted line) (a) r=0.3, s=9; (b) r=0.2, s=9; (c) r=0.2, s=18; (d) r=0.1, s=27.

of modes. Finally, its variability is smaller than the BIC criterion one. We enter now into more specific comments.

In case (a), modes have a large probability mass and they are easily detected since there are few modalities. Thus, both criteria have the same behavior since they find the true number of modes with a probability close to one even for small samples.

When the mode probabilities decrease (case (b)), it is more difficult to identify them. In such a case, the BIC criterion better finds the true number of modes than the proposed approach, for the small samples (size lower than 150). However, the BIC criterion has a moderated risk to overestimate the number of modes while

the proposed approach underestimates this number when it is wrong. When the sample size is larger than 200, the proposed approach obtains better results. Indeed, it finds a true number of modes almost always while the BIC criterion keeps an overestimation risk.

If the number of modalities increases (case (c)), then the problem becomes harder and the proposed approach also shows its interest since the BIC criterion is strongly biased in such case. The BIC criterion keeps this bias even for a large data set while the proposed approach almost always finds the true number of modes when the sample size is larger than 100.

Finally, note that in the more complex situations like in case (d) (few probability mass for the modes and large number of modalities), the proposed approach underestimates the number of modes when the sample size is small then converges to the true mode values when the sample size increases. Note that, in such a case, the bias of the BIC criterion stays significant even for a large data set.

Based on this experiment, the proposed criterion appears as the most relevant since its asymptotic behavior is better than the asymptotic behavior of the BIC criterion. Indeed, it never overestimates the number of modes and its variability is smaller than the variability of the BIC criterion.

4.5.2 Simulation with well specified model

Aim During this experiment, we highlight the good behavior of the algorithms (EM algorithm and Metropolis-within-Gibbs sampler) for performing the estimation of the MLE and the model selection. So, data are generated according to a CMM model, then the model and the MLE are estimated. The quality of the estimation is determined by the Kullback-Leibler divergence. We show that this quantity converges to zero when the sample size increases. So, we conclude to the good behavior of both algorithms.

Data generation A data set of six variables with three modalities is generated according to a bi-component CMM model with the following parameters:

$$\boldsymbol{\sigma} = (\{1,2\},\{3,4\},\{5,6\}),\ \ell_{kj} = 2,\ \boldsymbol{\pi} = (0.5, 0.5),\ \boldsymbol{\alpha}_{kj} = (0.4, 0.4, 0.2/7).$$

The modes are located at different modality crossings for both classes.

Results For different values of $n = (50, 100, 200, 400, 800)$, 100 samples are generated. The Kullback-Leibler divergence is computed between the true and the estimated parameters. Table 4.1 presents the mean of this divergence.

As the Kullback-Leibler divergence converges to zero, when the sample size increases, we claim that the estimated distribution converges to the true one. Thus, we conclude to the good behavior of the estimation algorithm.

4.5.3 Simulation with misspecified model

Aim During this experiment, we underline that the flexibility of the CMM model allows it to keep good results even if the model is misspecified. Thus, we simulate

n	50	100	200	400	800
mean	0.656	0.117	0.061	0.028	0.015
sd	0.636	0.052	0.018	0.007	0.003

Table 4.1 – Mean and standard deviation of the Kullback-Leibler divergence computed between the true parameters of the specified model and the maximum likelihood estimates associated to the model selected by the Metropolis-within-Gibbs algorithm for different sample sizes.

samples according to a bi-component mixture model where the intra-class dependencies are different for both components. A tuning parameter allows us to modify the strength of the intra-class dependencies and the class overlapping. The results of the CMM model are compared to those of the CIM model.

Data generation A data set of size 100 is sampled from the following bi-component mixture model of dimension six

$$p(\boldsymbol{x};\boldsymbol{\theta}) = 0.5\prod_{h=1}^{3} p(\boldsymbol{x}^{2h-1},\boldsymbol{x}^{2h};\boldsymbol{\theta}) + 0.5\; p(\boldsymbol{x}^{1};\boldsymbol{\theta})p(\boldsymbol{x}^{6};\boldsymbol{\theta})\prod_{h=1}^{2} p(\boldsymbol{x}^{2h},\boldsymbol{x}^{2h+1};\boldsymbol{\theta}), \quad (4.36)$$

with $p(\boldsymbol{x}^{j},\boldsymbol{x}^{j+1};\boldsymbol{\theta}) = p(\boldsymbol{x}^{j};\boldsymbol{\theta})\big(\lambda \mathbb{1}_{\{\boldsymbol{x}^{j}=\boldsymbol{x}^{j+1}\}} + (1-\lambda)p(\boldsymbol{x}^{j+1};\boldsymbol{\theta})\big)$ and with $p(\boldsymbol{x}^{j};\boldsymbol{\theta}) = \sum_{h=1}^{3}(1/3)^{x^{jh}}$. Thus, when $\lambda = 0$, the sample is generated by a uniform distribution and classes are confused. The larger is the tuning parameter λ, the larger are the intra-class dependencies and the class separation. Note that CMM is not the true model since the conditionally correlated variables are not the same in both classes.

Results For different values of $\lambda = (0.2, 0.4, 0.6, 0.8)$, 100 samples are generated. The Kullback-Leibler divergence associated to the model with the best number of classes (selected by the BIC criterion among $g = 1, ..., 4$) is computed. Table 4.2 presents the results obtained by the CMM and the CIM models.

λ	0.2	0.4	0.6	0.8
CMM	0.09 (1.00)	0.25 (1.16)	0.53 (2.08)	0.87 (2.10)
CIM	0.11 (1.00)	0.27 (1.00)	1.67 (1.12)	5.79 (1.40)

Table 4.2 – Kullback-Leibler divergence and mean of the number of classes obtained by CMM and CIM.

The larger is λ, the larger is the Kullback-Leibler divergence for both models. However, the flexibility of the CMM model allows to keep an acceptable value of the Kullback-Leibler divergence while this divergence grows dramatically faster with the CIM model. Furthermore, when the classes are well separated (large value of λ), the CMM model finds more often the true number of classes than the CIM model.

4.6 Analysis of two real data sets

For both applications, the estimation of the CMM model was performed by the R package `CoModes`. Both data sets are available in `CoModes` developed by the authors.

4.6.1 Seabirds clustering

Data We study a biological data set describing 153 puffins (seabirds) by five plumage and external morphological characteristics presented in Table 4.3 [Bre07]. These seabirds are divided into three subspecies *dichrous* (84 birds), *lherminieri* (34 birds) and *subalaris* (35 birds).

variables	m_j	modalities				
collar	5	none	...	...	...	continuous
eyebrows	4	none	...	...		very pronounced
sub-caudal	4	white	black	black and white		BLACK and white
border	3	none	...	many		
gender	2	male	female			

Table 4.3 – Presentation of the five plumage and external morphological variables describing the puffins.

Experimental settings The subspecies memberships of the individuals are blinded. For $g = 1, \ldots, 6$, the MLE of the CIM model is obtained by 25 initializations of an EM algorithm while 25 chains of 3000 iterations are performed for the model selection of the CMM model followed by 25 initializations of EM algorithm to find the MLE.

Results Table 4.4 presents the values of the BIC criterion for both models and different numbers of classes. Even if both models select two components, the values of the BIC criterion are better for the CMM model than for the CIM model for all the numbers of classes. Thus, the CMM model better fits the data than the CIM model.

g	1	2	3	4	5	6
CMM	-711	**-691**	-701	-709	-721	-727
CIM	-711	**-706**	-722	-745	-775	-805

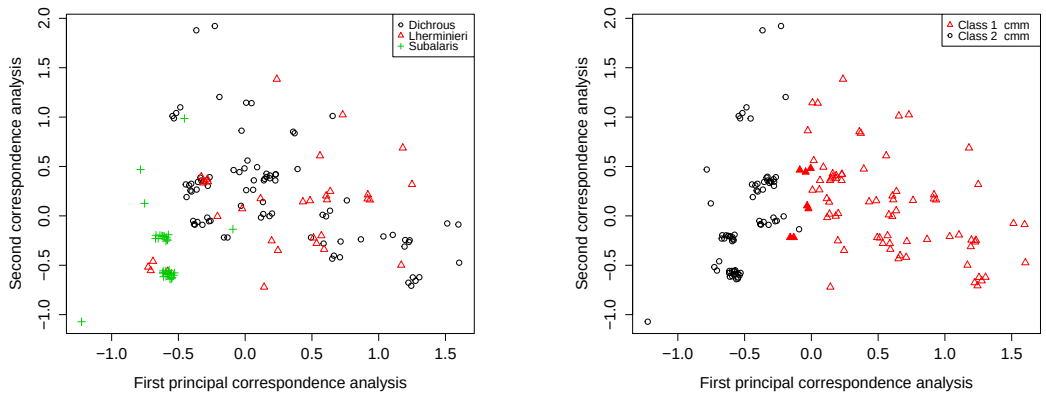
Table 4.4 – Values of the BIC criterion for different numbers of classes obtained by the CMM and the CIM models. Boldface indicates the best values of this criterion.

According to Table 4.5 displaying the confusion matrix between the estimated partitions and the subspecies, we claim that the Subalaris are more different than the two other subspecies. Indeed, both models affect all the Subalaris in class 2. If the estimated partitions by both models are similar, we remark that the CMM model affects less other subspecies in this class than the CIM model.

	CMM		CIM	
	class 1	class 2	class 1	class 2
Dichrous	52	32	48	36
Lherminieri	23	11	22	12
Subalaris	0	35	0	35

Table 4.5 – Confusion tables between the subspecies and estimated partition into two classes.

Figure 4.2(a) displays the seabirds scatterplot on the first correspondence analysis plan and indicates the subspecies. We note that all the Subalaris are in the same location (bottom left) for the first principal correspondence map. We display the partition corresponding to the best model (the CMM model with two components) in Figure 4.2(b). Note that, for both models, the first principal correspondence axis allows to define a classification rule.



(a) (b)

Figure 4.2 – Seabirds on the first principal correspondence analysis map (a) with the subspecies and (b) with the best CMM model estimated partition. The bold triangles indicate the individuals affected in class 1 for the CMM model and in class 2 for the CIM model. An i.i.d. uniform noise on $[0, 0.1]$ has be added on both axes for each individual in order to improve visualization.

We now describe the best bi-component CMM model. Even if the estimated model assumes conditional independence between variables, this model is of interest because of its sparsity. Indeed, it is more parsimonious than the CIM model since a small number of modes is estimated as shown by the summary proposed by κ_{kj} and ρ_{kj} defined in (4.5) and presented in Table 4.6. Thus, the first variables are characterized by few modalities with a high probability. As the variables are conditionally independent, κ_{kj} indicates the number of modalities having a probability upper than the uniform distribution. For example the multinomial distribution of the variable sub-caudal has two modes for both classes (so $\kappa_{kj} = 2/3$).

	collar	eyebrows	sub-caudal	border	gender
class 1	0.75 (0.93)	0.67 (0.91)	0.67 (0.88)	1.00 (1.00)	1.00 (0.55)
class 2	0.75 (0.98)	0.67 (0.77)	0.67 (0.99)	0.50 (0.97)	1.00 (0.57)

Table 4.6 – Summary of the CMM model with three classes: κ_{kj} is displayed in plain and ρ_{kj} is displayed in parenthesis.

The maximum likelihood estimates of the component parameters are presented by Figure 4.3. Each sub-figure corresponds to a block of variable, thus we note again that the estimated model assumes the conditional independence. For each block of variables, the modality crossings where one mode is estimated for at least one component are focused. For these modality crossings, we display their cumulated probability masses for each component (the component are identified by different colors). These modality crossings are presented by decreasing order of cumulated probability mass.

Note that the mode locations are discriminative since the modality black (resp. white) has a probability of 0.64 (resp. 0.24) for class 1 while the modality white (resp. BLACK and white) has a probability of 0.94 (resp. 0.05).

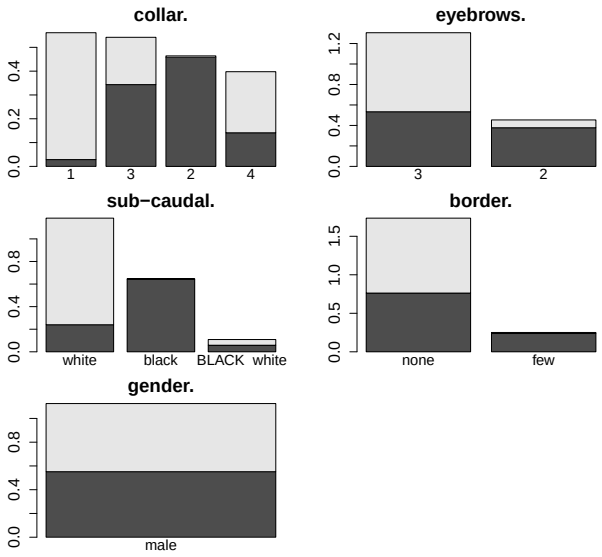


Figure 4.3 – Class parameters of the bi-components CMM model estimated on the Seabirds data. The black color (respectively the gray color) corresponds to the probability mass of the modes for class 1 (respectively to class 2).

Finally, the conditional independence assumption seems realistic since the conditional Cramer's V measures, presented in Table 4.7, are small. We also perform a bootstrap test of the global nullity of the Cramer's V by generating 1000 samples. We obtain a p-value of 0.91, so the conditional independence assumption is validated.

1	0.14	0.15	0.23	0.21
	1	0.36	0.20	0.13
		1	0.13	0.19
			1	0.01
				1

(a) Class 1

1	0.14	0.09	0.11	0.28
	1	0.24	0.21	0.26
		1	0.02	0.07
			1	0.17
				1

(b) Class 2

Table 4.7 – Matrix of the Cramer's V measures computed according to the estimated classes.

4.6.2 Acute inflammations clustering

Data We want to cluster 120 patients [CZ03] described by five binary variables (occurrence of nausea (Nau), lumbar pain (Lum), urine pushing (Pus), micturition pains (Mic) and burning of urethra (Bur)) and by one variable having three modalities (temperature of the patient (Tem): $T < 37C$, $37°C \leq T < 38°C$ and $38°C \geq T$). We know that some patients have one of the following diseases of the urinary system: inflammation of urinary bladder and Nephritis of renal pelvis origin.

Experimental conditions We use the same experimental conditions as the Seabirds clustering.

Results Table 4.8 presents the values of the BIC criterion for both models and different numbers of classes. For each number of classes, the BIC criterion value of the CMM model is better than those of the CIM model. Furthermore, the CMM model selects three classes while the CIM model selects four classes. This phenomenon can be due to the violated conditional independence assumption of the CIM model.

g	1	2	3	4	5	6
CMM	-510	-351	**-338**	-345	-399	-401
CIM	-527	-478	-439	**-407**	-412	-418

Table 4.8 – Values of the BIC criterion for different numbers of classes and for the CMM and the CIM models. Boldface indicates the best values of this criterion.

Note that the estimated distributions of the CIM and the CMM models are different. The obtained partition are also different. Table 4.9 displays the confusion matrices between the best CMM model and the CIM models with three and four classes. Thus, if 29 individuals constitute a group which is well separated from the other individuals (class 3) for the three models, the other individuals have a class membership determined by the selected model.

Figure 4.4 displays the individuals on the 1-5 principal correspondence analysis map where the estimated classes are well separated.

The CMM model with three classes has the following repartition of the variables into blocks: $\boldsymbol{\sigma} = (\{\text{Tmp, Pus, Mic, Bur}\}, \{\text{Nau}\}, \{\text{Lum}\})$. As shown by the sum-

	CMM		
	c1	c2	c3
CIM c1	40	0	0
CIM c2	10	41	0
CIM c3	0	0	29

	CMM		
	c1	c2	c3
CIM c1	40	0	0
CIM c2	10	20	0
CIM c3	0	21	0
CIM c4	0	0	29

Table 4.9 – Confusion matrices between the best CMM model and the CIM models with three and four classes.

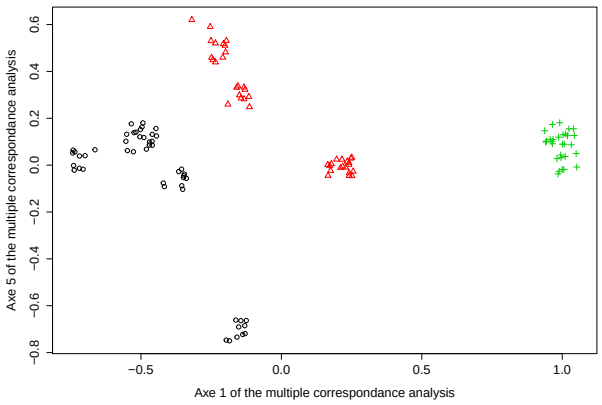


Figure 4.4 – Individuals on the 1-5 principal correspondence analysis map with the best CMM model estimated partition. An i.i.d. uniform noise on $[0, 0.1]$ has be added on both axes for each individual in order to improve visualization. Colors and symbols indicate the class membership.

mary ρ_{kj} and κ_{kj} displayed in Table 4.10, the three classes are concentrated in few modality crossings for the block one and in one location with a probability close to one for the two other blocks.

	Tmp, Nau, Lum, Mic	Pus	Bur
Class 1	0.41 (1.00)	1.00 (1.00)	1.00 (0.99)
Class 2	0.33 (0.99)	1.00 (1.00)	1.00 (1.00)
Class 3	0.25 (0.99)	1.00 (1.00)	1.00 (1.00)

Table 4.10 – Summary of the CMM model with three classes: κ_{kj} is displayed in plain and ρ_{kj} is displayed in parenthesis.

The following class interpretation is based on the class parameters displayed by Figure 4.5. Note that the variables *urine pushing* and *burning of urethra* are the most discriminative ones.

— The majority class (42%) groups individuals having no nausea and no lumber pain.

— The second class (34%) groups individuals having no nausea but lumber pain.
— The third class (24%) groups individuals having nausea and lumber pain. Furthermore, these individuals have some fiever and micturition pain.

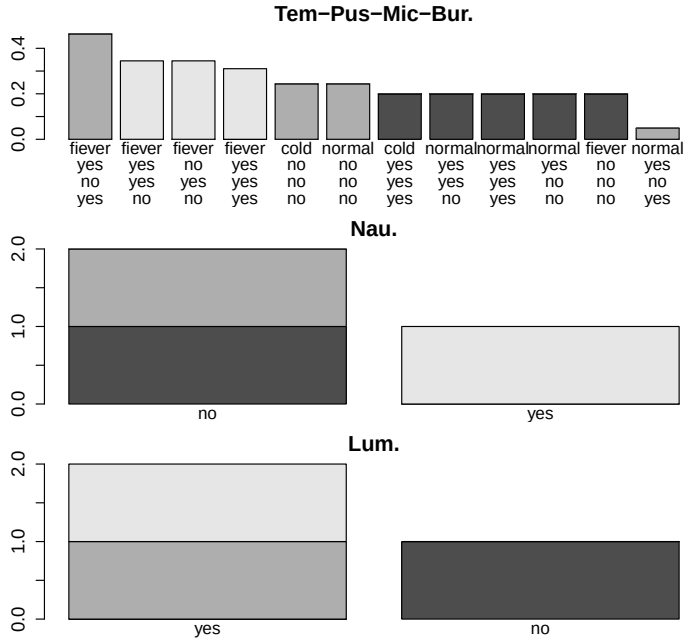


Figure 4.5 – Estimated parameters of the tri-component CMM model displayed by the barplot function of the package `CoModes`. Black color corresponds to class 1, black gray color corresponds to class 2 and pale gray color corresponds to class 3.

4.7 Conclusion

In this chapter, we have presented a new mixture model (CMM) to cluster categorical data. Its strength is to relax the conditional independence assumption and to stay parsimonious. A summary of the distribution is given by κ_{kj} and ρ_{kj} while each class can be summarized by the mode locations. As shown on the Seabirds application, the CMM model can outperform the classical latent class model even if the conditional independence assumption is true, thank's to its sparsity.

The combinatorial problems involved by the block detection and by the selection of the numbers of modes are avoided by a Metropolis-within-Gibbs algorithm. This algorithm can be used because the computation of the integrated complete-data likelihood can be efficiently approximated. Thus, this approach can be used to select the interactions of the log-linear mixture model per block.

However, the model is hardly estimated if the data set has a large number of variables. Some constraints on the block variables repartition could also be added (for instance the number of variables into blocks could be limited to three variables). Another solution could be to estimate the model by a forward/backward strategy but it is known that these methods are sub-optimal.

Finally, we imposed the equality of the repartition of the variables into blocks for all the classes. This property allows us to prove the generic identifiability of the CMM model. This lack of flexibility is counterbalanced by flexible block distribution. However, one could try to relax the class-equality of $\boldsymbol{\sigma}$ with the model no-identifiability risk.

Chapter 5

Model comparison performed by their R-packages

This chapter aim is to illustrate the R packages `Clustericat` *and* `CoModes` *which respectively perform the inference of the* MEDD *and the* CMM *models.*
In order to make a demonstration of both packages, we use them to perform the cluster analysis of the running example presented by Table 2.1 of Chapter 2. We remind that this data set displays the evaluation (sound or carious) of 3869 dental x-rays that may show incipient caries performed by five dentists.
Note that, this chapter can also be used as a tutorial of both packages. Indeed, it provides a presentation of their main functions and many scripts allowing to perform the cluster analysis.

All animals are equal, but some animals are more equal than others.
George Orwell — Animal Farm

5.1 Clustericat

5.1.1 Clustericat overview

Presentation The R package `Clustericat` performs the clustering of categorical data according to the MEDD model. Its main functions are implemented in C++. We remind that, in this model, variables are grouped into conditionally independent blocks in order to consider the main intra-class correlations. The intra-class dependency between variables grouped inside the same block is taken into account by mixing two extreme distributions, which are respectively independence and maximum dependency.

Download Clustericat The package is currently available on R-forge at the following url: https://r-forge.r-project.org/R/?group_id=1803. The installation and the loading of `Clustericat` can be performed by using the following R scripts.

Clustericat script 5.1 (Clustericat installation).

```
# R install command
> install.packages("Clustericat",
                   repos="http://R-Forge.R-project.org")
```

Clustericat script 5.2 (Clustericat loading).

```
# Clustericat loading
> require(Clustericat)
```

Estimation The parameter estimation by maximum likelihood is performed via a GEM algorithm while a Gibbs algorithm, used for the model selection, avoids the combinatorial problems induced by the block structure search.

5.1.2 Main functions

Five functions compose the `Clustericat` package. One function performs the cluster analysis of the data. Its tuning parameters can be specified by the user by calling a specific function. The three last functions are implemented in order to friendly present the parameters by providing numerical or graphical summaries.

The clustering function The cluster analysis can be performed with the function `clustercat()` taking four arguments.

Clustericat script 5.3 (The clustering function).

```
> clustercat(data, nb_cluster,
             modal= 0, strategy= strategycat(data))
```

This function has two mandatory arguments:

- A data frame `data` to cluster whose the columns are non-zero integers or factors.
- An integer vector `nb_cluster` specifying the number of classes.

Default values are taken for the arguments `modal` and `strategy`.

- The argument `modal` is a vector given the modality number for each variable.
- The argument `strategy` is an instance of the `strategycat` class which contains the adjustments inputs parameters related to the estimation algorithms.

The function `clustercat()` returns an instance of the `clustcat` class which contains all the outputs.

The tuning function The adjustments parameters of the estimation algorithms contained in the `strategycat` class can be specified by the function `strategycat()` taking four arguments.

Clustericat script 5.4 (The tuning function).

```
> strategycat(data, nb_init= 5, stop_criterion=
              20 * ncol(data), partition= partitioncat(data))
```

The input data matrix `data` is mandatory and the three others input parameters allow to tune the algorithms.

— The argument `nb_init` sets the number of times where a MCMC chain is started. By default 5 MCMC chains are initialized.
— The argument `stop_criterion` is the integer corresponding to the number of successive iterations of the MCMC chain where if no better model is fund then the algorithm is stopped. By default it takes the values of $20 \times d$.
— The argument `partition` is the initial value of the repartition of the variables into blocks ($\boldsymbol{\sigma}^{[0]}$). By default it is equal to the partition produced by the HAC minimizing the block number without block consisting of more than four variables.

Three tool functions `Clustericat` package also provides tool functions like `summary()`, `summary_dependencies()` and `plot()` respectively to summarize results, to present the main conditional dependencies and to visualize the parameters.

5.1.3 Clustericat to cluster the dentists data set

Clustering with MEDD We now present the results of the MEDD model obtained by the R package `Clustericat` by using the following script

Clustericat script 5.5 (Dentists data set clustering).

```
# Data set loading
> data("dentist")

# Definition of the tuning parameters for the estimation algorithm
> st <- strategycat(dentist, nb_init= 35, stop_criterion=
200)

# Estimation of the bi-component MEDD model.
> res <- clustercat(dentist, 2, modal= rep(2,5), strategy=
st)
```

Model selection The BIC criterion selects two classes with a value of -7473. It claims that the MEDD model better fits the data than the model presented in [QTK96] since its BIC criterion value is -7487. The BIC criterion values for the CIM and the MEDD models are displayed in Table 5.1. We indicate the computing time (in seconds), obtained with a processor Intel Core i5-3320M, to estimate the MEDD model where 20 MCMC chains were started with a stopping rules $q_{\max} = 100$ while the CIM model needs less than 0.1 sec with the R package RMixmod [LIL+12].

	g	1	2	3	4
CIM	BIC	-8766	-7511	**-7481**	-7503
MEDD	BIC	-7743	**-7473**	-7481	-7503
	time (sec)	1.7	4.9	6.1	7.7

Table 5.1 – BIC criterion values for the CIM and the MEDD models according to different numbers of classes for the dentistry data set. Best values are in bold.

We note that the MEDD model obtains better values for the BIC criterion than the CIM model when $g = 1, 2$. When the number of classes is larger ($g \geq 3$) the best MEDD model assumes the conditional independence between variables.

Comparison of the results The BIC criterion selects two classes for the MEDD model. This result is coherent with a splitting of the teeth between the sound and the carious ones. Furthermore, the two main characteristics of the log-linear mixture model imposed in [EH89] are automatically detected by the model: importance of the two modality crossings where all the dentists have the same diagnosis and a dependency between the diagnosis of the dentists 3 and 4. Thus, the estimated model is coherent with the imposed model presented in [EH89] while no information was given *a priori*.

Best model interpretation The model interpretation is facilitated by two tools functions. The function `summary()` provides a general overview of the model (information criteria, proportions, blocks of variables and intra-class dependencies ρ_{kb}).

Clustericat script 5.6 (Model overview).

```
# Function providing a model overview
> summary(res)


Number of classes: 2     BIC value: -7472.845
log-likelihood value: -7415.019
************************************
Proportions: 0.8550206 0.1449794
************************************
Blocks repartition of the variables for the class 1:
                    Variables          Rho
 Block 1   de1, de2, de3, de4, de5   0.3506754
Blocks repartition of the variables for the class 2:
              Variables      Rho
 Block 1       de3, de4    0.2481778
 Block 2   de1, de2, de5   0.0000000
```

The function `summary_dependencies()` focuses on the intra-class dependency parameters.

Clustericat script 5.7 (Summary of the intra-class dependencies).

```
# Function providing summary of the intra-class dependencies
> summary_dependencies(res)
Blocks repartition of the variables for the class 1:
Block 1 contains the variables: de1 de2 de3 de4 de5 with Rho= 0.3506754
     Tau       de1       de2       de3       de4       de5
 0.945732    sound     sound     sound     sound     sound
 0.054269   carious   carious   carious   carious   carious
****************************************************************
Blocks repartition of the variables for the class 2:
Block 1 contains the variables: de3 de4 with Rho= 0.2481778
     Tau       de3       de4
 0.653164    sound   carious
 0.346837   carious    sound
Block 2 contains the variables: de1 de2 de5 with Rho= 0
```

According to the previous outputs, the fitted model can be interpreted as follows

— The majority class ($\pi_1 = 0.86$) mainly gathers the sound teeth. There is a strong dependency between the five diagnoses ($\boldsymbol{\sigma}_1 = (\{1,2,3,4,5\})$ and $\rho_{11} = 0.35$). The dependency structure of the maximum dependency distribution indicates an over contribution of both modality interactions where the five dentists have the same diagnosis, especially when they claim that the tooth is sound ($\tau_{11}^{\text{all\_sound}} = 0.95$ and $\tau_{11}^{\text{all\_carious}} = 0.05$).

— The minority class ($\pi_2 = 0.14$) groups principally the carious teeth. There is a dependency between the dentists 3 and 4 which provides opposite diagnoses while the diagnoses of the other ones are independent given the class ($\boldsymbol{\sigma}_2 = (\{3,4\},\{1,2,5\})$, $\rho_{21} = 0.25$ and $\rho_{22} = 0$).

Best model visualization Finally, the function `plot()` provides a graphical summary of the parameters.

Clustericat script 5.8 (Graphical summary of the parameters).

```
# Function providing the graphical summary given by Figure 5.1
> plot(res)
```

On ordinates, the estimated classes are represented with respect to their proportion in decreasing order. Note that their corresponding area depends on their proportion. The cumulated proportions are indicated on the left side. On abscissa, 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 the strength of their dependencies (in decreasing order). The third one is the variables repartition per blocks. A black cell indicates that the variable is assigned into the block and a white cell indicates that, conditionally on this class, the variable is independent of the variables of this block. For example, this figure shows that the first class has a proportion of 0.86 and that all the variables are assigned into the same block.

5.2 CoModes

5.2.1 CoModes overview

Presentation The R package `CoModes` performs the clustering of categorical data according to the CMM model. We remind that, in this model, variables are grouped into conditionally independent blocks equal between classes and that each block follows a multinomial distribution per modes. All the functions of this package are implemented in R. So, they should be implemented in C++ in order to increase the computation speed.

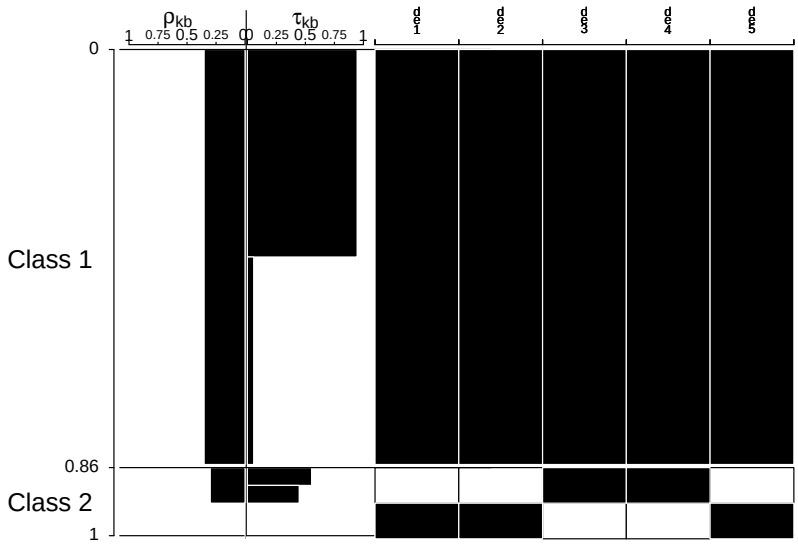


Figure 5.1 – Summary of the best MEDD according to BIC for the dentists data set.

Download CoModes The package is currently available on R-forge at the following url: https://r-forge.r-project.org/R/?group_id=1809. The installation and the loading of `CoModes` can be performed by using the following R scripts.

CoModes script 5.9 (CoModes installation).

```
# R install command
> install.packages("CoModes",
                   repos="http://R-Forge.R-project.org")
```

CoModes script 5.10 (CoModes loading).

```
# CoModes loading
> require(CoModes)
```

Estimation The parameter 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. Note that the Gibbs algorithm uses an efficient approximation of the integrated complete-data likelihood.

5.2.2 Main functions

Four functions compose the `CoModes` package: one function performs the cluster analysis and three functions help for the result interpretation.

The clustering function The cluster analysis can be performed with the function `CoModescluster()` taking seven arguments.

CoModes script 5.11 (Dentists data set clustering).

```
> CoModescluster(x, g, Gibbs_init= 2, Gibbs_iter= 50,
                 Gibbs_chauffe= 50, EM_init= 5, EM_tol= 0.001)
```

This function has two mandatory arguments:

— A data frame `x` to be analyzed whose columns are factors.

— An integer values `g` setting the number of classes.

Default values are taken for the five tuning arguments:

— The number of MCMC chains performed to select the best model is set by the argument `Gibbs_init` (default value is 2).

— The number of iterations of the Gibbs sampler is set by the argument `Gibbs_iter` (default value is 50).

— The number of iterations of the burn-in of the Gibbs sampler is set by the argument `Gibbs_chauffe` (default value is 50).

— The number of different initializations of the EM algorithm estimating the MLE for the best model according to the Gibbs sampler is set by the argument `EM_init` (default value is 5).

— The EM algorithm is stopped when the increase of the likelihood is smaller than `EM_tol` (default value is 0.001).

Three tool functions The `CoModes` package also provides tool functions like `summary()`, `barplot()` and `plot()` which respectively give a summary of the model, a graphical summary of the parameters and a scatter-plot of the individuals in the multiple correspondence map.

5.2.3 CoModes to cluster the dentists data set

Clustering with CMM We now display the results of the CMM model estimated with the R package `CoModes` by using the following script.

CoModes script 5.12 (Dentists data set clustering).

```
> res <- CoModescluster(dentist,3, Gibbs_init= 25,
                        Gibbs_iter= 500, EM_init= 25)
```

Model selection The values of the BIC criterion obtained by the three models (CIM, MEDD and CMM) are presented in Table 5.2.

The model fitting the best the data is the bi-component MEDD model. The CMM model fits the data better than the CIM model for a number of classes smaller than

g	1	2	3	4
CIM	-8766	-7511	**-7481**	-7503
MEDD	-7743	**-7473**	-7481	-7503
CMM	-8294	-7492	**-7481**	-7503

Table 5.2 – BIC criterion values for the CIM, the MEDD and the CMM models according to different numbers of classes for the dentistry data set. Best values are in bold.

three. Note that, when the class number is upper or equal to three, both MEDD and CMM models are equivalent to the CIM model.

A possible reason explaining the poor performance of the CMM model could be its constraint of the equality between class of the repartition of the variables into blocks. We remind that this assumption is not made by the MEDD model.

In order to illustrate the tools functions of `CoModes`, we now analyze the bi-component CMM model.

Bi-component CMM interpretation The model interpretation of the CMM model is facilitated by three tools function. The function `summary()` provides a general overview of the model (information criteria, numbers of modes, τ_{kb}, κ_{kb}).

CoModes script 5.13 (Dentists data set clustering).

```
# Function providing a model overview
> summary(res)


Number of variables: 5     Number of individuals: 3869
Number of modalities: 2 2 2 2 2
Class number: 2     log-likelihood: -7434.628     BIC: -7492.453

************************************
Mode number:
          de1-de2-de3   de4   de5
 Class 1          5     1     1
 Class 2          4     1     1
************************************
Tau index:
          de1-de2-de3       de4         de5
 Class 1   0.8495717   0.5477190   0.8945463
 Class 2   1.0000000   0.9798947   0.7185214
************************************
Kappa index:
          de1-de2-de3   de4   de5
 Class 1   0.7142857     1     1
 Class 2   0.5714286     1     1
```

The function `barplot()` provides a graphical summary of the parameters. Indeed, it plots a barplot reflecting the probability of the modes per class for each block by ordering the modality crossings according to their posterior probability.

CoModes script 5.14 (Dentists data set clustering).

```
# Barplot of the parameters presented by Figure 5.2
> barplot(res)
```

The majority class (displayed in gray) is mainly composed with the sound diagnoses. The second class (displayed in black) is composed with teeth diagnosed as carious by some dentists especially the fifth. Note that the dentist 4 mainly diagnoses the teeth as sound since its corresponding variable has a mode in this location for both classes.

Bi-component CMM visualization The function `plot()` provides a scatter-plot of the individuals, by indicating their class membership according to the MAP rule, in a correspondence analysis map where the axes are chosen by the user.

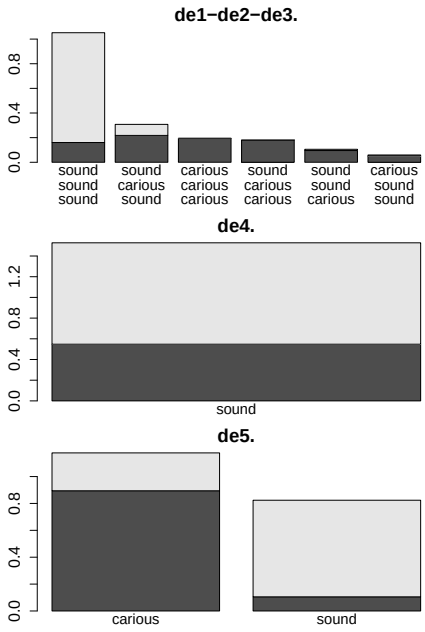


Figure 5.2 – Summary of the CMM parameters.

CoModes script 5.15 (Dentists data set clustering).

```
# Scatter-plot of the individuals presented by Figure 5.3
> plot(res, c(1,2))
```

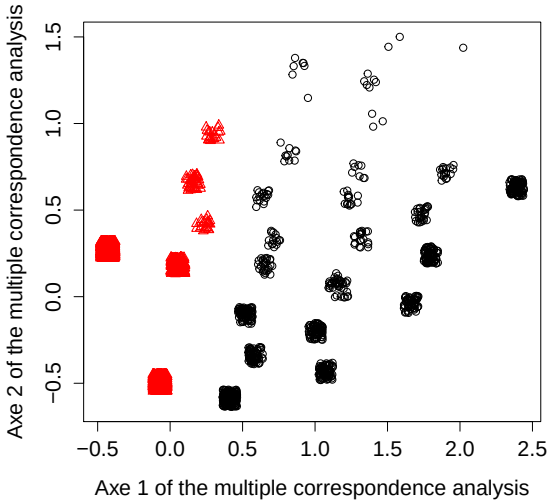


Figure 5.3 – Scatter-plot in the first correspondence analysis map. Individuals affected to the majority class are displayed by red triangles while those affected to the minority class are displayed by black circles. An i.i.d. uniform noise on $[0, 0.1]$ has been automatically added on both axes for each individual in order to improve visualization.

Conclusion of Part I

We have seen that the classical latent class model allows to fit well the small data sets thanks to its sparsity induced by its conditional independence assumption. For such data sets, more complex models taking into account the intra-class dependencies are irrelevant since the information of conditional dependency is not present.

When the number of individuals is sufficiently large according to the number of variables, the latent class model can be biased when its conditional independence assumption is violated. We have presented three main models of the bibliography relaxing this assumption but they are facing with different difficulties: model selection, instability or an interpretation of classes which is performed throughout another latent variable.

We have also proposed two mixture models which are specific versions of the log-linear mixture model. They consider the main intra-class dependencies thanks to a component distribution per independent blocks. Their main strength is that both models can be summarized by few meaningful parameters. Indeed, the MEDD model provides one coefficient and one dependency relationship by blocks of variables, while the CMM model provides few characteristic modes and two indicators of the dependency strength per block. Note that both models can consider interactions among more than two variables while the usual ones take into account the interactions of order one or two.

As all models of the log-linear mixture family, both proposed models are facing with a complex challenge for the model selection. This difficulty is double since the number of competing models can be huge. Furthermore, the information criteria are generally asymptotic, so they are failing when the number of models is large according to the sample size (for instance the BIC criterion is biased to select the mode number). We have proposed a MCMC algorithm performing a random walk among the models, in order to reduce this drawback. However, the computation time increases with the size of the model space. This phenomenon is a strong obstacle to the analysis by the proposed models of data sets with a large number of variables. Thus, the MEDD model is reserved for the cluster analysis of data sets with few variables. The CMM model is less complex and its model selection does not require any parameter estimate. So, the CMM model can cluster more complex data sets. However, our advice is to use these models on data sets having at most 20 variables. Indeed, when the number of variables is large, there are too many models in competition. So, the Gibbs algorithm requires too many iterations to sample according to its stationary distribution. In such a case, a pragmatic approach could consists in the estimation of a good model (but not the best one). Thus, the model

selection could be performed by a deterministic but sub-optimal approach like the forward method.

Finally, the approaches performing the model selection often require to infer the parameters for each candidate model while the only interpreted ones are those relative to the best model. The "Grail" would consists in performing the model selection without needing the parameters of the model candidates and after to infer the parameters only for the best model. Indeed, models having this property will simplify the challenge of the model selection.

Part II

Model-based clustering for mixed data

This part, devoted to the cluster analysis of mixed data, is split in three chapters.
The first one presents an overview of the clustering approach specific to the mixed data sets. We mainly focus on the two main mixture models which relax the conditional independence assumption and which fill in part of the lack of multivariate distributions for mixed data.
The second chapter presents a mixture model to cluster mixed data sets with continuous and categorical variables. This model derives from the multilevel latent class model developed for the categorical data analysis.
The third chapter presents one of the main contribution of this thesis. It consists in the mixture model of Gaussian copulas which allows to cluster data sets with any kind of variables admitting a cumulative distribution function. These results are part of a submitted article.

Une fois n'est pas coutume,
ni deux d'ailleurs,
ni trois.
À vrai dire,
on a jamais su
à partir de combien
c'était coutume
Stéphane De Groodt — Voyages en absurdie

Table of Contents

Chapter 6

Cluster analysis of mixed data sets: state of the art

The purpose of this chapter is to present the main approaches to cluster mixed data sets.
Firstly, we emphasize the specificity of the mixed data for the cluster analysis.
Secondly, we present some naive approaches to perform the cluster analysis of such data.
Finally, we detail the two most relevant approaches to cluster mixed data sets.

Génétique en bandoulière
Des chromosomes dans l'atmosphère
Des taxis pour les galaxies
Et mon tapis volant lui
Noir Désir — Le vent nous portera

6.1 Challenge of cluster analysis for mixed data

Introduction Nowadays, many data sets are often composed with mixed variables (different kinds of variables in the data set). So, it is essential to be able to cluster such data sets. The difficulty, inherent to the cluster analysis of mixed data performed by mixture models, is the lack of multivariate distribution for such data. Indeed, if the Gaussian (respectively the Poisson and the multinomial) distributions are reference to cluster continuous (respectively integer and categorical) data, there is no reference distribution for mixed data. So, the easiest approach consists in assuming the conditional independence between the variables and in selecting classical one-dimensional margin distributions for each component. However, this approach can lead to biases. Moreover, some models approach the distribution of intra-class correlated mixed data, but they are not easily meaningful. Indeed, the one-dimensional margins of each component do not follow classical distributions. As the mixture models are used to cluster, this objective appears to be crucial for us since it provides meaningful models.

Two crucial objectives This chapter proposes an overview of the clustering approaches for mixed data. It puts the light on the importance to provide multivariate mixture models which respect both following objectives:

1. *To provide classical one-dimensional margin distributions for each component.*
2. *To modelize the intra-class dependencies.*

These two objectives aim to simplify the model interpretation. Indeed, the practitioner is in a usual framework when the one-dimensional margins of the components are classical. Moreover, its class interpretation is more precise when the intra-class dependencies are modelized. Thus, the mixture models presented in the following chapters aim at respecting both objectives.

The data Throughout this part, we consider the e-variate vector of mixed variables denoted by $\boldsymbol{x}_i = (x_i^1, \ldots, x_i^c, \boldsymbol{x}_i^{c+1}, \ldots, \boldsymbol{x}_i^e)$. We denote by $\boldsymbol{x}_i^{\mathrm{C}} = (x_i^1, \ldots, x_i^c)$ its subset of the c continuous variables. In the same way, we denote by $\boldsymbol{x}_i^{\mathrm{D}} = (\boldsymbol{x}_i^{c+1}, \ldots, \boldsymbol{x}_i^e)$ its subset of the d discrete variables, where $c + d = e$. Note that the term discrete will be specified for each presented model. We denote by m the number of modality crossings of the set of the categorical variables of $\boldsymbol{x}_i^{\mathrm{D}}$.

Structure of this chapter Section 6.2 firstly presents three naive methods which are not relevant since they do not consider each variable in its native space. Secondly, this section formulates two extensions of classical methods to perform the cluster analysis of mixed variables. The two following sections present the two most relevant approaches to cluster mixed data. Section 6.3 details the mixture of location models and its extension per block. Section 6.4 introduces the underlined Gaussian mixture model. A conclusion is given in Section 6.5.

6.2 Overview of simple methods to cluster mixed data

6.2.1 Naive methods

In order to emphasize the difficulties inherent to the mixed data clustering, we enumerate three naive (but not efficient) methods which permit to cluster such data. However, all of these methods have a main drawback. Indeed, either they do not respect the kind of each variable, either they do not consider each variable in its native space.

Whole continuous method One may be tempted to cluster the discrete variables as if they were continuous. Thus, this method consists in converting the categorical and ordinal attribute values to numeric values. A method specific to the continuous data is then applied to perform the cluster analysis. This approach makes a very strong assumption for the ordinal variables. Indeed, it assumes that there is the same gap between all the couples of successive modalities. Moreover, this approach

has to be barred in the presence of categorical variables. Indeed, it assumes a meaningless order between the modalities. For instance, it is not possible to give numeric variables to categorical values like color or hobby.

Whole discrete method This method consists in a discretization of the continuous variables. Thus, a method specific to the categorical variables is used to perform the cluster analysis. The choices of the numbers of modalities and of the bound locations are delicate. However, these choices are crucial since they influence the results of the cluster analysis. Moreover, whatever are the numbers of categories, the discretization process leads to a loss of information. Finally, the intra-class dependencies are difficulty modelized since all the variables are considered as categorical (see Part I).

Factorial approach The Multiple Factorial Analysis method permits to project individuals described by categorical variables in a "continuous" space. Thus, by replacing the discrete variables by their factor coordinates, any method specific to the continuous variables can be used to cluster the data set. However, note that, even if this space is continuous, individuals can take only a finite number of values. This phenomenon also increases the risk of degeneracy. Moreover, the results are less meaningful since the variables are not clustered in their native space. Indeed, the interpretation of the classes is done by the parameters of the factorial space.

The three methods presented above are not efficient since they do not respect the nature of each variable. Thus, in the following, all the studied methods modelize the distribution of all the variables in their native space.

6.2.2 Extension of classical methods for mixed data

K-means algorithm The K-means algorithm only requires a definition of a distance between the individuals to cluster any kind of data. Different distance measures can be selected for mixed data (see, for instance, the suggestions of Z. Huang [Hua98] and of A. Ahmad and L. Dey [AD07]). Obviously, this approach keeps the drawbacks of the geometric methods discussed in Section 1.1.2.

Conditional independence mixture model The lack of reference distribution for mixed data is a problem to perform the cluster analysis with mixture models. This problem is easily avoided by the conditional independence model (see Section 1.2.3). Indeed, each component distribution is defined by the product of univariate distributions. Thus, classical distributions can be used as the one-dimensional margin distributions of the components as proposed by J. Bacher [Bac00] and by I. Moustaki and I. Papageorgiou [MP05]. The idea which consists in setting the one-dimensional margin distributions by classical distributions is a major notion. However, the conditional independence model is biased when its assumption is violated. Now, we present two main approaches which relax the conditional independence assumption.

6.3 Mixture of location models and its extension per blocks

The data The *location mixture model*, introduced by W.J. Krzanowski [Krz93], allows to cluster data sets with continuous and categorical variables.

Main idea It concatenates the whole categorical variables into a single one which follows a full multinomial distribution. Moreover, it assumes that the continuous variables follow a multivariate Gaussian distribution conditionally on the class and on each modality crossing. More precisely, its means depend on both class and categorical variables. Thus, the conditional dependency between the whole variables is taken into account.

6.3.1 Location model

Aim I. Olkin and R.F. Tate [OT61] note that data arisen from experimentations in psychology often contain both discrete and continuous variables. So, they introduce the location model to have measures of association between the variables of such data sets.

Main idea The *location model* defines the multinomial distribution on the whole categorical variables and a multivariate Gaussian distribution on the continuous variables conditionally on the categorical ones. More precisely, the set of the categorical variables $\boldsymbol{x}_i^{\mathrm{D}}$ is considered as one categorical variables which follows a free multinomial distribution $\mathcal{M}_m(\lambda_1, ..., \lambda_m)$. Thus, λ_h denotes the probability that $\boldsymbol{x}_i^{\mathrm{D}}$ takes the modality crossing h. Moreover, conditionally on $\boldsymbol{x}_i^{\mathrm{D}}$ taking the modality crossing h, the c-variate continuous variable $\boldsymbol{x}_i^{\mathrm{C}}$ follows a c-variate Gaussian distribution $\mathcal{N}_c(\boldsymbol{\mu}^h, \boldsymbol{\Sigma})$. Thus, the categorical variables influence the mean of the continuous variables but not their dispersion.

Notations As the set of the categorical variables $\boldsymbol{x}_i^{\mathrm{D}}$ is considered by the location model as one categorical variables, we use a complete disjunctive coding as such $x_i^{\mathrm{D}h} = 1$ if the individual takes the modality crossing h and $x_i^{\mathrm{D}h} = 0$ otherwise.

Definition 6.1 (Location model)**.** The vector of mixed variables $\boldsymbol{x}_i = (\boldsymbol{x}_i^{\mathrm{C}}, \boldsymbol{x}_i^{\mathrm{D}})$ is drawn by a location model if its pdf is written as follows

$$p(\boldsymbol{x}_i; \boldsymbol{\alpha}) = \prod_{h=1}^{m} \left(\lambda^h \phi_c(\boldsymbol{x}_i^{\mathrm{C}}; \boldsymbol{\mu}^h, \boldsymbol{\Sigma})\right)^{x_i^{\mathrm{D}h}}, \tag{6.1}$$

where $\boldsymbol{\alpha}$ groups the dispersion matrix $\boldsymbol{\Sigma}$ and the vector $(\lambda^h, \boldsymbol{\mu}^h; h = 1, \ldots, m)$.

6.3.2 Mixture of location models

The location model was extended to the mixture framework. Indeed, the mixture of location models was used in discriminant analysis by W.J. Krzanowski [Krz93]

and in cluster analysis by C.J. Lawrence and W.J. Krzanowski [LK96] in order to take into account the intra-class dependencies.

Definition 6.2 (Mixture of location models)**.** The vector of mixed data $\boldsymbol{x}_i = (\boldsymbol{x}_i^{\text{C}}, \boldsymbol{x}_i^{\text{D}})$ is drawn by a mixture of location models if the pdf of its component k is written as follows for $k = 1, \ldots, g$

$$p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) = \prod_{h=1}^{m} \left(\lambda_k^h \phi_c(\boldsymbol{x}_i^{\text{C}}; \boldsymbol{\mu}_k^h, \boldsymbol{\Sigma})\right)^{x_i^{\text{D}h}}, \tag{6.2}$$

where $\boldsymbol{\alpha}_k$ groups the dispersion matrix $\boldsymbol{\Sigma}$ and the vector $(\lambda_k^h, \boldsymbol{\mu}_k^h; h = 1, \ldots, m)$.

Intra-class dependencies The mixture of location models considers all the intra-class dependencies per couple of variables, as follows.

— If both variables are categorical, their intra-class dependencies are modelized by the full multinomial distributions.
— If both variables are continuous, their intra-class dependencies are modelized by the bivariate Gaussian distributions.
— If one variable is categorical and one variable is continuous, their intra-class dependencies are modelized by the influence of the categorical variable on the means of the Gaussian distributions of the continuous variable.

Identifiability As pointed-out by A. Willse and R.J. Boik [WB99], the mixture of location models is not identifiable because of the indeterminacy of class memberships at each location. In order to overcome this lack of identifiability, these authors add some constraints on the mean parameters of the Gaussian distributions.

One-dimensional margin distributions We have emphasized that the conditional independence model is meaningful, since its one-dimensional margin distributions of its components are classical (for instance, they consist in multinomial or a Gaussian distributions). For the mixture of location models, the one-dimensional margin distributions of the categorical variables for a component are classical since they are multinomial distributions. However, the one-dimensional margin distributions of the continuous variables for a component are not classical. Indeed, they consist in a mixture of homoscedastic Gaussian with m components.

Parameter estimation The inference can be easily performed in both frequentist and Bayesian frameworks even if the authors only presented it in the frequentist one. The MLE can be obtained by the following EM algorithm

Algorithm 6.3 (EM algorithm for the mixture of location models)**.**

Starting from an initial value $\boldsymbol{\theta}^{[0]}$, iteration $[r]$ is written as

— **E step**: calculate conditional probabilities

$$t_{ik}(\boldsymbol{\theta}^{[r]}) = \frac{\pi_k^{[r]} p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k^{[r]})}{p(\boldsymbol{x}_i; \boldsymbol{\theta}^{[r]})}. \tag{6.3}$$

— **M step**: maximization of the expectation of the complete-data log-likelihood

$$\pi_k^{[r+1]} = \frac{n_k^{[r]}}{n},\ \lambda_k^{h[r+1]} = \frac{1}{n_k^{[r]}} \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]}) x_i^{\mathrm{D}h}, \tag{6.4}$$

$$\boldsymbol{\mu}_k^{h[r+1]} = \frac{1}{n_k^{[r]}} \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]}) x_i^{\mathrm{D}h} \boldsymbol{x}_i^{\mathrm{C}}, \tag{6.5}$$

$$\boldsymbol{\Sigma}^{[r+1]} = \frac{1}{n^{[r]}} \sum_{k=1}^{g} \sum_{h=1}^{m_j} \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]}) x_i^{\mathrm{D}h} (\boldsymbol{x}_i^{\mathrm{C}} - \boldsymbol{\mu}_k^{h[r+1]})'(\boldsymbol{x}_i^{\mathrm{C}} - \boldsymbol{\mu}_k^{h[r+1]}), \tag{6.6}$$

where $n_k^{[r]} = \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]})$.

The MAPE can be easily obtained by selecting a usual prior. This prior assumes independence between the parameters and selects conjugate distributions for each parameters. It is also easy to build a Gibbs sampler since the parameters have explicit posterior distributions.

6.3.3 Mixture of blocks of location model

Main idea The number of parameters required by the mixture of location models increases with the number of categorical variables and with the number of their modalities. Thus, M. Jorgensen and L. Hunt [JH96, HJ99] propose an extension of this model. In their extension, the variables are split into conditionally independent blocks such that each block is composed with at most one categorical variable. Moreover, each block of variables follows a location model.

Definition 6.4 (Mixture of blocks of location model)**.** The vector $\boldsymbol{x}_i$ composed with continuous and categorical variables arises from a mixture of blocks of location model if the pdf of its component k is written as follows for $k = 1, \ldots, g$

$$p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) = \prod_{b=1}^{\mathrm{B}} p(\boldsymbol{x}_i^{\{b\}}; \boldsymbol{\alpha}_{kb}), \tag{6.7}$$

where $\boldsymbol{\alpha}_k = (\boldsymbol{\alpha}_{kb}; b = 1, \ldots, \mathrm{B})$ and if the pdf of block b for component k is written

as follows for all $b = 1, \ldots, \text{B}$

$$p(\boldsymbol{x}_i; \boldsymbol{\alpha}_{kb}) = \begin{cases} \phi_{d^{\{kb\}}}(\boldsymbol{x}_i^{\{b\}\text{C}}, \boldsymbol{\mu}_{kb}, \boldsymbol{\Sigma}_{kb}) & \text{if } \boldsymbol{x}_i^{\{b\}} \text{ is continuous} \\ \prod_{h=1}^{m^{\{b\}}} (\lambda_{kh})^{x_i^{\{b\}\text{D}h}} & \text{if } \boldsymbol{x}_i^{\{b\}} \text{ is categorical} \\ \prod_{h=1}^{m^{\{b\}}} \left(\lambda_{kb}^h \phi_{d^{\{kb\}}-1}(\boldsymbol{x}_i^{\{b\}\text{C}}; \boldsymbol{\mu}_{kb}^h, \boldsymbol{\Sigma}_{kb})\right)^{x_i^{\{b\}\text{D}h}} & \text{if } \boldsymbol{x}_i^{\{b\}} \text{ is mixed,} \end{cases} \tag{6.8}$$

where $\boldsymbol{x}_i^{\{b\}\text{C}}$ and $\boldsymbol{x}_i^{\{b\}\text{D}}$ are respectively the continuous and the categorical variables of block b.

Related models

— If there are only continuous variables (*i.e.* $c = e$ and $d = 0$) and if all the variables are affiliated into the same block (*i.e.* $\boldsymbol{x}_i^{\{1\}} = \boldsymbol{x}_i$), then the model is equivalent to the heteroscedastic Gaussian mixture model.

— If there are only continuous variables (*i.e.* $c = e$ and $d = 0$) and if each block is composed with only one variable (*i.e.* $\boldsymbol{x}_i^{\{b\}} = \boldsymbol{x}_i^b$, for $b = 1, \ldots, c$), then the model is equivalent to the Gaussian mixture model with conditional independence assumption (*i.e.* $\boldsymbol{\Sigma}_k$ is diagonal).

— If there are only categorical variables (*i.e.* $c = 0$ and $d = e$) then the model is equivalent to the latent class model (see Section 2.3.1).

Parameter estimation The inference is easily performed. Indeed, the conditional independence between the blocks allows to adapt the EM algorithm presented in Algorithm 6.3, in order to obtain the MLE of the mixture of blocks of location model. The Bayesian inference could be performed by a Gibbs sampler if the priors are assumed to be independent and follow conjugate distributions.

Model estimation of the repartition of the variables into blocks The authors estimate the repartition of the variables into blocks by an ascending method with a fixed number of classes. Indeed, an exhaustive approach is not doable (see Section 3.5). The aim of this ascending method is to optimize an information criterion. This method is initialized by the locally independent model. Then, different models are proposed by using the intra-class dependencies computed with the current model.

Performances of the mixture of blocks of location model As presented in [HJ11], the mixture of blocks of location model can outperform the locally independent model. However, this model has two main drawbacks. The first one is about the class interpretation, since the one-dimensional margin distribution of a component is not classical if the variable is continuous. The second one is about the difficulty to perform model selection. Indeed, the proposed approach can be sub-optimal to select the repartition of the variables into blocks. Furthermore, the choice of the correlation coefficient between a continuous variable and a categorical

one is subjective. However, this choice is crucial since it determines the candidates during the model estimation.

6.4 Underlined Gaussian mixture model

Main idea The underlined Gaussian mixture model, introduced by B.S. Everitt [Eve88], performs the cluster analysis of data sets with continuous and ordinal variables. Its main assumption is that the observed ordinal and binary variables are generated from underlying unobservable continuous variables according to the values of a set of thresholds.

Remark 6.5 (The categorical variables are not allowed)**.** The categorical variables (except the binary ones) cannot be modelized by the underlined Gaussian mixture model. Indeed, this model assumes an order between the modalities which is not present for such variables.

6.4.1 Description of the underlined Gaussian mixture model

Gaussian variable We consider the vector $\boldsymbol{y}_i = (\boldsymbol{x}_i^{\text{C}}, \boldsymbol{y}_i^{\text{D}})$ where $\boldsymbol{y}_i^{\text{D}}$ is a continuous vector of size d. The vector $\boldsymbol{y}_i$ is assumed to be drawn by the homoscedastic Gaussian mixture model whose the pdf is

$$p(\boldsymbol{y}_i; \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \phi_e(\boldsymbol{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}). \tag{6.9}$$

Gaussian latent variable In practice the variables $\boldsymbol{y}_i^{\text{D}}$ are not observed. However, they are related to the set of the observed discrete variables $\boldsymbol{x}_i^{\text{D}}$ as follows

$$\forall j = c+1, \ldots, e,\ \boldsymbol{x}_i^{jh} = 1 \text{ if } b_k^{jh} < y_i^{\text{D}j} \leq b_k^{jh+1}, \tag{6.10}$$

where $b_k^{jh} < b_k^{jh+1}$ for $j = c+1, \ldots, e$ and $h = 1, \ldots, m_j$ and where $b_k^{j1} = -\infty$ and $b_k^{jm_j+1} = \infty$. The bounds b_k^{jh} determine the observed discrete variables from the latent continuous ones. Thus, we obtain that the observed variables $\boldsymbol{x}_i = (\boldsymbol{x}_i^{\text{C}}, \boldsymbol{x}_i^{\text{D}})$ have the following pdf

$$p(\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \int_{\mathcal{S}_k(\boldsymbol{x}_i^{\text{D}})} \phi_e(\boldsymbol{y}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) d\boldsymbol{y}_i^{\text{D}}, \tag{6.11}$$

where $\mathcal{S}_k(\boldsymbol{x}_i^{\text{D}})$ is the domain of the integration of latent Gaussian variables $\boldsymbol{y}_i^{\text{D}}$ related to the observed discrete variables $\boldsymbol{x}_i^{\text{D}}$. More precisely, $\mathcal{S}_k(\boldsymbol{x}_i^{\text{D}}) = \mathcal{S}_k^{c+1}(\boldsymbol{x}_i^{c+1}) \times \ldots \times \mathcal{S}_k^{e}(\boldsymbol{x}_i^{e})$ where the interval $\mathcal{S}_k^{j}(\boldsymbol{x}_i^{j})$ is defined for $j = c+1, \ldots, e$ as such $\mathcal{S}_k^{j}(\boldsymbol{x}_i^{j}) =]b_k^{jh}, b_k^{jh+1}]$ if $x_i^{jh} = 1$.

Alternative form of the pdf An alternative (and more friendly) form of the pdf defined by (6.11) is obtained by noting that the conditional distribution of $\boldsymbol{x}_i^{\text{D}}$ given $\boldsymbol{x}_i^{\text{C}}$ is an underlined Gaussian one. This distribution has the mean $\boldsymbol{\mu}_k^{\text{D}|\text{C}}$ and the covariance matrix $\boldsymbol{\Sigma}^{\text{D}|\text{C}}$ which are defined by

$$\boldsymbol{\mu}_k^{\text{D}|\text{C}} = \boldsymbol{\mu}_k^{\text{D}} + \boldsymbol{\Sigma}_{\text{DC}}\boldsymbol{\Sigma}_{\text{CC}}^{-1}(\boldsymbol{x}_i^{\text{C}} - \boldsymbol{\mu}_k^{\text{C}}) \text{ and } \boldsymbol{\Sigma}^{\text{D}|\text{C}} = \boldsymbol{\Sigma}_{\text{DD}} - \boldsymbol{\Sigma}_{\text{DC}}\boldsymbol{\Sigma}_{\text{CC}}^{-1}\boldsymbol{\Sigma}_{\text{CD}}, \tag{6.12}$$

where $\boldsymbol{\mu}_k^{\text{C}}$ and $\boldsymbol{\mu}_k^{\text{D}}$ are respectively the means of $\boldsymbol{x}_i^{\text{C}}$ and of $\boldsymbol{x}_i^{\text{D}}$ and where the covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\text{CC}} & \boldsymbol{\Sigma}_{\text{CD}} \\ \boldsymbol{\Sigma}_{\text{DC}} & \boldsymbol{\Sigma}_{\text{DD}} \end{bmatrix}$ is decomposed into sub-matrices. For instance, $\boldsymbol{\Sigma}_{\text{CC}}$ is the sub-matrix of $\boldsymbol{\Sigma}$ composed by the rows and the columns related to the observed continuous variables. This alternative form of the pdf allows us to define the underlined Gaussian mixture model.

Definition 6.6 (Underlined Gaussian mixture model)**.** Let $\boldsymbol{x}_i$ the vector of e mixed variables drawn be the underlined Gaussian mixture model. Its pdf is written as follows

$$p(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \phi_c(\boldsymbol{x}_i^{\text{C}};\boldsymbol{\mu}_k^{\text{C}},\boldsymbol{\Sigma}_{\text{CC}}) \int_{\mathcal{S}_k(\boldsymbol{x}_i^{\text{D}})} \phi_d(\boldsymbol{y}_i^{\text{D}};\boldsymbol{\mu}_k^{\text{D}|\text{C}},\boldsymbol{\Sigma}^{\text{D}|\text{C}}) d\boldsymbol{y}_i^{\text{D}}. \tag{6.13}$$

Crucial condition for model identifiability Since the latent vector $\boldsymbol{y}_i^{\text{D}}$ is not observed, there is no information on its mean and variance. Thus, the model assumes that the elements of $\boldsymbol{\mu}_k^{\text{D}}$ are null and that the diagonal elements of $\boldsymbol{\Sigma}_{\text{DD}}$ are equal to one.

6.4.2 Estimation of the underlined Gaussian mixture model

The inference by maximization of the log-likelihood function is not easy because of the presence of d-dimensional integrals having no explicit form when $\boldsymbol{\Sigma}$ is not diagonal.

E.S. Everitt proposes to perform the inference by using simplex method. Note that its approach limits at four the number of discrete variables.

6.5 Conclusion

We have pointed out the importance to consider all the variables in their native space. However, the mixed data are not easily clustered by mixed models because of the lack of standard multivariate distribution for such variables. Thus, the aim is to propose relevant multivariate distribution for mixed data. We have put the light on the importance that the one-dimensional margins of this distributions are classical and that the dependencies are modelized.

The model presented in Chapter 7 allows to perform the cluster analysis of data sets with continuous and categorical variables by achieving these objectives. Note that this situation (data set with continuous and categorical variables) is the most studied one.

In the bibliography, the authors do not study the case of mixed variables with integer values. Thus, the model presented in Chapter 8 is of interest. Indeed, it is very general since it allows to cluster data sets with any kind of variables admitting a c.d.f.

Chapter 7

Model-based clustering of Gaussian and logistic distributions

This chapter introduces a sparse mixture model for data sets with continuous and categorical variables. The component distributions of the continuous variables are Gaussian. Moreover, the categorical variables are assumed to be independent conditionally on the class and on the continuous variables. Finally, conditionally on the continuous variables, the component distributions of the categorical variables are linear logistic distributions where few parameters are not zero in the regression.
The maximum likelihood inference and the model selection are simultaneously performed by a GEM *algorithm for a fixed number of classes.*
Numerical experiments illustrate the model relevance in a cluster analysis and in a semi-supervised classification when few individuals are labeled.

Si estirem tots, ella caurà
I molt de temps no pot durar
Segur que tomba, tomba, tomba
Ben corcada deu ser ja.
Si tu l'estires fort per acqui
I jo l'estiro fort per alla
Segur que tomba, tomba, tomba,
I ens podrem alliberar.
Lluis Llach — L'estaca

7.1 Introduction

We present a mixture model to cluster data sets with continuous and categorical variables. This model has a double objective: to provide classical one-dimensional margin distributions for each component and to modelize the intra-class dependencies.

For such a model, the continuous variables follow a multivariate Gaussian distribution for each component. Conditionally on the class and on the continuous variables, the categorical variables are assumed to be independently drawn by linear logistic distributions. The resulting model is also named: *mixture model of Gaussian and logistic distributions*.

The linear logistic regressions are classically used by mixture models for categorical data [For92]. Moreover, the multilevel latent class model [Ver03] uses latent continuous variables to modelize the intra-class dependencies between the categorical variables. So, it is natural to use a mixture model of Gaussian and logistic distributions when both of the continuous and categorical kinds of variables are observed.

A parsimonious version of this model is introduced by adding some constraints on the logistic parameter space. So, the resulting model is more easily interpretable and can perform a better trade off between the bias and the variance. For a fixed number of classes, the selection of the parsimonious model and the parameter estimation are simultaneously performed by a GEM algorithm which optimizes an information criterion. During our numerical experiments, we illustrate the relevance of the BIC criterion compared to the AIC criterion.

Finally, even if the model is introduced to perform a cluster analysis, it can be applied for a semi-supervised classification [CSZ+06]. Indeed, it is known that the generative approaches can outperform the methods specific of the classification challenge. This phenomenon is particularly observed when few observations are labeled [DMD06]. Indeed, these methods exploit the informations present in both labeled and unlabeled data while discriminative approaches take only into account the labeled data information. Indeed, they learn a classification rule only on the labeled data.

Structure of this chapter This article is organized as follows. Section 7.2 presents the mixture model of Gaussian and logistic distributions for data sets with continuous and categorical variables. This model performs the cluster analysis by providing classical one-dimensional distributions for each component and by modelizing the intra-class dependencies. In a clustering framework, Section 7.3 is devoted to the maximum likelihood estimation. Section 7.4 presents the GEM algorithm which simultaneously performs the estimation of both model and parameters by optimizing an information criterion. Section 7.5 presents different numerical experiments. They illustrate the relevance of the BIC criterion to perform the model selection. Moreover, they show the performances of the estimation algorithm and the model robustness. Section 7.6 presents one application in clustering and one application in semi-supervised classification on two real data sets. A conclusion is given in Section 7.7.

The data Let $\boldsymbol{x}_i = (x_i^1, \ldots, x_i^c, \boldsymbol{x}_i^{c+1}, \ldots, \boldsymbol{x}_i^e)$ be the e-variate vector of mixed variables. The first c variables are continuous and this subset of variables is denoted by $\boldsymbol{x}_i^{\mathrm{c}}$. The last d variables are categorical variables using a disjunctive coding and this subset of variables is denoted by $\boldsymbol{x}_i^{\mathrm{p}}$. Note that $c + d = e$.

7.2 Mixture model of Gaussian and logistic distributions

Aim The model performs the cluster analysis of continuous and categorical data set with a double objective: to provide classical one-dimensional margin distributions for each component and to modelize the intra-class dependencies.

Main idea 1 The pdf of each component is defined by the product between the pdf of the whole continuous variables and the pdf of the whole categorical variables conditionally on the whole continuous variables. More precisely, concerning the continuous variables, they follow a multivariate Gaussian distribution for each component. Concerning the categorical variables, the model assumes their independence conditionally on both the class membership and the continuous variables.

Definition 7.1 (Mixture model related to Main idea 1)**.** The vector $\boldsymbol{x}_i$ is drawn by a mixture model respecting Main idea 1 if the pdf of its component k is written as follows for $k = 1, \ldots, g$

$$\begin{aligned} p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) &= p(\boldsymbol{x}_i^{\mathrm{c}}; \boldsymbol{\alpha}_k) p(\boldsymbol{x}_i^{\mathrm{d}} | \boldsymbol{x}_i^{\mathrm{c}}; \boldsymbol{\alpha}_k) \\ &= \phi_c(\boldsymbol{x}_i^{\mathrm{c}}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \prod_{j=c+1}^{e} p(\boldsymbol{x}_i^j | \boldsymbol{x}_i^{\mathrm{c}}; \boldsymbol{\beta}_{kj}), \end{aligned} \tag{7.1}$$

where $\boldsymbol{\alpha}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\beta}_k)$ denotes the whole component parameters, where the vector $\boldsymbol{\mu}_k \in \mathbb{R}^d$ denotes the mean of the continuous variables for component k and where the matrix $\boldsymbol{\Sigma}_k$ denotes their covariance matrix. The vector $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{kj}; j = c+1, \ldots, e)$ denotes the whole parameters of component k which are related to the categorical variables. Indeed, the vector $\boldsymbol{\beta}_{kj}$ groups the parameters related to the distribution of the categorical variable $\boldsymbol{x}_i^j$ for component k.

Main idea 2 The model assumes that, for each component, the distribution of each categorical variable is a linear logistic regression whose the explanatory variables are the continuous ones.

Definition 7.2 (Mixture model of logistic regressions)**.** With the notation $x^0 = 1$, the component distributions of $\boldsymbol{x}_i^j$ (for $j = c+1, \ldots, e$) are defined, by the following pdf for $k = 1, \ldots, g$

$$p(\boldsymbol{x}_i^j | \boldsymbol{x}_i^{\mathrm{c}}; \boldsymbol{\beta}_{kj}) = \prod_{h=1}^{m_j} \left(\frac{\exp\left(\sum_{j'=0}^{c} \beta_{kj}^{j'h} x_i^{j'}\right)}{\sum_{h'=1}^{m_j} \exp\left(\sum_{j'=0}^{c} \beta_{kj}^{j'h'} x_i^{j'}\right)} \right)^{x_i^{jh}}, \tag{7.2}$$

where the parameters $\boldsymbol{\beta}_{kj} = (\beta_{kj}^{j'h}; j' = 0, \ldots, c; h = 1, \ldots, m_j) \in \mathbb{R}^{c+1}$ denotes the logistic parameters of the categorical variable $\boldsymbol{x}_i^j$ for class k. In order to insure the model identifiability, we put $\forall (k, j, j')$, $\beta_{kj}^{j'1} = 0$. Finally, note that the parameter β_{kj}^{0h} is the intercept of the logistic regression while the other parameters $\beta_{kj}^{j'h}$ (for $j' = 1, \ldots, c$) are the slope parameters.

Potential large number of parameters The model can require a large number of parameters. Indeed, the number of parameters involved by the model is equal to

$$(g-1)+g\left(\frac{c(c+3)}{2}\right)+g\sum_{j=c+1}^{e}(m_j-1)(c+1). \tag{7.3}$$

In order to obtain a better bias/variance tradeoff, we introduce a parsimonious version of the model by reducing the space of the logistic coefficients.

Sparse logistic functions for the mixed conditional dependencies The sparsity of the model is defined by the discrete parameters $\boldsymbol{\delta}_{kj}=(\delta_{kj}^{j'};j'=0,\ldots,c)$. Indeed, $\delta_{kj}^{j'}=1$ if categorical variable j is conditionally dependent on continuous variable j' for component k and $\delta_{kj}^{j'}=0$ otherwise. Note that $\delta_{kj}^{0}=0$ involves a null intercept in the logistic regression of categorical variable j for component k. Thus, $\boldsymbol{\delta}_{kj}$ fixes some logistic parameters to zero since $\boldsymbol{\beta}_{kj}\in S(\boldsymbol{\delta}_{kj})$ with

$$S(\boldsymbol{\delta}_{kj})=\left\{\boldsymbol{\beta}_{kj}:\ \forall(k,j,j') \text{ as such } \delta_{kj}^{j'}=1,\ \text{then } \forall h\ \beta_{kj}^{j'h}=0\right\}. \tag{7.4}$$

Controlled number of parameters The number of parameters required by the general model defined by $(g,\boldsymbol{\delta})$ is equal to

$$(g-1)+g\left(\frac{c(c+3)}{2}\right)+\sum_{k=1}^{g}\sum_{j=c+1}^{e}\sum_{j'=0}^{c}\delta_{kj}^{j'}(m_j-1). \tag{7.5}$$

Meaningful model The mixture model of Gaussian and logistic distributions provides meaningful classes. Indeed, each class can be summarized by few parameters: mean and variance for the continuous variables and probability of each modality for each categorical variable equal to

$$p(x^{jh}=1|z_{ik}=1)=\int_{\mathbb{R}^c}\phi_c(\boldsymbol{x}_i^{\mathrm{c}};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\frac{\exp\left(\sum_{j'=0}^{c}\beta_{kj}^{j'h}x_i^{j'}\right)}{\sum_{h'=1}^{m_j}\exp\left(\sum_{j'=0}^{c}\beta_{kj}^{j'h'}x_i^{j'}\right)}d\boldsymbol{x}_i^{\mathrm{c}}. \tag{7.6}$$

Although this integral is not explicit, it is easily approximated by a MCMC method. Furthermore, for each class, the dependencies between the continuous variables are modelized by the correlation matrix while the categorical variable $\boldsymbol{x}_i^j$ is conditionally dependent with the continuous one $x_i^{j'}$ if $\delta_{kj}^{j'}=1$.

Dependencies network Figure 7.1 gives an example of the dependencies between variables taken into account by the model. A link between variables denotes a dependency between variables and an absence of link denotes a conditional independence. Note that all the observed variables are linked with $\boldsymbol{z}_i$, there is a clique between the continuous variables, the categorical variables are not linked together and the intra-class dependency between the continuous and the categorical variables are defined by the discrete parameters $\boldsymbol{\delta}$.

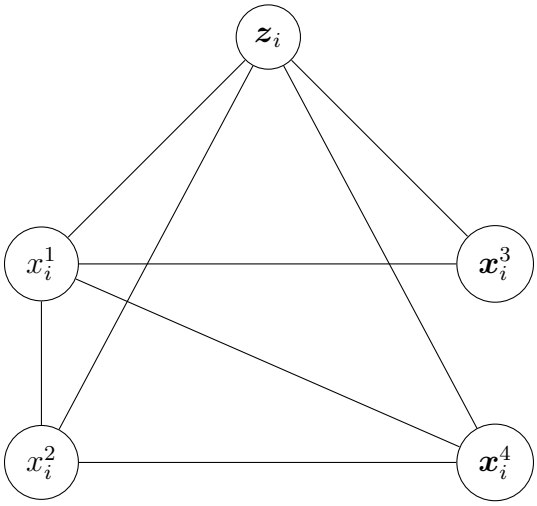


Figure 7.1 – Example of the dependencies taken into account by the model where $\boldsymbol{x}_i^{\mathrm{c}} = (x_i^1, x_i^2)$ and $\boldsymbol{x}_i^{\mathrm{d}} = (\boldsymbol{x}_i^3, \boldsymbol{x}_i^4)$ with $\delta_{k3}^1 = \delta_{k3}^2 = \delta_{k4}^1 = 1$ and $\delta_{k4}^2 = 0$.

Generic identifiability The mixture model of Gaussian and logistic distributions is generically identifiable. Details of the proof are given in Appendix B.1. The demonstration is split in two parts. Firstly, we sum over all the possible values of $\boldsymbol{x}_i^{\mathrm{d}}$ to obtain a mixture of Gaussian distributions and to use its identifiability results [Tei63, YS68]. Secondly, we show the identifiability of the parameters of each logistic function.

7.3 Maximum likelihood estimation via an EM algorithm

Main idea We consider the sample $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ which consists of n individuals assumed independently drawn by a mixture model of Gaussian and logistic distributions. The MLE is easily obtained by the following EM algorithm. This algorithm, that we detail below, is performed for a fixed model defined by the couple $(g, \boldsymbol{\delta})$.

Inference of the logistic parameters At the M step, the maximizations on the proportions and on the Gaussian parameters are easily performed. However, the estimation of the parameters related to the logistic functions involve to solve non-explicit equations. So, they are classically obtained by a Newton-Raphson method. Indeed, the aim is just to estimate the logistic regression parameters where individuals have different weights.

Algorithm 7.3 (EM algorithm).

Starting from an initial value $\boldsymbol{\theta}^{[0]}$, iteration $[r]$ is written as

— **E step**: calculate conditional probabilities

$$t_{ik}(\boldsymbol{\theta}^{[r]}) = \frac{\pi_k^{[r]} p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k^{[r]})}{p(\boldsymbol{x}_i; \boldsymbol{\theta}^{[r]})}. \tag{7.7}$$

— **M step**: maximization of the expectation of the complete-data log-likelihood

$$\pi_k^{[r+1]} = \frac{n_k^{[r]}}{n}, \ \boldsymbol{\mu}_k^{[r+1]} = \frac{1}{n_k^{[r]}} \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]}) \boldsymbol{x}_i,$$

$$\boldsymbol{\Sigma}_k^{[r+1]} = \frac{1}{n_k^{[r]}} \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]}) (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{[r+1]})'(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{[r+1]}),$$

$$\boldsymbol{\beta}_{kj}^{[r+1]} = \underset{\boldsymbol{\beta}_{kj} \in S(\boldsymbol{\delta}_{kj})}{\operatorname{argmax}} \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]}) \ln p(\boldsymbol{x}_i^j | \boldsymbol{x}_i^{\mathrm{c}}; \boldsymbol{\beta}_{kj}), \tag{7.8}$$

where $n_k^{[r]} = \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]})$.

7.4 Model selection via a GEM algorithm

Aim The model selection is a combinatorial problem because of the estimation of the discrete parameter $\boldsymbol{\delta}$. Thus, we want to obtain the parameter $\boldsymbol{\delta}$ which maximizes the information criterion for a fix number of classes.

Main idea It is generally impossible to find the MLE for each $\boldsymbol{\delta}$ since the number of competing models is $2^{g(c+1)d}$. As the BIC criterion is a penalization of the observed-data log-likelihood, we can use an EM algorithm maximizing the penalized observed-data log-likelihood [Gre90] (see Section 1.3.4). Thus, the M step consists in maximizing the expectation of the penalized complete-data likelihood.

Modification of the M step At iteration $[r]$, the M step of the EM algorithm aims at determining $(\boldsymbol{\delta}_{kj}^{[r+1]}, \boldsymbol{\beta}_{kj}^{[r+1]})$ as such

$$(\boldsymbol{\delta}_{kj}^{[r+1]}, \boldsymbol{\beta}_{kj}^{[r+1]}) = \underset{\boldsymbol{\delta}_{kj} \in \{0,1\}^{c+1}}{\operatorname{argmax}} \ \underset{\boldsymbol{\beta}_{kj} \in S(\boldsymbol{\delta}_{kj})}{\operatorname{argmax}} \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[r]}) \ln p(\boldsymbol{x}_i^j | \boldsymbol{x}_i^{\mathrm{c}}; \boldsymbol{\beta}_{kj}) - \frac{\nu_{kj}}{2} \ln n, \tag{7.9}$$

where $\nu_{kj} = \sum_{j'=0}^{c} \delta_{kj}^{j'}$ indicates the number of parameters required by the logistic regression related to categorical variable j for component k.

The GEM algorithm to avoid the combinatorial problems The space of $\boldsymbol{\delta}$ is large, so an exhaustive approach to determine $\boldsymbol{\delta}_{kj}^{[r+1]}$ according to (7.9) is not doable. Thus, we prefer to use a GEM version of this algorithm. In this algorithm, the expectation of the penalized complete-data likelihood is just increased at the M step. So, at iteration $[r]$ of the GM step, we use ascendant and descendant methods initialized by $(\boldsymbol{\delta}_{kj}^{[r]}, \boldsymbol{\beta}_{kj}^{[r]})$.

On the importance of several initializations Note that this approach keeps the classical properties of the EM algorithm. So, this deterministic algorithm converges to a local optimum of the penalized observed-data likelihood which depends of the initial value of $(\boldsymbol{\delta}^{[0]}, \boldsymbol{\theta}^{[0]})$. Thus, it is mandatory to perform this algorithm with different initializations to assure the convergence to a global maximum of the penalized observed-data likelihood.

7.5 Numerical experiments on simulated data sets

Aim We experimentally study the behavior of the GEM algorithm when it performs the model selection using both classical information criteria (AIC and BIC). The results attest to the good behavior of the GEM algorithm for the simultaneous estimation of the model and the parameters. Moreover, they show that the BIC criterion outperforms the AIC criterion. Indeed, this latter overestimates the number of components and the conditional dependencies between mixed data.

Structure of this section Firstly, data are simulated according to the mixture model of Gaussian and logistic distributions. Secondly, data are simulated according to other models.

7.5.1 Simulations by the well specified model

Known number of classes

Data generation Data are sampled according to the model of Gaussian and logistic distributions with two components and equal proportions. Individuals are described with two continuous variables and two binary ones. The model parameters are the following

$$\boldsymbol{\mu}_k = (\varepsilon(k-2), \varepsilon(k-1)),\ \boldsymbol{\Sigma}_k = \begin{bmatrix} 1 & k-1.5 \\ k-1.5 & 1 \end{bmatrix},\ \boldsymbol{\delta}_{k1} = (1,1,0),$$
$$\boldsymbol{\delta}_{k2} = (1,0,1),\ \beta_{kj}^{j'2} = \varepsilon, \quad (7.10)$$

where the parameter ε allows to fix the classes overlaps. The larger is ε, the better separated are the classes. For two classes overlaps and for different sizes of samples $(n = 50, 100, 200, 400)$, 25 data sets are generated.

Estimation conditions For each data set, three models are in competition: the "true" model, the model maximizing the AIC criterion and the model maximizing the BIC criterion. The estimation of the parameters related to the true model is performed by the EM algorithm maximizing the likelihood for a fix model (Algorithm 7.3). The two other parameter estimations are performed by the GEM algorithm maximizing an information criterion. In each case, 25 random initializations of the estimation algorithm are done. Note that, even in this simple case, it is not doable to perform an EM algorithm for each model since there are 2^{12} models in competition.

Results In Table 7.1, we present the Kullback-Leibler divergence of the true model $\boldsymbol{m}$ and of the best model according the AIC (respectively BIC) criterion denoted by $\boldsymbol{m}_{\text{AIC}}$ (respectively $\boldsymbol{m}_{\text{BIC}}$). We conclude to the well behavior of the algorithm used for the parameter estimation. Indeed, the Kullback-Leibler divergence tends to zero when n grows for the three approaches. Furthermore, for a finite sample size, the information criteria allow to reduce the Kullback-Leibler divergence by selecting less complex models.

Overlap	10%				20%			
n	50	100	200	400	50	100	200	400
$\boldsymbol{m}$	1.04	0.18	0.10	0.06	0.42	0.21	0.09	0.05
$\hat{\boldsymbol{m}}_{\text{AIC}}$	0.68	0.17	0.10	0.05	0.54	0.20	0.08	0.04
$\hat{\boldsymbol{m}}_{\text{BIC}}$	0.48	0.15	0.08	0.04	0.44	0.17	0.07	0.04

Table 7.1 – Means of the Kullback-Leibler divergences in a well specified model situation where the number of classes is known.

Number of classes unknown

Data generation Data are sampled according to model described in the previous simulation. For three overlapped classes and for different sample sizes ($n = 50, 100, 200, 400$), 25 data sets are generated.

Estimation conditions For each data set, the best models according to the AIC and the BIC criteria are estimated for $g = 1, \ldots, 4$. In each case, 25 random initializations of the algorithm are done.

Results Table 7.2 displays the mean of the best number of classes and the adjusted rand index [HA85] computed with the estimated partition of the best number of classes for both information criteria. The behavior of the BIC criterion is better since it underestimates the number of classes when the sample size is small and when classes overlap. Furthermore, its convergence to the true number of classes is faster than for the AIC criterion. Indeed, this latter overestimates the number of classes even if the data set is large. The adjusted Rand index related to the model which optimize the BIC criterion is small, since this index is equal to zero when

$g = 1$. Note that this index takes high values for the AIC criterion when this latter overestimates the number of classes. Thus, we can claim that when the AIC criterion overestimates the number of classes, it splits a "true" class into two classes.

Overlap	05%		10%		20%	
n	AIC	BIC	AIC	BIC	AIC	BIC
50	3.68 (0.76)	1.59 (0.40)	3.55 (0.60)	1.50 (0.16)	3.40 (0.41)	1.24 (0.04)
100	3.18 (0.76)	1.91 (0.62)	3.55 (0.63)	1.60 (0.30)	3.32 (0.53)	1.28 (0.05)
200	3.36 (0.76)	2.00 (0.76)	3.27 (0.69)	2.00 (0.56)	3.56 (0.56)	1.64 (0.16)
400	3.20 (0.83)	2.00 (0.80)	3.16 (0.66)	2.00 (0.28)	3.48 (0.59)	2.00 (0.28)

Table 7.2 – Means of the selected number of classes in plain and adjusted Rand indices in parenthesis computed for both information criteria.

Conditional independence situation

Data generation Data are sampled according to the bi-components model where the categorical variables are conditionally independent to the continuous ones. Its parameters are

$$\boldsymbol{\mu}_k = (k-2, k-1),\ \boldsymbol{\Sigma}_k = \begin{bmatrix} 1 & k-1.5 \\ k-1.5 & 1 \end{bmatrix},\ \boldsymbol{\delta}_{k1} = (1,0,0),$$
$$\boldsymbol{\delta}_{k2} = (1,0,0),\ \beta_{kj}^{02} = (-1)^k/2. \quad (7.11)$$

Estimation conditions For each data set, the best models according to the AIC and the BIC criteria are estimated by a GEM algorithm randomly initialized 25 times, with $g = 2$.

Results Table 7.3 displays the headcount where the logistic intercepts are not null and the headcount where the conditional dependency relationship between a categorical variable and a continuous one is zero.

n	AIC		BIC	
	$\delta_{kj}^0 = 1$	$\delta_{kj}^{(1,2)} = 1$	$\delta_{kj}^0 = 1$	$\delta_{kj}^{(1,2)} = 1$
50	0.58	0.32	0.50	0.29
100	0.58	0.31	0.61	0.21
200	0.63	0.27	0.62	0.20
400	0.73	0.23	0.73	0.14

Table 7.3 – Headcount where the logistic intercepts are not null ($\delta_{kj}^0 = 1$) and headcount where the conditional dependency relationship between a categorical variable and a continuous one is estimated ($\delta_{kj}^{(1,2)} = 1$) for both information criteria.

We note that the BIC criterion has a better behavior since the AIC criterion overestimates some relationships between variables when the conditional assumption is valid.

7.5.2 Simulations by misspecified models

Data generation Data sets of size 200 are sampled according to the following bi-components four-variate Gaussian mixture model

$$\pi_k = 0.5,\ \mu_{1j} = 0.5,\ \mu_{2j} = -\mu_{1j},\ \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \begin{bmatrix} 1 & \varepsilon & 0.5 & \varepsilon \\ \varepsilon & 1 & \varepsilon & 0.5 \\ 0.5 & \varepsilon & 1 & \varepsilon \\ \varepsilon & 0.5 & \varepsilon & 1 \end{bmatrix}, \qquad (7.12)$$

where ε is an adjustment parameter. If $\varepsilon = 0$, many variables are conditionally independent while, when ε is high, all the variables are conditionally dependent. The two last variables are discretized to obtain categorical data with four levels: $]-\infty, -1]$, $]-1, 0]$, $]0, 1]$ and $]1, \infty[$.

Estimation conditions The best bi-component model according to each information criterion is estimated by the EM algorithm initialized 25 times.

Results Table 7.4 displays the headcount of the no-null coefficients in the logistic regression for the 0.5 correlated variables and the ε-correlated variables according to both criteria for different values of ε. When $\varepsilon = 0$, the coefficient of the logistic regression between the ε-correlated variables have to be 0. Thus, the AIC criterion overestimates the conditional dependencies. However, it detects better the dependencies between the other variables.

d	AIC		BIC	
	0.5-correlated	ε-correlated	0.5-correlated	ε-correlated
0	0.77	0.22	0.62	0.08
0.2	0.74	0.26	0.63	0.08
0.4	0.77	0.34	0.52	0.13

Table 7.4 – Headcount where the conditional dependencies are modeled by both information criteria.

Because of the bias of the AIC criterion shown during all the numerical experiments, our advice is to use the BIC criterion to perform the model selection, even if this latter can neglect some conditional dependencies between mixed variables.

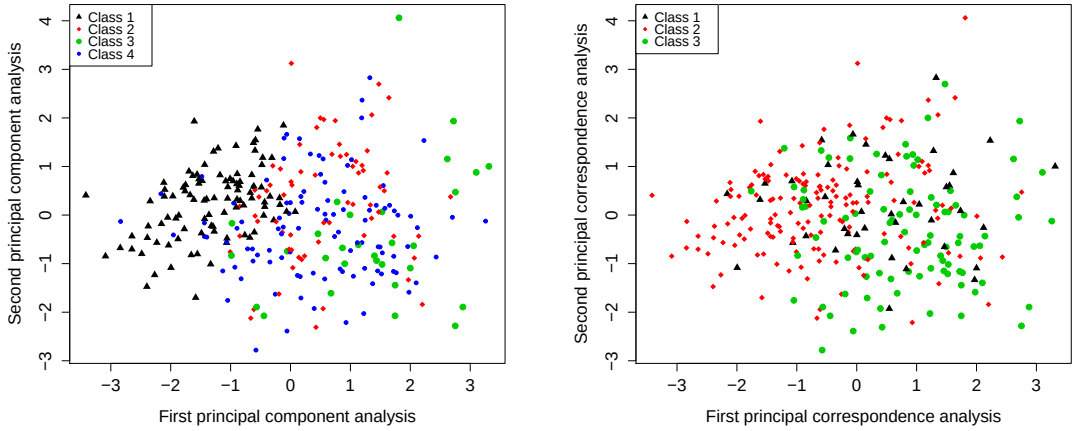
7.6 Analysis of two real data sets

7.6.1 Disease data clustering

Data set description The Cleveland Heart Disease data [Det88] described 303 patients per 14 variables (five continuous, eight categorical and one predicted attribute) and is available in the UCI machine learning repository. The predicted attribute is a binary variable indicating the presence of heart disease. We blind this

information during our clustering. Furthermore, the six individuals having missing values are omitted.

Homogeneous model-based clustering Note that both kinds of variables are important for the cluster analysis. Indeed, the analysis performed on the continuous variables by a mixture of Gaussian distributions selects four classes while the analysis performed on the categorical ones by the latent class model selects three classes. Figure 7.2 displays the estimated partition by the Gaussian mixture model in the first component map (Figure 7.2.a) and by the multinomial mixture model in the first correspondence map (Figure 7.2.b). Thus, we see that classes overlap in the first factorial maps and we do not find a factorial map where the estimated classes are well separated.



(a) (b)

Figure 7.2 – Partitions estimated by the homogeneous model: (a) partition of the Gaussian mixture model drawn in the first component map; (b) partition of the latent class model drawn in the first correspondence map.

As shown by the confusion matrix presented in Table 7.5, the partitions obtained by the model for homogeneous variables are very different from the predicted attribute. Indeed, the value of the adjusted rand index computed between the partition of the Gaussian mixture model and the predicted attribute is equal to 0.23 while it is equal to 0.10 when it is computed between the multinomial mixture model and the predicted attribute. If we set $g = 2$, the error rates are: 0.28 for the continuous case and 0.47 for the categorical case. Finally, both partitions of the homogeneous models are different since their adjusted rand index values are equal to 0.10.

Heterogeneous model-based clustering We perform the cluster analysis on the whole data set by using two models: the conditional independence model and the mixture of Gaussian and logistic distributions. The results displayed in Table 7.6 claim that the mixture model of Gaussian and logistic distributions better

Heart	Continuous variables				Categorical variables		
disease	class 1	class 2	class 3	class 4	class 1	class 2	class 3
absence	11	61	12	76	122	23	15
presence	36	14	43	44	39	20	78

Table 7.5 – Confusion tables between the predicted attributed and the two partitions estimated by the homogeneous mixture models.

approaches the data distribution. Indeed, this model selects two classes since its BIC criterion values are -7178.35 with one component and -7140.36 with three components. Note that the conditional independence model overestimates the number of classes since the BIC criterion selects three classes with a value of -7401. By considering the whole data set, the mixture model of Gaussian and logistic distributions obtains a more meaningful model since it has less classes and less parameters.

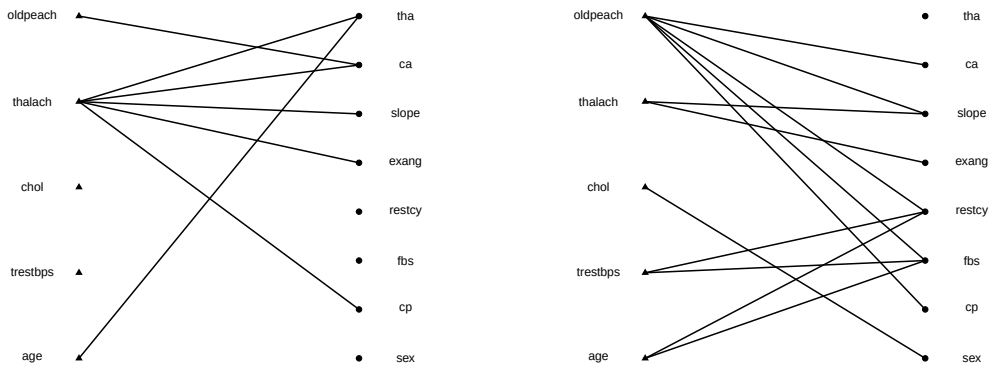
	Cond. indp.	proposed model
BIC criterion	-7449.71	-7122.74
Log-likelihood	-7310.21	-6871.22
Parameters	49	88

Table 7.6 – Values of the BIC criterion and of the log-likelihood function and number of parameters for both bi-component models in competition.

Best model interpretation The majority class (70%) groups individuals taking the smallest values for the continuous variables except for the variable *thalach*. Moreover, their variances are smaller than them of the minority class (30%), while the correlation between the continuous variables are stronger. As displayed in Figure 7.3, less continuous variables influence the categorical ones in class 1 (Figure 7.3.a) than in class 2 (Figure 7.3.b). The variable *thalance* impacts the most categorical variables in class 1 while it is the variable *oldpeachs* which impact the most categorical variable in class 2.

Figure 7.4 displays the partition obtained by the bi-component mixture model of Gaussian and logistic distributions in the first plan of the PCAmixte [CKSS12]. We can see that the second axis is discriminative between both estimated classes.

Comparison with other approaches The error rate obtained by the mixture model of Gaussian and logistic distributions is 38%. According to [HJ11], the conditional independence model and Multimix obtain an error rate of 23%. Finally, the traditional hierarchical clustering methods have a misclassification rate ranging between 22% and 46%. Note that the k-means approach [AD07] obtains an error rate of 15% but the selection of the number of classes is more delicate.



(a) Class 1. (b) Class 2.

Figure 7.3 – Dependencies between the continuous variables (triangles on the left) and the categorical variables (circle on the right) per class. A link involves a dependency (so a no null coefficient in the logistic regression in the class).

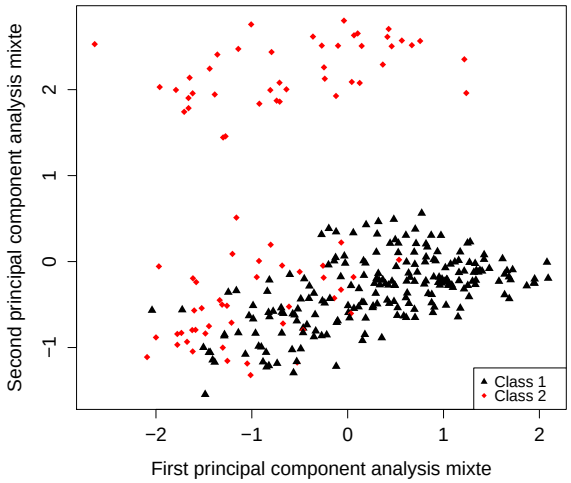


Figure 7.4 – Partition drawn in the first component map of the PCAmixte.

7.6.2 Melanoma semi-supervised classification

Data set description Melanoma is cancer of skin. The data set describes 205 patients [ABGK93] by four continuous variables (accompagned time in days, age in years, year of operation and the tumor thickness), by two binary variables (sex and presence/absence of ulcer) and by one status variable that we dichotomized (died from melanoma or not).

Discriminative and generative approaches comparison We illustrate that the mixture model of Gaussian and logistic distributions can outperform classical

methods of semi-supervised classification, especially when few observations are labeled. So, we randomly blind a part of the labels. We compare the error rate of the blinded partition obtained by mixture model of Gaussian and logistic distributions to the error rate obtained with a logistic regression. Table 7.7 presents the mean of the misclassifying rate computed on 100 randomly blinded partition for different percentages of missing labels.

% missing labels	20	40	60	70	80	90	95	97.5	100
proposed model	0.129	0.130	0.138	0.140	0.144	0.153	0.154	0.156	0.156
Logistic	0.128	0.129	0.134	0.145	0.169	0.220	0.254	0.274	NA

Table 7.7 – Mean of the misclassifying rate computed on the individuals having a missing membership.

Two different approaches for two different objectives We remind that the logistic regression aim is to directly modelize the border between the classes. Indeed, this methods was developed specifically for the semi-supervised classification. The aim of the mixture model of Gaussian and logistic distributions models is more ambitious. Indeed, it modelizes the whole distribution of the data.

Comments The presented results are as expected. When a majority of the individuals is labeled, the logistic regression obtains a lower misclassifying rate. However, when a majority of the individuals is unlabeled, the main information is contained by these individuals. Thus, during this experiment, we observe that the mixture model of Gaussian and logistic distributions outperforms the logistic model for an high rate of missing labels. Furthermore, this rate dramatically grows for the logistic regression when very few individuals are labeled while its stays stable for the mixture model of Gaussian and logistic distributions. Indeed, as 27.8% of the individuals was died from melanoma, the logistic regression is close to the worse error rate when 95% of the labels are missing.

Meaningful classes in cluster analysis We now interpret both classes obtained when all the individuals are unlabeled. The majority class (60%) groups individuals having a long accompany time and recently treated. The class is mainly composed by young women having a small tumor. In this class, the bigger is the tumor, the larger is the ulcer risk. In the minority class (40%), we find the patients where the accompany time is shorter (died or accompaniment stopped) and where the treatment is old. These patients are older, generally with a big tumor mainly present for the men and increasing the ulcer risk. The class interpretation is based on the margin parameters presented in Figure 7.5.

The confusion matrix displayed in Table 7.8 shows that the estimated partition is close to the survival status. The majority class involves a small risk of death from Melanoma while this risk is higher in the minority class.

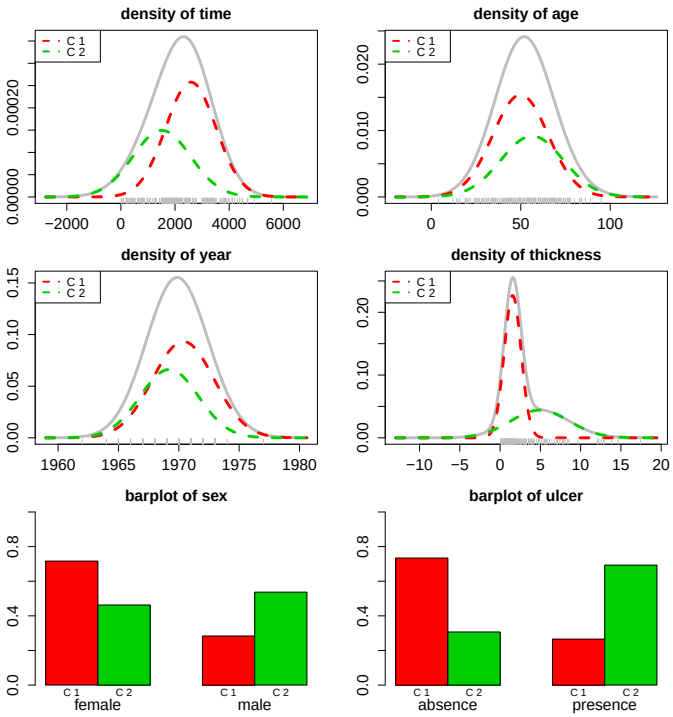


Figure 7.5 – Plotting of the margin parameters of the bi-component mixture model of Gaussian and logistic distributions estimated on the Melanoma data sets.

	mixture model of Gaussian and logistic distributions	
	majority class	minority class
not died form melanoma	121	5
died form melanoma	27	52

Table 7.8 – Confusion matrix between the partition estimated by the mixture model of Gaussian and logistic distributions and the survival status.

7.7 Conclusion

The mixture model of Gaussian and logistic distributions is an efficient approach to cluster data sets with continuous and categorical variables. So, it is a good challenger to the well-known models: *naive Bayes* and *mixture of location models* (and its derives). Its first advantage is to take into account the intra-class dependencies between all the variables. Thus, the proposed model avoids the biases involved by the conditional independence assumption. Its second advantage is to keep classical distributions for the one-dimensional margins of each component. Indeed, the practitioner easily summarizes each class by the parameters of the classical distributions and by the logistic functions.

The parsimonious versions of mixture model of Gaussian and logistic distributions allow to model the main conditional dependencies between mixed variables. Thus, the class interpretation is easier. The model selection and the parameter

estimation are simultaneously performed via a GEM algorithm maximizing an information criterion. According to our experiments, our advice is to use the BIC criterion as the information criterion.

The mixture model of Gaussian and logistic distributions can be used in a semi-supervised classification. According to our application, it can be a good challenger to the classical discrimination approaches, especially when few individuals are labeled.

Chapter 8

Model-based clustering of Gaussian copulas for mixed data

A mixture model of Gaussian copulas is presented to cluster mixed data where any kinds of variables are allowed if they admit a cumulative distribution function. This approach allows to straightforwardly define simple multivariate intra-class dependency models while preserving any one-dimensional margin distributions of each component of interest for the statistician. Typically in this work, the margin distributions of each component are classical parametric ones in order to facilitate the model interpretation. In addition, the intra-class dependencies are taken into account by the Gaussian copulas which provide one correlation coefficient, having good properties, per couple of variables and per class.
This model generalizes different existing models defined for homogeneous and mixed variables. The inference is performed via a Metropolis-within-Gibbs sampler in a Bayesian framework. Numerical experiments illustrate the model flexibility and its relevance.

Tu as vu Zorba
quand tu mets une loupe au soleil
et que tu rassembles
tous les rayons sur un seul point ?
Ce point-là prend bientôt feu.
Pourquoi ?
Parce que la force du soleil
ne s'est pas éparpillée,
elle s'est rassemblée
sur un seul point.
De même l'esprit de l'homme.
On fait des miracles
en concentrant son esprit
sur une seule et même chose.
Níkos Kazantzákis—Alexis Zorba.

8.1 Introduction

The aim of this chapter is to present a model-based clustering for mixed data of any kinds of variables admitting a cumulative distribution function. This model has a double objective: to preserve *classical distributions* for *all* its one-dimensional margin distributions of each component and to parsimoniously and meaningfully *modelize the intra-class dependencies*.

This objective can naturally be achieved by the use of copulas [Joe97, Nel99, GF07]. Indeed, copulas build a multivariate model by setting, on the one hand, the one-dimensional *margins*, and, on the other hand, the *dependency model* between variables. More precisely, the data distribution is approached by a full parametric *mixture model of Gaussian copulas* whose the margin distributions of each component are classical and whose the Gaussian copulas [Hof07, HNW11] modelize the intra-class dependencies. Note that [SK12, MDCL13] already use one Gaussian copula to define a distribution of mixed variables. The proposed model is also a generalization of this approach to the finite mixture model framework.

The new mixture model is meaningful since it permits a *three-level schema* which allows a friendly interpretation: the proportions indicate the class weights, the one-dimensional margin parameters of each components roughly describe the classes while the correlation matrices refine this description. Finally, by using the continuous latent structure of the Gaussian copulas, a PCA-type visualization per class allows to summarize the main intra-class dependencies and provides a scatterplot of the individuals according to the class parameters.

Note that I. Kosmidis and D. Karlis [KK14] have recently submitted an article which proposes to use a mixture of copulas to perform cluster analysis. The authors study different copulas, among them the Gaussian copulas are considered. Their model is close to the approach developed in this chapter. However, two important differences have to be mentioned. Firstly, we propose a Bayesian inference while the authors propose an approach by maximum likelihood under constraints. Secondly, some visualizations tools are presented here.

Structure of this chapter This paper is organized as follows. Section 8.2 presents the mixture model of Gaussian copulas introduced to cluster, its links with the existing models and its contribution to the visualization of mixed variables. Section 8.3 is devoted to the parameter estimation in a Bayesian framework since the maximum likelihood estimate is unattainable [PCK06]. Section 8.4 illustrates the behavior of the algorithm performing the inference and also the model robustness on numerical experiments. Section 8.5 presents three applications of the new mixture model by clustering three real data sets. Section 8.6 concludes this work. All these results are part of the article *Model-based clustering of Gaussian copulas for mixed data* [MBV14b].

8.2 Mixture model of Gaussian copulas

8.2.1 Finite mixture model

Data

The vector of e mixed variables is denoted by $\boldsymbol{x}_i = (x_i^1, \ldots, x_i^e) \in \mathbb{R}^c \times \mathcal{X}$, with $e = c + d$. Its first c elements are the set of the continuous variables, defined on the space $\mathbb{R}^c$ and further denoted by $\boldsymbol{x}_i^{\mathrm{c}}$. Its last d elements are the set of the discrete variables (integer, ordinal or binary), defined on the space $\mathcal{X}$ and further denoted by $\boldsymbol{x}_i^{\mathrm{d}}$. Note that if x_i^j is an ordinal variable with m_j modalities, then it uses a numeric coding $\{1, \ldots, m_j\}$.

Notation We remind that we use the generic notation $P(.;.)$ for the cumulative distribution functions (cdf) and $p(.;.)$ for the probability distribution function (pdf).

Probability distribution function

Definition 8.1 (Finite mixture model of parametric distributions)**.** Data $\boldsymbol{x}_i$ are supposed to be drawn by the mixture model of g parametric distributions whose the pdf is written as follows

$$p(\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k), \tag{8.1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha})$ denotes the whole parameters. The vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ groups the proportions of each class k denoted by π_k, and respecting the following constraints $0 < \pi_k \leq 1$ and $\sum_{k=1}^{g} \pi_k = 1$, while the vector $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_g)$ groups the parameters of each class k denoted by $\boldsymbol{\alpha}_k$.

Property 8.2 (Latent variable)**.** A finite mixture model can be expressed by using the latent variable $\boldsymbol{z}_i$. This categorical variable indicates the class membership by using a complete disjunctive coding and follows the multinomial distribution $\mathcal{M}_g(\pi_1, \ldots, \pi_g)$. Thus, (8.1) can be interpreted as the marginal distribution of $\boldsymbol{x}_i$ based on the distribution of the couple $(\boldsymbol{x}_i, \boldsymbol{z}_i)$.

8.2.2 Gaussian copula for mixed data

Component distributions following Gaussian copulas

Copulas allow to build a multivariate model by setting, on the one hand, the one-dimensional *margins*, and, on the other hand, the *dependency model* between variables. We now present the margin distribution of the components then we focus on the Gaussian copula which is of interest for us since it provides one correlation coefficient per couple of variables and since it allows an easy parameter estimation.

One-dimensional margins of the components

For each component, we assume that the margin distributions of each component belongs to the exponential family, in order to provide meaningful classes.

Definition 8.3 (One-dimensional margins of the components)**.** The margin distribution of the variable x_i^j, for component k, belongs to the exponential family and has $p(x_i^j;\boldsymbol{\beta}_{kj})$ for pdf and $P(x_i^j;\boldsymbol{\beta}_{kj})$ as cdf. More precisely,

— If x_i^j is *continuous*, its margin of component k follows a *Gaussian* distribution with mean μ_{kj} and variance σ_{kj}^2, *i.e.* $x_i^j|z_{ik}=1\sim\mathcal{N}_1(\mu_{kj},\sigma_{kj}^2)$ and $\boldsymbol{\beta}_{kj}=(\mu_{kj},\sigma_{kj}^2)\in\mathbb{R}\times\mathbb{R}^{+*}$.

— If x_i^j is *integer*, its margin of component k follows a *Poisson* distribution, *i.e.* $x_i^j|z_{ik}=1\sim\mathcal{P}(\boldsymbol{\beta}_{kj})$ and $\boldsymbol{\beta}_{kj}\in\mathbb{R}^{+*}$.

— If x_i^j is *ordinal*, its margin of component k follows a *multinomial* distribution, *i.e.* $x_i^j|z_{ik}=1\sim\mathcal{M}_{m_j}(\boldsymbol{\beta}_{kj})$, $\boldsymbol{\beta}_{kj}$ being defined on the simplex of size m_j.

Dependency model of the components

The mixture model of Gaussian copulas assumes that each component k follows a Gaussian copula whose the correlation matrix of size $e\times e$ is denoted by $\boldsymbol{\Gamma}_k$. We note $\Phi_e(.;\boldsymbol{\Gamma}_k)$ the cdf of the e-variate centred Gaussian distribution with correlation matrix $\boldsymbol{\Gamma}_k$, and $\Phi_1^{-1}(.)$ the inverse cumulative distribution function of $\mathcal{N}_1(0,1)$. Thus, we obtain the following definition of the component cdf.

Definition 8.4 (Cumulative distribution function of the components)**.** For the mixture model of Gaussian copulas, the cdf of component k is written as

$$P(\boldsymbol{x}_i;\boldsymbol{\alpha}_k)=\Phi_e(\Phi_1^{-1}(u_k^1),\ldots,\Phi_1^{-1}(u_k^e);\mathbf{0},\boldsymbol{\Gamma}_k), \tag{8.2}$$

where $u_k^j=P(x_i^j;\boldsymbol{\beta}_{kj})$ and where $\boldsymbol{\alpha}_k=(\boldsymbol{\beta}_k,\boldsymbol{\Gamma}_k)$ denotes the whole parameters of component k with $\boldsymbol{\beta}_k=(\boldsymbol{\beta}_{k1},\ldots,\boldsymbol{\beta}_{ke})$.

Property 8.5 (Standardized coefficient of correlation per class)**.** The Gaussian copula provides a coefficient of correlation per couple of variables which has good properties. Indeed, when both variables are continuous, it is equal to the upper bound of the coefficients of correlation obtained by all the monotonic transformations of the variables [KW97]. Furthermore, when both variables are discrete, it is equal to the polychoric coefficient of correlation [Ols79].

Property 8.6 (Second latent variable)**.** The mixture model of Gaussian copulas involves a second latent variable (added to the class membership) which consists in an e-variate continuous variable denoted by $\boldsymbol{y}_i=(y_i^1,\ldots,y_i^e)\in\mathbb{R}^e$. Conditionally on the class membership, this variable follows an e-variate centered Gaussian distribution. Indeed, if $\boldsymbol{y}_i|z_{ik}=1\sim\mathcal{N}_e(\mathbf{0},\boldsymbol{\Gamma}_k)$ and if

$$x_i^j=P^{-1}(\Phi_1(y^j);\boldsymbol{\beta}_{kj}),\ \forall j=1,\ldots,e, \tag{8.3}$$

then component k is a Gaussian copula whose the cdf is $P(\boldsymbol{x}_i;\boldsymbol{\alpha}_k)$.

Mixture model of Gaussian copulas for mixed data

We introduce the function $\Psi(\boldsymbol{x}_i^{\mathrm{C}}; \boldsymbol{\alpha}_k) = \left(\frac{x_i^j - \mu_{kj}}{\sigma_{kj}}; j = 1, \ldots, c\right)$ and the space of the antecedents of $\boldsymbol{x}_i^{\mathrm{D}}$ for class k is noted $\mathcal{S}_k(\boldsymbol{x}_i^{\mathrm{D}}) = \mathcal{S}_k^{c+1}(x_i^{c+1}) \times \ldots \times \mathcal{S}_k^e(x_i^e)$. The interval $\mathcal{S}_k^j(x_i^j) =]b_k^{\ominus}(x_i^j), b_k^{\oplus}(x_i^j)]$ is defined for $j = c+1, \ldots, e$ and its bounds are $b_k^{\ominus}(x_i^j) = \Phi_1^{-1}(P(x_i^j - 1; \boldsymbol{\beta}_{kj}))$ and $b_k^{\oplus}(x_i^j) = \Phi_1^{-1}(P(x_i^j; \boldsymbol{\beta}_{kj}))$. We now define the pdf of the components according to (8.2) as proposed in [SK12].

Definition 8.7 (Mixture model of Gaussian copulas)**.** Data $\boldsymbol{x}_i$ follows a mixture model of Gaussian copulas if its pdf is the finite mixture model defined in (8.1) whose the pdf of component k is written as

$$p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) = p(\boldsymbol{x}_i^{\mathrm{C}}; \boldsymbol{\alpha}_k) p(\boldsymbol{x}_i^{\mathrm{D}} | \boldsymbol{x}_i^{\mathrm{C}}; \boldsymbol{\alpha}_k) \tag{8.4}$$

$$= \frac{\phi_c(\Psi(\boldsymbol{x}_i^{\mathrm{C}}; \boldsymbol{\alpha}_k); \mathbf{0}, \boldsymbol{\Gamma}_{k\mathrm{CC}})}{\prod_{j=1}^c \sigma_{kj}} \int_{\mathcal{S}_k(\boldsymbol{x}_i^{\mathrm{D}})} \phi_d(\boldsymbol{u}; \boldsymbol{\mu}_k^{\mathrm{D}}, \boldsymbol{\Sigma}_k^{\mathrm{D}}) d\boldsymbol{u}, \tag{8.5}$$

where $\boldsymbol{\Gamma}_k = \begin{bmatrix} \boldsymbol{\Gamma}_{k\mathrm{CC}} & \boldsymbol{\Gamma}_{k\mathrm{CD}} \\ \boldsymbol{\Gamma}_{k\mathrm{DC}} & \boldsymbol{\Gamma}_{k\mathrm{DD}} \end{bmatrix}$ is decomposed into sub-matrices, for instance $\boldsymbol{\Gamma}_{k\mathrm{CC}}$ is the sub-matrix of $\boldsymbol{\Gamma}_k$ composed by the rows and the columns related to the observed continuous variables. Moreover, $\boldsymbol{\mu}_k^{\mathrm{D}}$ is the conditional mean of $\boldsymbol{y}_i^{\mathrm{D}}$ defined by $\boldsymbol{\mu}_k^{\mathrm{D}} = \boldsymbol{\Gamma}_{k\mathrm{DC}} \boldsymbol{\Gamma}_{k\mathrm{CC}}^{-1} \Psi(\boldsymbol{x}_i^{\mathrm{C}}; \boldsymbol{\alpha}_k)$ and $\boldsymbol{\Sigma}_k^{\mathrm{D}}$ is its conditional covariance matrix defined by $\boldsymbol{\Sigma}_k^{\mathrm{D}} = \boldsymbol{\Gamma}_{k\mathrm{DD}} - \boldsymbol{\Gamma}_{k\mathrm{DC}} \boldsymbol{\Gamma}_{k\mathrm{CC}}^{-1} \boldsymbol{\Gamma}_{k\mathrm{CD}}$.

Property 8.8 (Generative model)**.** The mixture model of Gaussian copulas involves the generative model split into the following three steps:

— Class membership *sampling*: $\boldsymbol{z}_i \sim \mathcal{M}_g(\pi_1, \ldots, \pi_g)$
— Gaussian copula *sampling*: $\boldsymbol{y}_i | z_{ik} = 1 \sim \mathcal{N}_e(\mathbf{0}, \boldsymbol{\Gamma}_k)$
— Observed data *deterministic computation*: $\boldsymbol{x}_i$ is obtained from (8.3).

Remarks

— *Homoscedastic models.* When the sample size is small, the trade off between the bias and the variance of the estimate may be better if some constraints on the parameter space are added. Thus, we propose a parsimonious version of the mixture model of Gaussian copulas by assuming the equality between the correlation matrices, so

$$\boldsymbol{\Gamma}_1 = \ldots = \boldsymbol{\Gamma}_g. \tag{8.6}$$

Note that this model is named homoscedastic since the covariance matrices of the latent Gaussian variables are equal between classes.

— *Number of parameters.* The heteroscedastic (respectively homoscedastic) mixture model of Gaussian copulas needs ν_{He} (respectively ν_{Ho}) parameters where

$$\nu_{\mathrm{He}} = (g-1) + g\left(\frac{e(e-1)}{2} + \sum_{j=1}^d \nu_j\right) \text{ and } \nu_{\mathrm{Ho}} = (g-1) + \frac{e(e-1)}{2} + g\sum_{j=1}^d \nu_j, \tag{8.7}$$

where ν_j denotes the number of parameters of the margin distribution of variable j for one component. More precisely, with the specific margin distribution of the components, ν_j is equal to

$$\nu_j = \begin{cases} 2 & \text{if } x^j \text{ is numeric} \\ 1 & \text{if } x^j \text{ is discrete} \\ m_j - 1 & \text{if } x^j \text{ is ordinal.} \end{cases} \tag{8.8}$$

— *Model identifiability.* The mixture model of Gaussian copulas is identifiable (in the sense defined in [Tei63, YS68]) if, at least, one variable is continuous or integer. The proof is given in Appendix B.2.

8.2.3 Strengths of the mixture model

Related models

The mixture model of Gaussian copulas allows to generalize many classical model-based clusterings, among them one can cite the following four.

— Obviously, if the correlation matrices are diagonal (*i.e.* $\boldsymbol{\Gamma}_k = \boldsymbol{I}$, $\forall k = 1, \ldots, g$), then the mixture model of Gaussian copulas is equivalent to the conditional independence mixture model.

— If all the variables are continuous (*i.e.* $c = e$ and $d = 0$), then the mixture model of Gaussian copulas becomes a multivariate Gaussian mixture model without constraint between the parameters [BR93].

— The mixture model of Gaussian copulas is linked to the binned Gaussian mixture model. For instance, it is equivalent, when data are ordinal, to the mixture model of [Gou06]. In such a case, this model is stable by fusion of modalities.

— When the variables are both continuous and ordinal, the mixture model of Gaussian copulas is a new parametrization of the mixture model proposed by Everitt [Eve88] (see Section 6.4). However, Everitt estimates directly the space $\mathcal{S}_k(\boldsymbol{x}_i^{\mathrm{D}})$ containing the antecedents of $\boldsymbol{x}_i^{\mathrm{D}}$ and not the margin parameters. Thus, the maximum likelihood inference is also performed via a simplex algorithm dramatically limiting the number of ordinal variables. Note that our approach for the inference avoids this drawback (see details in Section 8.3).

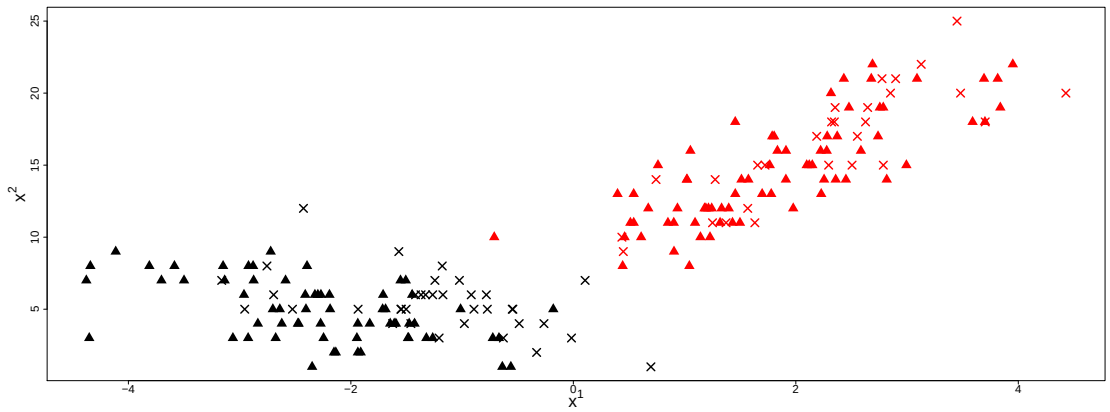
Data visualization per class: a by-product of Gaussian copulas

We can use the model parameters to obtain a *visualization* of the individuals *per class* and to bring out the main intra-class dependencies. Thus, for class k, we firstly compute the coordinates equal to $\mathbb{E}[\boldsymbol{y}_i|\boldsymbol{x}_i, z_{ik} = 1; \boldsymbol{\alpha}_k]$ and we secondly project them on the principal component analysis space of the Gaussian copula of component k, obtained by the spectral decomposition of $\boldsymbol{\Gamma}_k$.

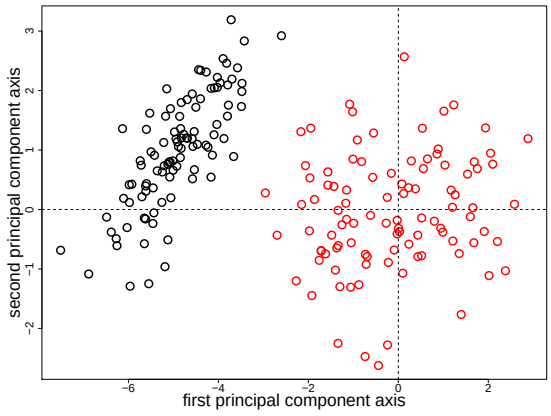
The individuals drawn by the component k follow a centred Gaussian distribution in the factorial map, so they are close to the origin. Those drawn by another component have an expectation different from zero, so they are farther to the origin. Finally, the correlation circle summarizes the intra-class correlations. The following example illustrates this phenomenon.

Example 8.9 (Mixture model of Gaussian copulas and visualization per class)**.** *Let the bi-component mixture model of Gaussian copulas composed with three variables (one continuous, one integer and one binary), in this order, with*

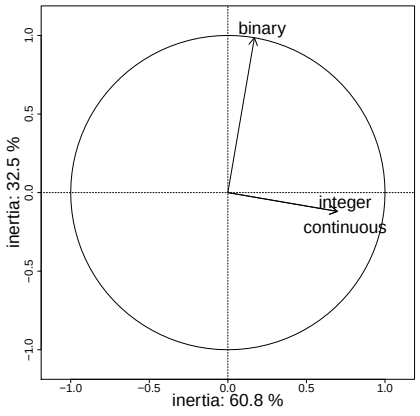
$$\boldsymbol{\pi} = (0.5, 0.5),\ \boldsymbol{\beta}_{11} = (-2, 1),\ \boldsymbol{\beta}_{12} = 5,\ \boldsymbol{\beta}_{13} = (0.5, 0.5), \boldsymbol{\beta}_{21} = (2, 1),\ \boldsymbol{\beta}_{22} = 15,$$
$$\boldsymbol{\beta}_{23} = (0.5, 0.5),\ \boldsymbol{\Gamma}_1 = \begin{pmatrix} 1 & -0.4 & 0.4 \\ -0.4 & 1 & 0.4 \\ 0.4 & 0.4 & 1 \end{pmatrix} \text{ and } \boldsymbol{\Gamma}_2 = \begin{pmatrix} 1 & 0.8 & 0.1 \\ 0.8 & 1 & 0.1 \\ 0.1 & 0.1 & 1 \end{pmatrix}.$$



(a)



(b)



(c)

Figure 8.1 – Example of visualization: (a) scatter-plot of the individuals described by three variables: one continuous (abscissa), one integer (ordinate) and one binary (symbol); (b) individuals scatter-plot in the first component map of class 2; (c) variables representation in the first component map of class 2. The color indicates the class memberships.

The visualization of class 2 is presented in Figure 8.1. Concerning the individuals, the scatter-plot shows a centered class (the red one) and a second class (the black one) located on the left side. Concerning the variables, the representation points out by a strong intra-class correlation between the continuous and the integer variables.

8.3 Bayesian inference via a Metropolis-within-Gibbs sampler

Aim We observe the sample $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ composed with n independent individuals $\boldsymbol{x}_i \in \mathbb{R}^c \times \mathcal{X}$ assumed to be drawn by a mixture model of Gaussian copulas. The aim is to infer the parameters according to the data.

Frequentist context The inference by maximum likelihood is a difficult problem for the full parametric copulas when the margin parameters are unknown. So, it is often replaced by the *Inference Function for Margins* method performing the inference in two steps (see Chapter 10 of [Joe97]). The first step estimates the margin parameters by maximizing each univariate likelihood while the second step estimates the correlation parameters by maximizing the likelihood conditionally on the margin parameters. This approach is used in [KK14]. However, the maximum likelihood estimate can be essentially obtained when the variables are continuous by using the fixed-point algorithm proposed by [SFK05]. Indeed, this approach can not be extended to the mixed data setting. Thus, an EM algorithm can not be implemented to obtain the maximum likelihood estimates of a mixture model of Gaussian copulas in the mixed data case. Furthermore, even if the M step would be explicit, the E step would be too much time consuming, if the discrete variables are numerous, because of the computation of the integral of dimension d defined in (8.5).

Bayesian context In order to avoid both previous problems, we prefer to work in a Bayesian framework. We firstly define the prior distributions and we secondly present the Gibbs sampler performing the inference.

8.3.1 Maximum *a posteriori* estimate

Prior distributions

Independence assumption A classical assumption is to suppose the independence between the prior distributions, thus

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}) \prod_{k=1}^{g} \left(p(\boldsymbol{\Gamma}_k) \prod_{j=1}^{d} p(\boldsymbol{\beta}_{kj}) \right). \tag{8.9}$$

Proportions The classical conjugate prior distribution of the proportion vector is the Jeffreys non informative one which is a Dirichlet distribution whose the parameters are equal to $1/2$

$$\boldsymbol{\pi} \sim \mathcal{D}_g \left(\frac{1}{2}, \ldots, \frac{1}{2} \right). \tag{8.10}$$

Margin parameters The prior distribution of the margin parameters are the classical conjugate ones. More precisely,

— if x^j is *continuous*, then $\boldsymbol{\beta}_{kj}$ denotes the parameters of a univariate Gaussian distribution so $p(\boldsymbol{\beta}_{kj}) = p(\mu_{kj}|\sigma^2_{kj})p(\sigma^2_{kj})$ with

$$\sigma^2_{kj} \sim \mathcal{G}^{-1}(c_0, C_0) \text{ and } \mu_{kj}|\sigma^2_{kj} \sim \mathcal{N}_1(b_0, \sigma^2_{kj}/N_0), \tag{8.11}$$

where $\mathcal{G}^{-1}(.,.)$ denotes the inverse gamma distribution. With an empirical Bayesian approach, the hyper-parameters (c_0, C_0, b_0, N_0) are fixed as proposed by [Raf96], so $c_0 = 1.28$, $C_0 = 0.36\text{Var}(\mathbf{x}^j)$, $b_0 = \frac{1}{n}\sum_{i=1}^n x_i^j$ and $N_0 = \frac{2.6}{\text{argmax } \mathbf{x}^j - \text{argmin } \mathbf{x}^j}$.

— if x^j is *integer*, $\boldsymbol{\beta}_{kj}$ denotes the parameter of a Poisson distribution and

$$\boldsymbol{\beta}_{kj} \sim \mathcal{G}(a_0, A_0). \tag{8.12}$$

According to [FS06], the values of hyper-parameters a_0 and A_0 are empirically fixed to $a_0 = 1$ and $A_0 = a_0 n / \sum_{i=1}^n x_i^j$.

— if x^j is *ordinal*, $\boldsymbol{\beta}_{kj}$ denotes the parameter of a multinomial distribution and its Jeffreys non informative conjugate prior involves that

$$\boldsymbol{\beta}_{kj} \sim \mathcal{D}_{m_j}\left(\frac{1}{2}, \ldots, \frac{1}{2}\right). \tag{8.13}$$

Correlation matrices The conjugate prior of a covariance matrix is the Inverse Wishart distribution denoted by $\mathcal{W}^{-1}(.,.)$. So, it is natural to define the prior of the correlation matrix $\boldsymbol{\Gamma}_k$ from the prior of the correlation matrix $\boldsymbol{\Lambda}_k$ since $\boldsymbol{\Gamma}_k|\boldsymbol{\Lambda}_k$ is deterministic [Hof07]. So,

$$\boldsymbol{\Lambda}_k \sim \mathcal{W}^{-1}(s_0, S_0) \text{ and } \forall 1 \leq h, \ell \leq e,\ \boldsymbol{\Gamma}_k[h,\ell] = \frac{\boldsymbol{\Lambda}_k[h,\ell]}{\sqrt{\boldsymbol{\Lambda}_k[h,h]\boldsymbol{\Lambda}_k[\ell,\ell]}}, \tag{8.14}$$

where (s_0, S_0) are two hyper-parameters. However, the classical approach consisting in fitting the hyper-parameters through an empirical Bayesian approach is not possible since $\boldsymbol{y}_i$ is not observed. We thus put $s_0 = e + 1$ and S_0 equal to the identity matrix, since in this case, the margin distribution of each correlation coefficient is uniform on $]-1, 1[$ [BMM00].

Posterior distribution

The Bayesian inference is performed by sampling a sequence of parameters from their posterior distribution. In practice, we use a Gibbs sampler which is the most popular approach to perform a Bayesian inference of mixture model since it uses the latent structure of the data. Indeed, it alternatively samples the class memberships conditionally on the parameters and on the data, and the parameters conditionally on the class memberships and on the data. Since its stationary distribution is $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x})$, the sequence of the generated parameters is drawn by the marginal posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$. This algorithm relies on two instrumental variables: the class membership of the individuals of $\mathbf{x}$ denoted by $\mathbf{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ and the Gaussian vector of the individuals denoted by $\mathbf{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$.

8.3.2 Gibbs sampler

Algorithm 8.10 (The Gibbs sampler)**.**

Starting from an initial value $\boldsymbol{\theta}^{[0]}$, its iteration $[r]$ is written as

$$\mathbf{z}^{[r]}, \mathbf{y}^{[r-1/2]} \sim \mathbf{z}, \mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{[r-1]} \tag{8.15}$$

$$\boldsymbol{\beta}_{kj}^{[r]}, \boldsymbol{y}_{[rk]}^{j[r]} \sim \boldsymbol{\beta}_{kj}, \boldsymbol{y}_{[rk]}^{j}|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}[r]}, \mathbf{z}^{[r]}, \boldsymbol{\beta}_{k\bar{j}}^{[r]}, \boldsymbol{\Gamma}_k^{[r-1]} \tag{8.16}$$

$$\boldsymbol{\pi}^{[r]} \sim \boldsymbol{\pi}|\mathbf{z}^{[r]} \tag{8.17}$$

$$\boldsymbol{\Gamma}_k^{[r]} \sim \boldsymbol{\Gamma}_k|\mathbf{y}^{[r]}, \mathbf{z}^{[r]}, \tag{8.18}$$

where $\boldsymbol{y}_{[rk]} = \boldsymbol{y}_{\{i:z_i^{[r]}=k\}}$, $\boldsymbol{y}_i^{\bar{j}[r]} = (y_i^{1[r]}, \ldots, y_i^{j-1[r]}, y_i^{j+1[r-1/2]}, \ldots, y_i^{e[r-1/2]})$ and $\boldsymbol{\beta}_{k\bar{j}}^{[r]} = (\boldsymbol{\beta}_{k1}^{[r]}, \ldots, \boldsymbol{\beta}_{kj-1}^{[r]}, \boldsymbol{\beta}_{kj+1}^{[r-1]}, \ldots, \boldsymbol{\beta}_{ke}^{[r-1]})$.

Remark 8.11 (Twice sampling of the Gaussian variable)**.** The Gaussian variable $\mathbf{y}$ is twice generated during one iteration of the Gibbs sampler but, obviously, its stationary distribution stays unchanged. This twice sampling is mandatory because of the strong dependency between $\mathbf{y}$ and $\mathbf{z}$, and between $\boldsymbol{y}_{[rk]}^{j}$ and $\boldsymbol{\beta}_{kj}$.

Remark 8.12 (On the Metropolis-within-Gibbs sampler)**.** If the samplings from (8.17) and (8.18) are classical, the two other ones are more complex. Indeed, the sampling from (8.15) involves to compute the conditional probabilities of the class memberships, so to compute the integral defined in (8.5). If the number of discrete variables is large, this computation is time consuming. However, the sampling from (8.15) can be efficiently performed by one iteration of a Metropolis-Hastings algorithm having $p(z_i, \boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{t}^{(r-1)})$ as stationary distribution. Concerning the sampling according to (8.16), it is performed in two steps. Firstly, the margin parameter is sampled by one iteration of a Metropolis-Hastings algorithm having $p(\boldsymbol{\beta}_{kj}|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\uparrow j(r)}, \mathbf{z}^{(r)}, \boldsymbol{\beta}_k^{\uparrow j(r)}, \boldsymbol{\Gamma}_k)$ as stationary distribution. Secondly, the latent Gaussian vector is sampled from its full conditional distribution.

Remark 8.13 (Twice sampling of the Gaussian variable)**.** The Gaussian variable $\mathbf{y}$ is twice generated during one iteration of the Gibbs sampler but, obviously, its stationary distribution stays unchanged. This twice sampling is mandatory because of the strong dependency between $\mathbf{y}$ and $\mathbf{z}$, and between $\boldsymbol{y}_{[rk]}^{j}$ and $\boldsymbol{\beta}_{kj}$.

We now detail the four steps of the Gibbs sampler and we point out the difficulties to sample from (8.15) and (8.16). Thus, both steps are modified to obtain the Metropolis-within-Gibbs sampler detailed in the next section.

Class membership and Gaussian vector sampling

The aim is to sample from (8.15). By using the independence between the individuals, the vectors $(\mathbf{z}, \mathbf{y})$ are easily sampled conditionally on $(\mathbf{x}, \boldsymbol{\theta}^{[r-1]})$ according to

$$p(\mathbf{z}, \mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{[r-1]}) = \prod_{i=1}^{n} p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{[r-1]}) p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta}^{[r-1]}). \tag{8.19}$$

We now detail both distributions of the right side of the above equation.

— Each $\boldsymbol{z}_i^{[r]}$ is independently sampled from the following multinomial distribution

$$\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{[r-1]} \sim \mathcal{M}_g(t_{i1}(\boldsymbol{\theta}^{[r-1]}), \ldots, t_{ig}(\boldsymbol{\theta}^{[r-1]})), \tag{8.20}$$

where $t_{ik}(\boldsymbol{\theta}^{[r-1]}) = \frac{\pi_k^{[r-1]} p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k^{[r-1]})}{p(\boldsymbol{x}_i; \boldsymbol{\theta}^{[r-1]})}$ is the posterior probability that $\boldsymbol{x}_i$ has been drawn by component k with the parameters $\boldsymbol{\theta}^{[r-1]}$.

— Each $\boldsymbol{y}_i^{[r-1/2]}$ is independently sampled by remarking that the first c elements of $\boldsymbol{y}_i$, denoted by $\boldsymbol{y}_i^{\mathrm{C}}$, are deterministic for a fix triplet $(\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta}^{[r-1]})$ with $z_{ik} = 1$ as such $\boldsymbol{y}_i^{\mathrm{C}} = \Psi(\boldsymbol{x}_i^{\mathrm{C}}; \boldsymbol{\alpha}_k^{[r-1]})$ while its last d elements, denoted by $\boldsymbol{y}_i^{\mathrm{D}}$, are sampled according to a d-variate Gaussian distribution $\mathcal{N}_d(\mathbf{0}, \boldsymbol{\Gamma}_k^{[r-1]})$ truncated on the space $\mathcal{S}_k(\boldsymbol{x}_i^{\mathrm{D}})$

$$p(\boldsymbol{y}_i^{\mathrm{D}}|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta}^{[r-1]}) \propto \prod_{k=1}^{g} \left(\phi_d(\boldsymbol{y}_i^{\mathrm{D}}; \boldsymbol{\mu}_k^{\mathrm{D}[r-1]}, \boldsymbol{\Sigma}_k^{\mathrm{D}[r-1]}) \mathbb{1}_{\{\boldsymbol{y}_i^{\mathrm{D}} \in \mathcal{S}_k(\boldsymbol{x}_i^{\mathrm{D}})\}}\right)^{z_{ik}}, \tag{8.21}$$

where $\boldsymbol{\mu}_k^{\mathrm{D}[r-1]} = \boldsymbol{\Gamma}_{k\mathrm{DC}}^{[r-1]} \boldsymbol{\Gamma}_{k\mathrm{CC}}^{-1[r-1]} \Psi(\boldsymbol{x}_i^{\mathrm{C}}; \boldsymbol{\alpha}_k^{[r-1]})$.

Remark 8.14 (Difficulties to compute $t_{ik}(\boldsymbol{\theta}^{[r-1]})$)**.** Note that the computation of $t_{ik}(\boldsymbol{\theta}^{[r-1]})$ involves to compute the integral defined in (8.5) which can be too much time consuming if d is large ($d > 6$). Thus, the sampling according to (8.19) is also performed by one iteration of a Metropolis-Hastings algorithm avoiding this difficulty and detailed in the next section.

Margin parameter and Gaussian vector sampling

The aim is the sampling from (8.16) which can be decomposed as follows

$$\begin{aligned} p(\boldsymbol{\beta}_{kj}, \boldsymbol{y}_{[rk]}^{j}|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}[r]}, \mathbf{z}^{[r]}, \boldsymbol{\beta}_{k\bar{j}}^{[r]}, \boldsymbol{\Gamma}_k^{[r-1]}) &= p(\boldsymbol{\beta}_{kj}|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}[r]}, \mathbf{z}^{[r]}, \boldsymbol{\beta}_{k\bar{j}}^{[r]}, \boldsymbol{\Gamma}_k^{[r-1]}) \\ &\quad \times p(\boldsymbol{y}_{[rk]}^{j}|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}[r]}, \mathbf{z}^{[r]}, \boldsymbol{\beta}_{k\bar{j}}^{[r]}, \boldsymbol{\beta}_{kj}, \boldsymbol{\Gamma}_k^{[r-1]}). \end{aligned} \tag{8.22}$$

We now detail both distributions of the right side of the above equation.

— The full conditional distribution of $\boldsymbol{\beta}_{kj}$ is defined with an unknown intercept such as

$$p(\boldsymbol{\beta}_{kj}|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}[r]}, \mathbf{z}^{[r]}, \boldsymbol{\beta}_{k\bar{j}}^{[r]}, \boldsymbol{\Gamma}_k^{[r-1]}) \propto p(\boldsymbol{\beta}_{kj}) \prod_{i=1}^{n} \left(p(x_i^j|\boldsymbol{y}_i^{\uparrow j[r]}, \boldsymbol{z}_i^{[r]}, \boldsymbol{\Gamma}_k^{[r-1]}, \boldsymbol{\beta}_{kj})\right)^{z_{ik}^{[r]}}. \tag{8.23}$$

The conditional distribution of $x_i^j|\boldsymbol{y}_i^{\uparrow j[r]}, \boldsymbol{z}_i^{[r]}, \boldsymbol{\Gamma}_k^{[r-1]}$ with $z_{ik}^{[r]} = 1$ used on the right side of the above equation is defined by

$$p(x_i^j|\boldsymbol{y}_i^{\uparrow j[r]}, \boldsymbol{z}_i^{[r]}, \boldsymbol{\Gamma}_k^{[r-1]}, \boldsymbol{\beta}_{kj}) = \begin{cases} \phi_1(\frac{x_i^j - \mu_{kj}}{\sigma_{kj}}; \tilde{\mu}_i, \tilde{\sigma}_i^2)/\sigma_{kj} & \text{if } 1 \leq j \leq c \\ \Phi_1(\frac{b^{\oplus}(x_i^j) - \tilde{\mu}_i}{\tilde{\sigma}_i}) - \Phi_1(\frac{b^{\ominus}(x_i^j) - \tilde{\mu}_i}{\tilde{\sigma}_i}) & \text{otherwise,} \end{cases} \tag{8.24}$$

where the real $\tilde{\mu}_i = \boldsymbol{\Gamma}_k^{[r-1]}[j, \bar{j}] \boldsymbol{\Gamma}_k^{[r-1]}[\bar{j}, \bar{j}]^{-1} \boldsymbol{y}_i^{\uparrow j[r]}$ is the full conditional mean of y_i^j, $\boldsymbol{\Gamma}_k[j, \bar{j}]$ being the row j of $\boldsymbol{\Gamma}_k$ deprived of the element j and $\boldsymbol{\Gamma}_k[\bar{j}, \bar{j}]$ being

the matrix $\mathbf{\Gamma}_k$ deprived of the row and the column j, and where $\tilde{\sigma}_i^2$ is the full conditional variance of y_i^j defined by $\tilde{\sigma}_i^2 = 1 - \mathbf{\Gamma}_k^{[r-1]}[j,\bar{j}]\mathbf{\Gamma}_k^{[r-1]}[\bar{j},\bar{j}]^{-1}\mathbf{\Gamma}_k^{[r-1]}[\bar{j},j]$.

— By the independence between the individuals, the full conditional distribution of $\boldsymbol{y}_{[rk]}^j$ is explicitly defined as

$$p(\boldsymbol{y}_{[rk]}^j|\mathbf{x},\boldsymbol{y}_{[rk]}^{\bar{j}[r]},\mathbf{z}^{[r]},\boldsymbol{\beta}_{k\bar{j}}^{[r]},\boldsymbol{\beta}_{kj},\mathbf{\Gamma}_k^{[r-1]}) = \prod_{i=1}^{n}\left(p(y_i^j|x_i^j,\boldsymbol{y}_i^{\uparrow j[r]},\boldsymbol{z}_i^{[r]},\boldsymbol{\beta}_{kj},\mathbf{\Gamma}_k^{[r-1]})\right)^{z_{ik}^{[r]}}. \quad (8.25)$$

If x^j is a continuous variable (*i.e.* $1 \leq j \leq c$), when $z_i^{[r]} = k$, the full conditional distribution of y_i^j is deterministic such as

$$y_i^{j[r]} = \frac{x_i^j - \mu_{kj}^{[r]}}{\sigma_{kj}^{[r]}}. \quad (8.26)$$

If x^j is a discrete variable (*i.e.* $c+1 \leq j \leq e$), when $\boldsymbol{z}_{ik}^{[r]} = 1$, the full conditional distribution of y_i^j is a truncated Gaussian distribution such as,

$$p(y_i^j|x_i^j,\boldsymbol{y}_i^{\uparrow j[r]},\boldsymbol{z}_i^{[r]},\boldsymbol{\beta}_{kj}^{[r]},\mathbf{\Gamma}_k^{[r-1]}) = \frac{\phi_1(y_i^j;\tilde{\mu}_i,\tilde{\sigma}_i^2)}{p(x_i^j;\boldsymbol{\beta}_{kj}^{[r]})}\mathbb{1}_{\{y_i^j\in[b_k^{\ominus[r]}(x_i^j),b_k^{\oplus[r]}(x_i^j)]\}}, \quad (8.27)$$

where $b_k^{\ominus[r]}(x_i^j) = P(x_i^j - 1;\boldsymbol{\beta}_{kj}^{[r]})$ and $b_k^{\oplus[r]}(x_i^j) = P(x_i^j;\boldsymbol{\beta}_{kj}^{[r]})$.

Remark 8.15 (Difficulties to sample the margin parameters)**.** The sampling of $\boldsymbol{\beta}_{kj}$ is not easily performed since the normalizing constant defined in (8.23) is unknown. This step is then replaced by one iteration of a Metropolis-Hastings algorithm as detailed in the next section. However, note that the sampling of $\boldsymbol{y}_{[rk]}^j$ from (8.27) is easily performed.

Vector of proportions sampling

The aim is the sampling from (8.17) which is classical for the mixture model. The conjugate Jeffreys non informative prior involves that

$$\boldsymbol{\pi}|\mathbf{z}^{[r]} \sim \mathcal{D}_g\left(\mathrm{n}_1^{[r]} + \frac{1}{2},\ldots,\mathrm{n}_g^{[r]} + \frac{1}{2}\right), \quad (8.28)$$

where $\mathrm{n}_k^{[r]} = \sum_{i=1}^n z_{ik}^{[r]}$.

Correlation matrix sampling

The aim is the sampling from (8.18). We use the approach proposed by [Hof07] in the case of semiparameteric Gaussian copula which is divided into two steps. Firstly, a covariance matrix is generated by its explicit posterior distribution, and secondly, the correlation matrix is deduced by normalizing the covariance matrix. When $(\mathbf{y},\mathbf{z})$ are known, we are in the well-known case of a multivariate Gaussian

mixture model with known means. Thus, the sampling according to $\mathbf{\Gamma}_k|\mathbf{y}^{[r]},\mathbf{z}^{[r]}$ is performed by the two following steps

$$\mathbf{\Lambda}_k|\mathbf{y}^{[r]},\mathbf{z}^{[r]} \sim \mathcal{W}^{-1}\left(s_0+\mathrm{n}_k^{[r-1]}, S_0+\sum_{\{i:z_i^{[r]}=k\}} \boldsymbol{y}_i^{[r]T}\boldsymbol{y}_i^{[r]}\right) \tag{8.29}$$

$$\forall 1\leq h,\ell\leq e,\ \mathbf{\Gamma}_k[h,\ell]=\frac{\mathbf{\Lambda}_k[h,\ell]}{\sqrt{\mathbf{\Lambda}_k[h,h]\mathbf{\Lambda}_k[\ell,\ell]}}. \tag{8.30}$$

Remark 8.16 (Sampling of the correlation matrices for the homoscedastic model)**.** As the homoscedastic model assumes the equality between the correlation matrices, in such a case we only sample one $\mathbf{\Lambda}$ so (8.29) is replaced by

$$\mathbf{\Lambda}|\mathbf{y}^{[r]},\mathbf{z}^{[r]} \sim \mathcal{W}^{-1}\left(s_0+n, S_0+\sum_{i=1}^{n} \boldsymbol{y}_i^{[r]T}\boldsymbol{y}_i^{[r]}\right), \tag{8.31}$$

and we put $\mathbf{\Lambda}_k=\mathbf{\Lambda}$ for $k=1,\ldots,g$.

According to both Remarks 8.14 and 8.15, the first two steps of the Gibbs sampler involve difficulties avoided by the following hybrid MCMC algorithm.

8.3.3 Metropolis-within-Gibbs sampler

When some steps of a Gibbs sampler cannot be easily simulated, it may be useful to perform the inference via a hybrid MCMC algorithm [RC04]. Thus, we use the Metropolis-within-Gibbs sampler which replaces both sampling from $\mathbf{z},\mathbf{y}|\mathbf{x},\boldsymbol{\theta}^{[r-1]}$ and $\boldsymbol{\beta}_{kj}|\mathbf{x},\boldsymbol{y}_{[rk]}^{\bar{j}[r]},\mathbf{z}^{[r]},\boldsymbol{\beta}_{k\bar{j}}^{[r]},\mathbf{\Gamma}_k^{[r-1]}$ (defined by (8.15) and (8.23)) by one iteration of two Metropolis-Hastings steps that we now detail.

Class membership and Gaussian vector sampling

The step (8.15) is performed via one iteration of the Metropolis-Hastings algorithm. This algorithm is independently performed to sample each couple $(\boldsymbol{z}_i,\boldsymbol{y}_i)$ since the individuals are independent. Its stationary distribution is

$$p(\boldsymbol{z}_i,\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}^{[r-1]}) \propto \prod_{k=1}^{g}\left(\pi_k^{[r-1]}\phi_e(\boldsymbol{y}_i;\mathbf{0},\mathbf{\Gamma}_k^{[r-1]})\mathbb{1}_{\{\boldsymbol{y}_i^{\mathrm{c}}=\Psi(\boldsymbol{x}_i^{\mathrm{c}};\boldsymbol{\alpha}_k^{[r-1]})\}}\mathbb{1}_{\{\boldsymbol{y}_i^{\mathrm{d}}\in\mathcal{S}_k(\boldsymbol{x}_i^{\mathrm{d}})\}}\right)^{z_{ik}}. \tag{8.32}$$

The Metropolis-Hastings algorithm samples a candidate $(\boldsymbol{z}_i^\star,\boldsymbol{y}_i^\star)$ by the instrumental distribution $q_1(.|\boldsymbol{x}_i,\boldsymbol{\theta}^{[r-1]})$ which uniformly samples $\boldsymbol{z}_i^\star$ then which samples $\boldsymbol{y}_i^\star|\boldsymbol{z}_i^\star$ as follows. Conditionally on $\boldsymbol{z}_{ik^\star}^\star=1$, this instrumental distribution is deterministic for the first c elements of $\boldsymbol{y}_i^\star$, denoted by $\boldsymbol{y}_i^{\star\mathrm{c}}$ such as $\boldsymbol{y}_i^{\star\mathrm{c}}=\Psi(\boldsymbol{x}_i^{\mathrm{c}};\boldsymbol{\alpha}_{k^\star}^{[r-1]})$, while it samples the last d elements of $\boldsymbol{y}_i^\star$ denoted by $\boldsymbol{y}_i^{\star\mathrm{d}}$ according to a *multivariate independent Gaussian* distribution truncated on $\mathcal{S}_{k^\star}(\boldsymbol{x}_i^{\mathrm{d}})$. Thus,

$$q_1(z_i,\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}^{[r-1]}) = \prod_{k=1}^{g}\left(\frac{1}{g}\frac{\phi_d(\boldsymbol{y}_i^{\mathrm{d}};\mathbf{0},\boldsymbol{I})}{\prod_{j=c+1}^{e}p(x_i^j;\boldsymbol{\beta}_{kj}^{[r-1]})}\mathbb{1}_{\{\boldsymbol{y}_i^{\mathrm{c}}=\Psi(\boldsymbol{x}_i^{\mathrm{c}};\boldsymbol{\alpha}_k^{[r-1]})\}}\mathbb{1}_{\{\boldsymbol{y}_i^{\mathrm{d}}\in\mathcal{S}_k(\boldsymbol{x}_i^{\mathrm{d}})\}}\right)^{z_{ik}^\star}. \tag{8.33}$$

The candidate is accepted with the probability

$$\rho_{1i}^{[r]} = \min\left\{ \frac{\prod_{k=1}^{g}\left(\pi_k \phi_e(\boldsymbol{y}_i^{\star}; \mathbf{0}, \boldsymbol{\Gamma}_k^{[r-1]})\right)^{z_{ik}^{\star}}}{\prod_{k=1}^{g}\left(\pi_k \phi_e(\boldsymbol{y}_i^{[r-1]}; \mathbf{0}, \boldsymbol{\Gamma}_k^{[r-1]})\right)^{z_{ik}^{[r-1]}}} \frac{q_1(\boldsymbol{z}_i^{[r-1]}, \boldsymbol{y}_i^{[r-1]} | \boldsymbol{x}_i)}{q_1(\boldsymbol{z}_i^{\star}, \boldsymbol{y}_i^{\star} | \boldsymbol{x}_i)}; 1 \right\}. \tag{8.34}$$

Thus, at iteration $[r]$ of the Algorithm 8.10, the sampling according to (8.15) is performed via one iteration of the following Metropolis-Hastings algorithm.

Algorithm 8.17 (Metropolis-Hastings)**.**

This algorithm has $p(\boldsymbol{z}_i, \boldsymbol{y}_i | \boldsymbol{x}_i, \boldsymbol{\theta}^{[r-1]})$ as stationary distribution. Its is written as follows

$$(\boldsymbol{z}_i^{\star}, \boldsymbol{y}_i^{\star}) \sim q_1(\boldsymbol{z}_i, \boldsymbol{y}_i | \boldsymbol{x}_i) \tag{8.35}$$

$$(\boldsymbol{z}_i^{[r]}, \boldsymbol{y}_i^{[r-1/2]}) = \begin{cases} (\boldsymbol{z}_i^{\star}, \boldsymbol{y}_i^{\star}) & \text{with probability } \rho_{1i}^{[r]} \\ (\boldsymbol{z}_i^{[r-1]}, \boldsymbol{y}_i^{[r-1]}) & \text{with probability } 1 - \rho_{1i}^{[r]}. \end{cases} \tag{8.36}$$

Margin parameter sampling

The step (8.16) is performed in two steps. Firstly the sampling of $\boldsymbol{\beta}_{kj}^{[r]}$ according to (8.23) is performed via one iteration of the Metropolis-Hastings algorithm whose the stationary distribution is $p(\boldsymbol{\beta}_{kj} | \mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}[r]}, \mathbf{z}^{[r]}, \boldsymbol{\beta}_{k\bar{j}}^{[r]}, \boldsymbol{\Gamma}_k)$. Secondly, the sampling of $\boldsymbol{y}_{[rk]}^{j[r]}$ is performed according to its conditional distribution given by (8.27). The instrumental distribution of the Metropolis-Hastings algorithm $q_2(.|\mathbf{x}, \mathbf{z})$ samples a candidate $\boldsymbol{\beta}_{kj}^{\star}$ according to the posterior distribution of $\boldsymbol{\beta}_{kj}$ under the conditional independence assumption (this distribution is explicit since the conjugate prior distributions are used). So,

$$q_2(.|\mathbf{x}, \mathbf{z}) = p(\boldsymbol{\beta}_{kj} | \mathbf{x}, \mathbf{z}, \boldsymbol{\Gamma}_k = \boldsymbol{I}). \tag{8.37}$$

Thus, according to (8.23), the candidate $\boldsymbol{\beta}_{kj}^{\star}$ is accepted with the probability

$$\rho_2^{[r]} = \min\left\{ \frac{p(\boldsymbol{\beta}_{kj}^{\star}) q_2(\boldsymbol{\beta}_{kj}^{[r-1]} | \mathbf{x}, \mathbf{z})}{p(\boldsymbol{\beta}_{kj}^{[r-1]}) q_2(\boldsymbol{\beta}_{kj}^{\star} | \mathbf{x}, \mathbf{z})} \prod_{\{i: z_i^{[r]} = k\}} \frac{p(\boldsymbol{y}_i^j | x_i^j, \boldsymbol{y}_i^{\uparrow j[r]}, \boldsymbol{z}_i, \boldsymbol{\beta}_{kj}^{\star}, \boldsymbol{\Gamma}_k^{[r-1]})}{p(\boldsymbol{y}_i^j | x_i^j, \boldsymbol{y}_i^{\uparrow j[r]}, \boldsymbol{z}_i, \boldsymbol{\beta}_{kj}^{[r-1]}, \boldsymbol{\Gamma}_k^{[r-1]})}; 1 \right\}.$$

Thus, at iteration $[r]$ of the Algorithm 8.10, the sampling from (8.16) is performed via one iteration of the following Metropolis-Hastings algorithm.

Algorithm 8.18 (Metropolis-Hastings).

This algorithm has $p(\boldsymbol{\beta}_{kj}|\boldsymbol{x}_{[rk]}, \boldsymbol{y}_{[rk]}^{\bar{j}[r]}, \mathbf{z}, \boldsymbol{\beta}_{k\bar{j}}^{[r]}, \boldsymbol{\Gamma}_k)$ as stationary distribution. It is written as follows

$$\boldsymbol{\beta}_{kj}^{\star} \sim q_2(\boldsymbol{\beta}_{kj}|\mathbf{x}, \mathbf{z}) \tag{8.38}$$

$$\boldsymbol{\beta}_{kj}^{[r]} = \begin{cases} \boldsymbol{\beta}_{kj}^{\star} & \text{with probability } \rho_2^{[r]} \\ \boldsymbol{\beta}_{kj}^{[r-1]} & \text{with probability } 1-\rho_2^{[r]}. \end{cases} \tag{8.39}$$

Remark 8.19 (Instrumental distributions)**.** Note that, the smaller are the intra-class dependencies of the variable $\boldsymbol{x}_i$, the closer of the stationary distributions are the instrumental distributions of both Metropolis-Hastings algorithms.

8.3.4 Label switching problem

The label switching problem is generally solved by specific procedures [Ste00b]. However, based on the argument developed in [JP14], these techniques are principally impacting when g is known.

When the model is used to cluster, the number of classes is unknown, and the model selection is performed by the BIC criterion which simultaneously avoids the label switching phenomenon. Indeed, on the one hand, this criterion selects quite separated classes when the sample size is small, so the label switching is not present in practice because of the class separability. On the other hand, even if it can select more classes when the sample size increases, the label switching problem is settled since this phenomenon vanishes asymptotically.

Obviously, when the number of classes is fixed and the size of sample is small, the label switching problem can occur. In such a case, our advice is naturally to use the procedures detailed in [Ste00b].

8.4 Numerical experiments on simulated data sets

In order to illustrate the properties of the model, two numerical experiments are performed. The first one consists in simulating data according to the proposed model and to study the convergence of the estimates. The second one consists in simulating data according to a mixture of Poisson distributions [KT08] in order to show the robustness of the proposed model. The estimate is computed by averaging the parameters sampled by the Gibbs algorithm.

Experiment conditions

For each situation, 100 samples are generated and the algorithm is initialized with the maximum likelihood estimate of the conditional independence model. A burn-in is performed during 1000 iterations even if the parameter initialization is relevant when the intra-class dependencies are small. The algorithm is stopped after

1000 iterations. The maximum *a posteriori* estimate is approximated by the mean of the sampled parameters. The Kullback-Leibler divergence is approximated via 10000 iterations of a Monte-Carlo method.

Simulation 8.20 (Mixed variables: one continuous, one integer and one binary)**.** We consider the mixture model of Gaussian copulas detailed in Example 8.9 and composed with one continuous variable, one integer variable and one binary variable. Figure 8.2 illustrates the decreasing behavior of the Kullback-Leibler divergence of the model with the maximum *a posteriori* estimate from the model with the true parameters according to the sample size in the mixed case. This simulation illustrates the good behavior of the Metropolis-within-Gibbs algorithm. Furthermore, the approximation of the maximum *a posteriori* estimate by the mean of the parameters sampled by this algorithm is efficient.

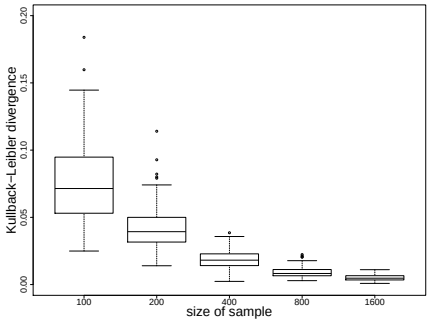


Figure 8.2 – Decrease of the Kullback-Leibler divergence of the model with the maximum *a posteriori* estimate from the model with the true parameter.

Simulation 8.21 (Robustness of the mixture model of Gaussian copulas)**.** During these experiments, data are sampled according to a bivariate Poisson mixture model [KT08] whose the margin parameters are denoted by $\boldsymbol{\alpha}_k = (\lambda_{k1}, \lambda_{k2}, \lambda_{k3})$. The simulation is performed with the following values of the parameters

$$\boldsymbol{\pi} = (1/3, 2/3),\ \lambda_{1h} = h \text{ and } \lambda_{2h} = 3 + h,\ \text{for } h = 1, 2, 3. \tag{8.40}$$

The error rate of this model computed with the Bayes' rule is equal to 9.5%. Results show that the flexibility of the mixture model of Gaussian copulas allows to efficiently fit these simulated data. Indeed, the Kullback-Leibler divergence becomes very small when the size of the sample increases. Furthermore, the error rate of the model seems to converge to a value just a little bit larger than the theoretical one (9.5%). We also note that the margin parameters of both components and the correlation coefficients seem to converge to their true values.

8.5 Analysis of three real data sets

We now cluster three real data sets by using the mixture model of Gaussian copulas. The parameters are estimated via the Metropolis-with-Gibbs algorithm initialized on the maximum likelihood estimate of the conditional independence

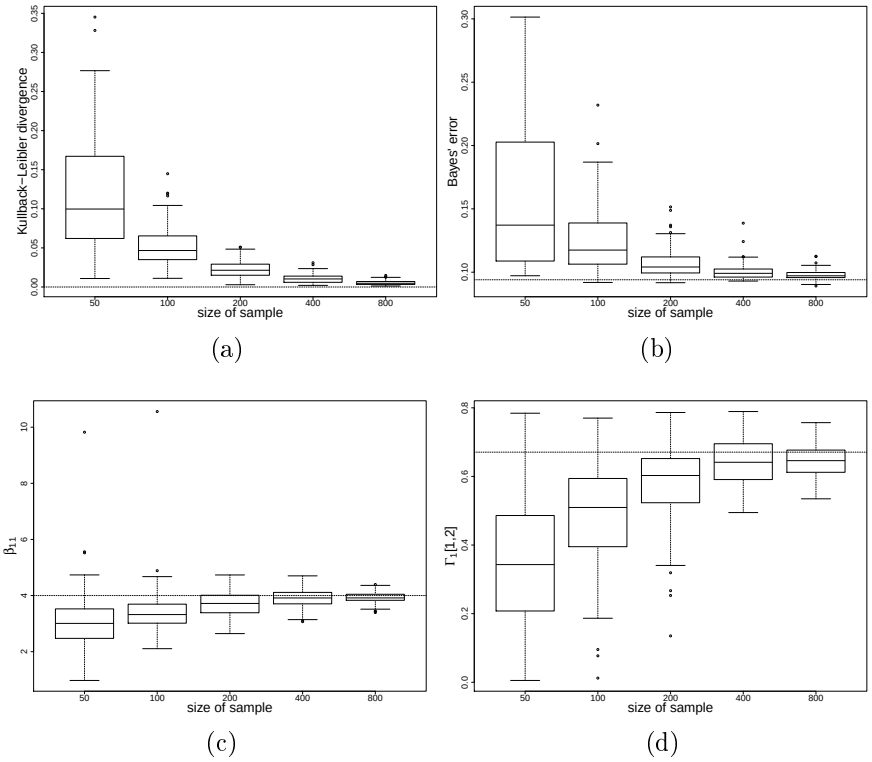


Figure 8.3 – Results of Simulation 4.2: (a) Kullback-Leibler divergence of the estimated model from the true one; (b) Error rate of the estimated model; (c) Value of the first margin parameter for the class 1; (d) Value of the correlation coefficient between both variables for class 1.

model. A burn-in is performed during 1000 iterations even if the parameter initialization is relevant when the intra-class dependencies are small. The algorithm is stopped after 1000 iterations and the estimate is obtained by taking the mean of the sampled parameters. The model selection is performed by using two information criteria (BIC criterion [Sch78], ICL criterion [BCG00]) computed on the maximum *a posteriori* estimate.

8.5.1 Liver disorder data set

The data

This data set [For90] describes 345 individuals by five blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption (five continuous variables) and by the number of quart-pint equivalents of alcoholic beverages drunk per day (one integer variable).

Model selection

We estimate the three mixture models (conditional independence one, heteroscedastic Gaussian copula mixture and homoscedastic Gaussian copula mixture) for different numbers of classes. Table 8.1 presents the values of both used information

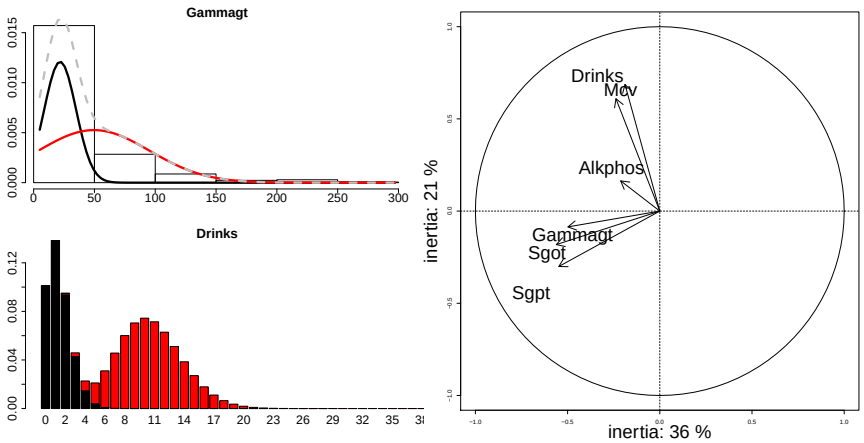
criteria. The values of both criteria obtained with the bi-component homoscedastic mixture model of Gaussian copulas are the best ones. However, note that the three models select two components.

	g	1	2	3	4	5	6
BIC	cond. indpt.	-8690	**-8017**	-8039	-8092	-8130	-8235
	hetero.	-8551	**-7935**	-8103	-8157	-8277	-8287
	homo.	-8551	**-7898**	-7999	-8032	-8050	-8123
ICL	cond. indpt.	-8690	**-8026**	-8060	-8117	-8208	-8341
	hetero.	-8551	**-7943**	-8120	-8171	-8322	-8306
	homo.	-8551	**-7907**	-8032	-8043	-8088	-8205

Table 8.1 – Values of the BIC and ICL criteria for the three mixture models estimated on the liver disorder data set.

Interpretation of the best model

We now describe the best model according to both criteria (the homoscedastic bi-component mixture model of Gaussian copulas) by using the margin parameters and the intra-class dependencies summarized by Figure 8.4. The model considers two classes whose the majority one ($\pi_1 = 0.60$) groups the individuals having a strong alcoholic consumption ($\boldsymbol{\beta}_{1\text{drinks}} = 10.6$) and large values of the five blood tests especially for the tests Sogt and Gammagt. The miniority class groups the individuals having a small alcoholic consumption ($\boldsymbol{\beta}_{2\text{drinks}} = 1.36$) and smaller values of the blood tests. For both classes, the three following blood tests are positively correlated with Sgpt, Sopt and Gammagt while the test Mcv is positively correlated with the number of alcoholic drinks.

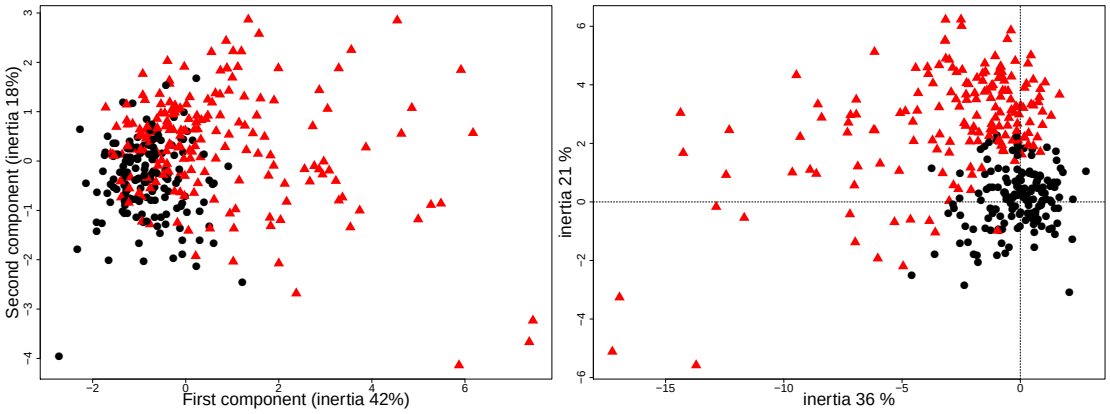


(a) $\pi_k p(x^j|\boldsymbol{\beta}_{kj}, z = k)$ (b) Variables in the first correlation circle given by $\boldsymbol{\Gamma}_k$

Figure 8.4 – Summary of the homoscedastic bi-component mixture model of Gaussian copulas for the liver disorder data set. Class 1 is displayed in black and Class 2 in red.

Partition study

As all the variables are numerical, Figure 8.5a can display the individuals and their class memberships in the first classical PCA map. However, as classes are not well separated in this map, the structure of the data is not brought out. Thus, Figure 8.5b displays the individuals in the first PCA map of class 1. In this map, classes are better separated since the first class (black circles) is centred while the second class (red triangles) is on the top part of the graphic. So, the second axis is discriminant. This summary is in agreement with the class interpretation since this axis is built by the variables *Mcv* and *drinks* which are themselves discriminant according to their margin parameters.



(a) First map of PCA (b) First map of the PCA of the class 1

Figure 8.5 – Visualization of the partition by the homoscedastic bicomponent mixture model of Gaussian copulas for the liver disorder data set (Class 1 is drawn by black circles and Class 2 by red triangles).

Note that the partitions obtained by the three bi-component models are similar but not identical as shown by Table 8.2.

	hetero.	
	c1	c2
c1-homo.	190	0
c2-homo.	5	150

(a)

	cond. indpt.	
	c1	c2
c1-homo.	190	0
c2-homo.	7	148

(b)

Table 8.2 – Confusion matrices between the partition obtained by the homoscedastic bi-component model and the partition obtained by: (a) the heteroscedastic bi-component model; (b) the conditional independence model.

Conclusion

On this data set, the mixture model of Gaussian copulas better fits the data according to the information criteria than the conditional independence model, even

if both models select the same number of classes. The PCA per class allows to summarize the intra-class dependencies and to bring out the separation of both classes hidden by a classical PCA.

8.5.2 Wine data set

The data

The data set [CCA+09] contains 6497 variants of the Portuguese "Vinho Verde" wine (1599 red wines and 4898 white wines) described by eleven physiochemical continuous variables (fixed acidity, volatile acidity, citric acidity, residual sugar, chlorides, free sulfur dioxide, total density dioxide, density, pH, sulphates, alcohol) and one integer variable (quality of the wine evaluated by experts). The kinds of the wines (red or white) are hidden and we cluster the data set with three different mixture models. Note that one white wine (number 4381) is excluded of the study since it is an outlier.

Model selection

We estimate the three mixture models (conditional independence one, heteroscedastic Gaussian copula mixture and homoscedastic Gaussian copula mixture) for different numbers of classes and we present the values of both used information criteria in Table 8.3. Both criteria distinctly select the bi-component heteroscedastic mixture model of Gaussian copulas. We now show that this model allows to well separate the white wines from the red ones then we give the model interpretation.

	g	1	2	3	4	5	6
BIC	cond. indpt.	-63516	-61069	-61010	**-55967**	-60250	-57163
	hetero.	-44675	**-34520**	-39724	-44692	-44484	-48349
	homo.	-44675	-39372	**-38289**	-45209	-43217	-42417
ICL	cond. indpt.	-63516	-61229	-61365	**-56310**	-60726	-58138
	hetero.	-44675	**-34688**	-40176	-44933	-44758	-48959
	homo.	-44675	-39607	**-38791**	-45380	-43345	-42667

Table 8.3 – Values of the BIC and ICL criteria for the three mixture models estimated on the wine data set.

Partition study

Table 8.4 presents the confusion matrices in order to compare the relevance of the estimated partitions according to the true one (wine color). These results strengthen the idea that the model best fitting the data is the bi-component heteroscedastic Gaussian copula mixture models. Indeed, its partition is the closest to the true one.

Figure 8.6 displays the individuals in a PCA map of both classes estimated by the bi-component heteroscedastic mixture model of Gaussian copulas. According to these scatter-plots, classes are well-separated. We now detail its parameters.

	white	red
c1	4359	9
c2	538	1590

(a) Adj. Rand: 0.68

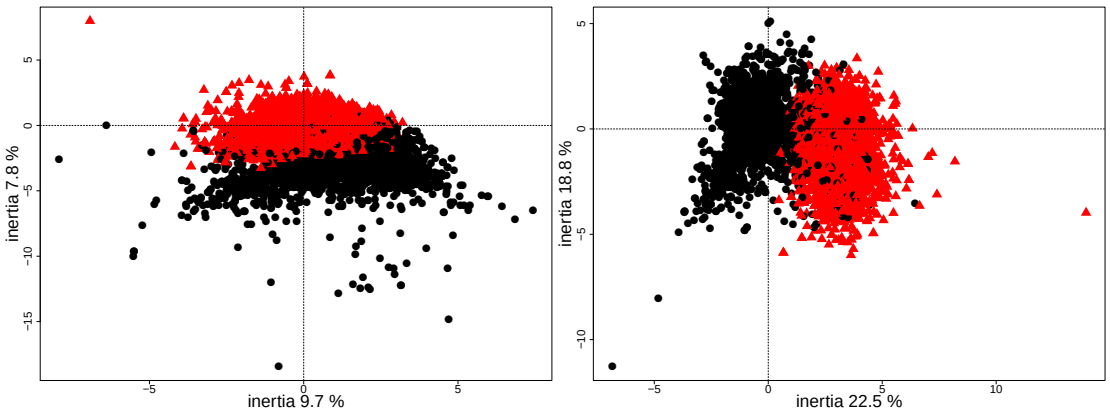
	white	red
c1	2441	12
c2	1911	7
c3	545	1580

(b) Adj. Rand: 0.30

	white	red
c1	2547	1561
c2	2007	35
c3	275	3
c4	68	0

(c) Adj. Rand: 0.00

Table 8.4 – Values of the adjusted Rand index and confusion matrices between the true partition and the estimated partition by: (a) the bi-component heteroscedastic Gaussian copula mixture; (b) the tri-component homoscedastic Gaussian copula mixture; (c) the four-component conditional independence mixture.



(a) PCA of the class 1, map(3,5)

(b) PCA of the class 2, map(1,2)

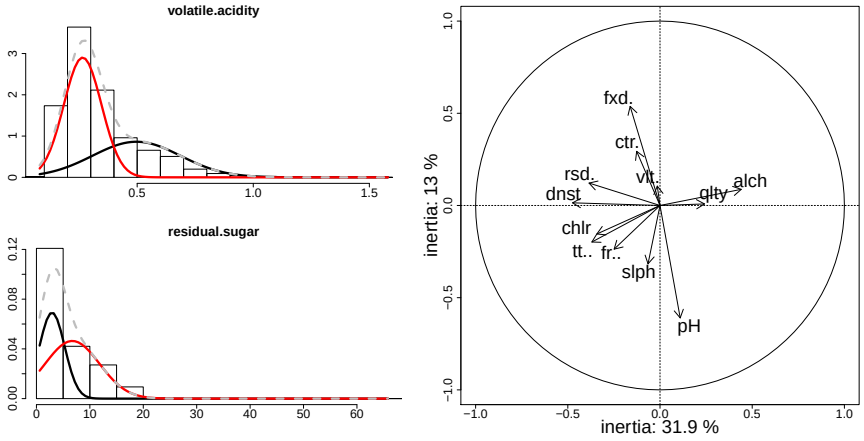
Figure 8.6 – Visualization of the partition by the heteroscedastic bicomponent mixture model of Gaussian copulas for the wine data set (Class 1 is drawn by black circles and Class 2 by red triangles).

Interpretation of the best model

The following interpretation is based on the margin parameters of the components and on the intra-class correlation matrices summarized by Figure 8.7. The majority class ($\pi_1 = 0.59$) is principally composed with white wines. This class is characterized by lower rates of acidity, pH, chlorides and sulphites than them of the minority class ($\pi_2 = 0.41$) which is principally composed by red wines. The majority class has larger values for both sulfur dioxide measures and the alcoholic rate. Note that the wine quality of both classes is similar ($\beta_{1\text{quality}} = 5.96$ and $\beta_{2\text{quality}} = 5.58$). The majority class is characterized by a strong correlation between both sulfur measures opposite to a strong correlation between the density and acidity measures. The minority class underlines that the wine quality is dependent with a larger alcoholic rate and small values for the chlorides and acidity measures.

Conclusion

On this data set, the Gaussian copula mixture models allows to reduce the number of classes and to better fit the data. Furthermore, its impact on the estimated



(a) $\pi_k p(x^j|\boldsymbol{\beta}_{kj}, z = k)$

(b) Variables in the first correlation circle given by $\boldsymbol{\Gamma}_2$

Figure 8.7 – Summary of the heteroscedastic bi-component Gaussian copula mixture model for the wine data set. Class 1 is drawn in black and Class 2 in red.

partition is significant. Based on the individual scatter-plots in the model PCA, the estimated classes are relevant since they are well-separated. Finally, the estimation of the intra-class dependencies helps the interpretation since its underlines the link between the wine quality of the minority class and its physiochemical properties.

8.5.3 Forest fire data set

The data

This data set describes 517 forest fires [CM07] in the north-east region of Portugal by using meteorological variables: seven continuous variables (four about the FWI system: FFMC, DMC, DC, ISI and two about the meteorology: temperature and relative humidity), two integer variables relative to the spatial coordinates and three binary ones indicating the presence of rain, the season (summer or not summer) and the day (week-end or not week-end).

Model selection

Table 8.5 presents the values of both used information criteria for the three mixture models. According to both criteria, the model better fitting the data is the homoscedastic mixture model of Gaussian copulas with three components.

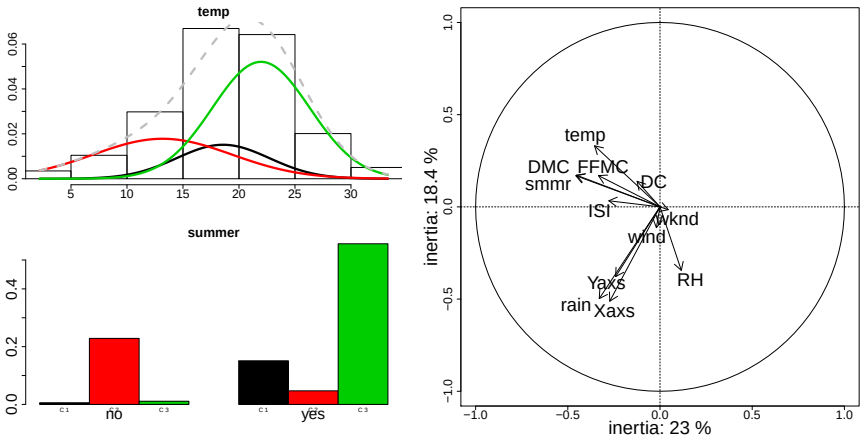
Interpretation of the best model

The following interpretation is based on the margin parameters on the intra-class correlation matrices summarized in Figure 8.8. The majority class ($\pi_1 = 0.57$) groups the fires developed with high temperature and small relative humidity. The measures of FMC, DMC and ISI are high. The second class ($\pi_2 = 0.26$) groups the winter fires. These fires are developed with a strong wind and no rain. All

	g	1	2	3	4	5	6
BIC	cond. indpt.	-16559	**-16296**	-16473	-17370	-17379	-17454
	hetero	-16559	**- 16002**	-16171	-16410	-16666	-16791
	homo.	-16559	-15899	**-15824**	-16300	-15946	-16034
ICL	cond. indpt.	-16559	**-16301**	-16494	-17401	-17400	-17527
	hetero	-16559	**-16014**	-16205	-16471	-16721	-16871
	homo.	–16559	-15907	**-15893**	-16352	-16020	-16137

Table 8.5 – Values of the BIC and ICL criteria for the three mixture models estimated on the forest fire data set.

the FWI measures take small values. The minority class ($\pi_3 = 0.17$) groups the summer fires developed with few values of FWI measures except the DC one. The temperature is median but the relative humidity is high. The intra-class correlation matrix underlines the dependencies between the summer and high temperature and values of FFMC and DMC. Finally, note that the space coordinates roughly follow the same distribution in the three classes.



(a) $\pi_k p(x^j|\boldsymbol{\beta}_{kj}, z = k)$

(b) Variables in the first correlation circle given by $\boldsymbol{\Gamma}_k$

Figure 8.8 – Summary of the homoscedastic bi-component mixture model of Gaussian copulas for the forest fire data set. Class 1 is displayed in green, Class 2 in red and Class 3 in black.

Partition study

Note that the partitions obtained by the three models are similar but not identical as shown by Table 8.6.

	hetero.	
	c1	c2
c1-homo.	244	23
c2-homo.	1	127
c3-homo.	122	0

(a)

	cond. indpt.	
	c1	c2
c1-homo.	265	2
c2-homo.	7	121
c3-homo.	111	11

(b)

Table 8.6 – Confusion matrices between the partition obtained by the homoscedastic tri-component model and the partition obtained by: (a) the heteroscedastic bi-component model; (b) the conditional independence model.

Conclusion

The model points out three classes of forest fires. It is more precise than the conditional independence model which roughly separates the summer fires from the other ones. Indeed, the homoscedastic mixture model of Gaussian copulas considers two kinds of summer fires. The restrictions done on the parameters spaces allow to better fit the data than the heteroscedastic Gaussian copula mixture model according to both criteria. Its impact is significant since the numbers of classes selected by both models are different.

8.6 Conclusion

The mixture model of Gaussian copulas uses the properties of copulas: independent choice of the margin distributions and of the dependency relations. Thus, this mixture allows to fix classical distributions belonging to the exponential family for the one-dimensional margin distributions of each component. Moreover, it takes into account the intra-class dependencies. An approach based on a PCA per class of the Gaussian latent variable allows also to summarize the main intra-class dependencies and to visualize the data by using the model parameters.

During both numerical experiments and applications, we pointed out that this model is sufficiently flexible to fit data drawn by an other one. Furthermore, it can reduce the biases of the conditional independence model (for instance the reduction of the number of classes).

The number of parameters increases with the numbers of classes and variables especially because of the correlation matrices of the Gaussian copulas. To avoid this drawback, we propose a homoscedastic version of the model assuming the equality between the correlation matrices. This model may better fit the data than the heteroscedastic Gaussian copula mixture model. However, it can be large when the number of variables increases. So, more parsimonious correlation matrices could be proposed to avoid this drawback in future works.

Finally, the model can not cluster non-ordinal categorical variables having more than two modalities. Indeed, in such case, the cumulative distribution function is not defined. An artificial order between the modalities could be added to define a cumulative distribution function but this method has three potential difficulties for which attention has to be paid: it assumes regular dependencies between the

modalities of two variables, its estimation would slow down the estimation algorithm and its stability would have to be studied.

Conclusion of Part II

We have seen that it is important to perform the cluster analysis in the native space of the variables, in order to provide meaningful results. Even if the method-based on mixture models is relevant, it suffers from a lack of multivariate distributions for mixed data.

The assumption of the conditional independence between the variables gives a meaningful model since it provides classical distributions for the one-dimensional margins of the components. This model is relevant especially when the sample size is small according to the number of variables. Indeed, in such a case, the information on the intra-class dependency is not present in the data set. However, it can be necessary to relax the conditional independence assumption.

The mixture of location models and its extension per blocks is also an alternative to the conditional independent model. Note that its extension per blocks appears to be more efficient since the number of parameters stays limited. However, the model interpretation can be difficult to perform by the practitioner since the one-dimensional margin distributions of the components are not classical for the continuous variables.

The second alternative consists in the underlined Gaussian mixture model which appears as more meaningful. However, its method used for the parameter estimation dramatically limits the number of discrete variables.

In this context, two main objectives appear to us as crucial: the model must to provide classical one-dimensional margin distributions for its components, and it must provide meaningful coefficients reflecting the intra-class dependencies. Based on both objectives, we have proposed two mixture models.

The first model allows to perform the cluster analysis of data sets with continuous and categorical variables. It derives from the multilevel latent class model developed for intra-class dependent categorical variables. Indeed, the components of this model are composed with a Gaussian and by logistic distributions. The specificity of our approach is to simultaneously perform the model selection and the parameter estimation in a GEM algorithm.

The second model is a mixture of Gaussian copulas. This model is very general since it performs the cluster analysis of mixed data sets with variables admitting a cumulative distribution functions. Moreover, it provides some visualization tools to summarize the intra-class dependencies and to display the individuals. However, the model complexity increases with the number of variables, even for the homoscedastic version. Thus, this model appears as unappropriated for data sets with large number of variables. So, more parsimonious versions could be considered to cluster such data sets.

General conclusion and perspectives

Conclusion

In this thesis, we have been interested in the cluster analysis of complex data. More precisely, we have focused on the categorical and mixed data sets. The objective was to introduce model-based approaches in order to cluster such data by modelizing the intra-class dependencies. Moreover, these models had to summarize the data distribution by few parameters to facilitate the interpretation.

Two models have been presented to perform the cluster analysis of categorical data sets. The main idea is to group the variables into conditionally independent blocks and to put a parsimonious distribution for each block. The combinatorial problems are ubiquitous when the categorical data sets with intra-class dependencies are analyzed. So, the presented models suffer from this problem during the model selection step. Even if two Bayesian approaches reduce this drawback, it is not realistic to perform, with these models, the cluster analysis of a data set with a lot of variables. However, when the variables are ordinal or binary, a possible answer to this problem can be given by the model-based copulas.

The mixture model of Gaussian copulas has been introduced to cluster mixed data sets. This model permits to obtain classical one-dimensional margins for each components and to modelize the intra-class dependencies. The general mixture model of Gaussian copulas does not suffer from combinatorial problems to perform the model selection. Thus, this model can be an efficient alternative to the model specific to the categorical data when the variables are binary or ordinal. Indeed, it avoids the combinatorial problems of the model selection.

Perspectives

Throughout this thesis, we have defined a class by the set of the individuals drawn by the same distribution. However, alternative definitions of a class could be used [BRC+10, Hen10].

The models have been introduced in a clustering framework. Obviously, they can be used in a semi-supervised or in a classification context. However, one can expect that these models outperform the discriminant approaches (like the logistic regression) only when few individuals are labeled. Indeed, their objective is more

ambitious than the discriminant approaches since they modelize the data distribution while the discriminant approaches focus on the boundaries between classes.

The models introduced in this thesis could manage data sets with missing values. Indeed, their estimation can be performed by an EM algorithm or by a Gibbs sampler which are known to manage such data.

The models which estimate a covariance matrix can require a large number of parameters. So, it is important to introduce some parsimonious versions of the mixture model of Gaussian copulas in order to manage data sets with a large number of variables. Based on the geometric approaches used for the Gaussian mixture models [CG95], some constrains could be added on the correlation matrix of the Gaussian copulas. However, this approach could make the estimation harder. Other parsimonious approaches inherited from the Gaussian framework could also be used (for instance the models for large data sets [BB14]). Finally, another research axis could consist in a generalization of the mixture model of dependency trees. Indeed, the copulas define the bivariate distribution for any couples of mixed variables (if they admit a cdf). Moreover, the model selection step of this method is classical. Finally, this approach could allow an inference by maximization of the likelihood by using methods inherited from [Eve88]. Thus, no *a priori* information would be added with this method of inference.

The correlation coefficient of the Gaussian copula has good properties when the margin distributions are well estimated. However, we have set the margin distributions of the components for the mixture model of Gaussian copulas. Thus, a semi-parametric approach (for instance, based on the works of [Hof07, HNW11]) since the properties of the correlation coefficient would be asymptotically guaranteed.

Finally, the mixture of the two extreme dependency distributions would be an alternative to the Gaussian copulas. This model can be defined by using copulas. Indeed, the maximum dependency distribution could be defined as the distribution which attains the Fréchet-Hoeffding upper bound. By adding some constraints (like the structure in tree), the model selection could be easily performed.

It hurts to set you free
But you'll never follow me
The end of laughter and soft lies
The end of nights we tried to die
This is the end
The Doors—The end.

Appendix A

Appendix of Part I

A.1 Generic identifiability of the mixture of the two extreme dependency distributions

The block distribution is generically identifiable when the block contains at least three variables or when the block contains at least two variables having at least three modalities. To prove this propoperty, we firstly show the generic identifiability of the model in both of the following simple cases: two variables with three modalities and three binary variables. Then, we conclude to the generic identifiability of the model.

Proposition A.1 (Two variables with three modalities)**.** *The mixture model of the two extreme dependency distributions is generically identifiable when* $d^{\{kb\}} = 2$, $m_1^{\{kb\}} = m_2^{\{kb\}} = 3$.

Proof. Suppose that there exists $\boldsymbol{\alpha}_{kj} = (\rho_{kj}, \boldsymbol{\xi}_{kb}, \boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb})$ and $\tilde{\boldsymbol{\alpha}}_{kj} = (\tilde{\rho}_{kj}, \tilde{\boldsymbol{\xi}}_{kb}, \tilde{\boldsymbol{\tau}}_{kb}, \tilde{\boldsymbol{\delta}}_{kb})$ as such

$$\forall \boldsymbol{x}_i^{\{kb\}}\ p(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\alpha}_{kb}) = p(\boldsymbol{x}_i^{\{kb\}}; \tilde{\boldsymbol{\alpha}}_{kb}). \tag{A.1}$$

We demonstrate that this equality involves that $\boldsymbol{\alpha}_{kj} = \tilde{\boldsymbol{\alpha}}_{kj}$. The demonstration is split in three parts which are determined by the three possibilities of $(\boldsymbol{\delta}_{kb}, \tilde{\boldsymbol{\delta}}_{kb})$ (equality, one relation equal for both parameters, no relation equal for both parameters). We show that (A.1) involves the equality between the dependency relations (*i.e.* $\boldsymbol{\delta}_{kb} = \tilde{\boldsymbol{\delta}}_{kb}$) and between the continuous parameters. Thus, (A.1) involves $\boldsymbol{\alpha}_{kb} = \tilde{\boldsymbol{\alpha}}_{kb}$.

• *Equality of the dependency relations* (*i.e.* $\boldsymbol{\delta}_{kb} = \tilde{\boldsymbol{\delta}}_{kb}$)
Without loss of generality, we assume that

$$\forall h, h' \in \{1, \ldots, 3\},\ h \neq h' :\ \delta_{kb}^{h2h} = 1 \text{ and } \delta_{kb}^{h2h'} = 0.$$

Then, the relation defined by (A.1) leads to the following system of nine equations for $h \in \{1, \ldots, 3\}$ and $h' \in \{1, \ldots, 3\} \setminus \{h\}$:

$$\left\{ \begin{array}{rl} (1-\rho_{kb})\xi_{kb}^{1h}\xi_{kb}^{2h} + \rho_{kb}\tau_{kb}^{h} & = (1-\tilde{\rho}_{kb})\tilde{\xi}_{kb}^{1h}\tilde{\xi}_{kb}^{2h} + \tilde{\rho}\tilde{\tau}_{kb}^{h} \\ (1-\rho_{kb})\xi_{kb}^{1h}\xi_{kb}^{2h'} & = (1-\tilde{\rho}_{kb})\tilde{\xi}_{kb}^{1h}\tilde{\xi}_{kb}^{2h'}. \end{array} \right. \tag{A.2}$$

We use the second line of the previous system with the following values of the couple (h, h'): (1,3), (2,3), (1,2) and (3,2). Thus, we obtain that

$$\frac{\xi_{kb}^{11}}{\xi_{kb}^{12}} = \frac{\tilde{\xi}_{kb}^{11}}{\tilde{\xi}_{kb}^{12}} \text{ and } \frac{\xi_{kb}^{11}}{\xi_{kb}^{13}} = \frac{\tilde{\xi}_{kb}^{11}}{\tilde{\xi}_{kb}^{13}}. \tag{A.3}$$

So, $\tilde{\xi}_{kb}^{11} = \xi_{kb}^{11}\frac{\tilde{\xi}_{kb}^{12}}{\xi_{kb}^{12}} = \xi_{kb}^{11}\frac{\tilde{\xi}_{kb}^{13}}{\xi_{kb}^{13}}$. There is a intercept $\varepsilon \in \mathbb{R}^+$ such that $\varepsilon = \frac{\tilde{\xi}_{kb}^{12}}{\xi_{kb}^{12}} = \frac{\tilde{\xi}_{kb}^{13}}{\xi_{kb}^{13}}$. We remind that $\sum_{h=1}^{3} \xi_{kb}^{1h} = \sum_{h=1}^{3} \tilde{\xi}_{kb}^{1h} = 1$. Moreover,

$$\sum_{h=1}^{3} \tilde{\xi}_{kb}^{1h} = \xi_{kb}^{11}\varepsilon + \xi_{kb}^{12}\varepsilon + \xi_{kb}^{13}\varepsilon = \varepsilon. \tag{A.4}$$

So, $\varepsilon = 1$. We conclude that $\xi_{kb}^{1h} = \tilde{\xi}_{kb}^{1h}$. The same reasoning is used to obtain that $\xi_{kb}^{2h} = \tilde{\xi}_{kb}^{2h}$. From this, we obtain the equality between $\rho_{kb} = \tilde{\rho}_{kb}$ and $\tau_{kb}^{h} = \tilde{\tau}_{kb}^{h}$. Finally, we obtain that $\boldsymbol{\alpha}_{kb} = \tilde{\boldsymbol{\alpha}}_{kb}$.

• *Only one relation is equal between both parametrizations*
Without loss of generality, we assume that $\delta_{kb}^{121} = \delta_{kb}^{222} = \delta_{kb}^{323} = 1$ and $\delta_{kb}^{h2h'} = 0$ otherwise while $\tilde{\delta}_{kb}^{122} = \tilde{\delta}_{kb}^{221} = \tilde{\delta}_{kb}^{323} = 1$ and $\tilde{\delta}_{kb}^{h2h'} = 0$.
From the system of nine equations defined by (A.1), we extract the following system

$$\begin{cases} (1-\rho_{kb})\xi_{kb}^{13}\xi_{kb}^{21} &= (1-\tilde{\rho}_{kb})\tilde{\xi}_{kb}^{13}\tilde{\xi}_{kb}^{21} \\ (1-\rho_{kb})\xi_{kb}^{13}\xi_{kb}^{22} &= (1-\tilde{\rho}_{kb})\tilde{\xi}_{kb}^{13}\tilde{\xi}_{kb}^{22} \\ (1-\rho_{kb})\xi_{kb}^{11}\xi_{kb}^{21} + \rho_{kb}\tau_{kb}^{1} &= (1-\tilde{\rho}_{kb})\tilde{\xi}_{kb}^{11}\tilde{\xi}_{kb}^{21} \\ (1-\rho_{kb})\xi_{kb}^{11}\xi_{kb}^{22} &= (1-\tilde{\rho}_{kb})\tilde{\xi}_{kb}^{11}\tilde{\xi}_{kb}^{22} + \tilde{\rho}_{kb}\tilde{\tau}_{kb}^{1}. \end{cases} \tag{A.5}$$

From the first two lines of the previous equation, we deduce that $\xi_{kb}^{22} = \xi_{kb}^{21}\frac{\tilde{\xi}_{kb}^{22}}{\tilde{\xi}_{kb}^{21}}$. We consider the last two lines where ξ_{kb}^{22} is replaced by $\xi_{kb}^{21}\frac{\tilde{\xi}_{kb}^{22}}{\tilde{\xi}_{kb}^{21}}$ and where the last line is multiplied by $\frac{\tilde{\xi}_{kb}^{21}}{\tilde{\xi}_{kb}^{22}}$. Thus,

$$\begin{cases} (1-\rho_{kb})\xi_{kb}^{11}\xi_{kb}^{21} + \rho_{kb}\tau_{kb}^{1} &= (1-\tilde{\rho}_{kb})\tilde{\xi}_{kb}^{11}\tilde{\xi}_{kb}^{21} \\ (1-\rho_{kb})\xi_{kb}^{11}\xi_{kb}^{21} &= (1-\tilde{\rho}_{kb})\tilde{\xi}_{kb}^{11}\tilde{\xi}_{kb}^{21} + \tilde{\rho}_{kb}\tilde{\tau}_{kb}^{1}\frac{\tilde{\xi}_{kb}^{21}}{\tilde{\xi}_{kb}^{22}}. \end{cases} \tag{A.6}$$

Thus, $\rho_{kb}\tau_{kb}^{1} + \tilde{\rho}_{kb}\tilde{\tau}_{kb}^{1}\frac{\tilde{\xi}_{kb}^{21}}{\tilde{\xi}_{kb}^{22}} = 0$. This result is in contradiction with the strict positivity of all the terms. So, it is not possible to respect (A.1) when only one relation is equal between both parametrizations.

• *No relation equal between both parametrizations*

Without loss of generality, we consider the following system

$$\begin{cases} (1-\rho_{kb})\xi_{kb}^{11}\xi_{kb}^{21}+\rho_{kb}\tau_{kb}^{1} &= (1-\tilde{\rho}_{kb})\tilde{\xi}_{kb}^{11}\tilde{\xi}_{kb}^{21} \\ (1-\rho_{kb})\xi_{kb}^{12}\xi_{kb}^{22}+\rho_{kb}\tau_{kb}^{2} &= (1-\tilde{\rho}_{kb})\tilde{\xi}_{kb}^{12}\tilde{\xi}_{kb}^{22} \\ (1-\rho_{kb})\xi_{kb}^{13}\xi_{kb}^{23}+\rho_{kb}\tau_{kb}^{3} &= (1-\tilde{\rho}_{kb})\tilde{\xi}_{kb}^{13}\tilde{\xi}_{kb}^{23} \\ (1-\rho_{kb})\xi_{kb}^{12}\xi_{kb}^{21} &= (1-\tilde{\rho}_{kb})\tilde{\xi}_{kb}^{12}\tilde{\xi}_{kb}^{21}+\tilde{\rho}_{kb}\tilde{\tau}_{kb}^{2} \\ (1-\rho_{kb})\xi_{kb}^{13}\xi_{kb}^{22} &= (1-\tilde{\rho}_{kb})\tilde{\xi}_{kb}^{13}\tilde{\xi}_{kb}^{22}+\tilde{\rho}_{kb}\tilde{\tau}_{kb}^{3} \\ (1-\rho_{kb})\xi_{kb}^{11}\xi_{kb}^{23} &= (1-\tilde{\rho}_{kb})\tilde{\xi}_{kb}^{11}\tilde{\xi}_{kb}^{23}+\tilde{\rho}_{kb}\tilde{\tau}_{kb}^{1} \\ (1-\rho_{kb})\xi_{kb}^{11}\xi_{kb}^{22} &= (1-\tilde{\rho}_{kb})\tilde{\xi}_{kb}^{11}\tilde{\xi}_{kb}^{22} \\ (1-\rho_{kb})\xi_{kb}^{12}\xi_{kb}^{23} &= (1-\tilde{\rho}_{kb})\tilde{\xi}_{kb}^{12}\tilde{\xi}_{kb}^{23} \\ (1-\rho_{kb})\xi_{kb}^{13}\xi_{kb}^{21} &= (1-\tilde{\rho}_{kb})\tilde{\xi}_{kb}^{13}\tilde{\xi}_{kb}^{21} \end{cases} \tag{A.7}$$

From the lines 1 and 4, we obtain that $\frac{\xi_{kb}^{11}}{\xi_{kb}^{12}} < \frac{\tilde{\xi}_{kb}^{11}}{\tilde{\xi}_{kb}^{12}}$. From the lines 7 and 2, we obtain that $\frac{\xi_{kb}^{11}}{\xi_{kb}^{12}} > \frac{\tilde{\xi}_{kb}^{11}}{\tilde{\xi}_{kb}^{12}}$. So, it is not possible to respect (A.1) when no relation is equal between both parametrizations. □

Proposition A.2 (Three binary variables)**.** *The mixture model of the two extreme dependency distributions is generically identifiable when* $d^{\{kb\}}=3$, $m_1^{\{kb\}}=m_2^{\{kb\}}=m_3^{\{kb\}}=2$.

Proof. Suppose that there exist $\boldsymbol{\alpha}_{kj}=(\rho_{kj},\boldsymbol{\xi}_{kb},\boldsymbol{\tau}_{kb},\boldsymbol{\delta}_{kb})$ and $\tilde{\boldsymbol{\alpha}}_{kj}=(\tilde{\rho}_{kj},\tilde{\boldsymbol{\xi}}_{kb},\tilde{\boldsymbol{\tau}}_{kb},\tilde{\boldsymbol{\delta}}_{kb})$ as such

$$\forall \boldsymbol{x}_i^{\{kb\}}\ p(\boldsymbol{x}_i^{\{kb\}};\boldsymbol{\alpha}_{kb})=p(\boldsymbol{x}_i^{\{kb\}};\tilde{\boldsymbol{\alpha}}_{kb}). \tag{A.8}$$

By writing the system with 8 equations related to (A.8), we obtain that $\forall j=1,\ldots,3:\ \xi_{kb}^{j1}(1-\tilde{\xi}_{kb}^{j1})=(1-\xi_{kb}^{j1})\tilde{\xi}_{kb}^{j1}$. Thus, $\forall j=1,\ldots,3:\ \xi_{kb}^{j1}=\tilde{\xi}_{kb}^{j1}$. We straightforwardly obtain the equality between the others parameters, so $\boldsymbol{\alpha}_{kb}=\tilde{\boldsymbol{\alpha}}_{kb}$. □

Conclusion The mixture model is stable by fusion of modalities and/or variables. So, we obtain the generic identifiability of all the models which can be written by fusion of modalities and/or variables as one of the following models: the three binary one and the two three-modalities one.

A.2 Generic identifiability of the mixture model of multinomiale distributions per modes

Generic identifiability of the CMM model with three blocks Let $k_0=\underset{k}{\operatorname{argmin}}\ \ell_{kb}$ and the matrix M_b where

$$M_b(k,h)=\alpha_{kb}^{\tau_{k_0b}(h)}. \tag{A.9}$$

By denoting by $\xi_b=\min_k \ell_{kb}+1$, generically, we have

$$\operatorname{rank}_K\ M_b=\min(g,\xi_b).$$

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

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

Generic identifiability of the CMM model with more than three blocks In the same way that [AMR09], we generalize the result with B blocks by observing that B blocks of categorical variables can be combined into three categorical variables. Thus, we can apply the Kruskal theorem.
COROLLARY 2 We consider a CMM model with B blocks where $\mathrm{B} \geq 3$. If there exists a tri-partition of the set $\{1, \ldots, \mathrm{B}\}$ into three disjoint non empty subsets S_1, S_2 and S_3, such that $\gamma_i = \prod_{j \in S_i} \xi_j$ with

$$\min(g, \gamma_1) + \min(g, \gamma_2) + \min(g, \gamma_3) \geq 2g + 2, \tag{A.10}$$

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

A.3 Computation of the integrate complete-data likelihood of the mixture model of multinomiale distributions per modes

In this Section, a proof of Proposition 4.12 is given. We firstly define a new parametrization of the block distribution facilitating the integrate complete-data likelihood computation. We secondly define the prior distribution of the new block parametrization according to the other parametrization. Thirdly, we underline the relation between the embedded models. We conclude by the integrate complete-data likelihood computation, which is the target result.

A.3.1 New parametrization of the block distribution

Without loss of generality, we assume that the elements of $\boldsymbol{\delta}_{kb}$ are ordered by decreasing values of the probability mass associated to them and we introduce the new parametrization of $\boldsymbol{a}_{kb}$ denoted $\boldsymbol{\varepsilon}_{kb}$ where $\boldsymbol{\varepsilon}_{kb} \in \mathcal{E}_{kb} = \left[\frac{1}{m^{\{b\}}}; 1\right] \times, \ldots, \times \left[\frac{1}{\mathrm{m}^{\{b\}} - \ell_{kb}}; 1\right]$ and where ε_{kbh} is defined by

$$\varepsilon_{kb}^h = \begin{cases} a_{kb}^{\delta_{kbh}} & \text{if } h = 1 \\ \frac{a_{kb}^{\delta_{kbh}}}{\prod_{h'=1}^{h-1}(1-\varepsilon_{kb}^{h'})} & \text{otherwise.} \end{cases}$$

Lemma A.3. *The conditional probability of* $\boldsymbol{x}^{\{b\}}$ *is*

$$p(\mathbf{x}^{\{b\}} | \mathbf{z}, \ell_{kb}, \tilde{\boldsymbol{\delta}}_{kb}, \boldsymbol{\varepsilon}_{kb}) = \prod_{h=1}^{\ell_{kb}} (\varepsilon_{kb}^h)^{\mathrm{n}_{kb}^{(h)}} (1 - \varepsilon_{kb}^h)^{\bar{\mathrm{n}}_{kb}^h}, \tag{A.11}$$

Proof.

$$\begin{aligned}
p(\mathbf{x}^{\{b\}}|\mathbf{z},\ell_{kb},\tilde{\boldsymbol{\delta}}_{kb},\boldsymbol{\varepsilon}_{kb}) &= p(\mathbf{x}^{\{b\}}|\mathbf{z},\ell_{kb},\boldsymbol{\alpha}_{kb}) \\
&= \prod_{h=1}^{\mathrm{m}^{\{b\}}}(\alpha_{kb}^{h})^{\mathrm{n}_{kb}^{h}} \\
&= \left[\prod_{h=1}^{\ell_{kj}}(\alpha_{kb}^{(h)})^{\mathrm{n}_{kb}^{(h)}}\right]\left(\alpha_{kb}^{(\ell_{kb}+1)}\right)^{\bar{\mathrm{n}}_{kb}^{\ell_{kb}}} \\
&= (\varepsilon_{kb}^{1})^{n_{kb}^{(1)}}\prod_{h=2}^{\ell_{kb}}\left[(\varepsilon_{kb}^{h})^{\mathrm{n}_{kb}^{(h)}}\left(\prod_{h'=1}^{h-1}(1-\varepsilon_{kb}^{h})^{n_{kb}^{(h)}}\right)\right]\prod_{h=1}^{\ell_{kb}}(1-\varepsilon_{kb}^{h})^{\bar{n}_{kb}^{\ell_{kb}}} \\
&= \prod_{h=1}^{\ell_{kb}}(\varepsilon_{kb}^{h})^{n_{kb}^{(h)}}(1-\varepsilon_{kb}^{h})^{\bar{n}_{kb}^{h}}.
\end{aligned}$$

□

A.3.2 Prior distribution

Lemma A.4. *The prior distribution of $\boldsymbol{\varepsilon}_{kb}$ is*

$$p(\boldsymbol{\varepsilon}_{kb}|\boldsymbol{\omega},\boldsymbol{\delta}_{kb}) = \frac{m^{\{b\}}}{m^{\{b\}}-\ell_{kb}}. \tag{A.12}$$

Proof. We remind that $\boldsymbol{a}_{kb}|\boldsymbol{\omega} \sim D^{t}_{\ell_{kb}+1}\left(1,\ldots,1;\mathrm{m}^{\{b\}}\right)$ and that

$$p(\boldsymbol{a}_{kb},\boldsymbol{\delta}_{kb}|\boldsymbol{\omega}) = p(\boldsymbol{\alpha}|\boldsymbol{\omega}) = p(\boldsymbol{\varepsilon}_{kb},\boldsymbol{\delta}_{kb}|\boldsymbol{\omega}). \tag{A.13}$$

So, we deduce the pdf of the prior distribution of $\boldsymbol{\varepsilon}_{kb}$

$$p(\boldsymbol{\varepsilon}_{kb}|\boldsymbol{\delta}_{kb},\boldsymbol{\omega}) = \frac{\prod_{h=1}^{\ell_{kb}}(\varepsilon_{kb}^{h})^{\gamma_{kb}^{h}-1}(1-\varepsilon_{kb}^{h})^{\sum_{h'=h+1}^{\ell_{kb}+1}(\gamma_{kb}^{h'}-1)}}{\int_{\boldsymbol{\varepsilon}_{kb}\in\mathcal{E}_{kb}}\prod_{h=1}^{\ell_{kb}}(\varepsilon_{kb}^{h})^{\gamma_{kb}^{h}-1}(1-\varepsilon_{kb}^{h})^{\sum_{h'=h+1}^{\ell_{kb}+1}(\gamma_{kb}^{h'}-1)}d\boldsymbol{\varepsilon}_{kb}}. \tag{A.14}$$

Thus, each ε_{kb}^{h} follows a truncated Beta distribution on the parameters space $\left[\frac{1}{\mathrm{m}^{\{b\}}-h+1},1\right]$ denoted by $\mathcal{B}e(\gamma_{kb}^{h},\sum_{h'=h+1}^{\ell_{kb}+1}(\gamma_{kb}^{h'}-1)+1)$. To assure the positivity of the parameters of the truncated Beta distributions, we put $\gamma_{kb}^{h}=1$, so

$$p(\boldsymbol{\varepsilon}_{kb}|\boldsymbol{\delta}_{kb},\boldsymbol{\omega}) = \frac{\mathrm{m}^{\{b\}}}{\mathrm{m}^{\{b\}}-\ell_{kb}}. \tag{A.15}$$

□

A.3.3 Relation between embedded models

Lemma A.5. *Let the model with $\ell_{kb}^{\ominus}$ modes and the parameters $(\tilde{\boldsymbol{\delta}}_{kb}^{\ominus},\boldsymbol{\varepsilon}_{kb}^{\ominus})$ and let the model with ℓ_{kb} modes and the parameters $(\tilde{\boldsymbol{\delta}}_{kb},\boldsymbol{\varepsilon}_{kb})$. Both modes are defined as such*

that $\ell_{kb}^{\ominus} = \ell_{kb} - 1$*, that the* $\ell_{kb}^{\ominus}$ *modes having the largest probabilities have the same locations* ($\forall h \in \boldsymbol{\delta}_{kb}^{\ominus},\ h \in \boldsymbol{\delta}_{kb}$) *and the same probability masses* ($\varepsilon_{kb}^{\ominus h} = \varepsilon_{kb}^{h},\ h < \ell_{kb}$)*. These embedded models follow this relation*

$$\frac{p(\mathbf{x}^{\{b\}}|\mathbf{z}, \ell_{kb}, \tilde{\boldsymbol{\delta}}_{kb}, \boldsymbol{\varepsilon}_{kb})}{p(\mathbf{x}^{\{b\}}|\mathbf{z}, \ell_{kb}^{\ominus}, \tilde{\boldsymbol{\delta}}_{kb}^{\ominus}, \boldsymbol{\varepsilon}_{kb}^{\ominus})} = \frac{(m^{\{b\}} - \ell_{kb} + 1)^{\bar{\mathrm{n}}_{kb}^{\ell_{kb}-1}-1}}{(m^{\{b\}} - \ell_{kb})^{\bar{\mathrm{n}}_{kb}^{\ell_{kb}}}} (\varepsilon_{\ell_{kb}})^{\mathrm{n}_{kb}^{(\ell_{kb})}} (1 - \varepsilon_{\ell_{kb}})^{\bar{\mathrm{n}}_{kb}^{\ell_{kb}}}. \quad \text{(A.16)}$$

Proof. We start by the following relation

$$\frac{p(\mathbf{x}^{\{b\}}|\mathbf{z}, \ell_{kb}, \boldsymbol{\alpha}_{kb})}{p(\mathbf{x}^{\{b\}}|\mathbf{z}, \ell_{kb}^{\ominus}, \boldsymbol{\alpha}_{kb}^{\ominus})} = \frac{(\alpha_{kb}^{\ell_{kb}})^{\mathrm{n}_{kb}^{(\ell_{kb})}} (\alpha_{kb}^{\ell_{kb}+1})^{\bar{\mathrm{n}}_{kb}^{\ell_{kb}}}}{(\alpha_{kb}^{\ominus \ell_{kb}})^{\bar{\mathrm{n}}_{kb}^{\ell_{kb}-1}}}. \quad \text{(A.17)}$$

Note that, $\varepsilon_{kb}^{h} = \varepsilon_{kb}^{\ominus h}$ when $(h = 1, \ldots, \ell_{kb} - 1)$, since $\alpha_{kb}^{(h)} = \alpha_{kb}^{\ominus(h)}$ and $\tilde{\tau}_{\ell_{kb}}(h) = \tilde{\tau}_{\ell_{kb}-1}(h)$ when $(h = 1, \ldots, \ell_{kb} - 1)$. Then, by using the reparamatrization in $\boldsymbol{\varepsilon}_{kb}$, the proof is completed. □

A.3.4 Integrated complete-data likelihood

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

$$p(\mathbf{x}^{\{b\}}|\mathbf{z}, \ell_{kb}) \approx \left(\frac{1}{m^{\{b\}} - \ell_{kb}}\right)^{\bar{\mathrm{n}}_{kb}^{\ell_{kb}}} \prod_{h=1}^{\ell_{kb}} \frac{Bi\left(\frac{1}{\mathrm{m}^{\{b\}}-h+1}; \mathrm{n}_{kb}^{(h)} + 1; \bar{\mathrm{n}}_{kb}^{h} + 1\right)}{\mathrm{m}^{\{b\}} - h}, \quad \text{(A.18)}$$

where $Bi(x; a, b) = B(1; a, b) - B(x; a, b)$, $B(x; a, b)$ being the incomplete beta function defined by $B(x; a, b) = \int_0^x w^a (1 - w)^b dw$. From the previous expression, its is straightforward to obtain $p(\mathbf{x}^{\{b\}}, \mathbf{z}|\boldsymbol{\omega})$.

Proof of Proposition 4.12. If, for the model with $\ell_{kb} - 1$ modes, the best modes locations are known and given by $\tilde{\boldsymbol{\delta}}_{kb}^{\ominus}$ then the conditional probability of $\mathbf{x}^{\{b\}}$ for a model with ℓ_{kb} modes is

$$p(\mathbf{x}^{\{b\}}|\mathbf{z}, \ell_{kb}, \tilde{\boldsymbol{\delta}}_{kb}^{\ominus}, \boldsymbol{\varepsilon}_{kb}) = \frac{1}{\mathrm{m}^{\{b\}} - \ell_{kb} + 1} \sum_{\tau \in \{1, \ldots, \mathrm{m}^{\{b\}}\} \setminus \{\tilde{\boldsymbol{\delta}}_{kb}^{\ominus}\}} p(\mathbf{x}^{\{b\}}|\mathbf{z}, \ell_{kb}, \{\tilde{\boldsymbol{\delta}}_{kb}^{\ominus}, \tau\}, \boldsymbol{\alpha}_{kb}^{\ominus}, \boldsymbol{\varepsilon}_{kb}), \quad \text{(A.19)}$$

Thus, by approximating this sum by its maximum element, we obtain that

$$p(\mathbf{x}^{\{b\}}|\mathbf{z}, \ell_{kb}, \tilde{\boldsymbol{\delta}}_{kb}^{\ominus}, \boldsymbol{\varepsilon}_{kb}) \approx \frac{1}{\mathrm{m}^{\{b\}} - \ell_{kb} + 1} p(\mathbf{x}^{\{b\}}|\mathbf{z}, \ell_{kb}, \tilde{\boldsymbol{\delta}}_{kb}, \boldsymbol{\alpha}_{kb}^{\ominus}, \boldsymbol{\varepsilon}_{kb}). \quad \text{(A.20)}$$

By using Lemma A.5, we obtain that:

$$\frac{p(\mathbf{x}^{\{b\}}|\mathbf{z}, \ell_{kb}, \tilde{\boldsymbol{\delta}}_{kb}^{\ominus}, \boldsymbol{\varepsilon}_{kb})}{p(\mathbf{x}^{\{b\}}|\mathbf{z}, \ell_{kb}^{\ominus}, \tilde{\boldsymbol{\delta}}_{kb}^{\ominus}, \boldsymbol{\varepsilon}_{kb}^{\ominus})} \approx \frac{(\mathrm{m}^{\{b\}} - \ell_{kb} + 1)^{\bar{\mathrm{n}}_{kb}^{\ell_{kb}-1}-1}}{(\mathrm{m}^{\{b\}} - \ell_{kb})^{\bar{\mathrm{n}}_{kb}^{\ell_{kb}}}} (\varepsilon_{kb}^{\ell_{kb}})^{\mathrm{n}_{kb}^{(\ell_{kb})}} (1 - \varepsilon_{kb}^{\ell_{kb}})^{\bar{\mathrm{n}}_{kb}^{\ell_{kb}}}. \quad \text{(A.21)}$$

As $p(\mathbf{x}^{\{b\}}|\mathbf{z}, \ell_{kb} = 0) = (m^{\{b\}})^{-n_k}$, by applying recursively the previous expression, we obtain that

$$p(\mathbf{x}^{\{b\}}|\mathbf{z}, \ell_{kb}, \boldsymbol{\varepsilon}_{kb}) \approx \left(\frac{1}{\mathrm{m}^{\{b\}} - \ell_{kj}}\right)^{\bar{\mathrm{n}}_{kb}^{\ell_{kb}}} \prod_{h=1}^{\ell_{kb}} \frac{(\varepsilon_{kb}^h)^{\mathrm{n}_{kb}^{(h)}} (1 - \varepsilon_{kb}^h)^{\bar{\mathrm{n}}_{kb}^h}}{\mathrm{m}^{\{b\}} - h + 1}. \tag{A.22}$$

□

Appendix B

Appendix of Part II

B.1 Identifiability of the mixture model of Gaussian and logistic distributions

Proposition B.1. *The mixture model of Gaussian and logistic distributions is generically identifiable.*

Proof. Suppose there are two mixture models of Gaussian and logistic distributions denoted by $p(\boldsymbol{x}_i;\boldsymbol{\theta})$ and $p(\boldsymbol{x}_i;\tilde{\boldsymbol{\theta}})$ such that

$$\forall \boldsymbol{x}_i,\ \sum_{k=1}^{g}\pi_k p(\boldsymbol{x}_i;\boldsymbol{\alpha}_k)=\sum_{k=1}^{\tilde{g}}\tilde{\pi}_k p(\boldsymbol{x}_i;\tilde{\boldsymbol{\alpha}}_k),\ 0<\pi_k,\tilde{\pi}_k\leq 1,\ \sum_{k=1}^{g}\pi_k=\sum_{k=1}^{\tilde{g}}\tilde{\pi}_k=1. \tag{B.1}$$

The aim is to prove that $\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}$. The demonstration is split in two parts. In the first one, we show the equality of the Gaussian distributions parameters and of the proportions. In the second one, we show the equality of the parameters of the logistic regressions.

— *Continuous parameters and proportions.*
We sum Equation (B.1) over all the possible values of $\boldsymbol{x}_i^{\mathrm{D}}$, so we obtain that

$$\forall \boldsymbol{x}_i^{\mathrm{C}},\ \sum_{k=1}^{g}\pi_k\phi(\boldsymbol{x}_i^{\mathrm{C}};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)=\sum_{k=1}^{\tilde{g}}\tilde{\pi}_k\phi(\boldsymbol{x}_i^{\mathrm{C}};\tilde{\boldsymbol{\mu}}_k,\tilde{\boldsymbol{\Sigma}}_k),\ 0<\pi_k,\tilde{\pi}_k\leq 1,$$
$$\sum_{k=1}^{g}\pi_k=\sum_{k=1}^{\tilde{g}}\tilde{\pi}_k=1. \tag{B.2}$$

The identifiability of the finite Gaussian mixtures models (see [Tei63] for the univariate case and [YS68] for the multivariate case) involves that $g=\tilde{g}$, $\pi_k=\tilde{\pi}_k$, $\boldsymbol{\mu}_k=\tilde{\boldsymbol{\mu}}_k$ and $\boldsymbol{\Sigma}_k=\tilde{\boldsymbol{\Sigma}}_k$.

— *Parameters of the logistic regressions.*
It is clear [Tei67] that if $\forall j=1+c,\ldots,e$, $\forall(\boldsymbol{x}_i^{\mathrm{C}},\boldsymbol{x}_i^{j})$:

$$\sum_{k=1}^{g}f_k(\boldsymbol{x}_i^{\mathrm{C}})p(\boldsymbol{x}_i^{j}|\boldsymbol{x}_i^{\mathrm{C}};\boldsymbol{\beta}_{kj})=\sum_{k=1}^{g}f_k(\boldsymbol{x}_i^{\mathrm{C}})p(\boldsymbol{x}_i^{j}|\boldsymbol{x}_i^{\mathrm{C}};\tilde{\boldsymbol{\beta}}_{kj}) \tag{B.3}$$

involves that $\boldsymbol{\beta}_{kj} = \tilde{\boldsymbol{\beta}}_{kj}$, where $f_k(\boldsymbol{x}_i^{\mathrm{c}}) = \pi_k \phi(\boldsymbol{x}_i^{\mathrm{c}}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, then mixture model of Gaussian and logistic distributions is identifiable.
Let the vector of size c denoted by $\boldsymbol{y}_i^{\mathrm{c}} = (y^1, \ldots, y^c)$ where all the elements are zero except the element j' which is equal to a. Without loss of generality, we consider that the $f_k(\boldsymbol{y}_i^{\mathrm{c}})$ are ordered such that $\boldsymbol{\Sigma}_k^{-1}(j', j') < \boldsymbol{\Sigma}_{k+1}^{-1}(j', j')$. From Equation (B.3), we deduce that

$$\sum_{k=1}^{g} f_k(\boldsymbol{y}_i^{\mathrm{c}}) \alpha_1(\boldsymbol{\beta}_{kj} | \boldsymbol{y}_i^{\mathrm{c}}) = \sum_{k=1}^{g} f_k(\boldsymbol{y}_i^{\mathrm{c}}) \alpha_1(\tilde{\boldsymbol{\beta}}_{kj} | \boldsymbol{y}_i^{\mathrm{c}}), \tag{B.4}$$

with $\left(\alpha_1(\boldsymbol{\beta}_{kj} | \boldsymbol{y}_i^{\mathrm{c}})\right)^{-1} = 1 + \sum_{h=2}^{m_j} \exp(\beta_{kj}^{0h} + \beta_{kj}^{j'h} a)$. We divided the above equation by $f_1(\boldsymbol{y}_i^{\mathrm{c}}) \alpha_1(\boldsymbol{\beta}_{1j} | \boldsymbol{y}_i^{\mathrm{c}})$, thus

$$1 + \sum_{k=2}^{g} \frac{f_k(\boldsymbol{y}_i^{\mathrm{c}}) \alpha_1(\boldsymbol{\beta}_{kj} | \boldsymbol{y}_i^{\mathrm{c}})}{f_1(\boldsymbol{y}_i^{\mathrm{c}}) \alpha_1(\boldsymbol{\beta}_{1j} | \boldsymbol{y}_i^{\mathrm{c}})} = \frac{\alpha_1(\tilde{\boldsymbol{\beta}}_{1j} | \boldsymbol{y}_i^{\mathrm{c}})}{\alpha_1(\boldsymbol{\beta}_{1j} | \boldsymbol{y}_i^{\mathrm{c}})} + \sum_{k=2}^{g} \frac{f_k(\boldsymbol{y}_i^{\mathrm{c}}) \alpha_1(\tilde{\boldsymbol{\beta}}_{kj} | \boldsymbol{y}_i^{\mathrm{c}})}{f_1(\boldsymbol{y}_i^{\mathrm{c}}) \alpha_1(\boldsymbol{\beta}_{1j} | \boldsymbol{y}_i^{\mathrm{c}})}. \tag{B.5}$$

Letting $a \to \infty$, $\sum_{k=2}^{g} \frac{f_k(\boldsymbol{y}_i^{\mathrm{c}}) \alpha_1(\boldsymbol{\beta}_{kj} | \boldsymbol{y}_i^{\mathrm{c}})}{f_1(\boldsymbol{y}_i^{\mathrm{c}}) \alpha_1(\boldsymbol{\beta}_{1j} | \boldsymbol{y}_i^{\mathrm{c}})} = 0$ and $\sum_{k=2}^{g} \frac{f_k(\boldsymbol{y}_i^{\mathrm{c}}) \alpha_1(\tilde{\boldsymbol{\beta}}_{kj} | \boldsymbol{y}_i^{\mathrm{c}})}{f_1(\boldsymbol{y}_i^{\mathrm{c}}) \alpha_1(\boldsymbol{\beta}_{1j} | \boldsymbol{y}_i^{\mathrm{c}})} = 0$ since the $f_k(\boldsymbol{y}_i^{\mathrm{c}})$ are ordered. Without loss of generality, if $m_j > 2$ we assume that $\beta_{1j}^{j'h} > \beta_{1j}^{j'h+1}$ if $1 < h < m_j$,

$$\lim_{a \to \infty} \frac{\alpha_1(\tilde{\boldsymbol{\beta}}_{1j} | \boldsymbol{y}_i^{\mathrm{c}})}{\alpha_1(\boldsymbol{\beta}_{1j} | \boldsymbol{y}_i^{\mathrm{c}})} = \lim_{a \to \infty} \exp\left((\beta_{kj}^{j'2} - \tilde{\beta}_{kj}^{j'2}) a + (\beta_{kj}^{02} - \tilde{\beta}_{kj}^{02})\right) = 1. \tag{B.6}$$

The above equation involves that $\beta_{kj}^{j'2} = \tilde{\beta}_{kj}^{j'2}$ and $\beta_{kj}^{02} = \tilde{\beta}_{kj}^{02}$. By repeating this argument for $h = 3, \ldots, m_j$, then for each $j = 1, \ldots, \mathrm{J_C}$ we conclude that $\boldsymbol{\beta}_{1j} = \tilde{\boldsymbol{\beta}}_{1j}$. By repeating this argument for $j = 1 + \mathrm{J_C}, \ldots, \mathrm{J_C} + \mathrm{J_D}$ then for $k = 2, \ldots, g$ we conclude that if Equation (B.1) is true then $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$.

□

B.2 Idenfitiability of the mixture model of Gaussian copulas

The model identifiability is proved by two propositions. The first proposition proves the model identifiability when the variables are continuous and/or integer. This proposition presents the reasoning in a simple case since it does not consider the ordinal variables. The second proposition proves that the model requires at least one continuous or integer variable to be identifiable.

Proposition B.2 (Identifiability with continuous and integer variables)**.** *The mixture model of Gaussian copulas is weakly identifiable [Tei63] if the variables are continuous and integer ones (*i.e. *the margin distributions of the components are Gaussian or Poisson distributions). Thus,*

$$\forall \boldsymbol{x} \in \mathbb{R}^c \times \mathbb{N}^d, \quad \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}; \boldsymbol{\alpha}_k) = \sum_{k=1}^{g'} \pi'_k p(\boldsymbol{x}; \boldsymbol{\alpha}'_k) \tag{B.7}$$

$$\Rightarrow \quad g = g', \ \boldsymbol{\pi} = \boldsymbol{\pi}', \ \boldsymbol{\alpha} = \boldsymbol{\alpha}'. \tag{B.8}$$

Proof. The identifiability of the multivariate Gaussian mixture models and of the univariate Poisson mixture model [Tei63, YS68] involves that (B.7) implies

$$g = g',\ \boldsymbol{\pi} = \boldsymbol{\pi}',\ \boldsymbol{\beta}_{kj} = \boldsymbol{\beta}'_{kj} \text{ and } \boldsymbol{\Gamma}_{k\text{CC}} = \boldsymbol{\Gamma}'_{k\text{CC}}. \tag{B.9}$$

We now show that $\boldsymbol{\Gamma}_{k\text{CD}} = \boldsymbol{\Gamma}'_{k\text{CD}}$ and $\boldsymbol{\Gamma}_{k\text{DD}} = \boldsymbol{\Gamma}'_{k\text{DD}}$.

Let $j \in \{1, \ldots, c\}$ and $h \in \{c+1, \ldots, e\}$. We denote by $\rho_k = \boldsymbol{\Gamma}_k(j,h)$, $\rho'_k = \boldsymbol{\Gamma}'_k(j,h)$, $v_k = \Phi_1^{-1}(P(x^j; \boldsymbol{\beta}_{kj}))$, $\varepsilon_k(x^j) = \pi_k \frac{\phi_1(v_k)}{\sigma_{kj}}$, $a_k = \frac{b_k^{\oplus}(x^j) - \rho_k v_k}{\sqrt{1-\rho_k^2}}$ and $a'_k = \frac{b_k^{\oplus}(x^j) - \rho'_k v_k}{\sqrt{1-\rho'^2_k}}$. Without loss of generality, we order the components as such $\sigma_{kj} > \sigma_{k+1j}$ and if $\sigma_{kj} = \sigma_{k+1j}$ then $\mu_{kj} > \mu_{k+1j}$, then (B.7) implies that

$$1 + \sum_{k=2}^{g} (\varepsilon_k(x^j)\Phi(a_k))/(\varepsilon_1(x^j)\Phi(a_1)) = \sum_{k=1}^{g} \varepsilon_k(x^j)\Phi(a'_k)/(\varepsilon_1(x^j)\Phi(a_1)).$$

Let $\gamma_t = \{(x^j, x^h) \in \mathbb{R} \times \mathbb{N} :\ a_1 = t\}$. Then, letting $x^h \to \infty$ as such $(x^j, x^h) \in \gamma_t$,

$$\forall t, \quad \frac{\int_t^{a'_1} \phi(u)du}{\Phi(t)} = 0. \tag{B.10}$$

Thus $a'_1 = a_1$, so $\rho'_1 = \rho_1$. Repeating this argument for $k = 2, \ldots, g$ and for all the couples (j,h), we conclude that $\boldsymbol{\Gamma}_{k\text{CD}} = \boldsymbol{\Gamma}'_{k\text{CD}}$.

When both variables are integer, we use the same argument with $\gamma_{(t,\xi)} = \{(x^j, x^h) \in \mathbb{N} \times \mathbb{N} :\ a_1 \in B(t,\xi)\}$. Note that if $\rho_1 \neq \rho'_1$ then $\exists n_0$ as such $\forall x^j > n_0\ a'_1 > t + \xi$. Letting $x^h \to \infty$ as such $(x^j, x^h) \in \gamma_{(t,\xi)}$, we obtain the following contradiction

$$\frac{\int_{t+\xi}^{a'_1} \phi(u)du}{\Phi(t-\xi)} = 0 \text{ and } \frac{\int_{t+\xi}^{a'_1} \phi(u)du}{\Phi(t-\xi)} > 0. \tag{B.11}$$

So, $a'_1 = a_1$ then $\rho_1 = \rho'_1$. Repeating this argument for $k = 2, \ldots, g$ and for all the couples (j,h), we conclude that $\boldsymbol{\Gamma}_{k\text{DD}} = \boldsymbol{\Gamma}'_{k\text{DD}}$. □

Proposition B.3 (Identifiability of the mixture model of Gaussian copulas)**.** *The mixture model of Gaussian copulas is weakly identifiable [Tei63] if at least one variable is continuous or integer.*

Proof. In this proof, we consider only one continuous variable and two binary variables. Obviously, the same reasoning can be extend to the other cases. We now show that $\boldsymbol{\Gamma}_{k\text{CD}} = \boldsymbol{\Gamma}'_{k\text{CD}}$ and $\boldsymbol{\Gamma}_{k\text{DD}} = \boldsymbol{\Gamma}'_{k\text{DD}}$.

Let $j = 1$ and let $h \in \{2, 3\}$. We note $\rho_k = \boldsymbol{\Gamma}_k(j,h)$, $\rho'_k = \boldsymbol{\Gamma}'_k(j,h)$, $v_k = \Phi_1^{-1}(P(x^j; \boldsymbol{\beta}_{kj}))$, $\varepsilon_k(x^j) = \pi_k \frac{\phi(v_k;0,1)}{\sigma_{kj}}$, $a_k = \frac{b_k^{\oplus}(x^j) - \rho_k v_k}{\sqrt{1-\rho_k^2}}$ and $a'_k = \frac{b'^{\oplus}_k(x^j) - \rho'_k v_k}{\sqrt{1-\rho'^2_k}}$. Without loss of generality, we order the components as such $\sigma_{kj} > \sigma_{[k+1]j}$ and if $\sigma_{kj} = \sigma_{[k+1]j}$ then $\mu_{kj} > \mu_{[k+1]j}$. Note that (B.7) implies that

$$1 + \sum_{k=2}^{g} (\varepsilon_k(x^j)\Phi(a_k))/(\varepsilon_1(x^j)\Phi(a_1)) = \sum_{k=1}^{g} \varepsilon_k(x^j)\Phi(a'_k)/(\varepsilon_1(x^j)\Phi(a_1)).$$

Letting $x^1 \to \infty$ and assuming that $\rho_k > 0$ then $\frac{\Phi(a'_k)}{\Phi(a_k)} = 1$. So, $\text{sign}(\rho_k) = \text{sign}(\rho'_k)$. By denoting $\kappa = \lim_{a\to\infty} \frac{\phi(a)}{\Phi(a)}$ and letting $x^1 \to \infty$ $\kappa \frac{1}{\kappa} \frac{\phi(a'_k)}{\phi(a_k)} = 1$. Thus $a'_1 = a_1$, so $\rho'_1 = \rho_1$ and $b_k^{\oplus}(x^j) = b_k'^{\oplus}(x^j)$ so $\boldsymbol{\beta}_{kh} = \boldsymbol{\beta}'_{kh}$.

Note that the same result can be obtain by tending x^1 to $-\infty$ is $\rho_k < 0$. Repeating this argument for $k = 2, \ldots, g$ and for all the couples (j, h), we conclude that $\boldsymbol{\Gamma}_{k\text{CD}} = \boldsymbol{\Gamma}'_{k\text{CD}}$ then $\boldsymbol{\Gamma}_{k\text{DD}} = \boldsymbol{\Gamma}'_{k\text{DD}}$. □

Bibliography

[AB90] A. Azzalini and A.W. Bowman. A look at some data on the Old Faithful geyser. *Applied Statistics*, (39):357–365, 1990.

[ABGK93] P.K. Andersen, O. Borgan, R.D. Gill, and N. Keiding. *Statistical models based on counting processes.* Springer Series in Statistics. Springer-Verlag, New York, 1993.

[AD07] A. Ahmad and L. Dey. A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, 63(2):503–527, 2007.

[Agr02] A. Agresti. *Categorical Data Analysis*, volume 359. John Wiley & Sons, 2002.

[Aka73] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[AMR09] E.S. Allman, C. Matias, and J.A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.

[AV07] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[Bac00] J. Bacher. A probabilistic clustering model for variables of mixed type. *Quality and Quantity*, 34(3):223–235, 2000.

[BB14] C. Bouveyron and C. Brunet. Model-based clustering of high-dimensional data: A review . *Computational Statistics and Data Analysis*, 71(0):52 – 78, 2014.

[BCG00] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7):719–725, 2000.

[BCG10] C. Biernacki, G. Celeux, and G. Govaert. Exact and Monte Carlo calculations of integrated likelihoods for the latent class model. *Journal of Statistical Planning and Inference*, 140(11):2991–3002, 2010.

[BCH09] T. Benaglia, D. Chauveau, and D.R. Hunter. An EM-like algorithm for semi- and nonparametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18(2):505–526, 2009.

[BCHY09] T. Benaglia, D. Chauveau, D.R. Hunter, and D.S. Young. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29, 2009.

[Ber06] P. Berkhin. A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer, 2006.

[BGLS06] J.F. Bonnans, J.C. Gilbert, C. Lemarechal, and C.A. Sagastizabal. Numerical Optimization: theoretical and practical aspects. 2006.

[Bie07] C. Biernacki. Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures for grouped data and behaviour of the EM algorithm. *Scandinavian Journal of Statistics*, 34(3):569–586, 2007.

[BL13] C. Biernacki and A. Lourme. Gaussian Parsimonious Clustering Models Scale Invariant and Stable by Projection. *Statistics and Computing*, page In press, 2013. RR-7932 RR-7932.

[BMM00] J. Barnard, R. McCulloch, and X.L. Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4):1281–1312, 2000.

[Boz87] H. Bozdogan. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

[BR93] J.D. Banfield and A.E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, pages 803–821, 1993.

[BR12] A.F. Berlinet and C. Roland. Acceleration of the EM algorithm: P-EM versus epsilon algorithm . *Computational Statistics and Data Analysis*, 56(12):4122 – 4137, 2012.

[BRC+10] J.P. Baudry, A.E. Raftery, G. Celeux, K. Lo, and R. Gottardo. Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, 19(2), 2010.

[Bre07] V Bretagnolle. Personal communication. *source: Museum*, 2007.

[CCA+09] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.

[CD+87] G. Celeux, J. Diebolt, et al. The EM and SEM algorithms for mixtures: Statistical and numerical aspects. *Rapport de Recherche Inria*, (641), 1987.

[CD92] G. Celeux and J. Diebolt. A stochastic approximation type EM algorithm for the mixture problem. *Stochastics: An International Journal of Probability and Stochastic Processes*, 41(1-2):119–134, 1992.

[CG91] G. Celeux and G. Govaert. Clustering criteria for discrete data and latent class models. *Journal of classification*, 8(2):157–176, 1991.

[CG92] G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3):315–332, 1992.

[CG95] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793, 1995.

[CHL10] D. Chauveau, D.R. Hunter, and M. Levine. Estimation for conditional independence multivariate finite mixture models. *HAL*, 2010.

[CKSS12] M. Chavent, V. Kuentz-Simonet, and J. Saracco. Orthogonal rotation in PCAMIX. *Advances in Data Analysis and Classification*, 6(2):131–146, 2012.

[CL68] C Chow and C Liu. Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, 14(3):462–467, 1968.

[CM07] P. Cortez and A. Morais. A data mining approach to predict forest fires using meteorological data. 2007.

[CS12] C. Choirat and R. Seri. Estimation in discrete parameter models. *Statistical Science*, 27(2):278–293, 2012.

[CSZ+06] O. Chapelle, B. Schölkopf, A. Zien, et al. *Semi-supervised learning*, volume 2. MIT press Cambridge, 2006.

[CZ03] J. Czerniak and H. Zarzycki. Application of rough sets in the presumptive diagnosis of urinary system diseases. *Artifical Inteligence and Security in Computing Systems, ACS'2002 9th International Conference Proceedings*, pages 41–51, 2003.

[Det88] R. Detrano. Cleveland Heart Disease Data. *Long Beach and Cleveland Clinic Foundation*, 1988.

[DLR77] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[DMD06] N. Dean, T.B. Murphy, and G. Downey. Using unlabelled data to update classification rules with applications in food authenticity studies. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 55(1):1–14, 2006.

[EH89] M.A. Espeland and S.L. Handelman. Using Latent Class Models to Characterize and Assess Relative Error in Discrete Measurements. *Biometrics*, 45(2):pp. 587–599, 1989.

[ELLS11] B.S. Everitt, S. Landau, M. Leese, and D. Stahl. *Cluster analysis*. London, wiley edition, 2011.

[Eve88] B.S. Everitt. A finite mixture model for the clustering of mixed-mode data. *Statistics & Probability Letters*, 6(5):305–309, 1988.

[For90] R.S. Forsyth. Pc/beagle user's guide, http://archive.ics.uci.edu/ml. *BUPA Medical Research Ltd*, 1990.

[For92] A.K. Formann. Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association*, 87(418):476–486, 1992.

[FR06] C. Fraley and A.E. Raftery. MCLUST version 3: an R package for normal mixture modeling and model-based clustering. Technical report, DTIC Document, 2006.

[FRW13] N. Friel, C. Ryan, and J. Wyse. Bayesian model selection for the latent position cluster model for social networks. *arXiv preprint arXiv:1308.4871*, 2013.

[FS06] S. Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer, 2006.

[FS08] S. Frühwirth-Schnatter. Finite Mixture and Markov Switching Models, 2008.

[FW12] N. Friel and J. Wyse. Estimating the evidence–a review. *Statistica Neerlandica*, 66(3):288–308, 2012.

[GF07] C. Genest and A.C. Favre. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering*, 12(4):347–368, 2007.

[GM13] I. Gollini and T.B. Murphy. Mixture of latent trait analyzers for model-based clustering of categorical data. *Statistics and Computing*, pages 1–20, 2013.

[GN08] G. Govaert and M. Nadif. Block clustering with bernoulli mixture models: Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6):3233–3245, 2008.

[GN10] G. Govaert and M. Nadif. Latent block model for contingency table. *Communications in Statistics—Theory and Methods*, 39(3):416–425, 2010.

[Goo74] L.A. Goodman. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231, 1974.

[Gou06] C. Gouget. *Utilisation des modèles de mélange pour la classification automatique de données ordinales*. PhD thesis, Université de Technologie de Compiègne, 2006.

[Gov10] G. Govaert. *Data analysis*, volume 136. Wiley, 2010.

[Gre90] P.J. Green. On use of the EM for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 443–452, 1990.

[Gre95] P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711, 1995.

[HA85] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[Hag88] J.A. Hagenaars. Latent structure models with direct effects between indicators local dependence models. *Sociological Methods & Research*, 16(3):379–405, 1988.

[Har72] D. Harper. Local dependence latent structure models. *Psychometrika*, 37(1):53–59, 1972.

[HDT06] H. Hwang, W.R. Dillon, and Y. Takane. An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents. *Psychometrika*, 71(1):161–171, 2006.

[Hen10] C. Hennig. Methods for merging Gaussian mixture components. *Advances in data analysis and classification*, 4(1):3–34, 2010.

[HJ99] L. Hunt and M. Jorgensen. Theory & Methods: Mixture model clustering using the MULTIMIX program. *Australian & New Zealand Journal of Statistics*, 41(2):154–171, 1999.

[HJ11] L. Hunt and M. Jorgensen. Clustering mixed data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4):352–361, 2011.

[HNW11] P.D. Hoff, X. Niu, and J.A. Wellner. Information bounds for Gaussian copulas. *arXiv preprint arXiv:1110.3572*, 2011.

[Hof07] P.D. Hoff. Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, pages 265–283, 2007.

[Hua98] Z. Huang. Extensions to the K-means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, 2(3):283–304, 1998.

[HWH07] D.R. Hunter, S. Wang, and T.P. Hettmansperger. Inference for mixtures of symmetric distributions. *The Annals of Statistics*, pages 224–251, 2007.

[HY01] D.J. Hand and K. Yu. Idiot's Bayes — Not So Stupid after All? *International Statistical Review*, 69(3):385–398, 2001.

[JB12] J. Jacques and C. Biernacki. Model-based clustering for multivariate partial ranking data. *Rapport de Recherche Inria*, RR8113, 2012.

[JH96] M. Jorgensen and L. Hunt. Mixture model clustering of data sets with categorical and continuous variables. In *Proceedings of the Conference ISIS*, volume 96, pages 375–384, 1996.

[Joe97] H. Joe. *Multivariate models and multivariate dependence concepts*, volume 73. CRC Press, 1997.

[JP14] J. Jacques and C. Preda. Model-based clustering of multivariate functional data. *Computational Statistics and Data Analysis*, 71:92–106, 2014.

[Ker00] C. Keribin. Consistent estimation of the order of mixture models. *Sankhya Serie A*, 62(1):49–66, 2000.

[KK14] I. Kosmidis and D. Karlis. Model-based clustering using copulas with applications. *ArXiv e-prints*, 2014.

[KL51] S. Kullback and R.A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[KM07] D. Karlis and L. Meligkotsidou. Finite mixtures of multivariate Poisson distributions with application. *Journal of statistical Planning and Inference*, 137(6):1942–1960, 2007.

[Kru56] Joseph B Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1):48–50, 1956.

[Krz93] W.J. Krzanowski. The location model for mixtures of categorical and continuous variables. *Journal of Classification*, 10(1):25–49, 1993.

[KT08] D. Karlis and P. Tsiamyrtzis. Exact Bayesian modeling for bivariate Poisson data and extensions. *Statistics and Computing*, 18(1):27–40, 2008.

[KW97] C.A.J. Klaassen and J.A. Wellner. Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli*, 3(1):55–77, 1997.

[LBA10] P. Latouche, E. Birmelé, and C. Ambroise. Bayesian methods for graph clustering. In *Advances in Data Analysis, Data Handling and Business Intelligence*, pages 229–239. Springer, 2010.

[LC98] E. L. Lehmann and G. Casella. *Theory of point estimation*, volume 31. Springer, 1998.

[LIL+12] R Lebret, S. Iovleff, F. Langrognet, C. Biernacki, G. Celeux, and G. Govaert. Rmixmod: The R Package of the Model-Based Unsupervised, Supervised and Semi-Supervised Classification Mixmod Library. *Preprint*, 2012.

[LK96] C.J. Lawrence and W.J. Krzanowski. Mixture separation for mixed-mode data. *Statistics and Computing*, 6(1):85–92, 1996.

[Llo82] S. Lloyd. Least squares quantization in PCM. *Information Theory, IEEE Transactions on*, 28(2):129–137, 1982.

[LLS00] T.S. Lim, W.Y. Loh, and Y.S. Shih. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning*, 40(3):203–228, 2000.

[MBV13a] M. Marbac, C. Biernacki, and V. Vandewalle. Model-based clustering for conditionally correlated categorical data. *Journal of Classification*, 2013.

[MBV13b] M. Marbac, C. Biernacki, and V. Vandewalle. Model-based clustering for conditionally correlated categorical data. Rapport de recherche RR-8232, INRIA, 2013.

[MBV14a] M. Marbac, C. Biernacki, and V. Vandewalle. Finite mixture model of conditional dependencies modes to cluster categorical data. *Submitted*, 2014.

[MBV14b] M. Marbac, C. Biernacki, and V. Vandewalle. Model-based clustering of Gaussian copulas for mixed data. *Submitted*, 2014.

[MDCL13] J.S. Murray, D.B. Dunson, L. Carin, and J.E. Lucas. Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, 108(502):656–665, 2013.

[MJ01] M. Meila and M.I. Jordan. Learning with mixtures of trees. *The Journal of Machine Learning Research*, 1:1–48, 2001.

[MK97] G.J. McLachlan and T. Krishnan. *The EM algorithm*. Wiley Series in Probability and Statistics: Applied Probability and Statistics, Wiley-Interscience, New York, 1997.

[MMR05] J.M. Marin, K. Mengersen, and C.P. Robert. Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics*, 25:459–507, 2005.

[MP00] G.J. McLachlan and D. Peel. *Finite mixutre models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics, Wiley-Interscience, New York, 2000.

[MP05] I. Moustaki and I. Papageorgiou. Latent class models for mixed variables with applications in Archaeometry. *Computational statistics & data analysis*, 48(3):659–675, 2005.

[Mut08] B. Muthén. Latent variable hybrids: Overview of old and new models. *Advances in latent variable mixture models*, 1:1–24, 2008.

[Nel99] R.B. Nelsen. *An introduction to copulas*. Springer, 1999.

[Ols79] U. Olsson. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4):443–460, 1979.

[OT61] I. Olkin and R.F. Tate. Multivariate correlation models with mixed discrete and continuous variables. *The Annals of Mathematical Statistics*, pages 448–465, 1961.

[Par62] E. Parzen. On estimation of a probability density function and mode. *Annals of mathematical statistics*, 33(3):1065–1076, 1962.

[PCK06] M. Pitt, D. Chan, and R. Kohn. Efficient Bayesian inference for Gaussian copula regression models. *Biometrika*, 93(3):537–554, 2006.

[Pea94] K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[PM00] D. Peel and G.J. McLachlan. Robust mixture modelling using the t distribution. *Statistics and computing*, 10(4):339–348, 2000.

[QTK96] Y. Qu, M. Tan, and M.H. Kutner. Random Effects Models in Latent Class Analysis for Evaluating Accuracy of Diagnostic Tests. *Biometrics*, 52(3):pp. 797–810, 1996.

[Raf96] A.E. Raftery. Hypothesis testing and model selection. In *Markov chain Monte Carlo in practice*, pages 163–187. Springer, 1996.

[Ral95] H. Ralambondrainy. A conceptual version of the K-means algorithm. *Pattern Recognition Letters*, 16(11):1147–1157, 1995.

[RC04] C.P. Robert and G. Casella. *Monte Carlo statistical methods*, volume 319. Citeseer, 2004.

[RG97] S. Richardson and P.J. Green. 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, 1997.

[RIW08] B.A. Reboussin, E.H. Ip, and M. Wolfson. Locally dependent latent class models with covariates: an application to under-age drinking in the USA. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(4):877–897, 2008.

[Rob07] C.P. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer, 2007.

[RSS+06] B.A. Reboussin, E.Y. Song, A. Shrestha, K.K. Lohman, and M. Wolfson. A latent class analysis of underage problem drinking: Evidence from a community sample of 16–20 year olds. *Drug and alcohol dependence*, 83(3):199–209, 2006.

[Sch78] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[SFK05] P. X-K. Song, Y. Fan, and J. D. Kalbfleisch. Maximization by parts in likelihood inference. *Journal of the American Statistical Association*, 100(472):1145–1158, 2005.

[SK12] M.S. Smith and M.A. Khaled. Estimation of copula models with discrete margins via Bayesian data augmentation. *Journal of the American Statistical Association*, 107(497):290–303, 2012.

[SRAT+06] S.M. Strauss, D.M. Rindskopf, J.M. Astone-Twerell, D.C. Des Jarlais, and H. Hagan. Using latent class analysis to identify patterns of hepatitis c service provision in drug-free treatment programs in the us. *Drug and alcohol dependence*, 83(1):15–24, 2006.

[Ste00a] M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

[Ste00b] M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

[Tei63] H. Teicher. Identifiability of finite mixtures. *Annals of Mathematical Statistics*, 34:1265–1269, 1963.

[Tei67] H. Teicher. Identifiability of mixtures of product measures. *Annals of Mathematical Statistics*, 38:1300–1302, 1967.

[Ver03] J.K. Vermunt. Multilevel latent class models. *Sociological methodology*, 33(1):213–239, 2003.

[Ver07] J.K. Vermunt. Multilevel mixture item response theory models: an application in education testing. *Proceedings of the 56th session of the International Statistical Institute. Lisbon, Portugal*, pages 22–28, 2007.

[VHH09] P. Van Hattum and H. Hoijtink. Market Segmentation Using Brand Strategy Research: Bayesian Inference with Respect to Mixtures of Log-Linear Models. *Journal of Classification*, 26(3):297–328, 2009.

[VR08] R. Varadhan and C. Roland. Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scandinavian Journal of Statistics*, 35(2):335–353, 2008.

[WB99] A. Willse and R.J. Boik. Identifiable finite mixtures of location models for clustering mixed-mode data. *Statistics and Computing*, 9(2):111–121, 1999.

[Wu83] C.F. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[YS68] S.J. Yakowitz and J.D. Spragins. On the identifiability of finite mixtures. *Annals of Mathematical Statistics*, 39:209–214, 1968.

[ZD12] A.Z. Zambom and R. Dias. A Review of Kernel Density Estimation with Applications to Econometrics. *arXiv preprint arXiv:1212.2812*, 2012.