HAL open science

Model-based clustering of Gaussian copulas for mixed data

Matthieu Marbac, Christophe Biernacki, Vincent Vandewalle

To cite this version:

Matthieu Marbac, Christophe Biernacki, Vincent Vandewalle. Model-based clustering of Gaussian copulas for mixed data. 2014. hal-00987760v1

HAL Id: hal-00987760

https://hal.science/hal-00987760v1

Preprint submitted on 6 May 2014 (v1), last revised 20 Dec 2016 (v4)

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

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

Model-based clustering of Gaussian copulas for mixed data

Matthieu Marbac - Christophe Biernacki - Vincent Vandewalle

May 6, 2014

Abstract

A mixture model of Gaussian copulas is presented to cluster mixed data (different kinds of variables simultaneously) 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 robustness 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 even if the data are simulated according to another model. Finally, three applications on real data sets strengthen the idea that the proposed model is of interest, since it reduces the biases of the locally independent model and since it provides a meaningful summary of the data.

Keywords. Clustering, Gaussian copula, Metropolis-within-Gibbs algorithm, Mixed data, Mixture models.
MSC 62H30, 62F15, 62-07, 62F07.

1 Introduction

Multivariate data sets are increasingly complex because of the informatics advent. Thus, they need to be summarized in order to extract the embedded information. *Clustering* provides an efficient solution of this challenge by grouping the individuals in few characteristic classes. It can be performed by probabilistic methods modelling the data generation whose the most popular and flexible one approaches the data distribution with a finite mixture model of parametric components [MP00]. In such a case, a class gathers together the individuals drawn by the same distribution.

Obviously, the choice of the component distribution depends on the kind of the variables at hand. For instance, one can use the Gaussian mixture model [BR93] to cluster continuous data, while a mixture of Poisson distributions [KT08] can be applied on integer data and a mixture of multinomial distributions [Goo74] can cluster ordinal data. However, many data sets contain mixed

variables (variables of different kinds) but few multivariate distributions exist for such data except three main models presented below. A more detailed survey is available in [HJ11].

The simplest way to cluster mixed variables consists in approaching the data distribution with a finite mixture model which assumes the independence conditionally on the class membership of each variable. This model, called *naive Bayes* or *locally independent model*, obtains good results in many real clustering problems [Lew98, HY01], especially when few individuals are described by several variables. Indeed, when its one-dimensional margins of each component follow classical distributions, it provides a meaningful summary of the data by its margin parameters. However, this model leads to biases when this assumption of conditional independence is violated (for instance, see the application of [VHH09]). In such a case, two methods can be envisaged. The first one performs a selection of the variables in order to cluster intra-class independent variables [MCMM09b, MCMM09a]. However, the risk of losing information is present, so it can be more efficient to use the second method which clusters the data with models relaxing the local independence. We now detail the two main models related to this second method.

The *location mixture model* [Krz93, WB99] concatenates the whole categorical variables in a single one following 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 depends on both the class and categorical variables while its covariance matrix is only set by the class membership. Thus, the conditional dependency between the whole variables is taken into account but this model needs too many parameters to obtain great success on real clustering problems. So, Jorgensen and Hunt [JH96, HJ99] propose an extension of the location mixture model. In their extension, the variables are split into conditionally independent blocks as such that each block is composed by at most one categorical variable. Moreover, each block of variables follows a location model. Indeed, for each block, the categorical variable follows a full multinomial distribution while the continuous variables follow a multivariate Gaussian distribution conditionally on the categorical one. However, this model has two main drawbacks. The first one is about the class interpretation, since the margin distribution of a component is not a classical distribution when the variable is continuous. Indeed, in such a case, it consists in a mixture of univariate Gaussian distributions. The second one is about the model selection, since the repartition of the variables into blocks is estimated by an ascendant method. Indeed, the estimation starts with the conditional independence assumption, then many models are proposed according to the correlation coefficients computed per class in order to improve an information criterion. However, this approach can be sub-optimal to perform the model selection. Furthermore, the choice of the correlation coefficient between a continuous variable and a categorical one is subjective but crucial since it determines the candidates during the model estimation.

The third main alternative approach, proposed by Everitt [Eve88], is the *underlying variables mixture model* which permits to cluster continuous and ordinal or binary variables. It is assumed that each discrete variable arises from a latent continuous variable. The probability distribution function of the whole

continuous variables (observed and unobserved) is a multivariate Gaussian mixture model. Thus, the probability distribution function of the observed variables is computed by integrating each Gaussian component on the unobserved continuous variable set. However, in practice, this computation is not doable when there are more than two discrete variables. More recently, in order to be able to cluster more numerous binary variables, Morlini [Mor12] has developed an extension of this model by estimating the scores of the latent variables from binary data. However, the class interpretation is more difficult since the parameters summarize the distributions of the scores and not the distribution of the native binary variables.

The aim of this paper 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 each class is summarized by its proportion, by the parameters of each marginal distributions and by the correlation matrix of the Gaussian copula which provides one coefficient per couple of variables measuring the intra-class dependency. In addition, a principal component analysis (PCA) computed per class is a straightforward by-product of the model. Indeed, it is directly computed from the correlation matrix of the class. It can be used to summarize the main intra-class dependencies and to provide a scatter-plot of the individuals according to the class parameters.

This paper is organized as follows. Section 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 3 is devoted to the parameter estimation in a Bayesian framework since the maximum likelihood estimate is unattainable [PCK06]. Section 4 illustrates the behavior of the algorithm performing the inference and also the model robustness on numerical experiments. Section 5 presents three applications of the new mixture model by clustering three real data sets. Section 6 concludes this work.

2 Mixture model of Gaussian copulas

2.1 Finite mixture model

Data

The vector of e mixed variables is denoted by $\boldsymbol{x} = (x^1, \ldots, x^e) \in \mathbb{R}^c \times \mathcal{X}$, 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 $\boldsymbol{x}^{\mathrm{C}}$. 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 $\boldsymbol{x}^{\mathrm{D}}$. Note that if x^j is an ordinal variable with m_j modalities, then it uses a numeric coding $\{1, \ldots, m_j\}$.

Remark 2.1 (Notations)**.** In this paper, we use the generic notation $P(.;.)$ for the cumulative distribution functions (cdf) and $p(.;.)$ for the probability distribution function (pdf).

Probability distribution function

Definition 2.2 (Finite mixture model of parametric distributions)**.** Data $\boldsymbol{x}$ are supposed to be drawn by the mixture model of g parametric distributions whose the pdf is written as follows

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

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 2.3 (Latent variable)**.** A finite mixture model can be expressed by using the latent variable $z \in \{1, \ldots, g\}$. This categorical variable indicates the class membership by using a condensed coding and follows the multinomial distribution $\mathcal{M}_g(\pi_1, \ldots, \pi_g)$. Thus, (1) can be interpreted as the marginal distribution of $\boldsymbol{x}$ based on the distribution of the couple $(\boldsymbol{x}, z)$.

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 2.4 (One-dimensional margins of the components)**.** The margin distribution of the variable x^j, for the component k, belongs to the exponential family and has $p(x^j; \boldsymbol{\beta}_{kj})$ for pdf and $P(x^j; \boldsymbol{\beta}_{kj})$ as cdf. More precisely,

- If x^j is *continuous*, its margin of the component k follows a *Gaussian* distribution with mean μ_{kj} and variance σ^2_{kj}, *i.e.* $x^j|z = k \sim \mathcal{N}_1(\mu_{kj}, \sigma^2_{kj})$ and $\boldsymbol{\beta}_{kj} = (\mu_{kj}, \sigma^2_{kj}) \in \mathbb{R} \times \mathbb{R}^{+*}$.
- If x^j is *integer*, its margin of the component k follows a *Poisson* distribution, *i.e.* $x^j|z = k \sim \mathcal{P}(\boldsymbol{\beta}_{kj})$ and $\boldsymbol{\beta}_{kj} \in \mathbb{R}^{+*}$.
- If x^j is *ordinal*, its margin of the component k follows a *multinomial* distribution, *i.e.* $x^j|z = k \sim \mathcal{M}_{m_j}(\boldsymbol{\beta}_{kj})$, $\boldsymbol{\beta}_{kj}$ 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 2.5 (Cumulative distribution function of the components)**.** For the mixture model of Gaussian copulas, the cdf of the component k is written as

$$P(\boldsymbol{x}; \boldsymbol{\alpha}_k) = \Phi_e(\Phi_1^{-1}(u_k^1), \ldots, \Phi_1^{-1}(u_k^e); \mathbf{0}, \boldsymbol{\Gamma}_k), \tag{2}$$

where $u_k^j = P(x^j; \boldsymbol{\beta}_{kj})$ and where $\boldsymbol{\alpha}_k = (\boldsymbol{\beta}_k, \boldsymbol{\Gamma}_k)$ denotes the whole parameters of the component k with $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{k1}, \ldots, \boldsymbol{\beta}_{ke})$.

Property 2.6 (Standardized coefficient of correlation per class)**.** The Gaussian copula provides a robust coefficient of correlation per couple of variables. Indeed, when both variables are continuous, it is equal to the upper bound of the coefficient 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 2.7 (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 $\boldsymbol{y} = (y^1, \ldots, y^e) \in \mathbb{R}^e$. Conditionally on the class membership, this variable follows an e-variate centred Gaussian distribution. Indeed, if $\boldsymbol{y}|z = k \sim \mathcal{N}_e(\mathbf{0}, \boldsymbol{\Gamma}_k)$ and if

$$x^j = P^{-1}(\Phi_1(y^j); \boldsymbol{\beta}_{kj}),\ \forall j = 1, \ldots, e, \tag{3}$$

then the component k is a Gaussian copula whose the cdf is $P(\boldsymbol{x}; \boldsymbol{\alpha}_k)$.

Mixture model of Gaussian copulas for mixed data

We introduce the function $\Psi(\boldsymbol{x}^{\mathrm{C}}; \boldsymbol{\alpha}_k) = \left(\frac{x^j - \mu_{kj}}{\sigma_{kj}}; j = 1, \ldots, c\right)$ and the space of the antecedents of $\boldsymbol{x}^{\mathrm{D}}$ for the class k noted $\mathcal{S}_k(\boldsymbol{x}^{\mathrm{D}}) = \mathcal{S}_k^{c+1}(x^{c+1}) \times \ldots \times \mathcal{S}_k^e(x^e)$.

The interval $\mathcal{S}_k^j(x^j) =]b_k^{\ominus}(x^j), b_k^{\oplus}(x^j)]$ is defined for $j = c+1, \ldots, e$ and its bounds are $b_k^{\ominus}(x^j) = \Phi_1^{-1}(P(x^j - 1; \boldsymbol{\beta}_{kj}))$ and $b_k^{\oplus}(x^j) = \Phi_1^{-1}(P(x^j; \boldsymbol{\beta}_{kj}))$. We now define the pdf of the components according to (2) as proposed by [SK12].

Definition 2.8 (Mixture model of Gaussian copulas)**.** Data $\boldsymbol{x}$ follows a mixture model of Gaussian copulas if its pdf is the finite mixture model defined in (1) whose the pdf of the component k is written as

$$p(\boldsymbol{x}; \boldsymbol{\alpha}_k) = p(\boldsymbol{x}^{\mathrm{C}}; \boldsymbol{\alpha}_k) p(\boldsymbol{x}^{\mathrm{D}} | \boldsymbol{x}^{\mathrm{C}}; \boldsymbol{\alpha}_k) \tag{4}$$

$$= \frac{\phi_c(\Psi(\boldsymbol{x}^{\mathrm{C}}; \boldsymbol{\alpha}_k); \mathbf{0}, \boldsymbol{\Gamma}_{k\mathrm{CC}})}{\prod_{j=1}^{c} \sigma_{kj}} \int_{\mathcal{S}_k(\boldsymbol{x}^{\mathrm{D}})} \phi_d(\boldsymbol{u}; \boldsymbol{\mu}_k^{\mathrm{D}}, \boldsymbol{\Sigma}_k^{\mathrm{D}}) d\boldsymbol{u}, \tag{5}$$

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 $\boldsymbol{y}^{\mathrm{D}}$ defined by $\boldsymbol{\mu}_k^{\mathrm{D}} = \boldsymbol{\Gamma}_{k\mathrm{DC}} \boldsymbol{\Gamma}_{k\mathrm{CC}}^{-1} \Psi(\boldsymbol{x}^{\mathrm{C}}; \boldsymbol{\alpha}_k)$ 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 2.9 (Generative model)**.** The mixture model of Gaussian copulas involves finally the generative model split into the following three steps:

- Class membership *sampling*: $z \sim \mathcal{M}_g(\pi_1, \ldots, \pi_g)$
- Gaussian copula *sampling*: $\boldsymbol{y} | z = k \sim \mathcal{N}_e(\mathbf{0}, \boldsymbol{\Gamma}_k)$
- Observed data *deterministic computation*: $\boldsymbol{x}$ is obtained from (3).

2.2.1 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

  $$\boldsymbol{\Gamma}_1 = \ldots = \boldsymbol{\Gamma}_g. \tag{6}$$

  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

  $$\nu_{\mathrm{He}} = (g-1) + g \left(\frac{e(e-1)}{2} + \sum_{j=1}^{d} \nu_j \right) \text{ and } \nu_{\mathrm{Ho}} = (g-1) + \frac{e(e-1)}{2} + g \sum_{j=1}^{d} \nu_j, \tag{7}$$

  where ν_j denotes the number of parameters of the margin distribution of the variable j for one component. More precisely, with the specific margin distribution of the components, ν_j is equal to

  $$\nu_j = \begin{cases} 2 & \text{if } x^j \text{ is numeric} \\ 1 & \text{if } x^j \text{ is discrete} \\ m_j - 1 & \text{if } x^j \text{ is ordinal.} \end{cases} \tag{8}$$

- *Model identifiability.* The mixture model of Gaussian copulas is identifiable (in the sense of [Tei63, YS$^+$68]) if, at least, one variable is continuous or integer. The proof is given in Appendix A.

2.3 Strengths of the mixture model

Related models

The Gaussian copula mixture model 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 locally independent 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]. However, Everitt estimates directly the space $\mathcal{S}_k(\boldsymbol{x}^{\mathrm{D}})$ containing the antecedents of $\boldsymbol{x}^{\mathrm{D}}$ 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 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 the class k, we firstly compute the coordinates equal to $\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}, z = k; \boldsymbol{\alpha}_k]$ and we secondly project them on the principal component analysis space of the Gaussian copula of the 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 to zero, so they are farther to the origin. Finally, the correlation circle summarizes the intra-class correlations. The following example illustrates this phenomenon.

Example 2.10 (Mixture model of Gaussian copulas and visualization per class)**.** Let the bi-component mixture model of Gaussian copulas composed by three variables (one continuous, one integer and one binary), in this order, with

$$\boldsymbol{\pi} = (0.5, 0.5),\ \boldsymbol{\beta}_{11} = (-2, 1),\ \boldsymbol{\beta}_{12} = 5,\ \boldsymbol{\beta}_{13} = (0.5, 0.5), \boldsymbol{\beta}_{21} = (2, 1),\ \boldsymbol{\beta}_{22} = 15,$$
$$\boldsymbol{\beta}_{23} = (0.5, 0.5),\ \boldsymbol{\Gamma}_1 = \begin{pmatrix} 1 & -0.4 & 0.4 \\ -0.4 & 1 & 0.4 \\ 0.4 & 0.4 & 1 \end{pmatrix} \text{ and } \boldsymbol{\Gamma}_2 = \begin{pmatrix} 1 & 0.8 & 0.1 \\ 0.8 & 1 & 0.1 \\ 0.1 & 0.1 & 1 \end{pmatrix}.$$

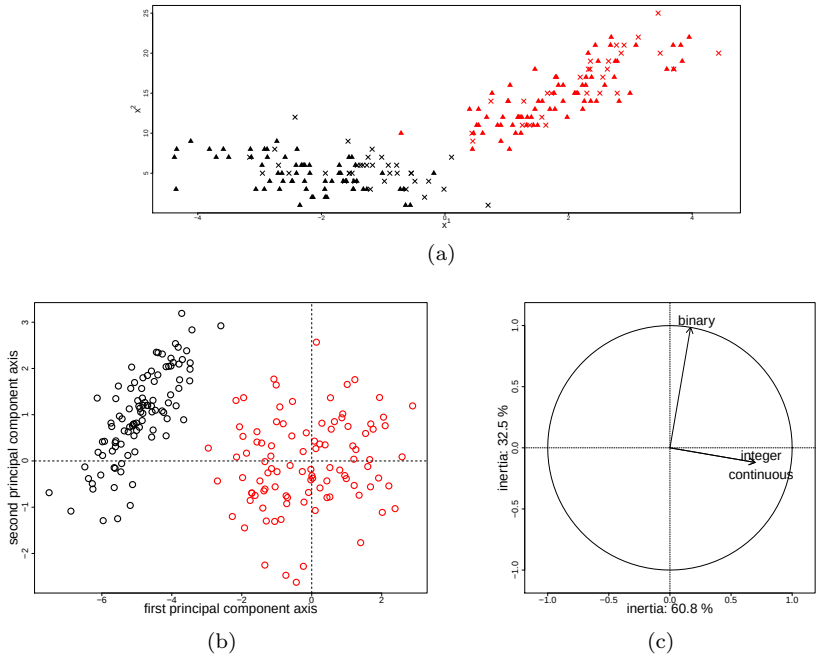


Figure 1: Example of visualization: (a) scatter-plot of the individuals described by three variables: one continuous (abscissa), one integer (ordinate) and one binary (symbol); (b) individuals scatter-plot in the first component map of class 2; (c) variables representation in the first component map of class 2. The color indicates the class memberships.

The visualization of the class 2 is presented in Figure 1. Concerning the individuals, the scatter-plot shows a centred 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.

3 Bayesian inference

Aim We observe the sample $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ composed by 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. 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 (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.

3.1 Maximum *a posteriori* estimate

Prior distributions

Independence assumption A classical assumption is to suppose the independence between the prior distributions, thus

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}) \prod_{k=1}^{g} \left(p(\boldsymbol{\Gamma}_k) \prod_{j=1}^{d} p(\boldsymbol{\beta}_{kj}) \right). \tag{9}$$

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

$$\boldsymbol{\pi} \sim \mathcal{D}_g \left(\frac{1}{2}, \ldots, \frac{1}{2} \right). \tag{10}$$

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

$$\sigma^2_{kj} \sim \mathcal{G}^{-1}(c_0, C_0) \text{ and } \mu_{kj}|\sigma^2_{kj} \sim \mathcal{N}_1(b_0, \sigma^2_{kj}/N_0), \tag{11}$$

  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

$$\boldsymbol{\beta}_{kj} \sim \mathcal{G}(a_0, A_0). \tag{12}$$

  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

$$\boldsymbol{\beta}_{kj} \sim \mathcal{D}_{m_j} \left(\frac{1}{2}, \ldots, \frac{1}{2} \right). \tag{13}$$

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 $\mathbf{\Gamma}_k$ from the prior of the correlation matrix $\mathbf{\Lambda}_k$ since $\mathbf{\Gamma}_k|\mathbf{\Lambda}_k$ is deterministic [Hof07]. So,

$$\mathbf{\Lambda}_k \sim \mathcal{W}^{-1}(s_0, S_0) \text{ and } \forall 1 \leq h, \ell \leq e,\ \mathbf{\Gamma}_k[h,\ell] = \frac{\mathbf{\Lambda}_k[h,\ell]}{\sqrt{\mathbf{\Lambda}_k[h,h]\mathbf{\Lambda}_k[\ell,\ell]}}, \tag{14}$$

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 $\boldsymbol{y}$ 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 $\mathbf{z} = (z_1, \ldots, z_n)$ and the Gaussian vector of the individuals denoted by $\mathbf{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$.

3.2 Gibbs sampler

Algorithm 3.1 (The Gibbs sampler)**.** Starting from an initial value $\boldsymbol{\theta}^{(0)}$, its iteration (r) is written as

$$\mathbf{z}^{(r)}, \mathbf{y}^{(r-1/2)} \sim \mathbf{z}, \mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{(r-1)} \tag{15}$$

$$\boldsymbol{\beta}_{kj}^{(r)}, \boldsymbol{y}_{[rk]}^{j(r)} \sim \boldsymbol{\beta}_{kj}, \boldsymbol{y}_{[rk]}^{j}|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}(r)}, \mathbf{z}^{(r)}, \boldsymbol{\beta}_{k\bar{j}}^{(r)}, \mathbf{\Gamma}_k^{(r-1)} \tag{16}$$

$$\boldsymbol{\pi}^{(r)} \sim \boldsymbol{\pi}|\mathbf{z}^{(r)} \tag{17}$$

$$\mathbf{\Gamma}_k^{(r)} \sim \mathbf{\Gamma}_k|\mathbf{y}^{(r)}, \mathbf{z}^{(r)}, \tag{18}$$

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 3.2 (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 (15) and (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 (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

$$p(\mathbf{z}, \mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{(r-1)}) = \prod_{i=1}^{n} p(z_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(r-1)}) p(\boldsymbol{y}_i|\boldsymbol{x}_i, z_i, \boldsymbol{\theta}^{(r-1)}). \tag{19}$$

We now detail both distributions of the right side of the above equation.

- Each $z_i^{(r)}$ is independently sampled from the following multinomial distribution
$$z_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(r-1)} \sim \mathcal{M}_g(t_{i1}(\boldsymbol{\theta}^{(r-1)}), \ldots, t_{ig}(\boldsymbol{\theta}^{(r-1)})), \tag{20}$$
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 the 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 $(\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta}^{(r-1)})$ as such $\boldsymbol{y}_i^{\mathrm{C}} = \Psi(\boldsymbol{x}_i^{\mathrm{C}}; \boldsymbol{\alpha}_{z_i}^{(r-1)})$ while its last d elements, denoted by $\boldsymbol{y}_i^{\mathrm{D}}$, are sampled according to a d-variate Gaussian distribution $\mathcal{N}_d(\mathbf{0}, \boldsymbol{\Gamma}_{z_i}^{(r-1)})$ truncated on the space $\mathcal{S}_{z_i}(\boldsymbol{x}_i^{\mathrm{D}})$
$$p(\boldsymbol{y}_i^{\mathrm{D}}|\boldsymbol{x}_i, z_i, \boldsymbol{\theta}^{(r-1)}) \propto \phi_d(\boldsymbol{y}_i^{\mathrm{D}}; \boldsymbol{\mu}_{z_i}^{\mathrm{D}(r-1)}, \boldsymbol{\Sigma}_{z_i}^{\mathrm{D}(r-1)}) \mathbb{1}_{\{\boldsymbol{y}_i^{\mathrm{D}} \in \mathcal{S}_{z_i}(\boldsymbol{x}_i^{\mathrm{D}})\}}, \tag{21}$$
where $\boldsymbol{\mu}_{z_i}^{\mathrm{D}(r-1)} = \boldsymbol{\Gamma}_{z_i \mathrm{DC}}^{(r-1)} \boldsymbol{\Gamma}_{z_i \mathrm{CC}}^{-1(r-1)} \Psi(\boldsymbol{x}_i^{\mathrm{C}}; \boldsymbol{\alpha}_{z_i}^{(r-1)})$.

Remark 3.3 (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 (5) which can be too much time consuming if d is large ($d > 6$). Thus, the sampling according to (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 (16) which can be decomposed as follows

$$\begin{aligned} p(\boldsymbol{\beta}_{kj}, \boldsymbol{y}_{[rk]}^{j}|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}(r)}, \mathbf{z}^{(r)}, \boldsymbol{\beta}_{k\bar{j}}^{(r)}, \boldsymbol{\Gamma}_k^{(r-1)}) &= p(\boldsymbol{\beta}_{kj}|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}(r)}, \mathbf{z}^{(r)}, \boldsymbol{\beta}_{k\bar{j}}^{(r)}, \boldsymbol{\Gamma}_k^{(r-1)}) \\ &\quad \times p(\boldsymbol{y}_{[rk]}^{j}|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}(r)}, \mathbf{z}^{(r)}, \boldsymbol{\beta}_{k\bar{j}}^{(r)}, \boldsymbol{\beta}_{kj}, \boldsymbol{\Gamma}_k^{(r-1)}). \end{aligned} \tag{22}$$

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 as such
$$p(\boldsymbol{\beta}_{kj}|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}(r)}, \mathbf{z}^{(r)}, \boldsymbol{\beta}_{k\bar{j}}^{(r)}, \boldsymbol{\Gamma}_k^{(r-1)}) \propto p(\boldsymbol{\beta}_{kj}) \prod_{\{i: z_i^{(r)} = k\}} p(x_i^j|\boldsymbol{y}_i^{\uparrow j(r)}, z_i^{(r)}, \boldsymbol{\Gamma}_k^{(r-1)}, \boldsymbol{\beta}_{kj}). \tag{23}$$

The conditional distribution of $x_i^j|\boldsymbol{y}_i^{\uparrow j(r)}, z_i^{(r)}, \boldsymbol{\Gamma}_k^{(r-1)}$ with $z_i^{(r)} = k$ used on the right side of the above equation is defined by

$$p(x_i^j|\boldsymbol{y}_i^{\uparrow j(r)}, z_i^{(r)}, \boldsymbol{\Gamma}_k^{(r-1)}, \boldsymbol{\beta}_{kj}) = \begin{cases} \phi_1(\frac{x_i^j-\mu_{kj}}{\sigma_{kj}}; \tilde{\mu}_i, \tilde{\sigma}_i^2)/\sigma_{kj} & \text{if } 1 \leq j \leq c \\ \Phi_1(\frac{b^{\oplus}(x_i^j)-\tilde{\mu}_i}{\tilde{\sigma}_i}) - \Phi_1(\frac{b^{\ominus}(x_i^j)-\tilde{\mu}_i}{\tilde{\sigma}_i}) & \text{otherwise,} \end{cases} \tag{24}$$

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 $\boldsymbol{\Gamma}_k$ 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 $\tilde{\sigma}_i^2 = 1 - \boldsymbol{\Gamma}_k^{(r-1)}[j,\bar{j}]\boldsymbol{\Gamma}_k^{(r-1)}[\bar{j},\bar{j}]^{-1}\boldsymbol{\Gamma}_k^{(r-1)}[\bar{j},j]$.

- By the independence between the individuals, the full conditional distribution of $\boldsymbol{y}_{[rk]}^j$ is explicitly defined as

$$p(\boldsymbol{y}_{[rk]}^j|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}(r)}, \mathbf{z}^{(r)}, \boldsymbol{\beta}_{k\bar{j}}^{(r)}, \boldsymbol{\beta}_{kj}, \boldsymbol{\Gamma}_k^{(r-1)}) = \prod_{\{i:z_i^{(r)}=k\}} p(y_i^j|x_i^j, \boldsymbol{y}_i^{\uparrow j(r)}, z_i^{(r)}, \boldsymbol{\beta}_{kj}, \boldsymbol{\Gamma}_k^{(r-1)}). \tag{25}$$

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 as such

$$y_i^{j(r)} = \frac{x_i^j - \mu_{kj}^{(r)}}{\sigma_{kj}^{(r)}}. \tag{26}$$

If x^j is a discrete variable (*i.e.* $c+1 \leq j \leq e$), when $z_i^{(r)} = k$, the full conditional distribution of y_i^j is a truncated Gaussian distribution as such,

$$p(y_i^j|x_i^j, \boldsymbol{y}_i^{\uparrow j(r)}, z_i^{(r)}, \boldsymbol{\beta}_{kj}^{(r)}, \boldsymbol{\Gamma}_k^{(r-1)}) = \frac{\phi_1(y_i^j; \tilde{\mu}_i, \tilde{\sigma}_i^2)}{p(x_i^j; \boldsymbol{\beta}_{kj}^{(r)})} \mathbb{1}_{\{y_i^j \in [b_k^{\ominus(r)}(x_i^j), b_k^{\oplus(r)}(x_i^j)]\}}, \tag{27}$$

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 3.4 (Difficulties to sample the margin parameters)**.** The sampling of $\boldsymbol{\beta}_{kj}$ is not easily performed since the intercept defined in (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 (27) is easily performed.

Vector of proportions sampling

The aim is the sampling from (17) which is classical for the mixture model. The conjugate Jeffreys non informative prior involves that

$$\boldsymbol{\pi}|\mathbf{z}^{(r)} \sim \mathcal{D}_g\left(n_1^{(r)} + \frac{1}{2}, \ldots, n_g^{(r)} + \frac{1}{2}\right), \tag{28}$$

where $n_k^{(r)} = \sum_{i=1}^n \mathbb{1}_{\{z_i^{(r)}=k\}}$.

Correlation matrix sampling

The aim is the sampling from (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

$$\mathbf{\Lambda}_k|\mathbf{y}^{(r)}, \mathbf{z}^{(r)} \sim \mathcal{W}^{-1}\left(s_0 + n_k^{(r-1)}, S_0 + \sum_{\{i:z_i^{(r)}=k\}} \boldsymbol{y}_i^{(r)T}\boldsymbol{y}_i^{(r)}\right) \tag{29}$$

$$\forall 1 \leq h, \ell \leq e,\ \mathbf{\Gamma}_k[h,\ell] = \frac{\mathbf{\Lambda}_k[h,\ell]}{\sqrt{\mathbf{\Lambda}_k[h,h]\mathbf{\Lambda}_k[\ell,\ell]}}. \tag{30}$$

Remark 3.5 (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 (29) is replaced by

$$\mathbf{\Lambda}|\mathbf{y}^{(r)}, \mathbf{z}^{(r)} \sim \mathcal{W}^{-1}\left(s_0 + n, S_0 + \sum_{i=1}^{n} \boldsymbol{y}_i^{(r)T}\boldsymbol{y}_i^{(r)}\right), \tag{31}$$

and we put $\mathbf{\Lambda}_k = \mathbf{\Lambda}$ for $k = 1, \ldots, g$.

According to both Remarks 3.3 and 3.4, the first two steps of the Gibbs sampler involve difficulties avoided by the following hybrid MCMC algorithm.

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 (15) and (23)) by one iteration of two Metropolis-Hastings steps that we now detail.

Class membership and Gaussian vector sampling

The step (15) is performed via one iteration of the Metropolis-Hastings algorithm. This algorithm is independently performed to sample each couple $(z_i, \boldsymbol{y}_i)$ since the individuals are independent. Its stationary distribution is

$$p(z_i, \boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(r-1)}) \propto \pi_{z_i} p(\boldsymbol{x}_i, \boldsymbol{y}_i|z_i, \boldsymbol{\theta}^{(r-1)}). \tag{32}$$

Note that $p(\boldsymbol{x}_i, \boldsymbol{y}_i|z_i, \boldsymbol{\theta}^{(r-1)}) = \phi_e(\boldsymbol{y}_i; \mathbf{0}, \mathbf{\Gamma}_{z_i}^{(r-1)})\mathbb{1}_{\{\boldsymbol{y}_i^{\mathrm{C}} = \Psi(\boldsymbol{x}_i^{\mathrm{C}}; \boldsymbol{\alpha}_{z_i}^{(r-1)})\}}\mathbb{1}_{\{\boldsymbol{y}_i^{\mathrm{D}} \in \mathcal{S}_{z_i}(\boldsymbol{x}_i^{\mathrm{D}})\}}$.

The Metropolis-Hastings algorithm samples a candidate $(z_i^\star, \boldsymbol{y}_i^\star)$ by the instrumental distribution $q_1(.|\boldsymbol{x}_i, \boldsymbol{\theta}^{(r-1)})$ which uniformly samples $z_i^\star$ then which samples $\boldsymbol{y}_i^\star|z_i^\star$ as follows. Conditionally on $z_i^\star$, this instrumental distribution is deterministic for the first c elements of $\boldsymbol{y}_i^\star$, denoted by $\boldsymbol{y}_i^{\star\mathrm{C}}$ as such $\boldsymbol{y}_i^{\star\mathrm{C}} =$

$\Psi(\boldsymbol{x}_i^{\mathrm{C}};\boldsymbol{\alpha}_{z_i^\star}^{(r-1)})$, 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 $\mathcal{S}_{z_i^\star}(\boldsymbol{x}_i^{\mathrm{D}})$. Thus,

$$q_1(z_i,\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}^{(r-1)}) = \frac{1}{g}\frac{\phi_d(\boldsymbol{y}_i^{\mathrm{D}};\mathbf{0},\boldsymbol{I})}{\prod_{j=c+1}^{e} p(x_i^j;\boldsymbol{\beta}_{z_i j}^{(r-1)})}\mathbb{1}_{\{\boldsymbol{y}_i^{\mathrm{C}}=\Psi(\boldsymbol{x}_i^{\mathrm{C}};\boldsymbol{\alpha}_{z_i}^{(r-1)})\}}\mathbb{1}_{\{\boldsymbol{y}_i^{\mathrm{D}}\in\mathcal{S}_{z_i}(\boldsymbol{x}_i^{\mathrm{D}})\}}. \tag{33}$$

The candidate is accepted with the probability

$$\rho_{1i}^{(r)} = \min\left\{\frac{\pi_{z_i^\star}\phi_e(\boldsymbol{y}_i^\star;\mathbf{0},\boldsymbol{\Gamma}_{z_i^\star}^{(r-1)})}{\pi_{z_i^{(r-1)}}\phi_e(\boldsymbol{y}_i^{(r-1)};\mathbf{0},\boldsymbol{\Gamma}_{z_i^{(r-1)}}^{(r-1)})}\frac{q_1(z_i^{(r-1)},\boldsymbol{y}_i^{(r-1)}|\boldsymbol{x}_i)}{q_1(z_i^\star,\boldsymbol{y}_i^\star|\boldsymbol{x}_i)};1\right\}. \tag{34}$$

Thus, at the iteration (r) of the Algorithm 3.1, the sampling according to (15) is performed via one iteration of the following Metropolis-Hastings algorithm.

Algorithm 3.6 (Metropolis-Hastings with $p(z_i,\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}^{(r-1)})$ as stationary distribution)**.**

$$(z_i^\star,\boldsymbol{y}_i^\star)\sim q_1(z,\boldsymbol{y}|\boldsymbol{x}_i) \tag{35}$$

$$(z_i^{(r)},\boldsymbol{y}_i^{(r-1/2)}) = \begin{cases} (z_i^\star,\boldsymbol{y}_i^\star) & \text{with probability } \rho_{1i}^{(r)} \\ (z_i^{(r-1)},\boldsymbol{y}_i^{(r-1)}) & \text{with probability } 1-\rho_{1i}^{(r)}. \end{cases} \tag{36}$$

Margin parameter sampling

The step (16) is performed in two steps. Firstly the sampling of $\boldsymbol{\beta}_{kj}^{(r)}$ according to (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 (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,

$$q_2(.|\mathbf{x},\mathbf{z}) = p(\boldsymbol{\beta}_{kj}|\mathbf{x},\mathbf{z},\boldsymbol{\Gamma}_k=\boldsymbol{I}). \tag{37}$$

Thus, according to (23), the candidate $\boldsymbol{\beta}_{kj}^\star$ is accepted with the probability

$$\rho_2^{(r)} = \min\left\{\frac{p(\boldsymbol{\beta}_{kj}^\star)q_2(\boldsymbol{\beta}_{kj}^{(r-1)}|\mathbf{x},\mathbf{z})}{p(\boldsymbol{\beta}_{kj}^{(r-1)})q_2(\boldsymbol{\beta}_{kj}^\star|\mathbf{x},\mathbf{z})}\prod_{\{i:z_i^{(r)}=k\}}\frac{p(\boldsymbol{y}_i^j|x_i^j,\boldsymbol{y}_i^{\uparrow j(r)},z_i,\boldsymbol{\beta}_{kj}^\star,\boldsymbol{\Gamma}_k^{(r-1)})}{p(\boldsymbol{y}_i^j|x_i^j,\boldsymbol{y}_i^{\uparrow j(r)},z_i,\boldsymbol{\beta}_{kj}^{(r-1)},\boldsymbol{\Gamma}_k^{(r-1)})};1\right\}.$$

Thus, at the iteration (r) of the Algorithm 3.1, the sampling according to (16) is performed via one iteration of the following Metropolis-Hastings algorithm.

Algorithm 3.7 (Metropolis-Hastings with $p(\boldsymbol{\beta}_{kj}|\boldsymbol{x}_{[rk]},\boldsymbol{y}_{[rk]}^{\bar{j}(r)},\mathbf{z},\boldsymbol{\beta}_{k\bar{j}}^{(r)},\boldsymbol{\Gamma}_k)$ as stationary distribution)**.**

$$\boldsymbol{\beta}_{kj}^\star\sim q_2(\boldsymbol{\beta}_{kj}|\mathbf{x},\mathbf{z}) \tag{38}$$

$$\boldsymbol{\beta}_{kj}^{(r)} = \begin{cases} \boldsymbol{\beta}_{kj}^\star & \text{with probability } \rho_2^{(r)} \\ \boldsymbol{\beta}_{kj}^{(r-1)} & \text{with probability } 1-\rho_2^{(r)}. \end{cases} \tag{39}$$

Remark 3.8 (Instrumental distributions)**.** Note that, the smaller are the intra-class dependencies of the variable $\boldsymbol{x}$, the closer of the stationary distributions are the instrumental distributions of both Metropolis-Hastings algorithms.

3.4 Label switching problem

The label switching problem is generally solved by specific procedures [Ste00]. However, based on the argument of [JB14], 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 dealing 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 of [Ste00].

4 Simulations

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] 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, the algorithm is initialized with the maximum likelihood estimate of the locally independent 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 4.1 (Mixed variables: one continuous, one integer and one binary)**.** We consider the mixture model of Gaussian copulas detailed in Example 2.10 and composed by one continuous variable, one integer variable and one binary variable. Figure 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 size of sample 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.

Simulation 4.2 (Robustness of the mixture model of Gaussian copulas)**.** During these experiments, data are sampled according to a bivariate Poisson mixture

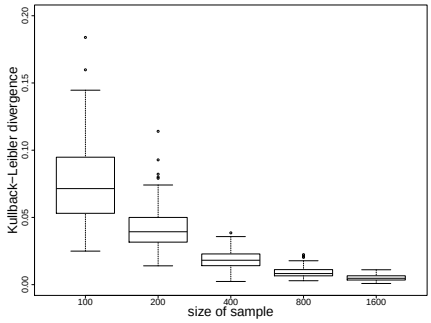


Figure 2: Decrease of the Kullback-Leibler divergence of the model with the maximum *a posteriori* estimate from the model with the true parameter.

model [KT08] whose the margin parameter are denoted by $\boldsymbol{\alpha}_k = (\lambda_{k1}, \lambda_{k2}, \lambda_{k3})$. The simulation is performed with the following values of the parameters

$$\boldsymbol{\pi} = (1/3, 2/3),\ \lambda_{1h} = h \text{ and } \lambda_{2h} = 3 + h,\ \text{for } h = 1, 2, 3. \tag{40}$$

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.

5 Applications

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 locally independent 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 he 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.

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).

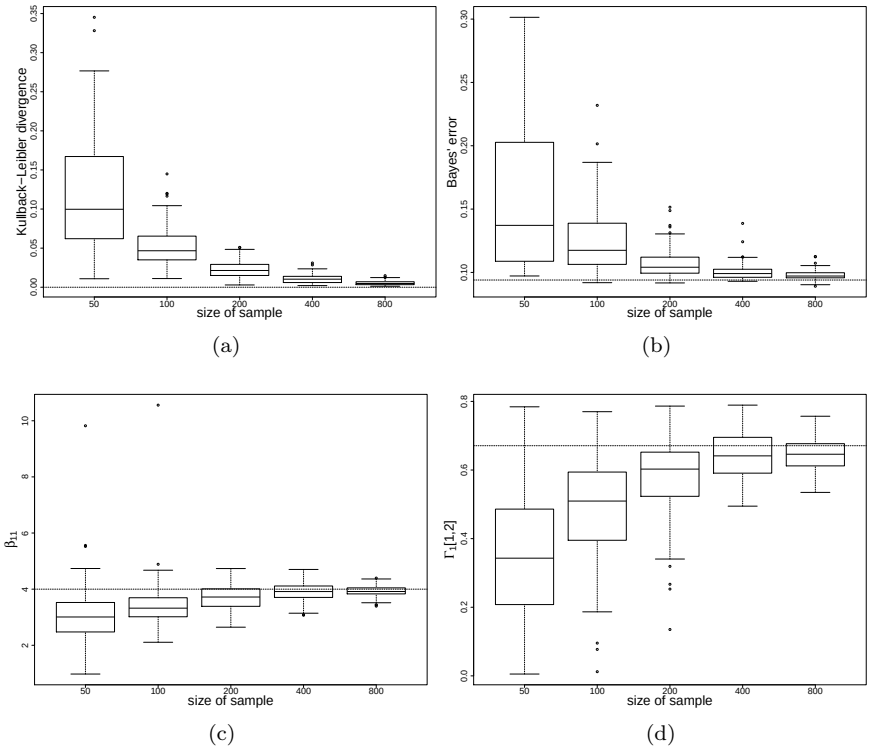


Figure 3: Results of Simulation 4.2: (a) Kullback-Leibler divergence of the estimated model from the true one; (b) Error rate of the estimated model; (c) Value of the first margin parameter for the class 1; (d) Value of the correlation coefficient between both variables for the class 1.

Model selection

We estimate the three mixture models (locally independent one, heteroscedastic Gaussian copula mixture and homoscedastic Gaussian copula mixture) for different numbers of classes. Table 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.

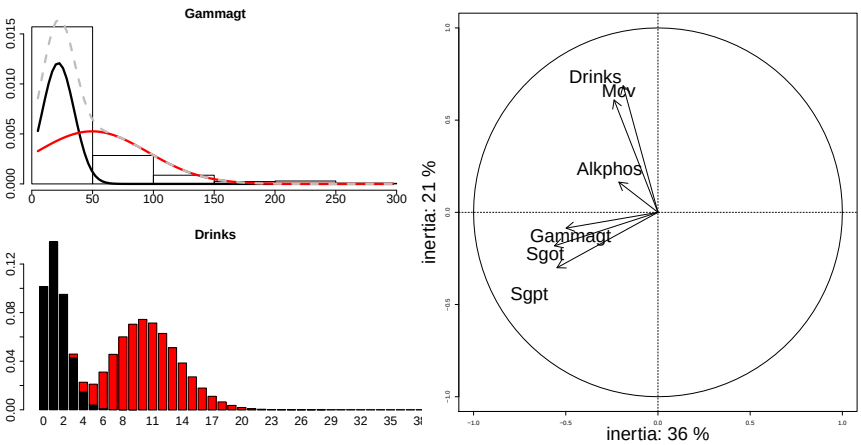
Interpretation of the best model

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 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

	g	1	2	3	4	5	6
BIC	loc. indpt.	-8690	**-8017**	-8039	-8092	-8130	-8235
	hetero.	-8551	**-7935**	-8103	-8157	-8277	-8287
	homo.	-8551	**-7898**	-7999	-8032	-8050	-8123
ICL	loc. indpt.	-8690	**-8026**	-8060	-8117	-8208	-8341
	hetero.	-8551	**-7943**	-8120	-8171	-8322	-8306
	homo.	-8551	**-7907**	-8032	-8043	-8088	-8205

Table 1: Values of the BIC and ICL criteria for the three mixture models estimated on the liver disorder data set.

is positively correlated with the number of alcoholic drinks.



(a) $\pi_k p(x^j|\boldsymbol{\beta}_{kj}, z = k)$

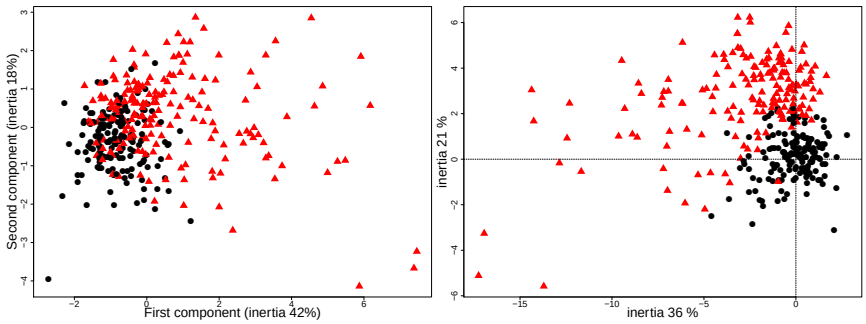
(b) Variables in the first correlation circle given by $\boldsymbol{\Gamma}_k$

Figure 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 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 5b displays the individuals in the first PCA map of the 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.

Note that the partitions obtained by the three bi-component models are similar but not identical as shown by Table 2.



(a) First PCA map

(b) First PCA of the class 1

Figure 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).

	hetero.	
	c1	c2
c1-homo.	190	0
c2-homo.	5	150

(a)

	loc. indpt.	
	c1	c2
c1-homo.	190	0
c2-homo.	7	148

(b)

Table 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 locally independent model.

Conclusion

On this data set, the mixture model of Gaussian copulas better fits the data according to the information criteria than the locally independent 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.

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 (locally independent 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 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	loc. indpt.	-63516	-61069	-61010	**-55967**	-60250	-57163
	hetero.	-44675	**-34520**	-39724	-44692	-44484	-48349
	homo.	-44675	-39372	**-38289**	-45209	-43217	-42417
ICL	loc. indpt.	-63516	-61229	-61365	**-56310**	-60726	-58138
	hetero.	-44675	**-34688**	-40176	-44933	-44758	-48959
	homo.	-44675	-39607	**-38791**	-45380	-43345	-42667

Table 3: Values of the BIC and ICL criteria for the three mixture models estimated on the wine data set.

Partition study

Table 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.

	white	red
c1	4359	9
c2	538	1590

(a)

	white	red
c1	2441	12
c2	1911	7
c3	545	1580

(b)

	white	red
c1	2547	1561
c2	2007	35
c3	275	3
c4	68	0

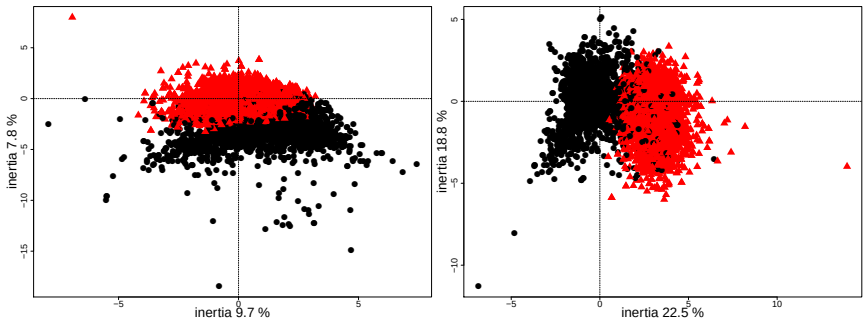
(c)

Table 4: 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 locally independent mixture.

Figure 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.

Interpretation of the best model

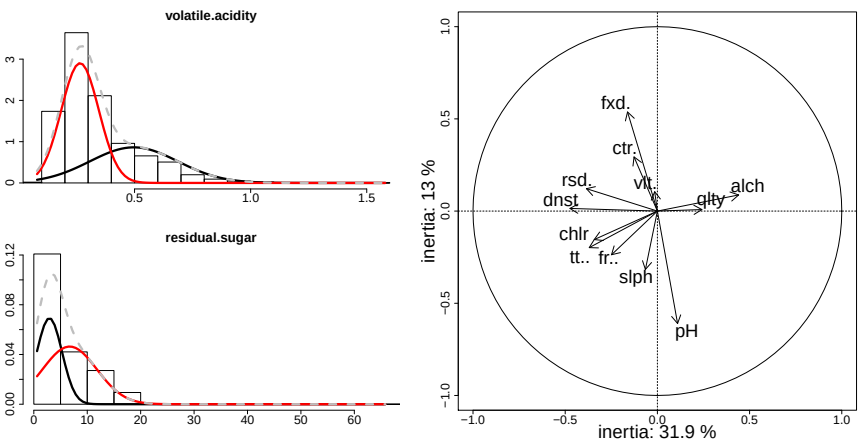
The following interpretation is based on the margin parameters and on the intra-class correlation matrices summarized by Figure 7. The majority class



(a) First PCA of the class 1, map(3,5) (b) First PCA of the class 2, map(1,2)

Figure 6: Visualization of the partition by the homoscedastic bicomponent mixture model of Gaussian copulas for the wine data set (Class 1 is drawn by black circles and Class 2 by red triangles).

($\pi_1 = 0.59$) is principally composed by 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.



(a) $\pi_k p(x^j|\boldsymbol{\beta}_{kj}, z = k)$ (b) Variables in the first correlation circle given by $\boldsymbol{\Gamma}_2$

Figure 7: Summary of the homoscedastic bi-component Gaussian copula mixture model for the wine data set. Class 1 is drawn in black and Class 2 in red.

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 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.

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

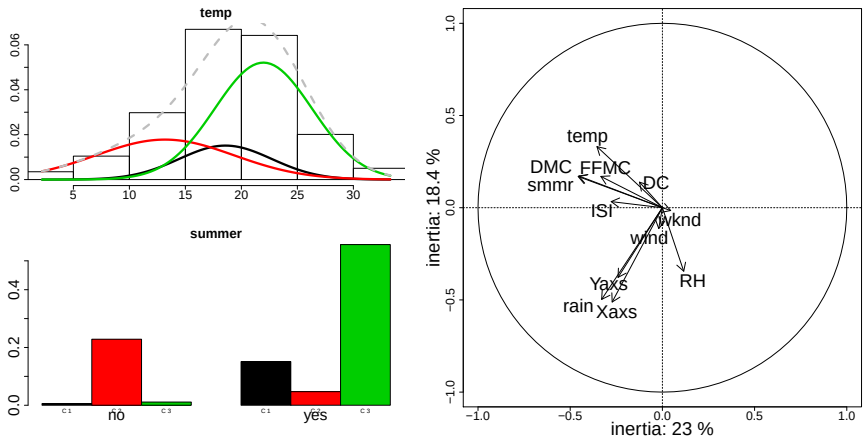
Table 5 presents the values of both used information criteria for the three mixture models. According to both criteria, the model better fitting the data is the homoscedastic mixture model of Gaussian copulas with three components.

	g	1	2	3	4	5	6
BIC	loc. indpt.	-16559	**-16296**	-16473	-17370	-17379	-17454
	hetero	-16559	**- 16002**	-16171	-16410	-16666	-16791
	homo.	-16559	-15899	**-15824**	-16300	-15946	-16034
ICL	loc. indpt.	-16559	**-16301**	-16494	-17401	-17400	-17527
	hetero	-16559	**-16014**	-16205	-16471	-16721	-16871
	homo.	–16559	-15907	**-15893**	-16352	-16020	-16137

Table 5: Values of the BIC and ICL criteria for the three mixture models estimated on the forest fire data set.

Interpretation of the best model

The following interpretation is based on the margin parameters on the intra-class correlation matrices summarized in Figure 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 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: 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 6.

	hetero.	
	c1	c2
c1-homo.	244	23
c2-homo.	1	127
c3-homo.	122	0

(a)

	loc. indpt.	
	c1	c2
c1-homo.	265	2
c2-homo.	7	121
c3-homo.	111	11

(b)

Table 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 locally independent model.

Conclusion

The model points out three classes of forest fires. It is more precise than the locally independent 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 allows 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.

6 Conclusion and future extensions

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 distribution 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 locally independent 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 study.

References

[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.

[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.

[BR93] J.D. Banfield and A.E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, pages 803–821, 1993.

[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.

[CM07] P. Cortez and A. Morais. A data mining approach to predict forest fires using meteorological data. 2007.

[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.

[FS06] S. Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer, 2006.

[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.

[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.

[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.

[HY01] D.J. Hand and K. Yu. Idiot's bayes—not so stupid after all? *International Statistical Review*, 69(3):385–398, 2001.

[JB14] J. Jacques and C. Biernacki. Model-based clustering for multivariate partial ranking data. *Journal of Statistical Planning and Inference*, 149:201–217, 2014.

[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.

[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.

[Lew98] D.D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Machine learning: ECML-98*, pages 4–15. Springer, 1998.

[MCMM09a] C. Maugis, G. Celeux, and M.L. Martin-Magniette. Variable selection for clustering with Gaussian mixture models. *Biometrics*, 65(3):701–709, 2009.

[MCMM09b] C. Maugis, G. Celeux, and M.L. Martin-Magniette. Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics & Data Analysis*, 53(11):3872–3882, 2009.

[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.

[Mor12] I. Morlini. A latent variables approach for clustering mixed binary and continuous variables within a Gaussian mixture model. *Advances in Data Analysis and Classification*, 6(1):5–28, 2012.

[MP00] G.J. McLachlan and D. Peel. *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics, Wiley-Interscience, New York, 2000.

[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.

[PCK06] M. Pitt, D. Chan, and R. Kohn. Efficient Bayesian inference for Gaussian copula regression models. *Biometrika*, 93(3):537–554, 2006.

[Raf96] A.E. Raftery. Hypothesis testing and model selection. In *Markov chain Monte Carlo in practice*, pages 163–187. Springer, 1996.

[RC04] C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Verlag, 2004.

[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.

[Ste00] 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. *The Annals of Mathematical Statistics*, pages 1265–1269, 1963.

[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.

[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.

[YS+68] S.J. Yakowitz, J.D. Spragins, et al. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.

A Proof of the model identifiability

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 A.1 (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,*

$$\forall \boldsymbol{x} \in \mathbb{R}^c \times \mathbb{N}^d, \quad \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}; \boldsymbol{\alpha}_k) = \sum_{k=1}^{g'} \pi'_k p(\boldsymbol{x}; \boldsymbol{\alpha}'_k) \tag{41}$$

$$\Rightarrow \quad g = g',\ \boldsymbol{\pi} = \boldsymbol{\pi}',\ \boldsymbol{\alpha} = \boldsymbol{\alpha}'. \tag{42}$$

Proof. The identifiability of the multivariate Gaussian mixture models and of the univariate Poisson mixture model [Tei63, YS+68] involves that (41) implies

$$g = g',\ \boldsymbol{\pi} = \boldsymbol{\pi}',\ \boldsymbol{\beta}_{kj} = \boldsymbol{\beta}'_{kj} \text{ and } \boldsymbol{\Gamma}_{k\mathrm{CC}} = \boldsymbol{\Gamma}'_{k\mathrm{CC}}. \tag{43}$$

We now show that $\boldsymbol{\Gamma}_{k\mathrm{CD}} = \boldsymbol{\Gamma}'_{k\mathrm{CD}}$ and $\boldsymbol{\Gamma}_{k\mathrm{DD}} = \boldsymbol{\Gamma}'_{k\mathrm{DD}}$.

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 (41) 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 $\gamma_t = \{(x^j, x^h) \in \mathbb{R} \times \mathbb{N} : \quad a_1 = t\}$. Then, letting $x^h \to \infty$ as such $(x^j, x^h) \in \gamma_t$,

$$\forall t, \quad \frac{\int_t^{a_1'} \phi(u)du}{\Phi(t)} = 0. \tag{44}$$

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 $\mathbf{\Gamma}_{k\text{CD}} = \mathbf{\Gamma}'_{k\text{CD}}$.

When both variables are integer, we use the same argument with $\gamma_{(t,\xi)} = \{(x^j, x^h) \in \mathbb{N} \times \mathbb{N} : \quad a_1 \in B(t, \xi)\}$. 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

$$\frac{\int_{t+\xi}^{a_1'} \phi(u)du}{\Phi(t-\xi)} = 0 \text{ and } \frac{\int_{t+\xi}^{a_1'} \phi(u)du}{\Phi(t-\xi)} > 0. \tag{45}$$

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 $\mathbf{\Gamma}_{k\text{DD}} = \mathbf{\Gamma}'_{k\text{DD}}$. □

Proposition A.2 (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 $\mathbf{\Gamma}_{k\text{CD}} = \mathbf{\Gamma}'_{k\text{CD}}$ and $\mathbf{\Gamma}_{k\text{DD}} = \mathbf{\Gamma}'_{k\text{DD}}$.

Let $j = 1$ and let $h \in \{2, 3\}$. We note $\rho_k = \mathbf{\Gamma}_k(j, h)$, $\rho_k' = \mathbf{\Gamma}'_k(j, h)$, $v_k = \Phi_1^{-1}(P(x^j; \boldsymbol{\beta}_{kj}))$, $\varepsilon_k(x^j) = \pi_k \frac{\phi(v_k;0,1)}{\sigma_{kj}}$, $a_k = \frac{b_k^{\oplus}(x^j) - \rho_k v_k}{\sqrt{1-\rho_k^2}}$ and $a_k' = \frac{b_k'^{\oplus}(x^j) - \rho_k' v_k}{\sqrt{1-\rho_k'^2}}$. 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 (41) 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 $\mathbf{\Gamma}_{k\text{CD}} = \mathbf{\Gamma}'_{k\text{CD}}$ then $\mathbf{\Gamma}_{k\text{DD}} = \mathbf{\Gamma}'_{k\text{DD}}$. □